Andreas Henrich
Hans-Christian Sperker (Eds.)

LWA 2013

Lernen, Wissen & Adaptivität
Workshop Proceedings
Bamberg, 7.–9. Oktober 2013

Preface

The joint workshop event LWA 2013 (Lernen, Wissen & Adaptivität) takes place in Bamberg, Germany on October 7th to October 9th 2013. Like in the years before the LWA hosts a broad scope of workshops of the German Informatics Society's special interest groups for

Information Retrieval (FG-IR)

Knowledge Discovery and Machine Learning (FG-KDML)

Knowledge and Experience Management (FG-WM)

In addition to workshops of the special interest groups we invited four talks covering current research questions in computer science:

Klaus-Dieter Althoff: Collaborative Multi-Expert-Systems: towards more flexibly acquiring, integrating, and processing case-specific and (more) general knowledge

Sung-Pil Choi: Systematic Approach to the Knowledge Extraction and Structuring for Scientific Big Data Analytics

Thorsten Staake: Smart Grid Data Analytics to Promote Energy Efficiency

Diedrich Wolter: Qualitative Representations of Space and Time: Lean Knowledge Representations for Efficient and Effective Reasoning

We are grateful for the support of the PC chairs for the single workshops and all the local staff at the University of Bamberg for the organizational and administrative work.
Furthermore, we thank all participants of the workshops for their contributions. Additionally, we want to thank all reviewers for their careful help in selecting and improving the provided submissions.

Bamberg, October 2013 Andreas Henrich and Hans-Christian Sperker

Keynotes of the LWA 2013

***Sung-Pil Choi*:**
Systematic Approach to the Knowledge Extraction and Structuring for Scientific Big Data Analytics

In this presentation, we introduce a methodical model for constructing and operating the software system of scientific big data analytics in order to support various R&D activities which are being performed by scientists and engineers for their research. The model includes two important technical aspects: information extraction model for extracting useful technological knowledge from scientific documents such as papers and patents; parallel execution model for maximizing the speed and volume of the information extraction model. We will explain each building block of the entire proposed system in detail while introducing a series of evaluation criteria for the components. Furthermore, the presentation will also cover some interesting research topics that will be necessary for us to enhance the introduced knowledge processing model in terms of performance as well as its functional completeness such as Textual Entailment Analysis, Advanced Information Retrieval using Paraphrases and Patent Analytics.

***Klaus-Dieter Althoff*:**
Collaborative Multi-Expert-Systems: towards more flexibly acquiring, integrating, and processing case-specific and (more) general knowledge

Case-based reasoning (CBR) and expert systems have a long tradition in artificial intelligence: CBR since the late 1970s and expert systems since the late 1960s. While expert systems are based on expertise and expert reasoning capabilities for a specific area of responsibility, CBR is an approach for problem solving and learning of humans and computers. Starting from different research activities, CBR and expert systems have become overlapping research fields. In this talk the relationships between CBR and expert systems are analyzed from different perspectives like problem solving, learning, competence development, and knowledge types. As human case-based reasoners are quite successful in integrating problem-solving and learning, combining different problem solving strategies, utilizing different kinds of knowledge, and becoming experts for specific areas of responsibility, computer based expert systems do not have the reputation to be successful at these tasks. Based on this, the talk will discuss the learning ability of expert systems on different levels and the role CBR may play here. A research project is introduced that aims at, among others, improving the learning ability of expert systems by systematically considering multiple expert(s) (systems) as well as the wisdom of the crowd. The corresponding software architecture integrates concepts from software engineering (experience factory, software product lines) and artificial intelligence (multi-agent systems, CBR). In the scope of this research CBR is used in various ways: for representing and processing the experience part of expertise, for supporting continuous knowledge evolution and increasing knowledge formalization, as well as for providing an open framework for constructing learning expert systems. The current state of implementation is presented as along with open challenges and an outlook on future research.

***Diedrich Wolter*:**
Qualitative Representations of Space and Time: Lean Knowledge Representations for Efficient and Effective Reasoning

Qualitative representations have been introduced to reduce overly detailed and comprehensive amounts of data to knowledge representations of manageable size at an appropriate level of detail. The idea underlying qualitative approaches is to abstract from all pieces of information that are not relevant for a task at hand. Qualitative representations are discrete symbolic representations of continuous, often infinite domains. In particular spatial and temporal domains feature a domain structure that allows for semantically rich but compact representations. After about two decades of research in the area of spatio-temporal representations, there are now manifold approaches to tackle various applications. In this talk I aim to characterize some of the key ingredients of successful qualitative representations. Looking at selected application domains I give an overview of how qualitative reasoning can support a variety of tasks.

***Thorsten Staake*:**
Smart Grid Data Analytics to Promote Energy Efficiency

Increasing energy efficiency and reducing carbon emissions are foremost objectives of our society. Information Systems (IS) that offer feedback on personal energy consumption can contribute to achieving these objectives as they can help to discover fields of optimization and motivate actors to use resources in a sustainable way. The presentation provides examples of such systems and details current research results on techniques that utilize patterns in household electricity consumption profiles in order to provide targeted saving advice.

Workshop "Information Retrieval" (IR-2013)

Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2013)

Classification and Learning

Graphs and Networks

Subgroups and Subspaces

Spatio-Temporal Aspects

Tagging and Labeling

Recommendations

Applications

Workshop "Knowledge and Experience Management" (FGWM-2013)

Business Processes

Wissensmanagement i.e.S.

Decision Support Systems

Decision Support Systems / Medien

Visualisierung

Workshop "Information Retrieval" (IR-2013)

The ubiquity of search systems has led to the application of information retrieval technology in many new contexts (e.g. mobile and international) and for new object types (products, patents, music, microblogs). To develop appropriate products, basic knowledge on information retrieval needs to be revisited and innovative approaches need to be applied, for example by allowing for more user interaction or by taking the user's situational context and the overall task into account. The quality of information retrieval needs to be evaluated for each context. Large evaluation initiatives respond to these challenges and develop new benchmarks.
The workshop Information Retrieval 2013 of the Special Interest Group on Information Retrieval within the German Gesellschaft für Informatik (GI) provides a forum for scientific discussion and the exchange of ideas. The workshop takes place in the context of the LWA "Learning, Knowledge and Adaptivity" workshop week (LWA, Oct 7-9, 2013) at the University of Bamberg in Germany. This workshop continues a successful series of conferences and workshops of the Special Interest Group on Information Retrieval (`http://www.fg-ir.de/`). The workshop addresses researchers and practitioners from industry and universities. Especially Doctorate and Master students are encouraged to participate and discuss their ideas with world renowned experts. An Industry Session will stimulate the exchange between information retrieval professionals and academics. The workshop is expected to include German as well as English presentations.

Topics of Interest

Submission should address current issues in Information Retrieval. They include (but are not limited to):

- Development and optimization of retrieval systems
- Information retrieval theory
- Retrieval with structured and multimedia documents
- Evaluation and evaluation research
- Text mining and information extraction
- Cross-lingual and cross-cultural IR
- Digital libraries
- User interfaces and user behavior, HCIR
- Interactive IR
- Machine learning in information retrieval
- Information retrieval and knowledge management
- Information retrieval and the semantic web
- Databases and information retrieval
- Social Search
- Task-based IR

- Web information retrieval (including blogs and microblogs)
- Clustering
- Patent retrieval
- Plagiarism detection
- Enterprise search
- Expert search
- Innovative concepts in IR teaching

We especially invite descriptions of running projects.

Program Chairs

- Dr.-Ing. Sascha Kriewel, University Duisburg-Essen, Germany
- Dr. Claus-Peter Klas, Fernuniversität in Hagen, Germany

Program Committee

- Ralf Schenkel, Universitaet Passau
- Christian Wolff, Regensburg University
- Thomas Mandl, University of Hildesheim
- Udo Kruschwitz, University of Essex
- Joachim Griesbaum, University of Hildesheim
- Ingo Frommholz, University of Bedfordshire
- Ulrich Thiel, Fraunhofer Gesellschaft
- David Zellhoefer, BTU Cottbus
- Johannes Leveling, Dublin City University (DCU)
- Benno Stein, Bauhaus-Universität Weimar
- Katja Hofmann, University of Amsterdam
- Reginald Ferber, Hochschule Darmstadt
- David Elsweiler, University of Regensburg
- Ingo Schmitt, Technical University Cottbus
- Michael Kluck, Stiftung Wissenschaft und Politik Berlin
- Vivien Petras, HU Berlin
- Christa Womser-Hacker, Universität Hildesheim

- Henning Müller, HES-SO Valais-Wallis
- Frank Hopfgartner, Technische Universität Berlin
- Norbert Fuhr, University of Duisburg-Essen
- Maximilian Eib, Chemnitz Technical University
- Thomas Roelleke, Queen Mary University of London

Supporting Collaborative Information Seeking and Searching in Distributed Environments

Thilo Böhm
University of Hagen
58084, Hagen, Germany
thilo.boehm@fernuni-hagen.de

Claus-Peter Klas
University of Hagen
58084, Hagen, Germany
claus-peter.klas@fernuni-hagen.de

Matthias Hemmje
University of Hagen
58084, Hagen, Germany
matthias.hemmje@fernuni-hagen.de

Abstract

In various work domains, the collaborative performance of a work-task by a team can lead to a shared information need required to fulfill this task. Many empirical studies identified collaborative information seeking and searching (IS&S) as everyday work patterns in order to solve a shared information need and to benefit from the diverse expertise and experience of the team members.

This paper presents first empirical results in an ongoing research project: We report on a pilot user study that investigates the collaborative IS&S practices of three work groups in academic and industrial research facilities. The aim of the conducted pilot study was to capture the use of software technologies for realizing collaboration, information seeking and sharing in real-world work settings. We discuss resulting design implications as guideline for extending the *ezDL*[1] system towards supporting collaborative IS&S activities.

1 Introduction

In various work domains, the collaborative performance of a work-task by a team can lead to a shared information need required to fulfill this task. Many empirical studies identified collaboration during information seeking and searching (IS&S) as everyday work patterns. Collaborative information seeking and searching (CIS&S) is characterized by parties that share the same information need and explicitly work together to satisfy that need and to benefit from the diverse expertise and experience of the team members. This collaboration involves synergetic interactions between individuals, negotiations, discussions and the adoption of other perspectives to produce a solution or strategy, which results from the different knowledge and backgrounds of the co-workers [30]. Effective and efficient collaboration in distributed environments requires a number of awareness information. In addition to information about the current activities in the group, gathering information about participants, their special skills and knowledge is necessary to allow for combination of expertise and efficient achievement of goals [28].

Previous research in the field of CIS&S has conceptualized, implemented and evaluated tools and systems for use at each stage of the information searching process: (1) query construction, (2) obtaining results, and (3) evaluating and using the results. These tools have been developed, to a large extent, in experimental settings. They provide an environment where collaboration is mediated at different layers (depth of mediation, [9]). Using frontend mediation, integrated functions in the UI allow communication, exchange of information, and provision of awareness cues. Using backend mediation, each person's activities can be combined algorithmically to produce the desired retrieval effects. However, recent empirical studies show that, despite the increasing availability of tools that are specifically designed to support CIS&S, these technologies are not used in practice [18]. Instead, simpler communications technologies that are part of everyday work are applied as means to realize CIS&S. In such environments, people communicate about the search process and the search products, but neither user interface nor utilized services (e.g. search engines and digital libraries) are aware that people intend to collaborate.

An arising research question we want to address is, how team members can be provided with information on the best suited collaboration partners and the collaborative activity to be performed in order to increase the efficiency and effectiveness of IS&S tasks in such environments. To approach this question, we conducted a pilot user study that aimed at capturing the tools and means in use by practitioners of different work groups in academia and industry to collaborate with their colleagues. From the results of this study, we derive implications for the design of an environment supporting CIS&S activities in team-based work-task situations.

The rest of the paper is structured as follows: Section 2 discusses related studies in the field of CIS&S and gives an overview of systems and techniques especially designed to support CIS&S. In section 3, we present the results of the conducted survey. Section 4 discusses resulting design implications and presents the application of these design implications to an extension of the *ezDL* system [1]. Finally, section 5 summarizes this paper and gives a brief outlook on the next project tasks.

2 Background and Related Work

Various empirical studies identified collaborative information retrieval as an everyday work pattern in order to solve a shared information need that occurs in the context of a work-task. The concept of the task has been defined by Byström and Hansen as an activity that is carried out to achieve a specific goal or has a specific item of work in focus [5]. A task may consist of several sub-tasks. A work-task represents a specific task that is carried out to fulfill a separable portion of a person's duties to his employer. As result of an identified information need, a

[1] *ezDL*: Easy Access to Digital Libraries, www.ezdl.de

work-task may consist of information seeking tasks that are further decomposed into information searching tasks [5]. Information seeking generally focuses on the satisfaction of a complex information need. It involves several sources and consultations of them. Information searching is particularly concerned with the satisfaction of a separable fraction of that complex information need.

2.1 Related User Studies

Bruce et al. [3] present an empirical study that investigates the collaborative information seeking behavior in two design teams. The authors found that collaborative information retrieval is an integral part of the daily work to solve shared information needs of the team. Identifying, analyzing and defining the information need, as well as the development of search strategies is performed collaboratively. This involves intra-team as well as extra-team collaboration [23].

CIS&S activities often involve information sharing. Talja [29] observed and classified different types of information sharing in an academic environment. These types are (1) strategic sharing, (2) paradigmatic sharing, (3) directive sharing, (4) social sharing, and (5) no sharing. Her investigations showed that in academia, collaborative information seeking is as common as individual information seeking. Scholars usually belong to different networks, i.e. social networks. According to Talja, these networks not only influence their choices of information seeking strategies, but are the place where information is sought, interpreted, used, and created.

A study conducted by Hansen and Jarvelin [10] analyzed the information seeking behavior of the employees of the Swedish patent office when engaged in the patent application process. They observed collaborative activities in all stages of the IS&S process: e.g., planning tasks, problem definition, search topic selection, query construction, and relevance assessments. The authors categorized the observed collaborative activities into document-related and human-related activities. Their study shows that collaborative activities are an important characteristic of IS&S tasks in professional settings.

Twidale et al. [31] observed collaboration between students at the computer terminals of the university library, although these systems weren't designed for collaborative usage. They identified several collaborative search strategies, such as asking for help, reusing searches. They categorized the observed activities into product-related and process-related activities. Morris [17] conducted a survey regarding web-search practices among the employees of a large IT company. She found that collaboration is largely accepted: over 97% of all users reported having used some form of collaboration when searching the web. Similar to Twidale et al., Morris identified activities regarding the search product and the search process.

Reddy and Spence [24] present a field study regarding the collaborative search behavior in multi-disciplinary teams in the context of medical care. The authors identified four triggers for collaborative IR activities: (1) complexity of information need, (2) fragmented information resources, (3) lack of domain expertise, and (4) lack of immediately accessible information.

2.2 Systems and Techniques

This section presents an analysis and classification of recent work in the area of CIS&S system support. As a basis for the classification of collaborative activities and techniques, we use the model developed by Landwich, Klas, and Hemmje [15] to describe an information searching task. Landwich et al. pursued the approach of an interactive information dialogue cycle as developed in [11]. They describe the information searching task as a dialogue between user and system consisting of six activities and assigned them to three stages (the so called interaction modes of the user):

(1) Access: Query construction and submission (Exploration),
(2) Orientation: Move within and refinement of the result set, change of focus (Focus, Navigation, Inspection),
(3) Assessment: Identification of relevant information objects (Evaluation, Store).

The dialogue cycle starts with a first query and ends after *n* cycles with a resolved or at least reduced information deficit. Figure 1 depicts this model integrated at the information searching level of the task model developed by Byström and Hansen [5].

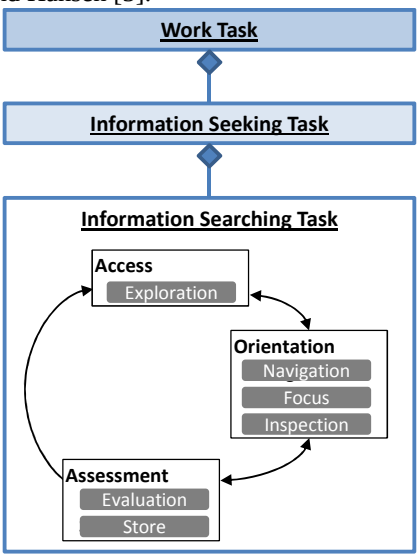


Figure 1: Task model of Byström and Hansen with integrated information searching activities as defined by Landwich et al.

Access

During *Access*, users are able to benefit from their co-workers by exchanging query definitions, modifying and executing them for their own purposes. This is realized in different ways. **Query Re-Use** refers to the activities that realize the exchange of (complete) query definitions between co-workers. The co-workers are able to perform the exchange interactively by

(1) choosing the query definition from a shared repository [25, 32],
(2) choosing the query definition from the query-history of another co-worker [20, 27], or
(3) exchanging the query definition as separate, persistently stored object [14, 31].

Group Feedback refers to a group based adoption of relevance feedback methods. This class of collaborative activities incorporates the - explicit or implicit - relevance judgments of the group members and modifies the query accordingly by adopting the weights of the query terms or expanding the query with additional query terms. This includes various approaches of query expansion techniques that typically extract search terms from highly ranked documents of previously issued queries [13]. Algorithmically extracting query terms based on relevance

judgments and suggesting them to the co-worker in a scenario with asymmetric user roles is presented in [22].

Orientation

During *Orientation*, division of labor strategies are implemented using **Result-Set Splitting,** i.e. the algorithmic division of a search result among the group members. The result set of a query is distributed algorithmically among the co-workers. These sub-sets are disjoint, i.e. the participants will only obtain documents that no other group member has seen before [8]. This splitting of search result sets can further be based on specific roles that are assigned to the participants, e.g. Prospector and Miner [22], or based on personal relevance, i.e. thematic focus and interests of the participant [19].

In addition to this, result sets can be enhanced algorithmically or manually using documents identified by other group members. **Result-Set Merging** is based on the similarity of the user profiles and the similarity of queries: Documents returned by previous queries and judged as relevant by co-workers will be added to the result set of a recently executed query [21]. **Document Recommendation** includes the interactive recommendation of documents or links. Information objects that have been identified by other participants and estimated as possibly interesting for another co-worker, are recommended and added to the work list of the co-worker [14, 27].

Assessment

During *Assessment*, collaboration addresses the diversity of knowledge across the group: **Combination of Judgments** refers to the combination of the different document assessments of the group members. The relevance of a document is determined by the opinions of multiple users through interactive voting: in [25] a scale-based approach is implemented, in [6] a traffic light based approach is used. **Re-Ranking** refers to the algorithmic re-ordering of the results. The ranks of the search results are determined not only by the relevance to the individual user, but also by the relevance to the entire group. This might be realized by using term frequencies in the stored objects or bookmarks of group members [19].

2.3 Discussion

Research at information searching level has conceptualized, implemented and evaluated collaborative activities for use at each stage of information searching process. Figure 2 depicts the classes of activities available for a team member to collaborate with the rest of the team. Previous research has focused on further improving collaborative tools by algorithmic optimization, e.g. improving similarity measures, as well as on improving human-human and human-computer interaction by facilitating communication, control and awareness mechanisms. However, these systems do not provide information on the best suited collaboration partners and the collaborative activity to be performed to increase the effectiveness and efficiency of the collaborative performance of IS&S tasks.

3 Pilot User Study

This section reports on a pilot user study that investigated the CIS&S practices of three work groups in academic and industrial research facilities. The conducted pilot study did not aim at analyzing the CIS&S processes in detail but rather at capturing the use of software technologies for realizing collaboration, information seeking and sharing in real-world settings.

Similar to the online survey conducted by Crescenzi and Capra [7], we made implicit assumptions about the components involved in the collaborative processes. Those were (1) a search component in which co-workers conduct searches to look for information, (2) a communication component in which co-workers coordinate their activities and communicate regarding the search process, and (3) an information sharing component in which collaborators share their search products.

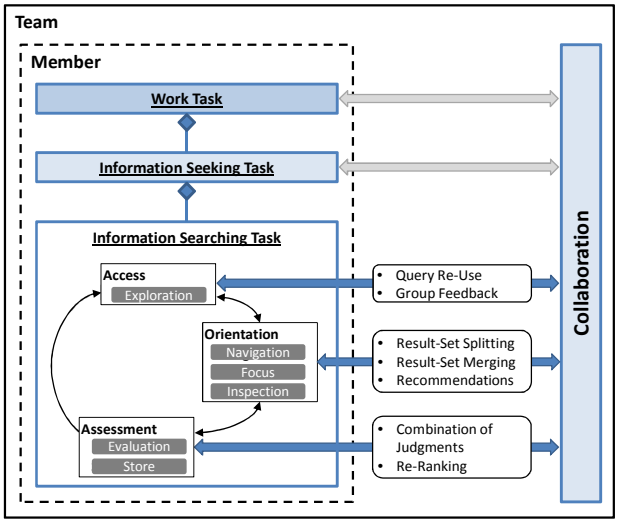


Figure 2: Summary of collaborative activities of a member within a team at information searching level

3.1 Method

Nowadays, scientists have a wide variety of software tools available to meet the daily work demands. To identify which technologies and means constitute the collaborative environment used by researchers to perform collaborative work-tasks, an online survey (implemented with Google Drive) has been conducted. We invited researchers to answer questions regarding the acquisition of required information for the collaborative performance of their work-tasks. In addition to questions regarding demographics, we were particularly interested in how they (1) collaborate with colleagues when performing a search task, (2) communicate with their colleagues and share information, and (3) how they identify colleagues who could be most helpful in regard to answering their questions and solving problems.

We asked members of two work groups of a university research facility (each in the field of life sciences) and the members of an industrial research department (in the field of information technology). The survey has been provided via e-mail distribution lists addressing (in sum) 52 people. 24 completed the entire survey, yielding a 46.2% response rate. The survey consisted of both free-text and multiple-choice questions

3.2 Results

Demographics

The age of the participants ranged from 24 to 43, with an average age of 33.25 years (s.d. = 5.14). 75% of respondents were male. Respondents were specialized in different fields of study. We clustered them into two groups: 37.5%

of respondents are specialized in the field of life sciences (including biology, molecular biology, biochemistry, and medicine), 62.5% of respondents are specialized in the field of information technology (including computer science, computer engineering, mathematics, and physics).

We wanted to estimate the degree of experience the respondents have in collaborating with colleagues. The number of articles published by multiple authors is often seen as a measure of research collaboration [4]. We decided to use this measure although not every research collaboration results in a publication and not all co-authored papers are result of collaborative research [4]. Participants were asked for the number of co-authored writings (papers of all types, grant application, project reports, etc.) they have contributed to. The given figures cover a broad range of values and thus yielding a large standard deviation (s.d.) of 23.8. The average number of co-authored writings is 18.73.

Additionally, we asked for the highest academic degree: 9% of the respondents hold a Bachelor's Degree (or equivalent), 26% of the respondents hold a Master's Degree (or equivalent), and 61% of the respondents hold a Doctor's Degree (or equivalent). The remaining 4% were Students before their first academic degree. Participants were asked to self-rate their search experience. On a five-point Likert scale, 4% rated themselves as inexperienced, 13% as moderately experienced, 67% as experienced, and 13% as expert. No respondent self-rated as "very inexperienced" user. Results show that, in addition to the high level of familiarity in search practices, the group of respondents is characterized by high degree of education, research and collaboration experience.

Search Habits and Result Management

Participants were asked about the (electronic) information sources they frequently use (figure 3) as well as tools utilized to organize and manage their search results (figure 4), i.e. scientific literature. Respondents could select electronic sources of information in a multiple choice box.

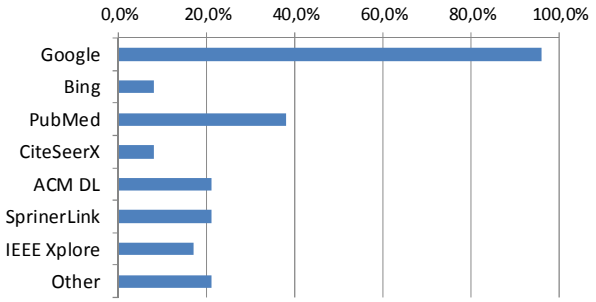


Figure 3: Electronic information sources used by respondents

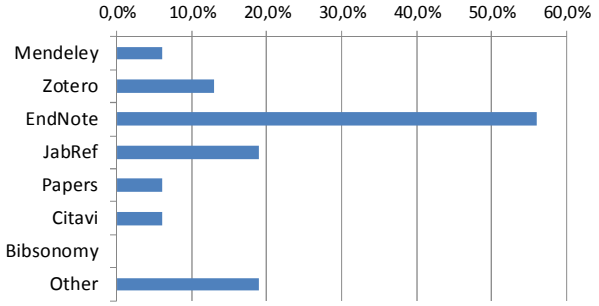


Figure 4: Literature management tools used by respondents

Additionally, they were able to extend this list by naming further tools (Other).

Figure 3 summarizes the selected sources of information. Other included Microsoft Academic Search, "Zentralblatt MATH", DBLP, and Ecosia. The results show Google as a common favorite choice, but they also provide evidence of the diversity of electronic information sources consulted during work-task performance. Figure 4 summarizes the selected literature management tools. Others are: www.citemaster.net, BibTeX, Citavi, and the Windows Explorer. In total 10 distinct tools have been named by the respondents. This too points to a broad variety of tools in operation.

Collaboration during Search

To learn more about practices of collaboration during search, we asked the participants in which stages of the search process they consult their colleagues or have been consulted. We asked about collaboration during data source selection (Q1) and query formulation (Q2a and Q2b). According to Marchionini [16], query formulation involves (a) an action mapping of the information seeker's search strategies and tactics onto the features the system interface provides, and (b) a semantic mapping of the information seeker's vocabulary onto the system's vocabulary. Therefore, we included questions on collaboration with respect to the interface and its functions (action mapping, Q2a) as well as collaboration regarding the query formulation (semantic mapping, Q2b). Furthermore, Q3 and Q4 address the result refinement as well as the result evaluation. Figure 5 depicts the results.

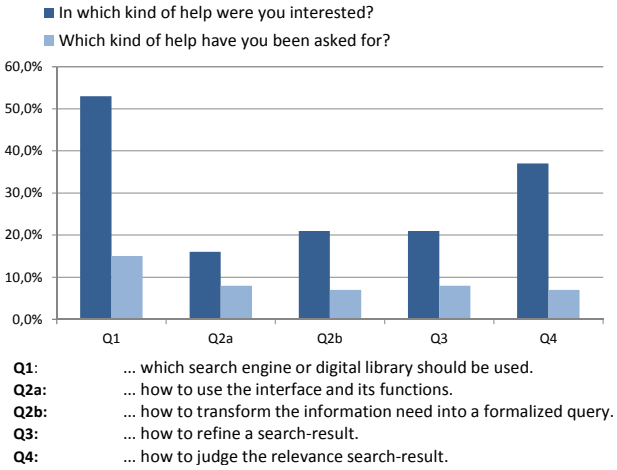


Figure 5: Percentages of respondents who collaborated during different search stages

Collaboration was found to be at its highest during the information source selection stage as well as during the assessment stage. However, collaboration can be identified in each stage of the search process.

Communication and Information Sharing Tools

We were also interested in communication (figure 6) and information sharing (figure 7) habits. As result of the growing prevalence of social networking [2, 18], we wanted to determine the degree to which such technologies are utilized for daily work routines. In a multiple choice grid, respondents could select (on a 5-point Likert scale) the frequency of technology usage in times per day. In addition, respondents were able to extend the provided list by adding tools not listed yet.

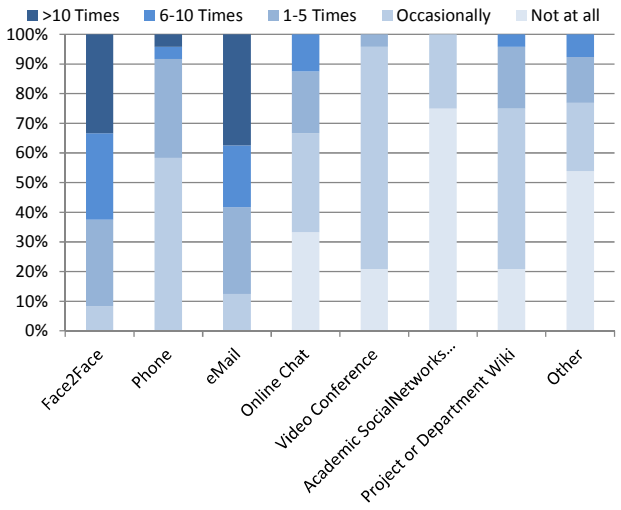


Figure 6: Frequency of use of various communication technologies among respondents

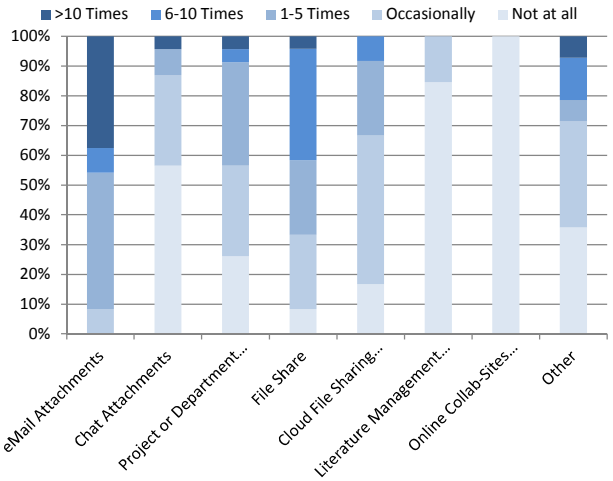


Figure 7: Frequency of use of different technologies for data and information sharing among respondents

The results in figure 6 show the importance of face-to-face communication and established remote communication technologies, i.e. phone and email. This is in line with other studies that identified communication technologies that are part of the everyday work as means to realize CIS&S [18]. It is noticeable that academic social networks seem to play only a small role in enabling communication between colleagues. Figure 7 depicts technologies for realizing data and information sharing utilized by the respondents. A predominance of e-mail attachments and the usage of file shares (local and cloud based) can be found. In contrast to this, integrated group support in literature management systems as well as online collaboration sites are rarely in use. A large list of additional tools (Others) has been named by respondents, which includes Google Drive, version control systems (namely GIT), Streamworks, and SAPmats (each specified twice). Furthermore, AeroFS, Teambox, and Adobe Connect have been added. This particularly large number of technologies used for realizing collaborative activities indicates a very heterogeneous collaboration environment where each co-worker uses his personally preferred tools.

Finding a Partner

We wanted to learn more about how respondents identify colleagues that are expected to be helpful in answering their questions. We asked two questions (1. *How did you know who might be able to help you*? 2. *How did you contact the person you asked for help?*) and provided an optional free-text field for answers. Twelve respondents provided insights on this process. Some answers show that colleagues are predominantly approached only after a first clarification using web-based search wasn't satisfying or helpful:

> *"I try to Google the issue* [...]. *If that's unsuccessful, I personally contact the colleagues who have experience with that [topic]. I explain my concrete scenario and ask them for help. Sometimes, they don't know the solution but give some new input where to look for."*

After analyzing all answers, we identified three categories of approaching colleagues when looking for help:

(1) Random contacting: Respondents ask colleagues without knowing whether they can provide the required information or not (e.g. "*Asking around in the team*", "[*asking*] *whoever is closest*").

(2) Specific contacting: Based on a personal network and an awareness of the qualifications of their team mates, colleagues are directly approached (e.g. *"I asked another biologist who is well versed with [the topic] and has demonstrated that in many fields."*, "[*I asked*] *colleagues who have a longer research experience and/or better background knowledge* [...]").

(3) Expert searching: An attempt is made to identify potentially helpful colleagues by looking at the University/research group websites.

Typical ways of contacting colleagues include e-Mail, Chat or personal contact with face-to-face communication (i.e. "went to their office").

Limitations

The demographic targeted by this survey is characterized by high academic degrees and a high experience in research collaboration. Respondents were residents in Germany. Additionally, the relatively small number of respondents might limit the significance of this study. The data we report can probably not be generalized beyond this demographic.

4 Conclusions

4.1 Design Implications

The results of our pilot study indicate that nowadays, collaboration is performed in a heterogeneous environment: It must be assumed that team members use their own personal configuration of software tools for the different information activities (i.e. communication, data and information sharing, seeking and searching, and result management). This configuration is based on personal preferences, work habits, and the special needs (e.g. thematically specialized DLs). The results indicate that a coupling of tools used in everyday work routines represents a necessity for an environment supporting CIS&S. Instead of providing communication and information sharing means integrated in one system, connecting to external tools and mediating between the co-workers seems to be a promising way. This might also allow for evaluating the mediated information to infer awareness-cues to facilitate group performance.

In line with other research [12], our results show that CIS&S often involves looking for informed people. We identified three approaches of identifying a potentially helpful colleague: expert search, random and specific contacting. The results indicate that collaboration could become more efficient, if team members could better identify co-workers who might be most helpful regarding their questions and problems. Also in line with previous studies (see section 2.1), collaboration can be observed in all stages in the search process. Our results indicate that collaboration during search preparation and result evaluation seems to be predominant. Providing group support for these aspects could most likely increase the efficiency and effectives of the CIS&S tasks.

4.2 Project Aim

The working hypothesis of our ongoing research project is that effective and efficient CIS&S requires the integration and coupling of various software tools which form the heterogeneous collaboration environment. This environment harbors knowledge in form of link-potential between the IR activities in the group and the data accessible through these tools. Besides the textual content created and managed by the tools in use, co-workers also interact with each other in many ways and build a collaboration network (CN):

- People are connected to other people as result of being a frequent communication partner or friend, or by being co-workers.
- Information objects are related to each other as result of citations or common attributes, like domain categories or keywords. Additionally, relations between documents may have been maintained manually or semi-automatically by users in form of tagging or clustering.
- Information objects are directly associated with people by the authorship relation, but also as result of reading, storing, assessing of and commenting on an information object.

We exploit this environment by collecting semantic knowledge about the individuals and their relation information: By tracking and storing this semantically linked data, i.e. information objects, user and their activities, a graph-based representation of the CN can be obtained. This representation is than analyzed and evaluated by means of semantic link analysis to generate situational support for the co-workers in each stage of the search process. Based on specific rules R_S for each stage s of the information searching process, the user support aims at increasing the group performance by (1) encouraging query diversity, (2) providing already discovered information, and (3) facilitating the alignment of result assessment.

In a first project phase, the CN shall be evaluated with the aim of identifying synergetic potential in the group (e.g. identify redundant activities or assessment conflicts). In a second phase, the CN shall be evaluated with the aim of activity suggestions to facilitate the effectiveness of the CIS&S tasks (e.g. query term suggestion, result set merging or splitting, result re-ranking). The objective of these suggestions is to increase the search performance of the group based on proposed measures for CIS&S [26].

Figure 8 depicts this concept (from bottom to top): CIS&S tasks are performed in a heterogeneous environment that connects co-workers and information objects via the utilized tools. The activities are tracked and stored in the CN. Rules extract awareness-cues during the CIS&S tasks for each stage of the search process.

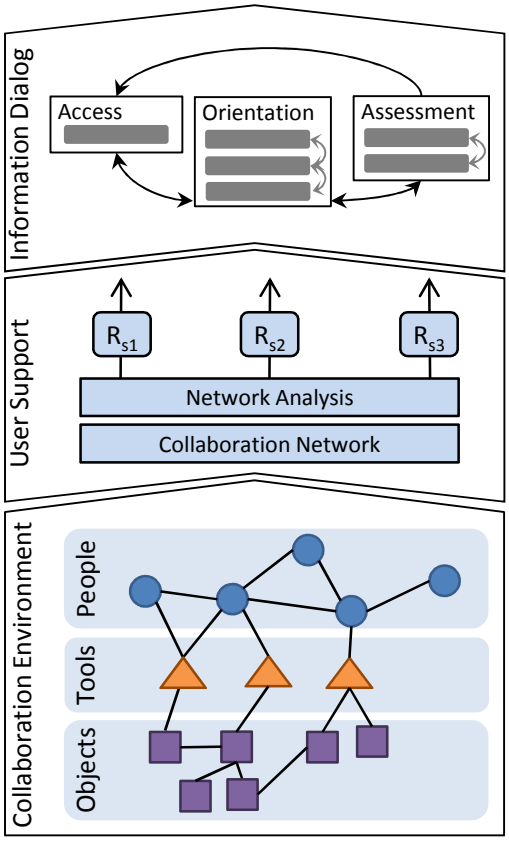


Figure 8: Layer model for supporting a group during IS&S tasks

4.3 Architecture

ezDL is the continuation of the Daffodil [14] project and implements meta-search in digital libraries and strategic support for users. The upper half of figure 9 shows the structure of the system. *ezDL* consists of a set of agents providing different aspects of the system functionality. Agents use a common communication bus for transferring messages between each other. Beginning on the left, a client connects to the MTA (Message Transfer Agent), which represents a connection point to the backend. The connection to remote search services (e.g., digital libraries) is managed by wrapper agents. A search request from the client is forwarded via the MTA to the Search Agent (SA). The SA collects all answers from all the remote DLs, merges the result lists and re-ranks them.

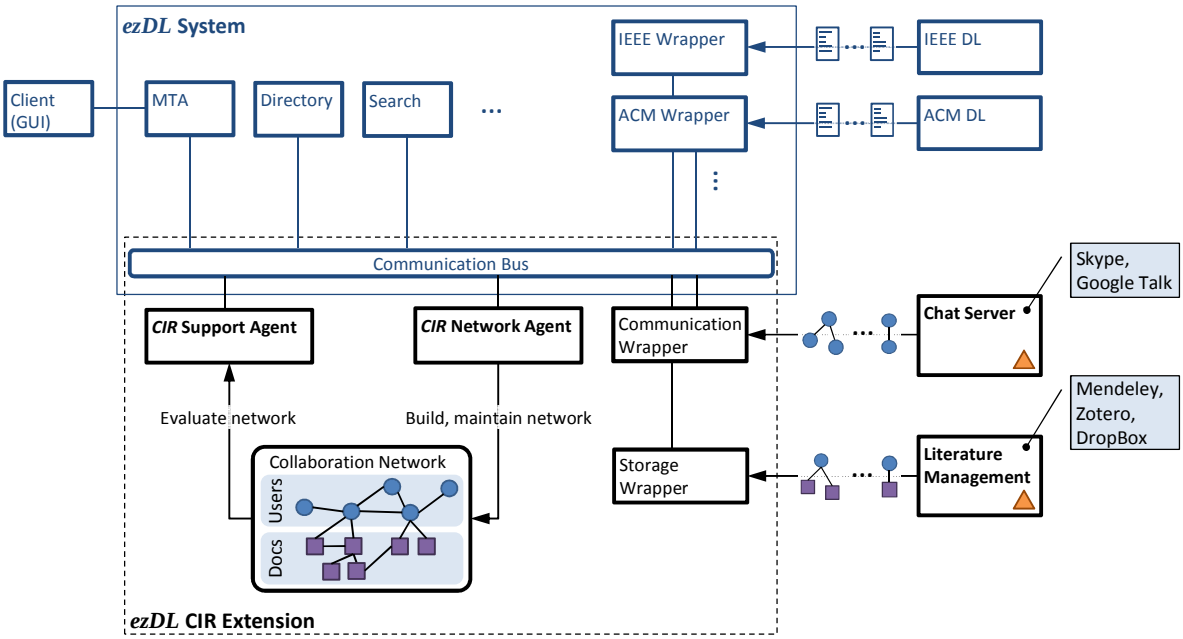


Figure 9: Structure of *ezDL* and intended extension for collaboration support at IS&S tasks

An extension of the *ezDL* system for supporting a group of collaborators is shown in lower half of figure 9. This extension addresses the identified design implications by connecting external tools utilized by the co-workers (CIR Network Agent), and by evaluating the gained information in order to generate awareness-cues (CIR Support Agent). External tools are connected by appropriate wrappers that connect, for example, to chat-servers. The objective of the CIR Network Agents is to gain information from the services about the users and its activities. These might be the communication frequency with co-workers or the stored documents. This information constitutes the collaboration network which is evaluated by the CIR Support Agent. Each time a search is performed, the client may request awareness-cues from the CIR Support Agent. The CIR Support Agent will handle support request regarding the stage *Access* (e.g. identify redundant and similar past queries), *Orientation* (e.g. bring up already discovered information), and *Assessment* (e.g. point to previous assessments and conflicting assessments of documents).

5 Summary and Outlook

In this paper, we reported on a pilot user study that investigates the CIS&S practices of three work groups in academic and industrial research facilities. The conducted pilot study captured the use of software technologies for realizing collaboration, information seeking and sharing in real-world work settings.

The results of our pilot study indicate that nowadays, collaboration is performed in a heterogeneous environment: It must be assumed that team members use their own personal configuration of software tools for the different information activities (i.e. communication, data and information sharing, seeking and searching, and result management). The results further indicate that a coupling of tools used in everyday work routines represents a necessity for the development of an environment supporting CIS&S.

We presented the design of an extension of the *ezDL* system that addresses the identified design implications by connecting external tools utilized by the co-workers, and by evaluating the gained information in order to generate awareness-cues. The aim of this is to provide group members with information on the best suited collaboration partners and the collaborative activity to be performed in order to increase the efficiency and effectiveness of IS&S tasks in such environments. The presented *ezDL* extension is currently being implemented. We plan an extensive evaluation of this system to address our initial research question: To which extend can group support in form of suggested activities improve the efficiency and effectiveness of CIS&S tasks in heterogeneous collaboration environments.

References

[1] Beckers, T. et al. 2012. ezdl: An interactive search and evaluation system. *SIGIR 2012 Workshop on Open Source Information Retrieval* (2012), 9–16.

[2] Brenner, J. 2013. Pew Internet: Social Networking (full detail). *Pew Internet & American Life Project*. (Feb. 2013).

[3] Bruce, H. et al. 2003. A comparison of the collaborative information retrieval behaviour of two design teams. *The New Review of Information Behaviour Research*. 4, 1 (2003), 139–153.

[4] Bukvova, H. 2010. Studying Research Collaboration: A Literature Review. (2010).

[5] Byström, K. and Hansen, P. 2005. Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology*. 56, 10 (2005), 1050–1061.

[6] Capra, R. et al. 2012. The ResultsSpace collaborative search environment. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (2012), 435–436.

[7] Crescenzi, A. and Capra, R. 2013. Collaboration in Search Processes and Results Sharing. *CSCW 2013*

Workshop on collaborative information seeking (2013).

[8] Foley, C. and Smeaton, A.F. 2010. Division of labour and sharing of knowledge for synchronous collaborative information retrieval. *Information processing & management*. 46, 6 (2010), 762–772.

[9] Golovchinsky, G. et al. 2009. A taxonomy of collaboration in online information seeking. *Arxiv preprint arXiv:0908.0704*. (2009).

[10] Hansen, P. and Järvelin, K. 2005. Collaborative information retrieval in an information-intensive domain. *Information Processing & Management*. 41, 5 (2005), 1101–1119.

[11] Hemmje, M. et al. 1996. A multidimensional categorization of information activities for differential design and evaluation of information systems. (1996).

[12] Hertzum, M. and Pejtersen, A.M. 2000. The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing & Management*. 36, 5 (2000), 761–778.

[13] Hust, A. 2004. Introducing query expansion methods for collaborative information retrieval. *Reading and Learning*. (2004), 252–280.

[14] Klas, C.P. et al. 2008. An experimental system for adaptive services in information retrieval. *Proceedings of the 2nd International Workshop on Adaptive Information Retrieval (AIR 2008)* (2008).

[15] Landwich, P. et al. 2009. Catching the user–logging the information retrieval dialogue. *SIGIR 2009 Workshop: Understanding the User* (Boston, USA, 2009).

[16] Marchionini, G. 1997. *Information seeking in electronic environments*. Cambridge Univ Pr.

[17] Morris, M.R. 2008. A survey of collaborative web search practices. *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (2008), 1657–1660.

[18] Morris, M.R. 2013. Collaborative search revisited. *Proceedings of the 2013 conference on Computer supported cooperative work* (New York, NY, USA, 2013), 1181–1192.

[19] Morris, M.R. et al. 2008. Enhancing collaborative web search with personalization: groupization, smart splitting, and group hit-highlighting. *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (2008), 481–484.

[20] Morris, M.R. and Horvitz, E. 2007. SearchTogether: an interface for collaborative web search. *Proceedings of the 20th annual ACM symposium on User interface software and technology* (2007), 3–12.

[21] Naderi, H. et al. 2007. An efficient collaborative information retrieval system by incorporating the user profile. *Adaptive Multimedia Retrieval: User, Context, and Feedback*. (2007), 247–257.

[22] Pickens, J. et al. 2008. Algorithmic mediation for collaborative exploratory search. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008), 315–322.

[23] Poltrock, S. et al. 2003. Information seeking and sharing in design teams. *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work* (2003), 239–247.

[24] Reddy, M.C. and Spence, P.R. 2008. Collaborative information seeking: A field study of a multidisciplinary patient care team. *Information processing & management*. 44, 1 (2008), 242–255.

[25] Romano Jr, N.C. et al. 1999. Collaborative information retrieval environment: Integration of information retrieval with group support systems. *System Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on* (1999), 10–pp.

[26] Shah, C. and González-Ibáñez, R. 2011. Evaluating the synergic effect of collaboration in information seeking. *SIGIR'11: Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval, July 24* (2011), 24–28.

[27] Shah, C. and Marchionini, G. 2010. Awareness in collaborative information seeking. *Journal of the American Society for Information Science and Technology*. 61, 10 (2010), 1970–1986.

[28] Sonnenwald, D.H. et al. 2004. Designing to support situation awareness across distances: an example from a scientific collaboratory. *Information Processing & Management*. 40, 6 (2004), 989–1011.

[29] Talja, S. 2002. Information sharing in academic communities: types and levels of collaboration in information seeking and use. *New Review of Information Behavior Research*. 3, 1 (2002), 143–159.

[30] Taylor-Powell, E. et al. 1998. Evaluating collaboratives: reaching the potential. *Madison, WI: University of Wisconsin—Cooperative Extension*. (1998).

[31] Twidale, M.B. et al. 1997. Browsing is a collaborative process. *Information Processing & Management*. 33, 6 (1997), 761–783.

[32] Walkerdine, J. and Rodden, T. 2001. Sharing searches: developing open support for collaborative searching. *Human-computer Interaction: INTERACT'01: IFIP TC. 13 International Conference on Human-Comupter Interaction, 9th-13th July 2001, Tokyo, Japan* (2001), 140.

Strategische Suchunterstützung auf Makro- und Mikroebene

Andreas Tacke, Sascha Kriewel
Universität Duisburg-Essen, Deutschland
{tacke | kriewel }@is.inf.uni-due.de

Abstract

Immer mehr Verbraucher greifen bei der Beschaffung medizinischer Informationen auf das Internet zurück. Dem durchaus komplexen Informationsbedürfnis stehen dabei zumeist sowohl mangelnde Kenntnis des Themengebiets als auch Probleme bei der Formulierung geeigneter Suchstrategien gegenüber. Des Weiteren können unvollständige und ungenaue Informationen medizinische Entscheidungen negativ beeinflussen. Bereits existierende Systeme adressieren diese Probleme nur teilweise. In dieser Arbeit wird die prototypische Umsetzung eines umfassenden Unterstützungskonzepts präsentiert, das Benutzern auf strategischer und taktischer Ebene angemessene Hilfestellungen zur Bearbeitung komplexer Suchaufgaben zum Thema medizinische Verbraucherinformationen anbietet und diese auf Software-Ebene integriert. Anschließend werden die Ergebnisse der Evaluation diskutiert, die zur Überprüfung der Tauglichkeit des Konzepts durchgeführt wurde.

1 Einführung

Verschiedene Studien haben gezeigt, dass immer mehr Verbraucher das Internet zur Beschaffung von Gesundheitsinformationen nutzen (vgl. [Fox, 2011]). Dabei fehlt den meisten Benutzern sowohl das nötige prozedurale Suchwissen als auch das Domänenwissen zur Bearbeitung komplexer Suchaufgaben, beispielsweise der Recherche nach Komplikationen und Behandlungsmethoden einer seltenen Krankheit. Dies wurde unter anderem von Studien bestätigt, die im Rahmen des KHRESMOI-Projekts durchgeführt wurden (vgl. [Boyer *et al.*, 2012]). Demnach stellen Benutzer trotz des oft sehr komplexen Informationsbedürfnisses größtenteils sehr allgemeine Anfragen an das Suchsystem. Als Folge wird häufig auf ineffiziente oder ineffektive Suchtaktiken und -strategien zurückgegriffen, was zu Frustration oder unvollständigen beziehungsweise fehlerhaften Suchergebnissen führen kann.

Ein weiteres wichtiges Problem bei der Suche nach medizinischen Verbraucherinformationen im Internet ist die Unvollständigkeit und Unzuverlässigkeit der von herkömmlichen Suchmaschinen zurückgelieferten Informationen. Die hohe Popularität beispielsweise von Selbsthilfeforen führt dazu, dass diese im Ranking oft sehr weit oben angezeigt werden. Deren Konsultation kann im schlimmsten Fall zu gefährlichen Selbstdiagnosen führen, die negative Auswirkungen auf den Gesundheitszustand eines Patienten haben können (vgl. [White and Horvitz, 2009]). Portalseiten können diese Probleme zwar abfedern, bieten dem Benutzer jedoch in den meisten Fällen keinerlei Unterstützung bei der Bearbeitung komplexer Suchaufgaben an.

2 Forschungsstand

Es existieren bereits mehrere Systeme, die Benutzer bei der Informationssuche unterstützen. Da diese jedoch entweder nur strategische oder taktische Unterstützung anbieten, soll hier nach Makro- und Mikro-Ebene unterschieden werden.

Marcia Bates [1990] stellte eine Hierarchie mit vier Ebenen vor, nach denen Suchaktivitäten von Benutzern klassifiziert werden können. Aktivitäten auf den höheren Ebenen setzen sich dabei jeweils aus mehreren Aktivitäten der darunterliegenden Ebenen zusammen.

- Ein **Move** bezeichnet die kleinste, atomare Suchaktivität eines Benutzers, beispielsweise die Eingabe eines Suchbegriffs. Moves können aber auch einfache Gedanken repräsentieren.
- Eine **Taktik** umfasst mehrere zusammenhängende Moves, mit dem Ziel, Problemsituationen während der Suche aufzulösen.
- Ein **Strategem** kann aus mehreren Taktiken und Moves bestehen und kann beispielsweise die effiziente Durchsuchung von Zitationsnetzen oder Verzeichnissen von Journalen umfassen.
- Eine **Strategie** steht für einen umfassenden Plan zur Durchführung einer Suche unter Verwendung von Strategemen, Taktiken und Moves.

Unterstützung auf der Makro-Ebene leisten also solche Systeme, die dem Benutzer Hilfe bei der Formulierung einer *Strategie* anbieten, während Systeme, die Unterstützung auf der Mikro-Ebene leisten auf der Ebene der *Taktiken* anzusiedeln sind.

2.1 Makro-Ebene

Bhavnani et al. [2003] entwickelten mit den *Strategy Hubs* ein Domänenportal für Mediziner. Dabei wurde der Suchprozess in zwei Phasen eingeteilt. Zunächst spezifizieren Benutzer mit Hilfe einer Klassifikation ihr Informationsbedürfnis. Anschließend werden zum ausgewählten Thema aufgabenbegleitend explizite Suchprozeduren mit geordneten Unterzielen und Beispielquellen angezeigt.

Argelagos und Pifarré [2011] erweiterten die webbasierte Lernumgebug WebQuest[1] um eingebettete Hilfestellungen, die die Problemlösungsfähigkeiten von Schülern im Rahmen der Bearbeitung authentischer Aufgaben

[1] http://webquest.org/

verbessern sollten. Dabei kamen Techniken des software-realisierten Scaffolding zum Einsatz (vgl. [Guzdial, 1994]).

2.2 Mikro-Ebene

Jansen [2004] entwickelte ein System zur automatisierten Unterstützung, das das Informationsbedürfnis von Benutzern aus einer Sequenz von Aktionen extrapoliert und auf dieser Basis situationsgerechte Vorschläge, wie zum Beispiel das erkennen und Verbessern von syntaktischen Fehlern in der Anfrage oder das Vorschlagen von Synonymen, anbietet. Die Hilfestellungen wurden dabei nur auf Anfrage des Benutzers angezeigt.

Kriewel [2010] erarbeitete ein automatisiertes Vorschlagssystem für den EZDL-Vorgänger DAFFODIL[2] (vgl. [Fuhr *et al.*, 2002]), das ähnlich wie bei Jansen [2004] die Situation des Benutzers analysiert, um automatisiert taktische Vorschläge zu generieren, die auf Anfrage abgerufen werden können.

3 EZDL und KHRESMOI

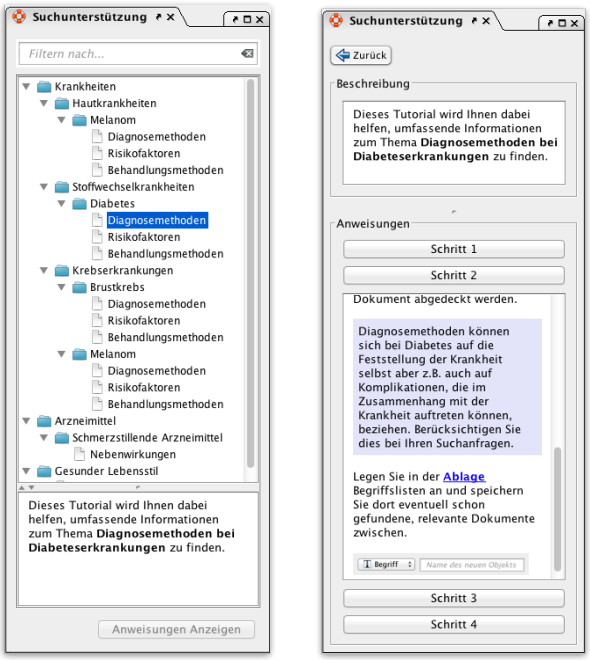
Bei KHRESMOI handelt es sich um ein von der Europäischen Union gefördertes Projekt mit dem Ziel sowohl professionellen Benutzern als auch Verbrauchern eine multilinguale und multimodale Suche nach medizinischen Informationen anbieten zu können (vgl. [Hanbury *et al.*, 2011]). Durch eine Reihe von verschiedenen spezialisierten Suchschnittstellen soll den Zielgruppen der Zugriff auf die für sie relevanten Informationen erleichtert werden. Eine dieser Schnittstellen wird beispielsweise aufbauend auf dem Projekt EZDL[3] von der Arbeitsgruppe Informationssysteme[4] an der Universität Duisburg-Essen entwickelt und gepflegt. EZDL wurde ebenfalls von der Arbeitsgruppe entworfen und bietet ein objektorientiertes Java-Framework zur Erstellung interaktiver Retrievalsysteme für die Suche in heterogenen digitalen Bibliotheken (vgl. [Beckers *et al.*, 2012]). Um die im KHRESMOI-Projekt festgelegten Anwendungsfälle abzudecken, wurde ein auf der EZDL-Referenzimplementierung basierender Client entwickelt, der unter anderem Abhilfe für die eingangs geschilderten Problemen bei der Suche nach medizinischen Verbraucherinformationen verschaffen sollte. Dieser Client diente als Grundlage für die Umsetzung des Unterstützungskonzepts, das im Rahmen dieser Arbeit erstellt wurde.

4 Umsetzung des Unterstützungskonzepts

Für das Unterstützungskonzept wurden zwei Werkzeuge entworfen, die jeweils die Makro- und die Mikro-Ebene abdecken sollten. Als Grundlage diente dabei der für KHRESMOI angepasste EZDL-Client. Die Unterstützung auf der Makro-Ebene sollte dabei von einem *Suchunterstützungswerkzeug* übernommen werden, das dem Benutzer aufgabenspezifische strategische Hilfestellungen anbietet, die Unterstützung auf der Mikro-Ebene von einem *Suchvorschlagswerkzeug*, das situationsabhängige taktische Vorschläge bereitstellt.

4.1 Suchunterstützungswerkzeug

Als Vorbild für das Suchunterstützungswerkzeug dienten die Strategy Hubs von Bhavnani et al. [2003]. Dem Benutzer sollte auch hier die Möglichkeit gegeben werden, zunächst sein Informationsbedürfnis zu spezifizieren. Dazu wurde innerhalb des Werkzeugs eine eigene Ansicht implementiert, die mit Hilfe einer Klassifikation die Auswahl eines bestimmten Themas erlauben soll (siehe Abbildung 1(a)). Hat der Benutzer ein Thema ausgewählt, wechselt



(a) Auswahlansicht (b) Bearbeitungsansicht

Abbildung 1: Suchunterstützungswerkzeug

er in die Bearbeitungsansicht. Hier soll mit Methoden des software-realisierten Scaffolding prozedurales Suchwissen durch gezielte strategische Hilfestellungen, beispielsweise durch Präsentieren konkreter Suchstrategien, vermittelt werden. Eine Akkordion-Ansicht erlaubt es, die Hilfestellungen in diskrete Schritte zu unterteilen, um so das Suchproblem in Teilprobleme zu zerlegen (siehe Abbildung 1(b)).

4.2 Suchvorschlagswerkzeug

Beim Suchvorschlagswerkzeug handelt es sich im wesentlichen um eine Neuimplementierung des in Abschnitt 2.2 vorgestellten Vorschlagssystems von Kriewel [2010]. Um dem Benutzer situationsabhängige Vorschläge präsentieren zu können, wird nach einer Suchanfrage die aktuelle Suchsituation anhand bestimmter Parameter im Backend von EZDL ausgewertet. Anschließend werden mittels *Case-based Reasoning* (vgl. [Marir and Watson, 1994]) die angemessensten Vorschläge aus einer Datenbank abgerufen und absteigend nach Relevanz geordnet an den Client zurückgeliefert, wo sie vom Vorschlagswerkzeug angezeigt werden (siehe Abbildung 2). Einige der Vorschläge sind ausführbar und können über das Werkzeug direkt auf die Suchsituation angewendet werden. Des Weiteren haben Benutzer die Möglichkeit, Vorschläge zu bewerten und so das zukünftige Ranking zu verbessern.

5 Evaluierung

Zur Untersuchung der Tauglichkeit des entwickelten Prototypen wurde im Rahmen einer Abschlussarbeit [Tacke, 2013] eine Benutzerstudie durchgeführt, durch die folgende Forschungsfragen geklärt werden sollten:

[2] **D**istributed **A**gents **f**or User-**F**riendly Access **o**f **Di**gital **L**ibraries

[3] **e**asy access to **D**igital **L**ibraries

[4] http://www.is.inf.uni-due.de/

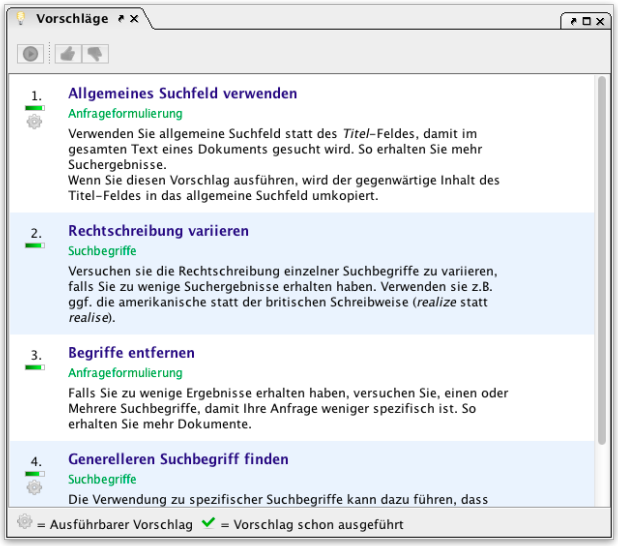


Abbildung 2: Suchvorschlagswerkzeug

1. Können Benutzer durch ein Suchsystem bei der Bildung einer umfassenden Suchstrategie zu einem Informationsbedürfnis unterstützt werden?
2. Kann durch die Integration von strategischer (Makro-Ebene) und taktischer Unterstützung (Mikro-Ebene) auf Software-Ebene erreicht werden, dass Benutzer die Funktionen eines Suchsystems zielführender einsetzen und dadurch erfolgreicher suchen?

Zur Beantwortung der ersten Frage wurde im Anschluss an jedes Benutzerexperiment jeweils eine Befragung durchgeführt, bei der die Probanden subjektiv bewerten sollten, wie gut sie sich bei der Bearbeitung der gestellten Suchaufgaben unterstützt fühlten. Bei der zweiten Frage sollte zunächst überprüft werden, ob Probanden mit integrierter Unterstützung die weiterführenden Funktionen des Systems und die angebotenen taktischen Hilfestellungen häufiger in Anspruch nehmen als Probanden, denen nur taktische Unterstützung angeboten wird. Im Anschluss sollte festgestellt werden, inwieweit sich dies unter Betrachtung der *Task Completion Rate* (TCR) auf den tatsächlichen Sucherfolg der Teilnehmer auswirkt.

An der Studie nahmen insgesamt 22 Personen teil (17 männlich und 5 weiblich), wobei das Durchschnittsalter 31 Jahre (Standardabweichung $8,0$) betrug. Die einzige obligatorische Voraussetzung bei der Auswahl der Probanden war, dass es sich um medizinische Laien handeln sollte. Vorgezogen wurden außerdem Teilnehmer, die keine Expertenkenntnisse bei der Informationssuche besaßen, die diese die Zielgruppe der Unterstützungsmaßnahmen sind.

5.1 Versuchsaufbau

Für die Experimente wurden die Teilnehmer zunächst in eine Experimental- und eine Kontrollgruppe aufgeteilt. Beide Gruppen arbeiteten jeweils mit einer angepassten Version der KHRESMOI-Variante von EZDL, wobei die Experimentalgruppe sowohl das Suchunterstützungs- als auch das Suchvorschlagswerkzeug erhielt, während das Suchunterstützungswerkzeug bei der Kontrollgruppe deaktiviert war.

Für das Experiment wurden zwei komplexe Suchaufgaben zum Thema medizinische Verbraucherinformationen erarbeitet, die von den Probanden zu bearbeiten waren (siehe Tabelle 1). Die Themen wurden bei der Erstellung der Aufgaben möglichst unterschiedlich gewählt, um eventuelle Vorkenntnisse einzelner Teilnehmer nicht so stark ins Gewicht fallen zu lassen. Darüber hinaus wurden die Aufgaben abwechselnd rotiert, um Lerneffekte abzumildern, die sich möglicherweise nach Bearbeitung der ersten Aufgabe einstellen könnten.

Als Kollektion wurde der Index von *Khresmoi Search for Everyone* über einen EZDL-*Wrapper* eingebunden. Dabei handelt es sich um eine vertikale Suchmaschine, die im Rahmen des KHRESMOI-Projekts entwickelt und gepflegt wird und Verbrauchern Zugang zu geprüften Gesundheitsinformationen verschaffen soll.

5.2 Ergebnisse

Neben jeweils einem Fragebogen vor und nach dem Experiment wurden außerdem mit Hilfe des EZDL-Logging-Frameworks die Aktionen der Benutzer im Umgang mit dem System bei der Bearbeitung der Aufgaben aufgezeichnet. Zu den erhobenen Daten zählten:

- Die Bearbeitungszeit in Minuten
- Die Zahl der gestellten Anfragen
- Häufigkeit bei der Verwendung der erweiterten Suchfelder
- Anzahl der betrachteten Dokumente
- Anzahl der gespeicherten Dokumente
- Die Häufigkeiten bei der Verwendung von Systemfunktionen wie Filter, Extraktion, oder Klassifikation
- Wie oft Suchvorschläge angefordert und ausgeführt wurden

Um die beiden Gruppen zu vergleichen, wurden die Parameter unter Verwendung des One-way-ANOVA-Tests (Konfidenzniveau 95%) gegenübergestellt, wobei die Signifikanz aufgrund der Stichprobengröße mit Hilfe des Welch-Tests überprüft wurde.

Bei der Bearbeitungszeit für beide Aufgaben ergab sich mit $36,36$ Minuten bei der Experimentalgruppe (Standardabweichung 6.61) und $35,73$ Minuten bei der Kontrollgruppe (Standardabweichung $6,81$) kein nennenswerter Unterschied. Die Auswertung der von den Benutzern durchgeführten Aktionen ließ signifikante Unterschiede erkennen (siehe Tabelle 2). Besonders ins Auge fiel dabei, dass Probanden der Experimentalgruppe mit $6,91$ wesentlich mehr fortgeschrittene Aktionen ausführten als Probanden der Kontrollgruppe ($3,18$, $p = 0,01 < 0,05$).

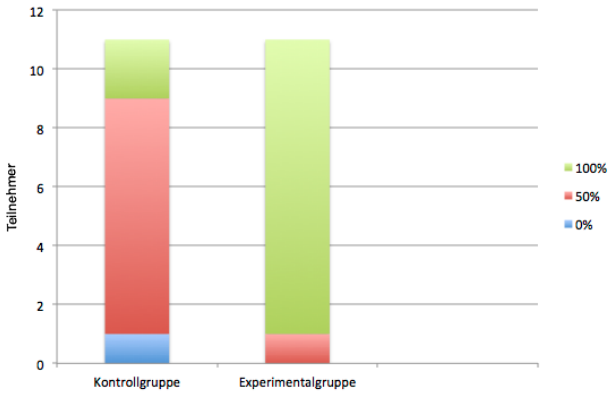


Abbildung 3: Task Completion Rates (kombiniert)

Unter Betrachtung der Task Completion Rate ließ sich feststellen, dass Teilnehmer der Experimentalgruppe wesentlich mehr Aufgaben erfolgreich im Sinne der Aufgabenstellung bearbeiteten. Die TCR in Experimentalgruppe lag bei $0,95$, während die Kontrollgruppe auf einen Wert von $0,54$ kam (siehe Abbildung 3). Bei der Auswertung des Abschlussfragebogens im Anschluss an das Experiment ergab sich für die Kontrollgruppe, dass von 11 der Befragten 3 die angebotenen Hilfestellungen als sehr hilfreich bei der Bearbeitung der Aufgabe bewerteten. Weitere 6 Probanden fanden die Unterstützungsfunktion noch hilfreich, 2 bewerteten sie neutral. Bei der Frage, ob die Hilfestellungen den Teilnehmern auch bei zukünftigen Suchen helfen würden, stimmten 3 der 11 Teilnehmer voll zu, 6 stimmten zu und 2 stimmten weniger zu. Lediglich einer der Probanden stimmte nicht zu.

5.3 Fazit

Die erste Forschungsfrage zielte darauf ab, ob Benutzer vom System bei der Bildung einer umfassenden Suchstrategie unterstützt werden können. Die Ergebnisse der Befragung lassen darauf schließen, dass die strategischen Hilfestellungen von den Probanden gut aufgenommen wurden. Ebenfalls vielversprechend war die hohe Zustimmung bei der Frage, ob die vermittelten Strategien auch in Zukunft bei der Suche nach Informationen hilfreich sein könnten. Bedingt durch den Versuchsaufbau wurden allerdings nur 11 der Probanden befragt. Daher sollte zu dieser Fragestellung ein eigenes Experiment mit mehr Teilnehmern durchgeführt werden.

Bei der zweiten Forschungsfrage sollte geklärt werden, inwieweit die Integration von strategischer und taktischer Unterstützung Benutzern dabei helfen kann, die Funktionen eines Suchsystems zielführender und effizienter einzusetzen und dadurch erfolgreicher zu suchen. Die Auswertung der Ergebnisse bestätigte die Annahme, dass Benutzer mit integrierter Unterstützung signifikant häufiger die Suchvorschläge und die weiterführenden Funktionen wie des Systems in Anspruch nehmen.

Unter Betrachtung der Task Completion Rate ließ sich außerdem feststellen, dass Benutzer, denen strategische und taktische Hilfestellungen geboten werden, wesentlich mehr Aufgaben erfolgreich bearbeiteten. Kritisch zu beurteilen ist dabei, dass nicht in die Bewertung mit eingeflossen ist, zu welchem Anteil ein Benutzer eine Aufgabe bearbeitet hat. Diese methodischen Schwächen könnten ebenfalls in einer weiteren Benutzerstudie berücksichtigt werden.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde eine prototypische Umsetzung für ein umfassendes Unterstützungskonzept auf Basis von EZDL vorgestellt. Das System sollte im Rahmen des KHRESMOI-Projekts Verbrauchern auf der Suche nach Gesundheitsinformationen bei der Bearbeitung komplexer Suchaufgaben angemessene strategische und taktische Hilfestellungen anbieten. Zu diesem Zweck wurden zwei spezielle Werkzeuge entwickelt, die im Anschluss durch eine Benutzerstudie auf ihre Tauglichkeit hin evaluiert wurden.

Die Ergebnisse dieser Studie ließen darauf schließen, dass die Integration von strategischer und taktischer Unterstützung auf Software-Ebene besonders unerfahrenen Benutzern, denen außerdem das nötige Domänenwissen zur Bearbeitung einer bestimmten Suchaufgabe fehlt, bei der Bildung einer umfassenden Suchstrategie unterstützen und den Erfolg bei der Suche steigern kann.

Zwar lag der Fokus dieser Arbeit auf der Suche nach medizinischen Verbraucherinformationen, die entwickelten Werkzeuge lassen sich jedoch ohne weiteres auf die Bedürfnisse von Ärzten anpassen. Einige der der von EZDL gebotenen Funktionen, die beispielsweise für die Suche nach Fachliteratur relevant sind und für den oben beschriebenen Anwendungsfall deaktiviert wurden, könnten so sinnvoll eingebunden werden.

7 Danksagung

Diese Arbeit wurde zum Teil durch das European Union Seventh Framework Programme (FP7/2007-2013), grant agreement 257528 (KHRESMOI[5]) gefördert.

Literatur

[Argelagós and Pifarré, 2011] Esther Argelagós and Manoli Pifarré. Improving information problem solving skills in secondary education through embedded instruction. *Computers in Human Behavior*, 28:515–526, 2011.

[5] http://khresmoi.eu/

Aufgabe	Beschreibung
Diabetes Typ I	**Szenario**: Sie oder eine Ihnen bekannte Person wurden vor Kurzem mit Diabetes Typ 1 diagnostiziert. Der behandelnde Arzt hat daraufhin davor gewarnt, dass im Zusammenhang mit Diabetes häufig Komplikationen auftreten. **Suchaufgabe:** Sie möchten zunächst herausfinden, welche Arten von Komplikationen bei Diabetikern (mit Ihrer Art der Diabetes) überhaupt auftreten können. Außerdem wollen Sie wissen, welche Untersuchungen oder Tests ein Arzt für die Diagnose dieser Komplikationen durchführen müsste. **Ziel:** Sie sind zufrieden, wenn Sie vier wichtige Komplikationen identifiziert haben und zu mindestens einer die nötigen Untersuchungen kennen.
Brustkrebs Stadium I	**Szenario:** Sie wurden vor Kurzem mit Brustkrebs im ersten Stadium diagnostiziert. **Suchaufgabe:** Sie möchten zunächst herausfinden, welche Stadien es bei Brustkrebs überhaupt gibt. Anschließend möchten Sie sich über mögliche Behandlungsmethoden informieren. **Ziel:** Sie sind zufrieden, wenn Sie die Stadien identifiziert haben und mindestens vier Behandlungsmethoden kennen.

Tabelle 1: Aufgabenbeschreibungen für die Benutzerstudie

Aktion	Experimentalgr. $\varnothing$ (σ)	Kontrollgr. $\varnothing$ (σ)	Welch-Test p
Grundlegende Aktionen			
Anfrage ausführen	$13,64$ $(3,44)$	$13,54$ $(6,73)$	$0,969$
Dokument betrachten	$17,45$ $(4,95)$	$19,45$ $(3,42)$	$0,283$
Dokument speichern	$8,45$ $(1,37)$	$8,18$ $(2,27)$	$0,737$
Fortgeschrittene Aktionen			
Filter verwenden	$4,18$ $(2,13)$	$1,64$ $(2,73)$	$0,025$
Klassifikation verwenden	$1,82$ $(0,87)$	$1,18$ $(1,08)$	$0,144$
Begriffe extrahieren	$0,91$ $(1,45)$	$0,36$ $(0,67)$	$0,270$
Suchvorschläge			
Vorschläge anfordern	$2,36$ $(0,67)$	$0,64$ $(1,12)$	$0,000$
Vorschläge ausführen	$1,91$ (1.04)	$0,36$ $(0,67)$	$0,001$

Tabelle 2: Aktionen während der Aufgaben

[Bates, 1990] Marcia J. Bates. Where should the person stop and the information search interface start? *Information Processing Management*, 26(5):575–591, 1990.

[Beckers *et al.*, 2012] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, and Sascha Kriewel. ezdl: An interactive search and evaluation system. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, 2012.

[Bhavnani *et al.*, 2003] Suresh K. Bhavnani, Christopher K. Bichakjian, Timothy M. Johnson, Roderick J. Little, Frederick A. Peck, Jennifer L. Schwartz, and Victor J. Strecher. Strategy hubs : Next-generation domain portals with search procedures. In *Proceedings of the CHI 2003 Conference*, volume 5, pages 393–400, 2003.

[Boyer *et al.*, 2012] Celia Boyer, Manfred Gschwandtner, Allan Hanbury, Marlene Kritz, Natalia Pletneva, Matthias Samwald, and Alejandro Vargas. Use case definition including concrete data requirements. Deliverable D8.2, Khresmoi, 2012.

[Fox, 2011] Susannah Fox. Health topics. Technical report, Pew Research Center, 2011.

[Fuhr *et al.*, 2002] Norbert Fuhr, Claus-Peter Klas, André Schäfer, and Peter Mutschke. Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In *Proceedings of the ECDL 2002 Conference*, pages 597–612. Springer-Verlag, 2002.

[Guzdial, 1994] Mark Guzdial. Software-realized scaffolding to facilitate programming for science learning. *Interactive Learning Environments*, 4(1):1–44, 1994.

[Hanbury *et al.*, 2011] Allan Hanbury, Celia Boyer, Manfred Gschwandtner, and Henning Müller. Khresmoi: Towards a multi-lingual search and access system for biomedical information. In *Med-e-Tel*, pages 412–416, 2011.

[Jansen, 2004] Bernard J. Jansen. Seeking and implementing automated assistance during the search process. *Information Processing Management*, 41:909–928, 2004.

[Kriewel, 2010] Sascha Kriewel. *Unterstützung beim Finden und Durchführen von Suchstrategien in Digitalen Bibliotheken*. PhD thesis, Universität Duisburg-Essen, 2010.

[Marir and Watson, 1994] Farhi Marir and Ian Watson. Case-based reasoning: A review. *The Knowledge Engineering Review*, 9(4):327–354, July 1994.

[Tacke, 2013] Andreas Tacke. Integration von strategischer Suchunterstützung auf Makro- und Mikro-Ebene. Diplomarbeit, Universität Duisburg-Essen, Duisburg, Deutschland, 2013.

[White and Horvitz, 2009] Ryen W. White and Eric Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, 27(4), 2009.

Interactive Query Expansion in Meta Search Engines

Wolfgang Köhler and **Daniel Backhausen** and **Claus-Peter Klas** and **Matthias Hemmje**
Distance University in Hagen, Germany
DE-58084, Hagen, Germany
Wolfgang.Koehler1, Daniel.Backhausen, Claus-Peter.Klas, Matthias.Hemmje@FernUni-Hagen.de

Abstract

For meta search systems like digital library solutions, techniques like recommendation and especially query expansion are complex to realize because often the content of the information objects is not present or directly accessible. This approach takes new roads by integrating suggestion terms from two distinct sources in an interactive hybrid recommendation system. The terms are acquired through lexical-syntactical analysis using WordNet, as well as through association rule mining among the query logs.

1 Introduction

Recent research in query expansion techniques for information retrieval systems has seldom taken the particular situation of meta search engines into account. There are several reasons for this: Meta search engines usually do not provide the content of information objects that is often necessary for expanding user queries. The well-known recommender systems based on similarity measures do not work here. Because of the lack of data, the cold start problem is even more aggravated. An additional challenge is the long time of processing a query, since the engine needs to wait on other search engines. Therefore, it is advisable to involve the user more in the process of query formulation rather than relying on blind feedback techniques. Our related works section illustrates the problems of recommendation in meta search engines and shows the approaches that we developed further to tackle those issues. The implementation section explains the concrete steps we try to take in implementing a hybrid interactive recommendation system based on lexical-syntactical analysis and association rule mining. We tested our system in EzDL, a digital library meta search engine. After discussing the user evaluation of the system, we conclude by lining out the possibilities for future research.

2 Related Work

There are several techniques to find suitable terms for query expansion. One of them is the so called automatic query expansion. Well-known approaches have been developed by Mita et. al in [Mitra *et al.*, 1998], Xu et. al. [Xu and Croft, 1996], and Qiu and Frei in [Qiu and Frei, 1993]. They use either a refined form of blind feedback, local context analysis based on a concept database, or a similarity thesaurus to increase the effectiveness of this procedure significantly.

However, none of the presented approaches is useful for query expansion in meta search engines. They all require some kind of information object content, which is typically not present in meta search engines. Instead, other additional content needs to be provided, for example a lexicon or a word net. Yet another way would be the use of user interaction data, which is stored in some kind of activity log.

In addition to the first mentioned way to expand queries, different attempts have been made to automatically expand queries using WordNet [Fellbaum, 1998] by exploiting lexical-semantic relations [Voorhees, 1994]. Even though these experiments did not show a significant improvement in query performance by just linking WordNet to an Information Retrieval system, this effect can indeed by reached with a more refined approach. [Kim *et al.*, 2004] demonstrated that performance can be enhanced by disambiguating the query terms first before expanding them. The authors suggest to disambiguate query terms by determining their root sense according to their context. Obviously, in a query for a meta search engine, there is not much context to draw from, so the usefulness of this approach would be limited in our scenario.

A more promising approach has been made by Liu et al. All synonyms that have a similar meaning are saved in a synset in WordNet. The correct meaning of a given term can be found by determining the most appropriate synset. According to [Liu *et al.*, 2004], this can be done effectively by using the information in the synset definition. For two words that are part of a nominal phrase, a check is made whether their synsets contain any information that helps to determine the correct meaning in this context. The synset definition might provide terms useful for query expansion. The approach has lead to a precision improvement of 15.6 to 21.5 % on the TREC 9, 10 and 12 datasets.

Recently, some more effort has been made to analyze query logs in order to identify good query expansion terms. As proposed by [Cui *et al.*, 2002], this can be even more effective than local context analysis. Here the authors use query logs as a basis for query expansion. However, the disadvantage for meta search systems regarding the need of content remains. One idea to address this issue is the approach proposed by Fonseca et al. in [Fonseca *et al.*, 2005]. Here the query logs are mined for association rules by inspecting which queries frequently co-occur in a user session. The brilliant idea coming from this approach is to equate itemsets and transaction sets known from association rule mining with sets of queries and sets of user sessions. For a given query, a relation graph is built, starting from the user query and showing the transitive associations between the queries. Circuits in the graph are called concepts. The concepts are candidates for query expansion. Compared to other approaches presented in this paper, this

one involves the interaction of the user. That means that a suitable concept needs to be explicity chosen by the user among the given options. Furthermore, the user can also specify the kind of relation between query and chosen concept, leading to a different Boolean connection between query and concept. Synonyms and specializations are connected to the query via the `OR` operator. Generalizations and associations are connected via the `AND` operator. According to Fonseca et al. this approach leads to an increase in precision of 53 % on average when tested with a web search engine.

Within our research work we implemented the approach presented by Fonseca et al. in [Fonseca *et al.*, 2005] and evaluated its usefulness for meta search engines. To use this approach in the context of meta search systems, different modifications have to be made, which we will present in this paper.

Before elaborating on this, however, we need to discuss how an effective interactive recommender system for query expansion should look like. Harman shows through experiments in the Cranfield 1400 test collection in [Harman, 1988] that the effectiveness of the system significantly increases, if it draws on two distinct sources for the expansion terms. This is what we want to call the two-window approach. A third source did not bring as much improvement, possibly because the terms presented are already included in the first two sources.

3 Implementation

This leads us to our concept of building a hybrid recommendation system that is based on 1) a lexical-semantic analysis and 2) query log analysis using a mixture of the approaches proposed by [Liu *et al.*, 2004] and [Fonseca *et al.*, 2005]. In presenting the suggested query terms, we want to follow the two-window approach by [Harman, 1988].

For our research prototype we used EzDL[1], a meta search system for digital libraries. EzDL has already proven its usefulness in different research activities such as the implementation and evaluations of interactive information retrieval scenarios [Klas *et al.*, 2004; Klas and Hemmje, 2009] at the University of Duisburg-Essen and the DistanceUniversity of Hagen. As described by Beckers et al. in [Beckers *et al.*, 2012], EzDL is a service-oriented system that can be used as a meta-search system for heterogeneous sources or digital libraries. In addition it provides an evaluation framework with already existing tools and rich user logs. EzDL consists of a backend agent based system and a rich user client, giving access to the services.

To address the cold start problem present in collaborative recommender systems, we initially use WordNet. It is used here as a database for recommendations based on lexical-semantical relationships. First, after performing stemming on the query term, the term is looked up in WordNet for all parts of speeches. Potential neighboring words in the query are used to disambiguate the synset following the approach of [Liu *et al.*, 2004]. The most appropriate synset is given the highest rank in the list. Then, all synsets related to the best synset are added to the list of suggestions as related synsets. The suggestions are displayed in a drop down box which opens when the user clicks on the term within the query (see figure 1). This disambiguation helps to avoid information overload.

[1] http://www.ezdl.de

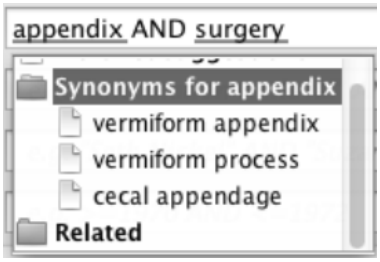


Figure 1: Popup box with search suggestions for the word *appendix*.

Since a meta search system often integrates different heterogeneous data sources, information search takes longer than in local retrieval solutions. In this case, users are encouraged to carefully formulate their query. For this purpose, EzDL offers the possibility to specify query terms more precisely by using various more meaningful input fields like title, author, and year of publication. We argue that when using meta search systems, queries are more sophisticated than in other systems like web search solutions where users usually submit very short and ambiguous queries. This means that the probability that two distinct users will enter exactly the same query is not as high as in many web retrieval systems.

As a consequence to this, expanding the query as it is suggested by [Fonseca *et al.*, 2005] is not sufficient. Rather, the particular query terms should be used as the basis for an association rule mining among the query logs. Association rule mining is a computationally expensive process. Therefore, the computation is performed during the start of the client application of EzDL. The results are stored in a database so that they are immediately available upon the next start of the client. In this way, the client is instantly supplied with working data.

The ChARM algorithm introduced by [Zaki and Hsiao, 2005] provides a very efficient method for computing association rules. The big advantage of the algorithm is that it only acquires the closed frequent itemsets, which avoids a lot of redundancy found in other association rule mining algorithms. The algorithm does so by taking a "round trip" over the sets of items and transactions through a Galois connection. Considering also the set of transactions avoids having to solve an NP-complete problem, namely, finding all frequent itemsets. The ChARM algorithm, instead, only finds the closed frequent itemsets. In the next step, ChARM mines non-redundant association rules by utilizing the concept of minimal generators which is applied to the closed frequent itemsets, as explained in [Zaki, 2004]. Reducing redundancy is key in making this algorithm so efficient. In comparison to Apriori, ChARM reduces the generation of redundant rules up to a factor of 66.

The ChARM algorithm can be configured by setting the values of minimum support and minimum confidence. For testing purposes, we used a minimum support of 2 and a minimum confidence of 30%.

Using the association rules stored in the database, a query relation graph is built considering each term in the query. For each term, binary association rules containing the term are considered for expanding the tree. Binary association rules are rules that identify a mapping between exactly two terms. Each term is represented by a node, but no term is represented by more than one node. This way of building the graph reflects transitive relationships between

the query terms. The transitive relationships can be identified by finding all elementary circuits in the graph. After all of those are found, it could be decided whether the graph is Hamiltonian, i.e. if the graph consists of one circuit containing all the nodes of the graph. This question in itself is not of our concern, but it is interesting to note that the question whether a graph is Hamiltonian is an NP-complete problem.

The procedure of building a query graph is induced every time the query is changed, and thus it needs to be executed very fast. Clearly, messing with NP-complete problems would not be something we would like to deal with on a regular basis. The algorithm by Tarjan presented in [Tarjan, 1973] finds all elementary circuits with a complexity of $O((|V| \cdot |E|)(|C| + 1))$ for $|V|$ nodes, $|E|$ edges, and $|C|$ circuits. Thus, by using this algorithm we can reach polynomial complexity as long as we do not have to deal with a huge number of circuits.

If all the nodes of a given circuit are already contained in another circuit, the first circuit is redundant, and it will be removed from the set of circuits. The remaining circuits are presented to the user as concepts for query expansion. In the presentation, words of the concept that are also part of the current user query are omitted. The user can choose the kind of relationship between concept and current query, which determines how the concept is linked to the current Boolean query (see figure 2).

4 Evaluation

The modified client was evaluated with six undergradute and graduate students, but none were experts in computer science or literature. The group included representatives of both genders. The students were asked to search for information objects about deadlocks. They were instructed to find relevant sources for writing a research paper about this topic. Before they started, they were briefly introduced about the concept of deadlocks in computer science. While searching, they verbalized their thoughts. In addition, the screen was recorded. In the beginning, the students had trouble identifying the recommendation tool at all, due to misplacement on the screen or because they did not realize the functionality of the tool. The first three had to be encouraged to take a look. As a response to this, the user interface was changed to highlight the query expansion tool.

The result regarding the suggestions given by WordNet were only noticed by some users, and they were quickly dismissed as not relevant to the query. The suggestions created from the query logs were treated differently. While the query expansion tool integrated in EzDL (figure 2) was often treated with initial scepticism, the suggestions proposed lead at the end to relevant search results in most of the cases. Since the users were not familiar with the concepts of deadlocks before the evaluation, the suggestions helped them to see which other concepts might be related to the topic, and which are the key authors on this topic. Even if the suggestions were not used via the tool, the students read and reused the suggested terms in new queries.

As a side finding, the Boolean expressions of the suggestions were mostly not understood. In fact, the users expected queries to be linked exactly the other way around than how Fonseca et al. did it in their recommendation system, i.e., they expected generalizations to be linked by the `OR` operator. Another evaluation on a larger scale needs to show if the sample in this case was too small, if the evaluation by Fonseca et al. was somehow faulty, or if the situation of this case affects the evaluation to turn out differently.

5 Discussion and Next Steps

In this paper we proposed, implemented and evaluated a two step recommendation system for query expansion in meta search engines. The system adapts to all users, as the query base increases.

The next steps will be manyfold. First, further evaluative research will show whether the linking of Boolean expression needs to be done in a different way, and whether the values that we have used for minimum support and minimum confidence are appropriate. Secondly, the integration of other services like Wikipedia for a better disambiguation and suggestion of query terms will be tried. Thirdly, we will investigate, from the human-computer interaction point of view, how to better highlight the suggestions components and how to make them more recognizable without disturbing the work flow of the user. And fourthly, we will make the system task aware, as described in [Backhausen, 2012], in order to learn task based and not with respect to all user logs. This way, the suggestions should be more focused. Finally, we will investigate query formulation which can be assisted by building suggestions using the meta information of objects that are marked as relevant by other users or stored in their personal library, as described in [Landwich *et al.*, 2009].

References

[Backhausen, 2012] Daniel T. J. Backhausen. Adaptive ir for exploratory search support. In *SIGIR: Doctoral Consortium*, page 992, 2012.

[Beckers *et al.*, 2012] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, and Sascha Kriewel. ezdl: An interactive search and evaluation system. In *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 9–16, Dunedin, New Zealand, August 2012. Department of Computer Science, University of Otago.

[Cui *et al.*, 2002] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web*, WWW '02, pages 325–332. ACM, 2002.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet*. Language, speech and communication. MIT Press, Cambridge, MA, 2 edition, 1998.

[Fonseca *et al.*, 2005] Bruno M. Fonseca, Paulo Golgher, Bruno Pôssas, Berthier Ribeiro-Neto, and Nivio Ziviani. Concept-based interactive query expansion. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 696–703. ACM, 2005.

[Harman, 1988] D. Harman. Towards interactive query expansion. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '88, pages 321–331. ACM, 1988.

[Kim *et al.*, 2004] Sang-Bum Kim, Hee-Cheol Seo, and Hae-Chang Rim. Information retrieval using word senses: root sense tagging approach. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 258–265. ACM, 2004.

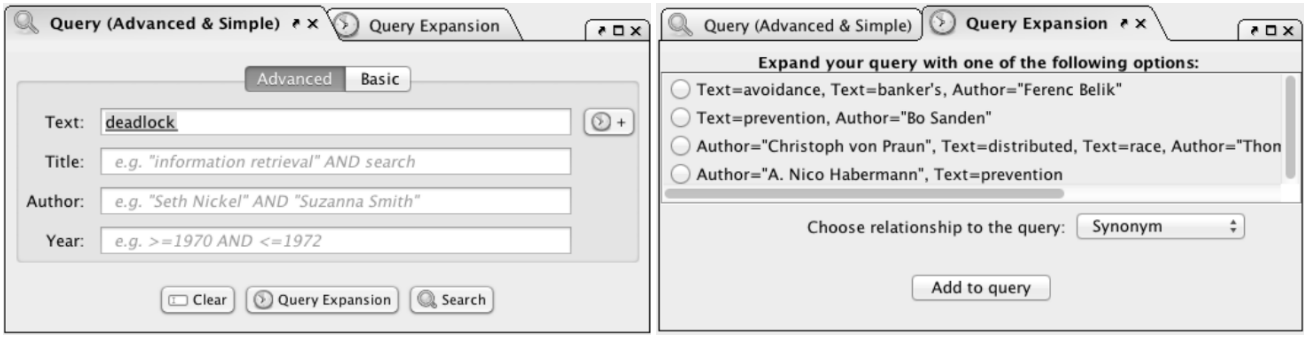


Figure 2: Concepts for query expansion suggested for the query "deadlock".

[Klas and Hemmje, 2009] Claus-Peter Klas and Matthias Hemmje. Catching the user - user context through live logging in daffodil. In *SIGIR 2009 Workshop: Understanding the User, Boston, USA*, 2009.

[Klas *et al.*, 2004] Claus-Peter Klas, Norbert Fuhr, and André Schaefer. Evaluating strategic support for information access in the DAFFODIL system. In *Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2004)*, 2004.

[Landwich *et al.*, 2009] Paul Landwich, Tobias Vogel, Claus-Peter Klas, and Matthias Hemmje. Model to support patent retrieval in the context of innovation-processes by means of dialogue and information visualisation. *Electronic Journal of Knowledge Management*, 7:87–98, 1 2009.

[Liu *et al.*, 2004] Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 266–272. ACM, 2004.

[Mitra *et al.*, 1998] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 206–214. ACM, 1998.

[Qiu and Frei, 1993] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 160–169. ACM, 1993.

[Tarjan, 1973] Robert Tarjan. Enumeration of the elementary circuits of a directed graph. *SIAM Journal on Computing*, 2(3):211–216, 9 1973.

[Voorhees, 1994] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 61–69. Springer-Verlag New York, Inc., 1994.

[Xu and Croft, 1996] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11. ACM, 1996.

[Zaki and Hsiao, 2005] M.J. Zaki and C.-J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):462–478, April 2005.

[Zaki, 2004] Mohammed Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223–248, 2004.

The D2Q2 Framework: On the Relationship and Combination of Language Modelling and TF-IDF

Thomas Roelleke, Hany Azzam, Marco Bonzanini, Miguel Martinez-Alvarez and Mounia Lalmas

Abstract

Language Modelling (LM) and TF-IDF are two retrieval models with different foundations. There have been efforts aiming at establishing the relationship between these models, and whether one includes the other. Whether their combination could yield a third and better model is an open research question. This paper revisits the foundations of LM and TF-IDF and explores how these models' *bare* structures relate and how these structures can be combined. We begin with the premise that TF-IDF is the $P(d|q)/P(d)$ side of retrieval, which complements the common view that LM is $P(q|d)/P(q)$. Next, a hybrid framework based on the decomposition of the product of the two sides, $P(d|q)/P(d) \cdot P(q|d)/P(q)$, is developed. This leads to the *D2Q2* family of models, which joins the inner components of LM and TF-IDF instead of combining their scores. This paper provides new insights into the relationship between LM and TF-IDF, and experimental results show that the D2Q2 models perform comparably to competitive baselines.

1 Introduction

There has been significant research into how to combine retrieval models and how to relate them. Approaches such as [Bartell *et al.*, 1994; Croft *et al.*, 1990; Lee,] have shown the importance of combining different retrieval models through, for example, score fusion. Other approaches have proposed how to *analyse* different retrieval models' components and compare them [Fang and Zhai, 2005]. Both research directions have furthered the development of more effective models.

Two types of retrieval models that have been closely analysed and compared are language modelling (LM), and those based on term frequency (TF) and inverse document frequency (IDF). These models have different foundations. Variants of the former are based on the mixtures (smoothing) [Zhai and Lafferty, 2004; Zaragoza *et al.*, 2003]. TF-IDF models differ with regard to the TF quantification and normalisations employed [Robertson *et al.*, 1994; Singhal *et al.*, 1996; He and Ounis, 2005; Kwok, 1996; Taylor *et al.*, 2006]. Efforts to establish the relationship between these models and whether or not the former includes the features of the latter include [Zhai and Lafferty, 2001]. By examining the foundations of these retrieval models we learn that LM directly derives from the conditional probability $P(q|d)$ (q is the query, d is the document)

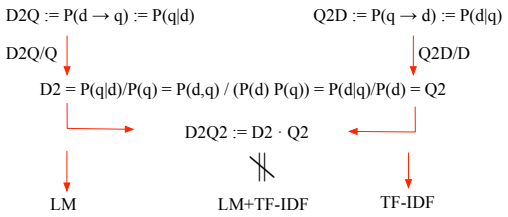


Figure 1: The D2Q2 framework.

[Ponte and Croft, 1998; Hiemstra, 2000; Lafferty and Zhai, 2003]. TF-IDF, on the other hand, is viewed as a *heuristic* model [Salton *et al.*, 1976; Croft, 2000; Metzler and Croft, 2004], and its probabilistic and information-theoretic interpretation is an ongoing debate [Church and Gale, 1995a; Church and Gale, 1995b; Aizawa, 2003; Robertson, 2004; Wu *et al.*, 2008; Roelleke and Wang, 2008]. Drawing from and furthering this type of deeper analysis allows us to better understand and relate these models' components.

This paper contributes several theoretical findings. We showcase a side-by-side derivation of LM and TF-IDF that helps to clarify the relationship between LM and TF-IDF. This derivation goes so far as to show that, just as LM has a TF-IDF nature, before the decomposition of document and query probabilities, *TF-IDF has an LM nature* as well. Next ,we develop a hybrid framework, leading to the D2Q2 family of models, that joins the inner components of LM and TF-IDF.

Figure 1 outlines the connections between D2Q2, its two subcomponents D2 and Q2, LM and TF-IDF. Essentially, D2Q2 rests on two ways to decompose the document-query independence measure DQI $:= P(d,q)/(P(d) \cdot P(q))$. D2Q2 combines the *inner* parts of TF-IDF and LM, trying to push the best of each into a *hybrid model*. Continuing with the derivation, the inner components of LM ($P(q|d)/P(q) = D2$) and TF-IDF ($P(d|q)/P(d) = Q2$) are combined. This provides an integrative framework that incorporates the characteristics of both LM and TF-IDF. Moreover, the instances of this framework, D2Q2, are retrieval models in their own right, which can be compared with the traditional models LM and TF-IDF. Interestingly, although D2Q2 "combines" models, it is different from the aforementioned fusion approaches. While fusion combines scores, D2Q2 incorporates the characteristics of both LM and TF-IDF into one probabilistic framework, and therefore we refer to D2Q2 as a "hybrid" model, as opposed to a model that fuses scores.

This paper is structured as follows. Section 2 consolidates the preliminaries necessary to appreciate the contri-

bution of this paper. Section 3 shows the relationship between LM and D2Q2. More precisely, it shows that LM corresponds to D2. The relationship between TF-IDF and D2Q2 is shown in Section 4 (TF-IDF corresponds to Q2). Section 5 discusses the relationship between LM and TF-IDF. From these follows in Section 6 the description of the D2Q2 framework, a theoretically sound combination of the LM and TF-IDF models into a family of hybrid models. Section 7 shows that the D2Q2 retrieval models perform comparably to competitive baselines.

2 Background & Preliminaries

2.1 LM and TF-IDF

We present the LM and TF-IDF models[1]. Note that TF-IDF is also referred to as a weighting scheme in the context of the vector space model. This paper emphasises that TF-IDF *is* a retrieval model at the same level as LM, as in [Hiemstra, 2000].

Let d be a document, q a query, c a collection and t a term. The standard definition of the retrieval status value associated with the LM model can be written as follows:

$$RSV_{\text{LM}}(d,q,c) := \sum_{t \in q} TF(t,q) \cdot \log\left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)}\right) \quad (1)$$

$TF(t,q)$ is the within-query term frequency, $P(t|d)$ is the within-document (foreground) term probability, $P(t|c)$ is the collection-wide (background) term probability, and λ_d is the document-dependent mixture parameter.

In the Dirichlet-based LM [Zhai and Lafferty, 2004], λ_d is proportional to the document length. Let $\lambda_d := \frac{dl}{dl+\mu}$, where *dl* is the document length and μ is a parameter. This setting of λ_d reflects trust in probabilities estimated from *long* documents.

$$RSV_{\text{Dirich-LM}}(d,q,c) := \sum_{t \in q} TF(t,q) \cdot \log\left(\frac{\mu}{\mu+dl} + \frac{dl}{dl+\mu} \cdot \frac{P(t|d)}{P(t|c)}\right) \quad (2)$$

The retrieval status value associated with the TF-IDF model can be written as follows:

$$RSV_{\text{TF-IDF}}(d,q,c) := \sum_{t \in d \cap q} TF(t,d) \cdot TF(t,q) \cdot IDF(t,c) \quad (3)$$

$TF(t,d)$ is the within-document term frequency quantification; $TF(t,q)$ is for the query. For independence of term occurrences, the setting is $TF(t,d) := tf_d$ where tf_d is the *total* within-document term frequency. This setting is known to be inferior to $TF(t,d) := tf_d/(tf_d + K_d)$, the setting known from BM25 [Robertson *et al.*, 1994], where K_d is a normalisation factor proportional to the pivoted document length, $pivdl(c) := dl/avgdl(c)$. We refer to this TF quantification as BM25-TF, and we also denote it as $TF_K(t,d)$, to make explicit the parameter K. For *IDF*, the common setting is $IDF(t,c) := -\log P_D(t|c)$, where $P_D(t|c) = df(t,c)/N_D(c)$ is the Document-based term probability (based on the set of Documents, hence, the subscript capital D), and $df(t,c)$ is the collection-wide document frequency of term t.

[1] Similar investigation was carried out for the BM25 model; however in this paper we focus on LM and TF-IDF.

Note that IDF is based on a *Document-based* term probability ($P(t|c) := P_D(t|c)$), whereas LM is *Location-based* ($P(t|c) := P_L(t|c)$) [Hiemstra, 2000]. We return to these two event spaces (Documents vs. Locations) in Section 4.5, where an essential assumption is made to establish the connection between TF-IDF and D2Q2.

2.2 Document-Query (In)dependence (DQI)

An common measure in probabilistic models is the document-query independence, formalised as follows:

$$DQI(d,q) := \frac{P(d,q)}{P(d) \cdot P(q)} \quad (4)$$

The DQI measures the document-query (in)dependence. $DQI = 1$ means that document and query intersect as if they were independent; $DQI < 1$ means that the intersection is less; and $DQI > 1$ means that the intersect is greater than if they were independent.

The *DQI* is a concept related to information theory. It is the inner component of the "mutual information" $MI(X,Y) := \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x) \cdot P(y)}$. The DQI is the argument of the log. The relationship of DQI to MI (and hence to conditional entropy) backs DQI as an information-theoretic measure [Gale and Church, 1991]. It also shows the theoretical justification of D2Q2, which leverages the DQI measure in its derivation. Lastly, DQI is related to exhaustiveness and specificity (another foundation of D2Q2).

2.3 Exhaustiveness and Specificity

The product $P(q|d) \cdot P(d|q)$ can be interpreted as *exhaustiveness* $\cdot$ *specificity*, where $P(q|d)$ is set to measure exhaustiveness and $P(d|q)$ specificity. These concepts were used in logic-based retrieval *frameworks* [Nie, 1992; Wong and Yao, 1995]. We retain the idea, and define an exhaustiveness-specificity measure:

$$ES(d,q) := P(q|d) \cdot P(d|q) \quad (5)$$

From this definition, it immediately follows the relationship between ES and DQI, which can be expressed in as follows:

$$ES(d,q) = \frac{P(d,q) \cdot P(d,q)}{P(d) \cdot P(q)} = P(d,q) \cdot DQI(d,q) \quad (6)$$

The role of $ES(d,q)$ and $DQI(d,q)$ is explained in Section 5. Mainly, the combination of exhaustiveness and specificity, plus the meaning of DQI, give a meaning to D2Q2.

To estimate $P(q|d)$ and $P(d|q)$, the query q and document d are viewed as *sequences* of independent term events. However, the independence assumption can be seen as sub-optimal. Hence, many approaches such as [Gao *et al.*, 2004; Hou *et al.*, 2011] capture dependence when estimating the document and query probabilities. Similarly, D2Q2 considers dependence by using the notion of semi-subsumed events. The next section reviews this assumption and relates it to the BM25-TF; this justifies why the BM25-TF is later used in D2Q2.

2.4 Semi-subsumed Events

The superior retrieval quality achieved by the BM25-TF is evidence for the dependence of the multiple occurrences of the same term [Robertson *et al.*, 1994]. For instance, [Wu and Roelleke, 2009] pointed out that the BM25-TF

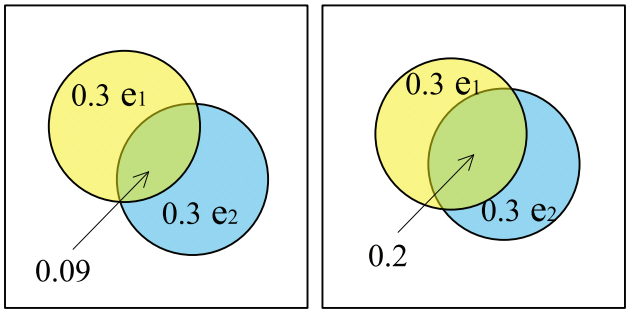


Figure 2: Independent and Semi-subsumed.

can be explained by assuming term occurrences to be *semi-subsumed* events, an important concept for making the proposed hybrid D2Q2 framework a with solid and probabilistic foundations.

In general, the decomposition of event d into term events can be written as:

$$P(d|q) = \prod_{t \in d} P(t|q)^{TF(t,d)} \tag{7}$$

The setting of $TF(t,d)$ reflects probabilistic assumptions:

$$TF(t,d) := \begin{cases} tf_d & \text{independent} \\ 2 \cdot tf_d/(tf_d+1) & \text{semi-subsumed} \\ 1 & \text{subsumed} \end{cases}$$

$TF(t,d) = tf_d$ (total term frequency) views the occurrences as *independent*, whereas $TF(t,d) = 1$ views them as *subsumed* events. Semi-subsumed is between the two. Figure 2 illustrates the computation of $P(e_1, e_2)$ for the case of independent and semi-subsumed events. For IR, event e_i corresponds to the multiple occurrence of a term t_i. For independent events, we obtain $P(e_1, e_2) = 0.3^2 = 0.09$; and for semi-subsumed events, $P(e_1, e_2) = 0.3^{2 \cdot 2/(2+1)} \approx 0.2$. The conjunctive probability of semi-subsumed events is larger than that of independent events. The success of the BM25-TF proves that the multiple occurrences of a term are not independent. The notion of semi-subsumed events assigns a sound semantics to the BM25-TF, making it a well-defined ingredient of D2Q2.

We have discussed the preliminaries of LM and TF-IDF, document-query-(in)dependence (DQI), exhaustiveness and specificity, and semi-subsumed events. The next two sections use these to show the connection between LM and D2, and TF-IDF and Q2.

3 LM as the D2 side of D2Q2

We demonstrate that LM corresponds to the D2 side of D2Q2. We start with reviewing the probabilistic roots of LM as explored in [Hiemstra, 2000; Zhai and Lafferty, 2004]. The notation D2Q stands for $P(q|d)$, and D2 for $P(q|d)/P(q)$, which we denote as D2Q/Q.

$$\text{D2Q} := P(q|d), \qquad \text{D2} := \text{D2Q}\,/Q := \frac{P(q|d)}{P(q)} \tag{8}$$

This section addresses the estimation of $P(q|d)$, or more precisely, of $P(q|d,c)$, where the notation makes explicit the collection "c" used to estimate the background term probability.

3.1 Term (In)dependence Assumption

To estimate $P(q|d,c)$, the query is decomposed into terms:

$$P(q|d,c) = \prod_{t \in q} P(t|d,c)^{TF(t,q)} \tag{9}$$

The conditional d, c makes it explicit that the query and term probabilities depend on both the document d (foreground) *and* the collection c (background). The setting of *TF* reflects two common assumptions made for term events:

$$TF(t,q) := \begin{cases} tf_q & \text{independent} \\ 1 & \text{subsumed} \end{cases} \tag{10}$$

For $P(q|d,c)$, and therefore, D2Q, which assumption is followed is not crucial since often $tf_q = 1$ for short queries. Next we discuss the estimation of $P(t|d,c)$.

3.2 Term Probability Mixture

$P(t|d,c)$ is estimated using a mixture of foreground and background probabilities, essentially to avoid the so-called "zero-probability problem" [Zhai and Lafferty, 2004]. The within-document term probability $P(t|d)$ is mixed with the collection term probability $P(t|c)$ to obtain $P(t|d,c)$:

$$P(t|d,c) = \lambda_d \cdot P(t|d) + (1 - \lambda_d) \cdot P(t|c) \tag{11}$$

The parameter λ_d may be set constant (Jelinek/Mercer mixture, for example, $\lambda_d \approx 0.8$, [Hiemstra, 2000]). Alternatively, $\lambda_d := \frac{dl}{dl+\mu}$ (Dirichlet mixture, *dl* is document length) means that the estimate of $P(t|d)$ is more trusted for longer documents.

We discussed the estimation of $P(q|d,c)$, including the term (in)dependence assumption, leading to the formulation of D2Q. We also referred to D2 as D2Q/Q, that is D2 is equal to D2Q normalised by Q. We discuss the normalisation step next, which leads us to the formulation of D2.

3.3 Normalisation

Applying Equation 9 to Equation 8, making the collection c explicit, and decomposing $P(q|c)$ in the same way as $P(q|d,c)$ (Equation 9), D2 can be decomposed as follows:

$$\text{D2} = \text{D2Q}\,/Q = \frac{P(q|d,c)}{P(q|c)} = \prod_{t \in q} \left(\frac{P(t|d,c)}{P(t|c)} \right)^{TF(t,q)} \tag{12}$$

Using the term probability mixture estimation of $P(t|d,c)$ (Equation 11), we arrive at the following form of D2, which we denote D2-linear, where the subscript indicates the type of the mixture (here a linear mixture):

$$\text{D2}_{\text{linear}} := \prod_{t \in q} \left[(1 - \lambda_d) + \frac{\lambda_d \cdot P(t|d)}{P(t|c)} \right]^{TF(t,q)} \tag{13}$$

We define a second form of D2, denoted D2-extreme, to capture the case of $\lambda_d = 1$ if $t \in d$, and $\lambda_d = 0$ otherwise:

$$\text{D2}_{\text{extreme}} := \prod_{t \in d \cap q} \left[\frac{P(t|d)}{P(t|c)} \right]^{TF(t,q)} \tag{14}$$

We discuss in more detail the extreme mixture when we present Q2, as it establishes the relationship between Q2 and TF-IDF.

3.4 Retrieval Status Value

For each of the D2 forms above, we define an associated retrieval status value (RSV), which can serve as a ranking function. Essentially, the RSV's apply the logarithm.

$$RSV_{\text{D2-linear}}(d,q,c) := \log \text{D2}_{\text{linear}} \tag{15}$$

$$RSV_{\text{D2-extreme}}(d,q,c) := \log \text{D2}_{\text{extreme}} \tag{16}$$

In decomposed form, the RSV's become:

$$RSV_{\text{D2-linear}}(d,q,c) = \sum_{t \in d \cap q} TF(t,q) \cdot \log\left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)}\right) \tag{17}$$

$$RSV_{\text{D2-extreme}}(d,q,c) = \sum_{t \in d \cap q} TF(t,q) \cdot \log \frac{P(t|d)}{P(t|c)} \tag{18}$$

We next make the connection between LM and D2 explicit, namely that D2=LM for the linear form of LM.

3.5 D2 and LM

The following theorem (proof omitted) shows the exact relationship between D2 (the linear form) and LM:

Theorem 1 D2*-linear is an interpretation of LM:*

$$\text{RSV}_{LM}(d,q,c) = \text{RSV}_{D2\text{-}linear}(d,q,c) \tag{19}$$

Showing that D2=LM does not reveal a new result; the estimation of D2 (leading to D2-linear) was carefully chosen to lead to LM. We nonetheless presented the above steps to prepare for the more complex case demonstrating the relationship between Q2 and TF-IDF.

4 TF-IDF as the Q2 side of D2Q2

We have shown that LM is the D2 $:= P(q|d)/P(q)$ side of D2Q2. Next, we show that TF-IDF is the Q2 $:= P(d|q)/P(d)$ side of D2Q2. This section on TF-IDF is organised analogously to the previous one on LM. For TF-IDF, $P(d|q)$ is the starting point, from where we mirror the steps followed in Section 3. Q2D stands for $P(d|q)$, Q2 for $P(d|q)/P(d)$, where Q2 is Q2D normalised by D denoted Q2D/D.

$$\text{Q2D} := P(d|q), \qquad \text{Q2} := \text{Q2D}\,/D := \frac{P(d|q)}{P(d)} \tag{20}$$

Equation 20 (Q2D) corresponds to Equation 8 (D2Q). Next, we estimate $P(d|q)$.

4.1 Term (In)dependence Assumption

Again we explicate the collection c. To estimate $P(d|q,c)$, the document is decomposed into terms:

$$P(d|q,c) = \prod_{t \in d} P(t|q,c)^{TF(t,d)} \tag{21}$$

Equation 21 corresponds to Equation 9 ($P(q|d,c)$). There are three assumptions encoded in the TF quantification:

$$TF(t,d) := \begin{cases} tf_d & \text{independent} \\ 2 \cdot tf_d/(tf_d + K_d) & \text{semi-subsumed} \\ 1 & \text{subsumed} \end{cases} \tag{22}$$

The semi-subsumed assumption (BM25-TF) led to superior retrieval performance [Robertson *et al.*, 1994]. The parameter K_d is proportional to the pivoted document length $pivdl = dl/avgdl$. The parameter adjusts the semi-subsumption assumption.

4.2 Term Probability Mixture

We again use a mixture model to estimate $P(t|q,c)$:

$$P(t|q,c) = \lambda_q \cdot P(t|q) + (1-\lambda_q) \cdot P(t|c) \tag{23}$$

Equation 23 corresponds to Equation 11 ($P(t|d,c)$).

4.3 Normalisation

Normalisation leads to Q2 (as Q2D/D).

$$\text{Q2} = Q2D/D = \frac{P(d|q,c)}{P(d|c)} = \prod_{t \in d} \left(\frac{P(t|q,c)}{P(t|c)}\right)^{TF(t,d)} \tag{24}$$

Equation 24 corresponds to Equation 12 (D2). As for D2, we define two forms of Q2, linear and extreme. Q2-linear derives directly from applying the term probability mixture to estimate $P(t|q,c)$.

$$\text{Q2}_{\text{linear}} := \prod_{t \in d} \left[(1-\lambda_q) + \frac{\lambda_q \cdot P(t|q)}{P(t|c)}\right]^{TF(t,d)} \tag{25}$$

Equation 25 corresponds to Equation 13 (D2-linear).

The extreme mixture comes from setting $\lambda_q = 1$ if $t \in q$, and $\lambda_q = 0$ otherwise.

$$\text{Q2}_{\text{extreme}} := \prod_{t \in d \cap q} \left(\frac{P(t|q)}{P(t|c)}\right)^{TF(t,d)} \tag{26}$$

Equation 26 corresponds to Equation 14 (D2-extreme). Section 4.6 will show that it is the extreme form of Q2 that is related to TF-IDF.

4.4 Retrieval Status Value

We take the log to define the corresponding retrieval status value for both forms of Q2, and obtain the following:

$$RSV_{\text{Q2-linear}}(d,q,c) = \sum_{t \in d \cap q} TF(t,d) \cdot \log\left((1-\lambda_q) + \lambda_q \cdot \frac{P(t|q)}{P(t|c)}\right) \tag{27}$$

$$RSV_{\text{Q2-extreme}}(d,q,c) = \sum_{t \in d \cap q} TF(t,d) \cdot \log \frac{P(t|q)}{P(t|c)} \tag{28}$$

Note the symmetry between Equation 27 and 17, and between Equation 28 and 18.

We continue with Q2-extreme, showing that it corresponds to TF-IDF. Equation 28 has a factor $1/P(t|c)$, the inverse term probability, which reminds of $IDF(t,c) := \log(1/P_D(t|c))$, which we recall is based on the space of *Documents*. However, all the probabilistic estimates so far are based on the space of *Locations* (terms occur at locations). The next section reviews the assumption that allows to transfer the *Location*-based probability $P_L(t|c)$ into the *Document*-based probability $P_D(t|c)$. The transformation between event spaces is necessary to demonstrate since it is one of the pillars between Q2 and TF-IDF.

4.5 Query Term Probability Assumption

We review first the query term probability assumption discussed in [Roelleke and Wang, 2006], which allows the transfer of the Location-based probabilities, $P_L(t|q)/P_L(t|c)$ in Equation 28, to the Document-based probabilities, $1/P_D(t|c)$.

To illustrate the difference between the two spaces, Documents and Locations, consider the following example.

Let term t occur in $tf_c = n_L(t,c) = 1,000$ *Locations* of collection c. Let it occur in $df(t,c) = n_D(t,c) = 200$ *Documents* of collection c. The notation conforms with traditional formulation, and indicates the duality between counting Locations and counting Documents. Then, the average (expected) within-document term frequency is: $avgtf(t,c) = tf_c/df(t,c) = 1,000/200 = 5$. Now let the collection c have $N_L(c) = 10^9$ Locations, and $N_D(c) = 10^6$ Documents. The Location-based probability is $P_L(t|c) = n_L(t,c)/N_L(c) = 1,000/10^9$, the Document-based one is $P_D(t|c) = n_D(t,c)/N_D(c) = 200/10^6$. The average document length is $avgdl(c) = N_L(c)/N_D(c) = 10^3$.

Then, for the fraction of term probabilities, we obtain:

$$\frac{P_L(t|c)}{P_D(t|c)} = \frac{n_L(t,c)/N_L(c)}{n_D(t,c)/N_D(c)} = \frac{avgtf(t,c)}{avgdl(c)} \quad (29)$$

This equation has been referred to as Poisson bridge [Roelleke and Wang, 2006], since it is related to a Poisson probability (we do not need to detail for this paper).

This relationship between Location-based and Document-based term probability enables us to establish the relationship between Q2 and TF-IDF. The relationship is based on the following query term probability assumption:

$$P_L(t|q) = avgtf(t,c)/avgdl(c) \quad (30)$$

What does this assumption express? In the example above, the average document length is $avgdl(c) = 1,000$ and the average within-document term frequency is $avgtf(t,c) = 5$; therefore, $P_L(t|q) = 5/1,000$. With this assumption bursty terms obtain *higher* probabilities than less bursty ones: the query term probability is *proportional* to the burstiness of the term, a reasonable assumption to make.

This assumption leads to $P_L(t|c) = P_L(t|q) \cdot P_D(t|c)$. In turn, this transform the fraction $P_L(t|q)/P_L(t|c)$ (see Equation 28) into an expression based on the Document-based term probability as in IDF:

$$\frac{P_L(t|q)}{P_L(t|c)} = \frac{P_L(t|q)}{P_L(t|q) \cdot P_D(t|c)} = \frac{1}{P_D(t|c)} \quad (31)$$

This establishes the relationship between Q2 and TF-IDF.

$$\log \mathrm{Q2}_{\text{extreme}} = \sum_{t \in d \cap q} TF(t,d) \cdot \log \frac{1}{P_D(t|c)} \quad (32)$$

Next we give the formal proof that shows Q2 (extreme form) is the probabilistic interpretation of TF-IDF.

4.6 Q2 and TF-IDF

In Section 3.5, the relationship between D2 and LM was a direct one. The relationship between Q2 and TF-IDF is less direct, as it relies as above shown on the "query term probability assumption". In addition, whereas showing the relationship between D2 and LM, i.e. LM=D2, relied on a *linear* mixture, showing the relationship between TF-IDF and Q2, i.e. Q2=TF-IDF, relies on the *extreme* mixture.

Given the query term probability assumption, the relationship between Q2 and TF-IDF is expressed as follows.

Theorem 2 $\mathrm{Q2}_{extreme}$ *is an interpretation of* TF-IDF*, if* $P_L(t|q) = P_L(t|c)/P_D(t|c)$*:*

$$P_L(t|q) = \frac{P_L(t|c)}{P_D(t|c)} \Longrightarrow$$
$$\mathrm{RSV}_{TF\text{-}IDF}(d,q,c) = \mathrm{RSV}_{\mathrm{Q2}_{extreme}}(d,q,c) \quad (33)$$

Proof Inserting Equation 3 for $RSV_{\text{TF-IDF}}$ and Equation 28 for $RSV_{\mathrm{Q2}_{\text{extreme}}}$ yields:

$$\sum_t TF(t,d) \cdot TF(t,q) \cdot IDF(t,c) = \sum_t TF(t,d) \cdot \log \frac{P_L(t|q)}{P_L(t|c)}$$

The assumption for $P_L(t|q)$ yields:

$$\frac{P_L(t|q)}{P_L(t|c)} = \frac{P_L(t|c)}{P_D(t|c) \cdot P_L(t|c)} = \frac{1}{P_D(t|c)}$$

Therefore, $\mathrm{Q2}_{\text{extreme}}$ is an interpretation of TF-IDF (for a binary query TF quantification $TF(t,q)$).

We have shown that $\mathrm{D2}_{\text{linear}}$ corresponds to LM, and that $\mathrm{Q2}_{\text{extreme}}$ corresponds to TF-IDF. In the next section, we focus on the relationship between D2 and Q2.

5 On the relationship between D2 (LM) and Q2 (TF-IDF)

Section 2.2 introduced the Document-Query (In)dependence (DQI) measure: $DQI(d,q) := P(d,q)/(P(d) \cdot P(q))$. From the definitions of D2 and Q2, we obtain that $\mathrm{D2} = DQI = \mathrm{Q2}$. This means that D2 and Q2 are equivalent:

$$\mathrm{D2} = \frac{P(q|d,c)}{P(q|c)} = \frac{P(d,q|c)}{P(d|c) \cdot P(q|c)} = \frac{P(d|q,c)}{P(d|c)} = \mathrm{Q2} \quad (34)$$

In other words, before decomposing events into term events and until term (in)dependence assumption made, D2 and Q2 measure the same, that is, LM and TF-IDF aim at measuring the same. The decomposition of d and q into terms breaks the equivalence of D2 and Q2.

$$\sum_{t \in q} TF(t,q) \cdot \log \frac{P(t|d,c)}{P(t|c)} \neq \sum_{t \in d} TF(t,d) \cdot \log \frac{P(t|q,c)}{P(t|c)} \quad (35)$$

For D2, $P(t|d,c)$ is estimated as the linear mixture $\lambda_d \cdot P(t|d) + (1-\lambda_d) \cdot P(t|c)$, establishing that "D2=LM". For Q2, an extreme mixture for $P(t|q,c)$ is applied and we assumed that $P_L(t|q) = P_L(t|c)/P_D(t|c)$, which led to "Q2=TF-IDF".

The following inequality stresses the difference between LM (D2-linear) and TF-IDF (Q2-extreme).

$$TF(t,q) \log \left[(1-\lambda_d) + \lambda_d \cdot \frac{P_L(t|d)}{P_L(t|c)} \right] \neq TF(t,d) \log \frac{1}{P_D(t|c)}$$

We have shown the steps from the equality D2=Q2=DQI that holds before decomposition into term events to the inequality LM≠TF-IDF that comes from the term (in)dependence assumption. This not only shows a relationship between LM and TF-IDF, but explains what connects them, and what separates them.

6 The D2Q2 Framework

We have shown the relationships between LM and D2, between TF-IDF and Q2, between D2 and Q2, and between LM and TF-IDF. The preliminaries introduced concepts (i.e. DQI measure), recalled IR pillars (i.e. exhaustiveness times specificity measure) and relatively recent theory such as semi-subsumed events. Together, the relationships and preliminaries form the theoretical ground of D2Q2.

Our starting point is $ES(d,q) = P(q|d) \cdot P(d|q)$, the exhaustiveness-times-specificity measure commonly used as the basis to justify retrieval models. By analogy, we

define D2Q2 as the product of D2 (Equation 8) and Q2 (Equation 20):

$$\text{D2Q2} := \text{D2} \cdot \text{Q2} \tag{36}$$

where D2 relates to D2Q (exhaustiveness) and Q2 relates to Q2D (specificity). We also know that D2 (linear) corresponds to LM and Q2 (extreme) corresponds to TF-IDF. In other words, D2Q2 "joins" LM and TF-IDF.

We show now that D2Q2 corresponds to DQI^2, where one of the *DQI* relates to LM and the other relates to TF-IDF. This is expressed as follows:

$$\text{D2Q2} = DQI^2$$

By inserting Equation 36 for D2Q2 and Equation 4 for *DQI*, we obtain the decomposed form:

$$\frac{P(q|d)}{P(q)} \cdot \frac{P(d|q)}{P(d)} = \frac{P(d,q)}{P(d) \cdot P(q)} \cdot \frac{P(d,q)}{P(d) \cdot P(q)} \tag{37}$$

We continue now with the two forms of D2Q2, namely, $\text{D2Q2}_{\text{extreme}}$ and $\text{D2Q2}_{\text{linear}}$, which we further decompose:

$$\text{D2Q2}_{\text{extreme}} = \prod_{t \in d \cap q} \left[\left(\frac{P(t|d)}{P(t|c)} \right)^{TF(t,q)} \cdot \left(\frac{P(t|q)}{P(t|c)} \right)^{TF(t,d)} \right] \tag{38}$$

$$\text{D2Q2}_{\text{linear}} = \prod_{t \in d \cap q} \left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)} \right)^{TF(t,q)} \cdot \left((1-\lambda_q) + \lambda_q \cdot \frac{P(t|q)}{P(t|c)} \right)^{TF(t,d)} \tag{39}$$

Equations 38 and 39 contain the core contribution of this paper: *the seamless and symmetric composition of probabilistic parameters into a score that embeds LM and TF-IDF*. The main properties of D2Q2 are:

1. A symmetric pattern of the two models' components: for LM these are $P(t|d)$ and $TF(t,q)$, and for TF-IDF these are $P(t|q)$ and $TF(t,d)$; the collection-wide term probability $P(t|c)$ is common to both. The term frequency $TF(t,d)$ and $TF(t,q)$ can be set as in BM25: $TF_K(t,x) := tf_x/(tf_x + K_x)$, which corresponds to assuming the occurrences of t to be semi-subsumed; alternatively, if assuming independence, then $TF(t,x) := tf_x$, where tf_x is the total term frequency count.
2. Derivation and interpretation based on conditional probabilities and document-query independence (*DQI*): $\text{D2} = P(q|d)/P(q) = DQI$ relates to LM, and $\text{Q2} = P(d|q)/P(d) = DQI$ relates to TF-IDF. To decompose D2 and Q2, the "extreme" or the "linear" mixture assumption is applied to both $P(t|d,c)$ and $P(t|q,c)$, leading to $P(t|d)/P(t|c)$ and $P(t|q)/P(t|c)$.
3. The two fractions $P(t|d)/P(t|c)$ and $P(t|q)/P(t|c)$ measure "divergence", i.e. they express that a term with $P(t|d) > P(t|c)$ and $P(t|q) > P(t|c)$ is a good term, where a term is good if its probability in d and q is greater than in collection c. Conditional entropy and Kullback-Leibler divergence incorporate such factors.
4. The "discriminativeness", expressed by $1/P(t|c)$, occurs twice, for the document side and for the query side; this is similar to the vector-space model, where the *idf* is in both the document and query vectors.

For each of D2 and Q2, there is the choice to apply either a linear or the extreme mixture. Our experiments, described next, focus on D2Q2-extreme, which does not involve any mixture parameter, and D2Q2-linear, the model with two mixture parameters (λ_d and λ_q). We define the D2Q2 retrieval status value using logs.

$$RSV_{\text{D2Q2}}(d,q) := \log \text{D2Q2} \tag{40}$$

The next equations show the decomposed, logarithmic form of $\text{D2Q2}_{\text{extreme}}$ (Equation 38) and $\text{D2Q2}_{\text{linear}}$ (Equation 39):

$$RSV_{\text{D2Q2-extreme}}(d,q,c) = \sum_{t \in d \cap q} \left[TF(t,q) \cdot \log \frac{P(t|d)}{P(t|c)} + TF(t,d) \cdot \log \frac{P(t|q)}{P(t|c)} \right] \tag{41}$$

$$RSV_{\text{D2Q2-linear}}(d,q,c) = \sum_{t \in d \cap q} TF(t,q) \cdot \log \left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)} \right) + TF(t,d) \cdot \log \left((1-\lambda_q) + \lambda_q \cdot \frac{P(t|q)}{P(t|c)} \right) \tag{42}$$

The above decomposed forms illustrates how D2Q2 joins the inner components of LM and TF-IDF, showing that D2Q2 is *hybrid*, i.e. a model beyond combining scores.

7 Experiments

Although the main contribution of this paper was the relationship between LM and TF-IDF, it remains interesting to investigate the experimental performance of D2Q2.

7.1 Set-up

We introduced two retrieval functions derived from D2Q2, $RSV_{\text{D2Q2-extreme}}$ and $RSV_{\text{D2Q2-linear}}$. We now investigate their retrieval performance on a range of collections, outlined in Table 1, of varying size and content.

	Documents $N_D(c)$	Topics $N_Q(c)$	Size
TREC-2	700,000+	50	1.3 GB
TREC-3	700,000+	50	1.3 GB
TREC-8	500,000+	50	834 MB
WT2g	247,000+	50	2 GB
Blogs06	3,200,000+	50	88.8GB

Table 1: Collection Statistics

Following TREC settings [Ounis *et al.*, 2006], for the Blog06 collection, we index only the permalinks (the blog posts and their associated comments). The Porter stemmer was used for stemming. No stopwords removal was applied. We only used the title topic field. We measure retrieval quality with Mean Average Precision (MAP) (topical MAP on Blog06 [Ounis *et al.*, 2006]) and P@10.

Model	Equation
LM_{Dirich}	Equation 2
TF_K-IDF	Equation 3
LM+TF_K-*IDF*	Combinations of retrieval scores
$\text{D2Q2}_{\text{extreme},TF_K}$	Equation 41
$\text{D2Q2}_{\text{linear},TF_K}$	Equation 42

Table 2: Retrieval Models.

Table 2 associates the retrieval models with their respective equations. The first two correspond to the LM and TF-IDF models, the third to the combination of scores of LM

and TF-IDF, and the last two are the two models derived from D2Q2. In TF_K-IDF and in D2Q2, the TF_K component is the BM25-TF, i.e. $TF_K(t,d) = tf_d/(tf_d + K_d)$, where the common setting is $K_d = k_1 \cdot (b \cdot dl/avgdl + (1-b))$. We also set $K_d = 1$ to observe the effect of the BM25-TF on performance.

We used $TF_{\text{K(b=0.25,k1=1.2)}} \cdot$ IDF (which corresponds to BM25 with no relevance information), LM with Dirichlet smoothing and the combination LM+TF_K-*IDF* as baselines. The parameters b, k_1 and μ_D were set to 0.25, 1.2 and 2000, respectively, while μ_Q was set to the average query length. The aforementioned settings were applied across all of the collections, i.e. the retrieval models were not tuned per collection.

For LM+TF_K-*IDF* we used two methods to combine LM and TF_K-IDF inspired by [Larkey and Croft, 1996], and for each method we use two normalisation scheme. The first method is based on *adding* the normalised scores of the documents retrieved by both LM and TF_K-IDF. The normalisation was done either by dividing each individual score by the maximum score for each retrieval model or by dividing by the sum of the scores for each model. The other combination was performed by *multiplying* the normalised scores which were retrieved by both retrieval models. The normalisations were applied in a similar fashion as for the first method.

7.2 Results and Analysis

Table 3 shows for selected models the MAP and P@10. The performance of the TF-IDF with independence assumption, where $TF(t,d) = tf_d$, was omitted since too poor to be considered as a baseline (MAP in average was one third of the MAP achieved by TF_K-IDF). Similar observations were made for D2Q2 with independence assumption, and as such the corresponding results are omitted.[2]

The setting $TF_K := tf/(tf + K)$ was instrumental in achieving competitive retrieval performance, and hence we report only results for this setting. We discussed the notion of "semi-subsumed" events which embeds the BM25-TF into D2Q2. In D2, $TF_K(t,q)$ is applied whereas in Q2, it is $TF_K(t,d)$. D2Q2-extreme has no mixture parameters, whereas for D2Q2-linear, the parameter μ_D controls the Dirichlet mixture parameter λ_d (and μ_Q controls λ_q). The overall result is expressed by the relative distance between models (last row of Table 3).

Overall, most candidates deliver about the same performance, with marginal differences among the top candidates. Only one score combination (multiplication of normalised LM and TF-IDF scores) is a poor outlier. The D2Q2 family of models has in half of the cases (5 of 10 benchmarks) the best performer. Some members of the D2Q2 family performed better than others, where in tendency, the linear mixtures are better than extreme mixtures. This is as expected, since the extreme mixtures rely on assumptions that neglect the Dirichlet mixture parameter.

We ran statistical significance tests based on Student's paired t-test with confidence levels $\alpha = 0.01$ and $\alpha = 0.05$. In all cases, the results for the best D2Q2 model and the best traditional model were not significantly different. On one hand, this confirms the reasonable performance of the D2Q2 models. On the other hand, if we had expected an improvement from devising a new model that consists of the inner organs of LM and TF-IDF, then we are disappointed, since the single models perform already relatively well on their own.

Overall, the experimental results show that the hybrid D2Q2 performs within the main-fold of the retrieval quality reported for the baselines. Regarding the comparison of the *score aggregation* LM+TF_K-*IDF* versus the *hybrid* D2Q2, the score aggregation is outperformed by the hybrid (except for TREC-2 where the difference is marginal). In the light of the aforementioned expectation that combining two models delivers the averaged quality, the performance of D2Q2 underlines the effect of hybridity. This supports the conclusion that D2Q2 combines the LM and TF-IDF features such that a *micro* combination of probabilities performs better than a *macro* combination of scores as expressed by LM+TF_K-*IDF*.

D2Q2 shows a stable performance that is marginally better than the baselines, but D2Q2 does not significantly outperform the baselines. The experiments confirm the rationale underlying D2Q2, a framework that encompasses LM and TF-IDF, and their combinations. In particular, D2Q2 truly combines the LM and TF-IDF features into a theory based on probabilities, exhaustiveness and specificity.

8 Conclusions

This research was motivated by investigating the relationship between LM and TF-IDF to attempt to provide answers to statements such as "we know *why* TF-IDF works, and we know *that* LM works, but we do not know *why* LM works". By developing a side-by-side derivation of LM and TF-IDF, a framework based on $P(q|d) \cdot P(d|q)$ emerged, which we named D2Q2. The main contribution of this paper is the theory that underpins the probabilistic framework D2Q2, where the D2 side is LM, and the Q2 side is TF-IDF. This theory reveals the link between LM and TF-IDF, and the D2Q2 framework shows how the features of both models can be combined in a theoretically sound manner. In addition, D2Q2 shows comparable retrieval performance to competitive baselines, making D2Q2 to be not just another unifying framework but a retrieval model in its own right.

Our emphasis was on LM and TF-IDF. Future work will elaborate on the relationship between BM25 and D2Q2. D2Q2 establishes a balanced view on LM and TF-IDF, and this can potentially lead to a consolidated anatomy of the models, viewing LM and TF-IDF as the models for missing relevance, and devising BM25-D2 (an LM-based BM25) and BM25-Q2 (TF-IDF-BM25) as relevance models.

References

[Aizawa, 2003] Akiko Aizawa. An information-theoretic perspective of TF-IDF measures. *IP&M*, 39:45–65, 2003.

[Bartell *et al.*, 1994] B. T. Bartell, G. W. Cottrell, and R. K. Belew. Automatic combination of multiple ranked retrieval systems. *ACM SIGIR*, pages 173–181, 1994.

[Church and Gale, 1995a] K. Church and W. Gale. Inverse document frequency (IDF): A measure of deviation from Poisson. *Workshop on Very Large Corpora*, pages 121–130, 1995.

[Church and Gale, 1995b] K. Church and W. Gale. Poisson mixture. *Natural Language Engineering*, 1(2):163–190, 1995.

[Croft *et al.*, 1990] W.B. Croft, R. Krovetz, and H. Turtle. Interactive retrieval of complex documents. *IP&M*, 26(5):593–613, 1990.

[2] We could however notice that the independence assumption was less detrimental for the D2 (LM) side than for the Q2 (TF-IDF) side. This is because for D2, the assumption is for the query ($TF(t,q)$), which usually contains only few multiple occurrences of terms.

	TREC-2		TREC-3		TREC-8		WT2g		Blog06	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
$LM_{\text{Dir},\mu=2000}$	18.02	41.20	22.87	48.20	21.48	40.00	29.85	46.20	29.21	60.80
$TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$	**18.90**	*42.80*	**25.05**	*50.20*	22.31	40.20	*31.42*	**49.20**	*30.27*	**63.40**
LM+TF-IDF										
$LM_{\text{Dir},\mu=2000} + TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Max norm)	18.65	**44.40**	24.46	49.20	*22.45*	*41.40*	31.15	47.40	29.56	61.20
$LM_{\text{Dir},\mu=2000} + TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Sum norm)	18.72	44.00	24.30	49.60	22.38	*41.40*	31.16	47.20	29.85	62.80
$LM_{\text{Dir},\mu=2000} \cdot TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Max norm)	13.56	43.20	18.80	47.60	19.35	41.00	26.93	47.00	27.10	60.00
$LM_{\text{Dir},\mu=2000} \cdot TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Sum norm)	6.58	5.80	6.39	5.20	6.65	4.60	6.54	5.60	21.97	32.80
D2Q2										
$\text{D2Q2}_{\text{extreme},TF_{\text{K=1}},TF(t,q)=1}$	17.59	42.80	23.00	47.00	**23.16**	42.40	31.92	45.20	29.44	55.80
$\text{D2Q2}_{\text{extreme},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=0.5}$	16.89	38.40	20.73	40.20	22.08	42.00	31.74	46.40	29.22	58.00
$\text{D2Q2}_{\text{extreme},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=1}$	17.24	40.00	24.17	49.80	22.65	**43.60**	28.80	44.80	28.23	54.60
$\text{D2Q2}_{\text{linear},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=1,\mu_D=2000}$	18.48	44.00	24.81	**51.00**	22.52	42.20	31.36	48.00	29.85	62.40
$\text{D2Q2}_{\text{linear},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=1,\mu_D=2000,\mu_Q=avgql}$	18.60	44.20	24.83	50.40	22.59	42.20	31.38	48.00	29.87	62.20
$\text{D2Q2}_{\text{linear},TF_{\text{K=1}},TF(t,q)=1,\mu_D=2000,\mu_Q=avgql}$	17.13	41.20	20.81	40.20	21.68	40.60	**32.03**	46.60	**30.40**	61.80
best D2Q2 - best traditional	-0.30	-0.20	-0.22	0.80	0.71	2.20	0.61	-1.20	0.13	-1.00
relative difference	-0.01	-0.004	-0.008	0.01	0.03	0.05	0.02	-0.02	0.004	-0.01

Table 3: MAP and P@10 (Best traditional model *italicised*, best overall model in **bold).**

[Croft, 2000] W. Bruce Croft. *Combining Approaches to Information Retrieval*, IR 1, pages 01–36. Kluwer Academic Publishers, 2000.

[Fang and Zhai, 2005] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. *ACM SIGIR*, pages 480–487, 2005.

[Frei *et al.*, 1996] H.P. Frei, D. Harmann, P. Schäuble, and R. Wilkinson, editors. *ACM SIGIR*, 1996.

[Gale and Church, 1991] William A. Gale and Kenneth Ward Church. Identifying word correspondences in parallel texts. *HLT*, 1991.

[Gao *et al.*, 2004] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. *ACM SIGIR*, pages 170–177, 2004.

[He and Ounis, 2005] Ben He and Iadh Ounis. A study of the dirichlet priors for term frequency normalisation. *ACM SIGIR*, pages 465–471, 2005.

[Hiemstra, 2000] Djoerd Hiemstra. A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139, 2000.

[Hou *et al.*, 2011] Y. Hou, L. He, X. Zhao, and D. Song. Pure high-order word dependence mining via information geometry. *Advances in Information Retrieval Theory ICTIR*, pages 64–76. 2011.

[Kwok, 1996] K. L. Kwok. A new method of weighting query terms for ad-hoc retrieval. I*ACM SIGIR*, pages 187–195, 1996.

[Lafferty and Zhai, 2003] John Lafferty and ChengXiang Zhai. *Probabilistic Relevance Models Based on Document and Query Generation*, chapter 1. Kluwer, 2003.

[Larkey and Croft, 1996] L. S. Larkey and W. B. Croft. Combining classifiers in text categorization. [1996], pages 289–297.

[Lee,] Joon Ho Lee. Analyses of multiple evidence combination. *SIGIR Forum*, 31(SI):267–276.

[Metzler and Croft, 2004] Donald Metzler and W. Bruce Croft. Combining the language model and inference network approaches to retrieval. *IP&M*, 40(5):735–750, 2004.

[Nie, 1992] J.J. Nie. Towards a probabilistic modal logic for semantic-based information retrieval. *ACM SIGIR*, pages 140–151, 1992.

[Ounis *et al.*, 2006] Iadh Ounis, Craig Macdonald, Maarten de Rijke, Gilad Mishne, and Ian Soboroff. Overview of the trec 2006 blog track. *TREC*, 2006.

[Ponte and Croft, 1998] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. *ACM SIGIR*, pages 275–281, 1998.

[Robertson *et al.*, 1994] S. Robertson, S. Jones S. Walker, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. *Text REtrieval Conference*, 1994.

[Robertson, 2004] S.E. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:503–520, 2004.

[Roelleke and Wang, 2006] Thomas Roelleke and Jun Wang. A parallel derivation of probabilistic information retrieval models. In *ACM SIGIR*, pages 107–114, 2006.

[Roelleke and Wang, 2008] Thomas Roelleke and Jun Wang. TF-IDF uncovered: A study of theories and probabilities. In *ACM SIGIR*, pages 435–442, 2008.

[Salton *et al.*, 1976] G. Salton, A. Wong, and C.T. Yu. Automatic indexing using term discrimination and term precision. *IP&M*, 12:43–56, 1976.

[Singhal *et al.*, 1996] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalisation. [1996], pages 21–39.

[Taylor *et al.*, 2006] Michael Taylor, Hugo Zaragoza, Nick Craswell, Stephen Robertson, and Chris Burges. Optimisation methods for ranking functions with multiple parameters. *ACM CIKM*, 2006.

[Wong and Yao, 1995] S. K. M. Wong and Y. Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst.*, 13(1):38–68, 1995.

[Wu and Roelleke, 2009] Hengzhi Wu and Thomas Roelleke. Semi-subsumed events: A probabilistic semantics for the BM25 term frequency quantification. *ICTIR (International Conference on Theory in Information Retrieval)*, 2009.

[Wu *et al.*, 2008] Ho Chung Wu, Robert Wing Pong Luk, Kam-Fai Wong, and Kui-Lam Kwok. Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3), 2008.

[Zaragoza *et al.*, 2003] Hugo Zaragoza, Djoerd Hiemstra, and Michael Tipping. Bayesian extension to the language model for ad hoc information retrieval. *ACM SIGIR*, pages 4–9, 2003.

[Zhai and Lafferty, 2001] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. *ACM SIGIR*, pages 334–342, 2001.

[Zhai and Lafferty, 2004] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

Expert search in semantically annotated enterprise data: integrating query dependent and query independent relevance factors

Felix Engel, Matthias Juchmes, Matthias Hemmje
Distance University Hagen
Felix.Engel, Matthias.Juchmes, Matthias.Hemmje@fernuni-hagen.de

Abstract

The documentation of processes or employee- and product-related-data in the enterprise does comprehensively contribute to the preservation and future access to acquired in-house knowledge. Sophisticated access to this data is an essential part of successful knowledge management. With the increasing use of semantic web recommendations and technologies in enterprise new challenges and opportunities concerning data access arise. Search for experts in documented enterprise data has been a famous research topic for years. A major reason for this is its beneficial impact on accessing existing enterprise potential. Thus search for experts is a valuable application for the enterprise advancement.

However, recent expert search systems largely implement relevance ranking on basis of topic relevance between a potential expert and the topic of the query. Nevertheless, even though the query topic as relevance evidence source has been proven as one of the most important factors in expert search; it only reflects the relevance between the query topic and the closeness of an expert to the topic. The analysis of further evidence sources is part of researches in the field of expertise seeking. Such research results are rarely taken into account in recent expert search implementations. The provision of comprehensive semantic annotations in enterprises opens new potential and challenges for the implementation of sophisticated expert search systems, taking into account not only topic closeness.

1 Introduction

SMART VORTEX[1] is an integrated project co-financed by the European Union within the 7th Framework Program. The project objective is the provision of an extensive set of intelligent and interoperable tools, methods and services for the management of massive data streams alongside the whole product lifecycle spanning from product idea generation, design, manufacturing and service to product disposal. Within this objective one focus is on supporting collaboration of people involved in the product lifecycle. Part of this objective is the identification of in house experts for collaboration initiation.

Recent expert search systems calculate expert relevance on basis of topic closeness. Beside topic closeness, various additional evidence sources are part of a holistic expert relevance ranking calculation. Those findings are part of studies in the field of expertise seeking. In SMART VORTEX enterprise data encompasses models of various entities such as for instance people, products and their interrelation in the enterprise. As a common basis for modeling and data representation in SMART VORTEX recent semantic web recommendations and tools are applied. In fact the present data represents a semantic information integration of various existing heterogeneous data sources such as ERP systems, PLM systems as well as employee- or product information data bases among others. Various person features are implicit part of this semantic enterprise graph and valuable for the representation of diverse expertise seeking evidence sources.

1.1 Problem statement

Finding experts within an enterprise for any kind of problem is a complex and time consuming task. Few isolated applications for expert finding exist, however the demand for comprehensive solutions keeps on being a non-trivial task. According to Balog et al. [Balog *et al.*, 2012] in the field of expert search a clear distinction between the two fields of expertise retrieval and expertise seeking could be made.

Expertise retrieval includes all content-related approaches that process a document database using information extraction and data mining techniques, among others. The processed data in this case could be searched subsequently with the aid of known information retrieval algorithms. In contrast to the content-related approaches, researches in the field of expertise retrieval analyze all further evidence sources that lead to the decision if the potential expert is relevant from a user perspective. Such evidence sources include for instance the freshness of knowledge, experience, reliability or social closeness. However, recent expert search applications widely realize expertise retrieval approaches and rarely take into account results from expertise seeking researches [Hoffmann *et al.* 2012 and Balog *et al.*, 2012]. Various features relevant for expertise seeking are implicit part of the SMART VORTEX semantic enterprise graph. In order to use these implicitly modeled features in ranking tasks they need to be computable. This could be realized through the application of known algorithms in the field of the semantic search or through simple functions basing on graph functions. Relevance calculations in the field of semantic web are for instance the calculation of popularity, rarity or association length approaches. Generally such features are independent from the query itself. Query dependent calculations similar to Albertonie et al. [Albertonie *et al.*, 2006] that base on the specification of a path in the semantic graph in contrast are query dependent. A query dependent

[1] http://smartvortex.eu/

evidence source could be for instance the number of connections to enterprise roles with specific constraints (e.g. only management or service roles) or the freshness of knowledge given a specific query topic.

However query dependent and independent calculations only make statements about the graph structure. The resulting assessment is dependent on the task at hand and has to be given retrospectively. Furthermore, the assumption that several features are part of the overall relevance assessment lead to the problem of meaningful aggregation of features into one ranking function. Aggregation of relevance calculation constituents is a common problem in retrieval tasks. Learning to rank is an approach that has recently been applied for similar problem statements.

1.2 Objectives

The aim of the work introduced in this short paper is the development of an expert search approach in a semantically annotated enterprise knowledge base. The approach should integrate various sources of expertise evidence beyond content-related proximity. To reach this goal the approach shall take into account the results of various expertise seeking investigations in order to enhance expert relevance calculation in the sense of expertise seeking findings. Calculation of evidence could be dependent- or independent from a query and should take into account existing relevance calculation approaches from research in the field of the semantic web. These various evidence calculations must in the end be aggregated and assessed according to the relevance aspects to be fulfilled.

2 Sources of evidence in expertise seeking tasks

The research areas of expertise seeking and information seeking are closely related. Expertise seeking investigations take a user centric perspective in an expert search task. The focus of these investigations is the analysis of those evidence sources that are crucial for choosing an expert from a user point of view.

Karunakaran et al. [Karunakaran *et al.*, 2012] emphasize the physical proximity of an expert, especially under the consideration of the degree of acquaintanceship. Woudstra et al. [Woudstra *et al.*, 2008] as well as Helms et al. [Helms *et al.*, 2013] consider this finding as part of an access related aspect. Especially the influence of social factors with varying characteristics is part of expertise seeking investigations. Yuan et al. [Yuan *et al.*, 2007] emphasize that social closeness between people in particular is valuable for expert search, because user and expert are unbiased in their communication. Woudstra et al. respond in their investigation to quality related factors like e.g. the actuality of acquired knowledge or the reliability of a potential expert.

Some of the mentioned aspects like e.g. the degree of acquaintanceship in a semantically annotated knowledge base could be calculated via famous semantic web techniques such as for instance popularity. Popularity calculates the degree of connectivity in the graph. Such calculations are independent from the query itself. Other sources of evidence cannot be calculated by these well-known relevance measures. In the case of approachability [Woudstra *et al.*, 2008] for instance, the relevance of an expert candidate can be calculated by the fact that he is part of the same working group, project or else. This condition is query dependent and could not be calculated by known semantic web relevance measures.

2.1 Query dependent relevance calculations

Query dependent calculations could be characterized by the fact that they could only be calculated based on the query itself. The calculated value in this respect describes a proportion to a query on base of specified basic conditions. Specification of such conditions in a semantically annotated knowledge base demands knowledge about the representation and relation between modeled entities. Since this knowledge is not explicitly part of the model itself, it is external. A considerable similar problem statement and approach has been published by Albertonie et al. [Albertonie *et al.*, 2006] in order to calculate the similarity between instances of a semantic knowledge base. Albertonie et al. have applied simple calculation units specifying paths and a similarity function. A query dependent calculation in this sense is for instance the amount of relations between an expert and the topics of the search query.

2.2 Query independent relevance calculations

Plenty of the applied relevance measures in the semantic web community are graph based algorithms. Such relevance measures are inspired by findings in the field of graph theory. A famous measure e.g. is *popularity*, which measures the amount of in- and outgoing links of an instance. Furthermore, the *association length* analyses the length between instances or *subsumption* which takes into account the taxonomic graph structure. The problem with such measures is that their result is depending on the task at hand. For instance a long *association length* could be interesting because it identifies an unobvious relation between instances. On the other hand shortest paths could be preferred, because they reflect a tight coupling of instances. Same holds true for the *popularity* measure. Here an instance with lots of relations could be relevant because of its high connectivity, but on the other hand an instance with few connections is specific and hence could be relevant. All of these measures are based on the graph structure itself and can be calculated independent from the search query.

3 Rules for configurations of interdependencies between relevance calculations

As stated above, the relevance degree of a query independent measure has to be assessed regarding the search task at hand. The same also holds true for the query dependent measures. In contrast to a general purpose entity search, in the scope of this work it is clear if a high or a low measure value indicates relevance or irrelevance. For instance, if the aim of the search is to find an expert as a course leader it might be of relevance if the potential expert already has course leader experience. This fact could be inferred by counting the number of course leader roles one has already taken. On the other hand it might be better to find a potential expert with few active roles to find someone with appropriate time capacities.

In order to illustrate the approach described above, the example search for a course leader is introduced. Following sources of evidence (SE) are part of the search:

- **SE 1:** How good is the potential experts (P) insight in enterprise processes? *Expertise* [Heath *et al*., 2006]
- **SE 2:** Does the potential expert match the query topic exactly or more specifically? E.g. in a query with the topic ObjectOrientedProgramming, an expert matching this topic exactly will be preferred over an expert with more specific knowledge (e.g. Java), because the course will introduce general concepts of object oriented programming as opposed to concepts specific to Java. *Topic of knowledge* [Woudstra *et al*., 2008]
- **SE 3:** A high number of connections of the potential expert in the enterprise should be preferred, because if the potential expert is well connected in the enterprise, it could be stated that he has a good standing. Nevertheless, besides good standing, a tight coupling between user and expert is of importance. Among others *Familiarity* [Woudstra *et al*., 2008]

In this example the search for an expert shall be evaluated as the sum of the above three evidence calculations. The calculation of these sources of evidence can be implemented as follows. Source of evidence 1 can be calculated by simply counting the enterprise roles a potential expert has already taken. This approach is pretty similar to the *count* function definition by Albertonie et al. The assumption is that the more roles a potential expert has taken the better he knows internal enterprise processes. Source of evidence 2 can be calculated by applying subsumption. In this application a more general result is be preferred. Source of evidence 3 spans two calculations. The degree of a potential expert connection can be calculated by the *popularity* measure. The tight coupling between user and expert is measured through application of the association length measure. In this application shortest paths are preferred.

Based on the above assumptions the search application needs a function to count how often a relation between potential expert and enterprise roles exist. Furthermore, the functions *subsumption, popularity* and *association lentgh* are part of the whole calculation. Hence, the above mentioned calculations are aggregated through the definition of the following person feature vector:

$$\begin{pmatrix} \text{feature } 1: \textit{count} \\ \text{feature } 2: \textit{subsumption} \\ \text{feature } 3: \textit{popularity} \\ \text{feature } 4: \textit{associationLength} \end{pmatrix}$$

Two sample instances of above feature vector could be as follows:

$$P_1 = \begin{pmatrix} 3 \\ 0.7 \\ 0.8 \\ 0.2 \end{pmatrix}; P_2 = \begin{pmatrix} 1 \\ 0.5 \\ 0.9 \\ 0.6 \end{pmatrix}$$

Given these sample instances of feature vectors, it is obvious that the calculated values just express the values of the applied functions. To fully support the source of evidence described above, rules have to be applied in order to make a statement about how well a calculated value supports the relevance of potential experts. In this sample application, a potential expert with a high value related to source of evidence 1 should be preferred. The following rule supports this statement: if $\{feature1_{p1} > feature1_{p2} \rightarrow P_1\}$ else $\{P_2\}$. However, an expert is even more relevant if the value of source of evidence 2 is low. This could be expressed by the rule: if $\{feature2_{p1} < feature2_{p2} \rightarrow P_1\}$ else $\{P_2\}$. The calculation of source of evidence 3 is more complex, because it is composed of two sub calculations. The following rule expresses the required statement: if $\{(feature3_{p1} > feature3_{p2})$ AND $(feature4_{p1} < feature4_{p2}) \rightarrow P_1\}$ else $\{P_2\}$.

The aggregation of these query dependent und independent features via rules apparently is a promising approach to express expertise seeking evidence sources. In fact the application of rules for the assessment of query dependent and independent feature calculation can be regarded as the description of a relevance pattern. To calculate a ranking model from a relevance pattern definition like that defined by above rules, the application of learning to rank is promising.

4 Application of learning to rank for relevance pattern learning

Learning to Rank (LTR) is an application in the research field of machine learning. LTR is used to learn a relevance ranking model of objects that are represented by relevance labeled feature vectors. In fact LTR learns a relevance pattern. Those learned ranking models are coefficients of a ranking function that calculates a relevance value for an object from its feature values. A machine learning algorithm like Support Vector Machines is applied to analyze the training data with the aim to find an appropriate model based on the data. Hence, a good model does not only match the rankings represented by the training data, but can be applied to general search queries not part of the training data set.

Liu [Liu, 2009] distinguishes between the three learning approaches *pointwise*, *pairwise* and *listwise*. The chosen approach influences the structure of the training data, and thus also the machine learning algorithms used to analyze this data. To date LTR is often applied in document retrieval tasks, like in Joachims, 2002 [Joachims, 2002]). Recently, some researches have been made that apply LTR in semantically annotated knowledge bases. Dali *et al.* [Dali *et al.*, 2012] use LTR to learn a ranking model for the aggregation of query-independent relevance measures in semantic databases. Features in this case include *popularity* related calculations. Labels for the test data are gathered by crowd sourcing among others. Fujita *et al.* [Fujita *et al.*, 2012] use LTR to recommend queries that are semantically similar to the original query. Chen et al. [Chen *et al*., 2011] apply LTR to rank relationships in RDF graphs. In this approach LTR is used in order to learn the user's preference based on various graph measures like *association length* or *popularity*. However, LTR-techniques include approaches which learn a ranking model based on labeled training data. Hence, critical requirement for each application that make use of an LTR approach is the existence of test data annotated with rele-

vance labels. Generally, relevance labeling is done by experts or collected through crowd sourcing. The disadvantages of these approaches are the high costs and high failure rates.

However, in the application described here the relevance pattern is already known and described through rules (c.f. section 3). Hence, test data labeling in this case doesn't have to be realized by experts or else but by the evaluation of rules.

The following approach is conceivable for test data labeling based on rules as introduced above, in a pairwise LTR application. In a pairwise LTR setting feature vector instances are treated in pairs. Each pair is sorted into one of two classes if possible, depending on which of the vectors is more relevant. If no such decision can be made, the pair is not classified. Thus, algorithms for this approach have to solve a binary classification problem. The above defined rules are evaluated for each pair of feature vector instances as follows: Each possible pair of feature vector instances has to be evaluated given the above described rules. The evaluation result for each rule votes for one of the two feature vector instances. Two results of this voting approach are possible. In the case that one of the two vectors has more votes than the other, the vector with more votes is labeled as more relevant. In case of a tie, both vectors are too similar and thus can't be taken into consideration for the learning process.

The example feature vector instances (P_1, P_2) are evaluated on basis of above rules as follows:

- SE 1: $3 > 1$, votes for P1
- SE 2: $0.7 < 0.5$, votes for P2
- SE 3: $(0.8 > 0.9)$ AND $(0.2 > 0.6)$, votes for P2

The result of the evaluation is one vote for P1 and two votes for P2. Hence in a pairwise LTR approach feature vector instance P2 is labeled as more relevant as P1. Given a reasonable amount of those test data LTR is able to construct a relevance ranking model that reflects the relevance aspects described through rules.

5 Summary and outlook

This short paper introduced an approach for the integration of query dependent and independent relevance measures in a semantically annotated knowledge base, for the integration of expertise seeking parameters in an expert search task. The described approach aggregates several sources of evidence for the task of expert search going behind pure topic based relevance ranking. The application of rules as specification of a relevance pattern to be learned is the input for an LTR approach that learns a ranking model for unseen queries.

Open questions among others are the evaluation of this approach and hence which expertise seeking parameters can be calculated. Which of these parameters are dependent on a registered user and which can be calculated without registered users? With respect to the LTR application it is crucial to evaluate the dependency between size of database, required amount of training data and dimension of the feature vector.

Acknowledgments

This work has been partly supported by the FP7 EU project SMART VORTEX.

References

[Albertonie *et al*., 2006] Albertoni, R., & De Martino, M. (2006). Semantic similarity of ontology instances tailored on the application context. On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE, 4275, 1020–1038.

[Balog *et al*., 2012] Balog, K. Fang Y., de Rijke, M., Serdyukov, P. und Si, L. (2012). Expertise Retrieval. Foundations and Trends® in Information Retrieval, 6(2-3), 127–256. doi:10.1561/1500000024

[Chen *et al*., 2011] Chen N. & Prasanna, V.K., 2011. Learning to Rank Complex Semantic Relationsships Technical Report., (November). Available at: http://www-scf.usc.edu/~nchen/paper/ltr.pdf.

[Dali *et al.*, 2012] Dali, L., Fortuna, B., Duc, T. and Mladenić, D. (2012). Query-Independent Learning to Rank for RDF Entity Search. In The Semantic Web: Research and Applications, 484-498

[Fujita *et al.*, 2012] Fujita, S., Dupret, G., & Baeza-Yates, R. (2012). Learning to Rank Query Recommendations by Semantic Similarities. *arXiv preprint arXiv:1204.2712*.

[Helms *et al.*, 2013] Helms, R., Diemer, D., & Lichtenstein, S. (2011, July). Exploring barriers in expertise seeking: why don't they ask an expert?. In PACIS (p. 77).

[Heath *et al*., 2006] Heath, T., Motta, E., & Petre, M. (2006). Person to person trust factors in word of mouth recommendation

[Hoffmann *et al*., 2010] Hofmann, K., Balog, K., Bogers, T., & de Rijke, M. (2010). Contextual Factors for Finding Similar Experts 1. Journal of the American society for information science and technology, 61(5), 994–1014.

[Joachims, 2002] Joachims, T. (2002). Optimizing Search Engines Using Clickthrough Data. In Proceedings of the Eighth ACM SIG KDD International Conference on Knowledge Discovery and Data Mining, 133-142

[Karunakaran *et al*., 2012] Karunakaran, A. & Reddy, M., 2012. Barriers to collaborative information seeking in organizations. Proceedings of the American Society for Information Science and Technology, 49(1), pp.1–10.

[Liu, 2009] Liu, Tie-Yan. "Learning to rank for information retrieval." *Foundations and Trends in Information Retrieval* 3.3 (2009): 225-331.

[Woudstra *et al*., 2008] Woudstra, Lilian and van den Hooff, Bart. Inside the source selection process: Selection criteria for human information sources. Inf. Process. Manage. May, 2008. Doi 10.1016/j.ipm.2007.07.004

[Yuan *et al*., 2007] Yuan, Y.C., Carboni, I. & Ehrlich, K., 2007. The impact of affective relationships and awareness on expertise retrieval: a multilevel network perspective on transactive memory theory

Towards the Semantification of Technical Documents

Sebastian Furth[1] **and Joachim Baumeister**[1,2]
[1] denkbares GmbH, Friedrich-Bergius-Ring 15, 97076 Würzburg, Germany
[2] University of Würzburg, Institute of Computer Science, Am Hubland, 97076 Würzburg, Germany
{firstname.lastname}@denkbares.com

Abstract

In the domain of engineering large corpora of technical documents are commonly created and used. Applications such as semantic search offer advantages in accessing those documents, but require them to be semantically annotated. Annotating these corpora manually is in most cases not feasible. In recent years a lot of machine learning methods have proved their ability to annotate documents automatically. The downside of these methods is their need for training data. We present a holistic approach for the semantification of technical documents without training data. The approach tackles different challenges such as terminology extraction, semantic annotation, and reviewing. Our approach has been successfully applied to the technical documents corpora of two German machine builders

1 Introduction

Large corpora of technical documents exist in the domain of engineering. In contrast to other corpora they are often multilingual and consist of large, contentually structured and illustrated documents. Examples are operation manuals, installation guides or repair manuals. One of the main characteristics of these documents is the standardized terminology in form of a controlled vocabulary.

Exploiting the information contained in such documents can be useful for a variety of application scenarios. An example of such a scenario is the fast and effective access of information, which can be useful when searching for the repair instructions of a special assembly. *Semantic Search* [Guha *et al.*, 2003] enables such an information access. In contrast to traditional search engines ontologies are used to connect textual content with semantic information which can then be exploited during the retrieval to improve search results.

The connections between text resources and semantic information are created in a process called *Ontology Population* [Buitelaar and Cimiano, 2008], where an ontology structure is filled with instances. These instances describe for example what the main subject (in terms of ontology concepts) of a document is. This is vaguely related to *Subject Indexing* [Hutchins, 1978; Albrechtsen, 1993] which in turn can be considered as part of the more general problem of *Document Classification* [Sebastiani, 2002]. Creating these instances manually requires an in-depth analysis of the underlying documents, which is time-consuming and often cost-intensive.

In the field of *Information Extraction* there exist established methods for the extraction of semantic information from natural language texts. Most of these methods are based on supervised *Machine Learning* approaches, which require a sufficient amount of traning data for good results. In real-world scenarios such training data is often not available and the creation under the cost-benefit ratio not economic. The absence of training data implies in most cases missing test data which leads to a challenge regarding the evaluation, as standard measures like precision, recall and f-measure can not be estimated.

In this paper we present an holistic approach for the automatic semantification of technical documents that does not require training data. We call our approach holistic, as it is an complete process that covers all steps necessary for the semantification of existing technical documents. In our context semantification means the identification and annotation of the main subjects for a given document. The contribution of this paper is a process that relying on well established methods tackles the problem of semantifying technical documents without training data. The remainder of this paper is structured as follows: In Section 2 we give an overview of our approach, Section 3 describes the semantic annotation in detail, Section 4 shows the applicability of our approach in an industrial case study, Section 5 gives an overview of related work while Section 6 shows some future directions regarding our approach before concluding.

2 Process Overview

In this section we give an overview of our approach as depicted in Figure 1. Starting with unstructured technical documents (mainly PDF files) we enrich, segment, and process them in order to reach our goal of semantification. The semantification requires the availability of terminology, which is extracted from various sources. We added an explicit review stage to the process, as the results are in most cases crucial for the performance of target applications and we are not able to evaluate them due to the absence of adequate test data. Reviewed documents are also used as sources for the terminology extraction stage. In a postprocessing stage the data is prepared for target applications.

2.1 Preprocessing

The first stage of our process consists of a series of preprocessing steps. The preprocessing is necessary to prepare the input documents for the semantic annotation. In detail

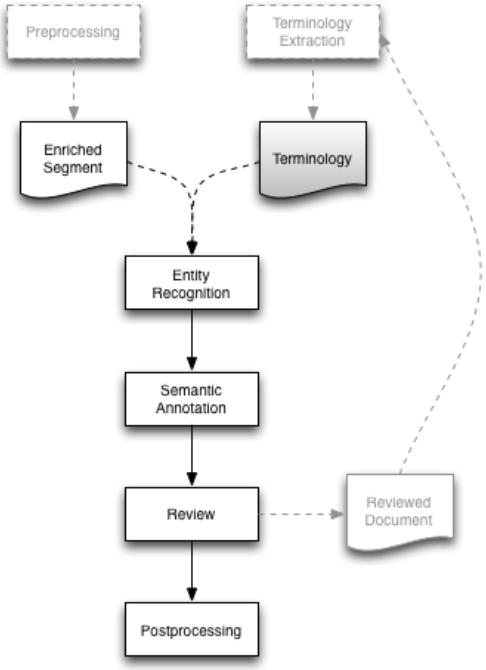


Figure 1: Overview of the semantification process.

the steps of this process stage are (1) the conversion and (2) the segmentation of the input documents as well as (3) the addition of structure to the segments.

As stated before we are mainly confronted with documents in the PDF format. To simplify the further processing we convert all documents to XML. Therefore we evaluated different PDF conversion tools and chose the Xpdf-based tool "pdf2xml"[1], as the generated XML provides a lot of exploitable information about the document's original structure. In order to achieve our goal of identifying the main subjects for each segment, we first need to split the input documents into segments. Depending on the data quality of the input documents different segmentation methods are used, e. g. structural segmentation based on the PDF outline (provided as PDF bookmarks), formatting or lexical analysis. An example for the latter one is the well established TextTiling [Hearst, 1997] approach. Each segment is enriched with structure using different methods from *Natural Language Processing* like *Tokenization*, *Part-of-Speech Tagging* or *Parsing*.

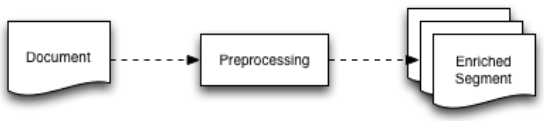


Figure 2: Converting documents to enriched segments.

2.2 Terminology Extraction

A characteristic of technical documents is the usage of a special and relatively fixed and controlled vocabulary. We exploit this characteristic by limiting the set of identifiable subjects to a given set of concepts. Together with related terms they form the *terminology* which is the basis for the semantic annotatation method presented in Section 3. The goal of this processing step is the extraction of the terminology from various sources (see Figure 3).

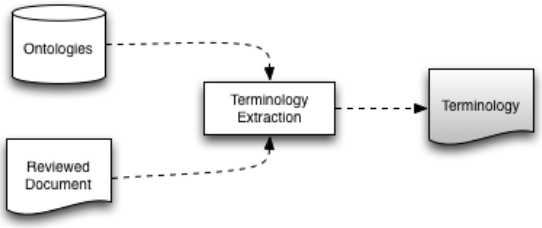


Figure 3: Extracting terminology from different sources.

The set of concepts is derived from the structural description of real world entities like machines. We assume that each concept has a human readable label. These labels are used as the most important element in the set of related terms. This set of terms is complemented by terms derived from concepts that have a relation to our given set of concepts but are not included in the set of identifiable subjects, e. g. assuming that our given set of concepts covers all assemblies of a machine, related concepts could be all parts the assemblies consist of.

A reasonable way to formalize knowledge is the definition of an ontology. There exist a couple of standardized languages for the formalization of ontologies, e. g. RDF(S) [Brickley and Guha, 2004] or OWL [Krötzsch *et al.*, 2012]. Hence it is not surprising that the structural description of real world entities like machines is often provided in the form of an ontology. When confronted with an ontology we use domain-specific SPARQL [Harris and Seaborne, 2012] queries to extract the terminology, i. e. in most cases the labels of concepts. As stated before, our process implies an explicit review step, producing reviewed documents. These documents can also be exploited in terms of terminology extraction.

2.3 Entity Recognition

For each segment we now need to identify occurrences of terminology terms, as our semantic annotation algorithm is based on these terms. So, the extracted terminology is the basis for an entity recognition step. As we are confronted with a controlled vocabulary and thus exactly know what entities (terms) we want to recognize, we use a dictionary-based entity recognition method to identify all occurrences of terminology terms in the segments. At the moment the lookup of terms is based on word stems produced by a standard Porter stemmer [Porter, 1980]. Regarding multi-word terms, we allow order independent matches, i. e. all permutations as well as non-contiguous matches, i. e. ignoring non-matching tokens between tokens belonging to a term.

2.4 Semantic Annotation

After the entity recognition step we are usually confronted with a lot of identified terms, indicating different concepts. For each segment the task is now the inference of the main concepts based on the recognized terms. We use an approach derived from Explicit Semantic Analysis proposed by [Gabrilovich and Markovitch, 2007]. This method will be described in detail in Section 3.

[1] https://sourceforge.net/projects/pdf2xml/

2.5 Review

Depending on the requirements regarding the data quality, we propose a manual review of the results of the semantic annotation by domain experts. As the availability of domain experts is a crucial element in this step, we propose the usage of an appropriate iteractive review tool (see Figure 4 for an example) that helps to decrease the review time.

For our task such a review tool needs to fulfill at least the following requirements: (1) Display the hierarchical segmentation of a specific document, (2) display the main subjects for each segment, (3) allow the addition and deletion of subjects. In order to minimize the review time for each document we additionally propose the usage of a visual component and the highlighting of critical annotations. The visual component should be able to display the semantic similarity of identified subjects, as in technical documents the subject in a sequence of segments often stays constant or at least semantically similar. An example for this claim is a technical document that covers the mounting and unmounting of assemblies. In such a document the probability is high that the corresponding segments of a specific assembly are in a sequence. In the visual component we then expect characteristic patterns like the steps displayed in Figure 4. Additionally, we propose that segments without any annotations or with a lot of semantically unrelated annotations should be automatically detected and highlighted.

There exist various metrics for the computation of semantic similarities. Examples for approaches based on WordNet [Fellbaum, 1998] were proposed among others by Jiang et al. [Jiang and Conrath, 1997] or Lin [Lin, 1998]. These metrics might be adapted due to the specifity of the used terminology.

Figure 4 shows a sample review tool. In the left the title of the current document is displayed and a status for the document (new, in progress, reviewed) can be specified by the reviewer. Below, the hierarchical segmentation of the document is displayed in a tree view element. The tree view can be used for checking and navigating through the segmentation. Clicking on an element in the tree view loads the information regarding the semantic annotations for the selected segment. The loaded information is displayed in the right part of the application. In the upper part a visual component (Visual Report) displays the results based on semantic similarity[2]. Missing annotations are indicated using a red placeholder. At the bottom of the right part detailed information (Details) about the semantic annotations are available. They can be accessed by scrolling the view or by clicking on a data point in the visual component. For a thorough review it may be necessary to look up the text of a segment, thus we provide direct access to the text in the original document. The detail view also provides possibilities for the addition and removal of concepts.

2.6 Postprocessing

The final step in the proposed process is concerned with postprocessing tasks. Such tasks typically handle the resource preparation for the target applications, evaluate the results or apply measurements to the extracted data.

[2]In the example we use taxonomic information for the computation of semantic similarity.

3 Semantic Annotation of Technical Documents

For the identification of the main subjects of a segment we use an approach derived from Explicit Semantic Analysis [Gabrilovich and Markovitch, 2007]. It was originally developed for the determination of semantic relatedness of texts and is based on a semantic interpreter which copes with a fixed set of concepts, representing each of them as an attribute vector of words. The concepts correspond to Wikipedia articles. The words are extracted from the article text and assigned weights using the TFIDF scheme [Salton and Buckley, 1988]. The semantic interpreter is realized as an inverted index that maps each word into a list of concepts in which it appears. When confronted with an input document, the relevance of the concepts contained in the index can be computed by using the semantic interpreter. For each word in the input document the inverted index is asked for the corresponding concepts and their TFIDF weights. The relevance of the concepts is computed by summing up the weights. The result is a weighted vector of concepts, where the top-ranked concept is the most relevant for the underlying document. The semantic relatedness of texts can then be determined by comparing the computed weighted concept vectors.

3.1 Building the Semantic Interpreter

In the presented approach we also use a semantic interpreter. However its purpose is not the determination of semantic relatedness of texts but the identification of the main subjects of a segment. Therefore terms and concepts are extracted from the terminology. Instead of TFIDF weights we use acquired domain knowledge to manually specify the weights, e. g. assuming we have a hierarchy of assemblies, then labels of the direct predecessors and successors of an assembly are weighted higher than the transitive ones. Another example are parts lists where we determine the weight of the parts' labels as a function of the components they are used in, i. e. parts that are used in only one component get the highest weight. In the following let $C = \{c_j\}$ be the set of concepts, $T = \{t_i\}$ be the set of terms, $\langle k_j \rangle$ be an inverted index entry for term t_i, where the weight k_j represents the strength of the association between term t_i and concept c_j.

3.2 Using Document Characteristics for Term Weighting

To determine the main subject of a segment, we first represent a segment as a list of terms. The terms correspond to annotations made by the dictionary-based entity recognition method used in a preceding process stage. In contrast to [Gabrilovich and Markovitch, 2007] we also take document characteristics into account by weighting the terms. We consider several document specific information like relevance in the document (segment frequency)[3], formatting (bold, italics, underscoring) or the position in the segment (headline). In the following let $S = \{t_i\}$ be the segment, and let $\langle v_i \rangle$ be its weight vector, where v_i is the weight of term t_i.

3.3 Ranking Concepts

For each segment we then use the semantic interpreter to get a ranked list of concepts. The ranking is done using the algorithm given as pseudo code in listing 1.

[3]As we split the document in segments, the segment frequency corresponds to the document frequency in other corpora.

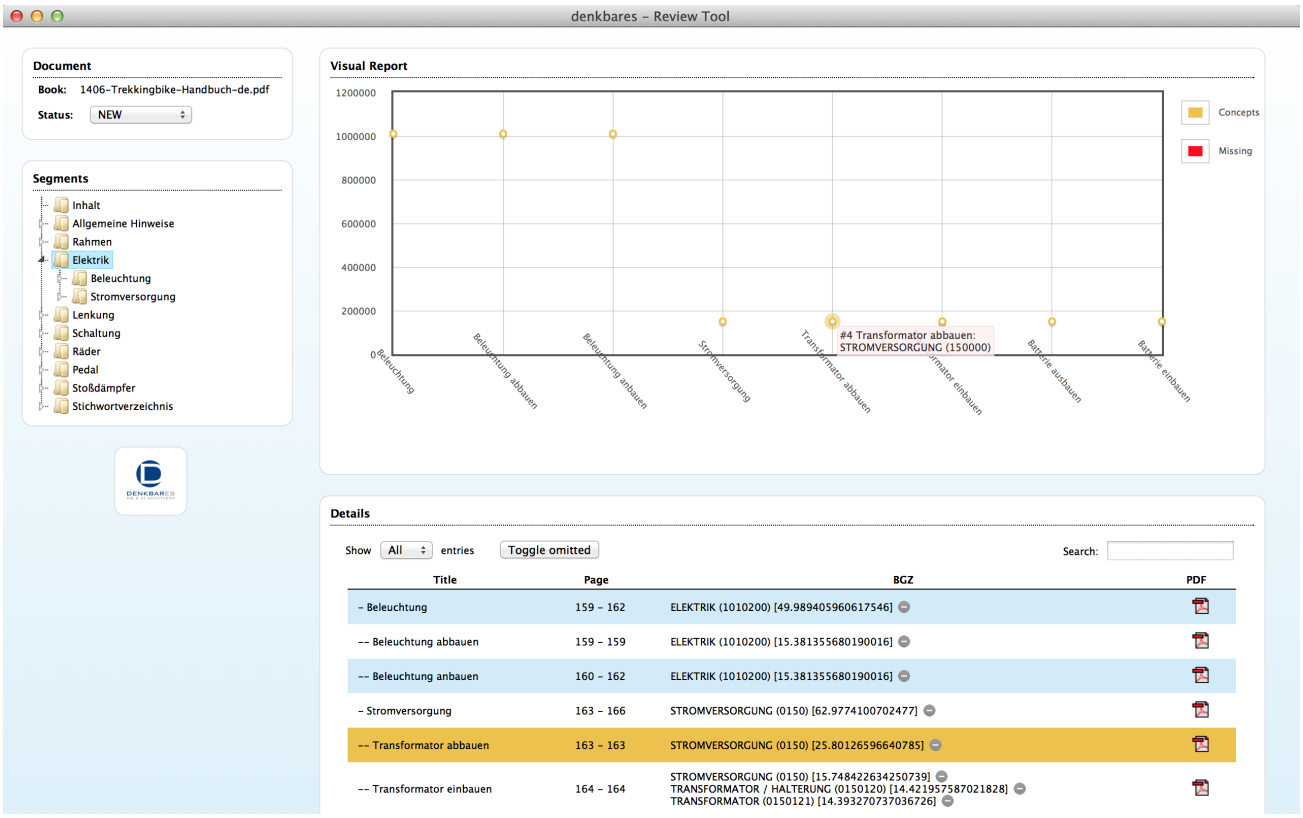


Figure 4: A tool for the manual review of semantic annotations, containing the hierarchical segmentation (left), a visual report (top right) and a detail view (bottom right).

```
getRankedConcecpts(S, ⟨v_i⟩)
 Map<Concept, Double> ranking
 for each t_i in S
  ⟨k_j⟩ = SemanticInterpreter.get(t_i)
  for each k_j
   wtdrelatedness = k_j * v_i
   ranking.update(c_j, wtdrelatedness)
 ranking.sort(WeightedRelatedness, DESC)
 return ranking
```

Listing 1: An algorithm for the term-based ranking of concepts.

The algorithm basically iterates through all terms t_i in a segment S, asks for the inverted index entry $\langle k_j \rangle$ of all concepts c_j related to term t_i and sums up the product of term weight v_i and relation strength k_j, we call it weighted relatedness. The temporary results are saved in a map which gets sorted for the final result in descending order on the weighted relatedness score. This score expresses the relevance of the concepts for the segment, i. e. a higher score means higher relevance.

3.4 Determining the Sprint Group

The algorithm described in the last section produces a ranking of relevant concepts. We now need to identify the most relevant concepts — we call it the sprint group[4].

For the determination of the sprint group we propose two different strategies. The first one simply uses a threshold, the second one is based on statistical outlier tests. So the basic approach for determining the sprint group is taking the score of the most relevant concept. Based on this score we add all concepts to the sprintgroup that are within a specified threshold, e. g. 90% of the highest score. Basically this yields good results, but there are scenarios where it does not fit. An example for such a scenario is when all concepts have low scores, i. e. no concept is really relevant for the segment. Using the basic approach the majority of the concepts would enter the sprint group. To tackle this issue we propose the usage of statistical outlier tests. Using such tests we can determine whether scores exist that offset from the rest. A simple test is for example to compute the interquartile range ($IQR = Q_{75} - Q_{25}$) and then to treat all scores that are higher than $Q_{75} + \alpha * IQR$ as outliers. There are more sophisticated outlier tests like Grubbs' test for outliers [Grubbs, 1969].

[4]Corresponding to the sprint group in cycling races, that offsets against the peloton.

4 Case Study

We have already applied our approach to corpora of two German mechanical engineering companies. In the following we describe the procedure for an engineering company for harvesting technology.

4.1 The data set

The corpus contains about 9000 technical PDF documents, covering different machines. Each document has up to 2000 pages and is of a certain type, e. g. repair manual, operation manual, circuit diagram or installation guide. The documents address different target groups ranging from maintenance staff to end users what influences the structure and the level of detail.

The terminology was mainly extracted from two ontologies. The first ontology describes relations of assemblies,

products, and machines, e. g. that the cylinder block assembly is part of the engine assembly, which itself is a part of a certain product or machine — in the following we will refer to this ontology as *core ontology*. The second ontology describes in detail which parts are build in a special assembly, e. g. that a certain valve is part of the cylinder head — we called this ontology the *parts ontology*. Assemblies and parts have had labels attached as literals using the RDFS property `rdfs:label` and language attributes. We used SPARQL to extract concepts (assemblies) and terms. Concepts were represented using their URI while the labels discribed above were used as terms.

4.2 Processing the corpus

The corpus of technical documents ran through the complete process as described in Section 2. The documents were provided in the PDF format and got converted to XML. Then a segmentation algorithm used the included PDF bookmarks to segment the documents. Structure was added to the produced segments, using a standard whitespace tokenizer and a maximum-entropy part-of-speech tagger. Then a dictionary-based entity recognition algorithm annotated all occurrences of terms extracted from the core and parts ontologies. A semantic intepreter with domain-specific weights (see next section) identified the main subjects of each segment. The results were reviewed using the review tool depicted in Figure 4. The reviewed results were finally converted into an XML format compatible with the target application.

4.3 Weighting term-concept relations

In the following we describe the weighting of the term-concept relations in detail. The $\langle k_j \rangle$ values indicating the strength of the association between term t_i and concept c_j were computed differently for assembly and part terms. For terms extracted from the core ontology we definded the weight as $k_j = \frac{1}{\#edges\ between\ concepts}$, i. e. the label of the concept in focus will get the maximum weight of 1, which means that this label indicates the concept best. Predecessors and successors in the assembly hierarchy got lower weights, e. g. the parents and children got the weight 0.5, grandparents and grandchildren the weight 0.33.

This approach was not feasible for terms from the parts ontology, because there are parts that are semantically different but have the same label (e. g. "valve" or "screw"). The more parts have the same label, the less suitable are they for the inference of a particular concept, i. e. their weight should be adapted accordingly. We decided to define the weight for terms from the parts ontolgy as $k_j = \frac{1}{concept\ frequency}$ where $concept\ frequency$ is the number of concepts that have a part represented by a particular label. This procedure shifts the focus from concepts to labels for terms from the parts ontology. The maximum weight of 1 is assigned to parts that have a unique label and are built in only one assembly. Parts with common labels that are used in a variety of assemblies get lower weights, e. g. parts with the label "screw" are built in more than 500 assemblies, so the weight is as low as 0.002.

4.4 Evaluation

The evaluation of our approach covers the performance regarding the semantic annotation of the segments, i. e. the identification of the main subject. As described above no training or test data were supplied, so we used documents that were reviewed by domain experts using the proposed review tool.

This allowed us to measure different key performance indicators, ranging from precision, recall, and f-measure to the number of corrections that needed to be made by the domain expert. We additionally measured the time needed for the correction of the automatically generated results for a couple of chapters using the proposed review tool.

For the evaluation we selected five documents from the corpus. These documents covered different machines, document types and languages. Table 1 shows the results, where the first three columns correspond to precision, recall, and f-measure and the forth and fifth column show the number of corrections made by a domain expert — the minus (-) indicates the removal of an assigned concept and the plus (+) the addition of a missing concept.

Document	P	R	F	-	+
d1-SYS-de	0,67	1,00	0,80	5	0
d2-RHB-de	0,85	0,87	0,86	16	13
d3-RHB-fr	0,81	0,74	0,77	4	6
d4-RHB-de	1,00	0,92	0,96	0	3
d5-RHB-de	0,77	0,77	0,77	31	30
Overall	0,82	0,83	0,82	56	52

Table 1: Precision, Recall, F-Measure and Number of Corrections.

The results show the overall applicability of our approach. Averaged over the five documents we yield a f-measure of 82%. In these documents 108 corrections needed to be done by the domain expert. As the availability of a domain expert is critical, we also estimated the time needed for a correction. The correction time was measured for randomly selected chapters from the documents above[5]. For each of the selected chapters we measured the number of corrections as well as the total time needed for applying them (see Table 2) — we measured an average correction time of 18 seconds per correction.

Document	# Corrections	∅ Time/Correction
d1-SYS-de (3)	2	22 s
d1-SYS-de (6)	1	20 s
d2-RHB-de (4)	5	8 s
d2-RHB-de (7)	10	16 s
d4-RHB-de (8)	1	28 s
d4-RHB-de (10)	1	16 s
d5-RHB-de (3)	10	20 s
d5-RHB-de (7)	12	14 s
Overall	42	18 s

Table 2: Measuring the correction effort: number of corrections and average time.

5 Related Work

To the best of our knowledge we are not aware of another holistic approach for the problem of the semantification of technical documents, although there exist alternative approaches for single steps of our approach. We use standard methods for the preprocessing, structural enrichment

[5]We left out the French document due to the absence of a French domain expert.

and entity recognition, so we will not consider them in this section, but focus on the terminology extraction, semantic annotation, and the review tool.

Regarding the term extraction, alternative approaches use combinations of statistic, linguistic, contextual or semantic information for the identification and selection of relevant terms, e. g. the C-Value/NC-Value approach [Frantzi *et al.*, 2000] or the TRUCKS-System [Maynard *et al.*, 2008]. As we are confronted with a limited and controlled vocabulary, our approach of extracting terms from ontologies is superior, because we have complete control over the results.

Regarding the semantic annotation, which in our case is the identification of the main subject for a segment or document, latent approaches exist, e. g. Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] or Latent Semantic Analysis (LSA) [Deerwester *et al.*, 1990]. We want to identify a concrete (or explicit) concept, so the latent approaches do not fit for our problem.

Regarding the review of semantic annotations we do not know of another tool for the review of the main subject of a segment or document. Ontosophie [Celjuska and Vargas-Vera, 2004] is a system for the population of an event ontology and uses supervised machine learning for learning extraction rules. These rules also compute a confidence value which is used to determine whether a human reviewer needs to accept an extracted information. The idea of our review tool is to guide a human reviewer through an entire book and highlight critical annotations for rapid correction.

6 Conclusion and Future Work

We proposed a holistic approach for the semantification of technical documents without training data. We defined a process for tackling a couple of challenges, such as terminology extraction, semantic annotation and reviewing. We use standard techniques for the preprocessing and the structural enrichment of the documents. The core of our approach is the semantic annotation which is based on Explicit Semantic Analysis and domain ontologies. This allows for the easy adaptation to new corpora.

We already applied our approach to the techical documents corpora of two mechanical engineering companies. We were able to achieve promising results on these corpora (average f-measure of 82%). We also developed a tool for the manual review and correction of semantic annotations. Experiments with domain experts showed that the average correction time is 18 seconds — which allows for the complete review of a large technical document in a couple of minutes.

For the future we plan to improve the weighting of the term-concept relations. We will investigate different directions: (1) a general applicable weighting scheme, (2) more sophisticated domain specific weighting schemes and (3) the adaption of the weights using the manually reviewed documents. Regarding the review tool we will test other visualization techniques in order to improve the review time and results. We also plan to improve the evaluation of our approach by (1) building or using a public available test corpus and (2) comparing our results to established supervised Machine Learning approaches, using manually reviewed documents as training data.

References

[Albrechtsen, 1993] Hanne Albrechtsen. Subject analysis and indexing: from automated indexing to domain analysis. *Indexer*, 18:219–219, 1993.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Brickley and Guha, 2004] Dan Brickley and Ramanathan V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. *W3C Recommendation*, 10, 2004.

[Buitelaar and Cimiano, 2008] Paul Buitelaar and Philipp Cimiano. *Ontology learning and population: bridging the gap between text and knowledge*, volume 167. Ios Press, 2008.

[Celjuska and Vargas-Vera, 2004] David Celjuska and Maria Vargas-Vera. Ontosophie: A semi-automatic system for ontology population from text. In *Proceedings of the 3rd International Conference on Natural Language Processing (ICON)*, 2004.

[Deerwester *et al.*, 1990] Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[Frantzi *et al.*, 2000] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[Gabrilovich and Markovitch, 2007] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artificial intelligence*, volume 6, page 12, 2007.

[Grubbs, 1969] Frank E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, February 1969.

[Guha *et al.*, 2003] Ramanathan Guha, Rob McCool, and Eric Miller. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709. ACM, 2003.

[Harris and Seaborne, 2012] Steve Harris and Andy Seaborne. SPARQL 1.1 query language. Proposed recommendation, W3C, 2012.

[Hearst, 1997] Marti A. Hearst. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[Hutchins, 1978] William J. Hutchins. The concept of 'aboutness' in subject indexing. In *Aslib Proceedings*, volume 30, pages 172–181. MCB UP Ltd, 1978.

[Jiang and Conrath, 1997] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[Krötzsch *et al.*, 2012] Markus Krötzsch, Peter F. Patel-Schneider, Sebastian Rudolph, Pascal Hitzler, and Bijan Parsia. OWL 2 Web Ontology Language Primer (Second Edition). Technical report, W3C, October 2012.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[Maynard *et al.*, 2008] Diana Maynard, Yaoyong Li, and Wim Peters. NLP techniques for term extraction and ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 107–127, 2008.

[Porter, 1980] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[Salton and Buckley, 1988] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[Sebastiani, 2002] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

NERSENG: Query Analysis and Indexing

Carsten Kropf, Bertram Schlecht
Institute of Information Systems
D-95028, Hof, Germany
{carsten.kropf, bertram.schlecht}@iisys.de

Abstract

This article describes the general setup of NERSENG, a search engine for named entity related web documents. The search engine is, in this case, mainly adopted towards analyzing the documents searching for person names occurring inside the textual parts of crawled documents. We explain the general search engine architecture as well as the occurrence and distribution of entities (person names) in queries and documents. The two major contributions of our work are on the one hand methods to automatically extract entities from unstructured queries and on the other hand an efficient indexing strategy for being able to deliver the search results fast to a query issuer.

1 Introduction

Search engines gained increasing interest over the last years. Major providers, like Google, Bing or Yahoo get billions of requests every day. These standard textual retrieval engines have to efficiently handle access to the fulltexts crawled and indexed before. Therefore, established technologies, like the inverted index can be used to efficiently explore the underlying data space.
These techniques have been investigated for a long time, now, and users can, nowadays, use a sophisticated set of methods to retrieve the desired results. Not only the deep web analysis, also the storage, retrieval and ranking parts of the search engines are very sophisticated, today.
However, in most cases, the standard search engines only search for occurrence of certain terms inside texts. These texts may also be searched for phrases. Yet, most queries are retrieved solely from an inverted index. Structured information, derived in advance, cannot be searched by most search engines, whereas there also exist approaches (like Google Knowledge Graph). However, these structured searches mostly rely on structured information and cannot intermix the data with searches for keywords. Thus, there is the lack of a possibility to search for, e.g., a person name and keywords related to this name.
There already exist methods for extracting structured information from unstructured texts, like named entity recognition (NER) technologies which are able to extract, e.g. a person or company name, from a fulltext part. Likewise, those information might also come from data stored as semantic annotations inside the texts. Based on this, there also exists the possibility to enhance the retrieval process by taking into account these portions of information also for searches.
Searches (except phrase searches) always focus a certain topic. This topic is then intermixed with certain keywords closely related to it. As an example, people searching for "Bill Cutting" (a role of the film Gangs of New York) might also want to know how much money the actor got for this role. In this case, the entity (person name) is mixed together with a query keyword to filter the articles about the person after this keyword, e.g. "Bill Cutting money". Most fulltext search engines would then retrieve data containing the keywords and not necessarily detect the entity name inside this query. For our example the user mostly gets results, which describe how to save money by reducing bills. Only a few of these results is linked to the entity "Bill Cutting".
This paper results from NERSENG (**N**amed **E**ntity **R**etrieval **S**earch **ENG**ine), a web search engine focussed towards exactly these objects of investigation is given. We try to build a search engine supporting simultaneous searches for named entities together with query keywords to construct a more sophisticated searching experience for users.
This paper describes current work carried out with focus on the following parts:

- Statistics of entities in search queries and documents (section 3)
- Description of the search engine (section 4)
- Detection of entities in search queries (section 5)
- Database and indexing scheme used for storing/retrieval of the documents (section 6)

2 Related Work

The index structure used for storing the data for enabling fast search operations is based on a B-Tree (B+-Tree) [Bayer and McCreight, 1972]. Hybrid index structures extending the functionality of base structures to enable fast access to heterogeneous data types have already been proposed for geo-textual application domains. Most of the structures focus towards Geographic Information Retrieval Systems. Examples for these structures are the (M)IR^2-Tree [Felipe *et al.*, 2008] or the bR*-Tree [Zhang *et al.*, 2009]. An overview of currently available and used techniques in the research area of spatial keyword query processing can be found in [Chen *et al.*, 2013]. Retrieval techniques which build the basics of the hybrid index structure, used here, can be found in [Göbel *et al.*, 2009] or [Göbel and Kropf, 2010] which also use Zipf's Law [Zipf, 1949] to distinguish between high and low frequently used terms. However, all of these structures are focussed on spatial in

combination with textual searches whereas this paper focusses on a combination between texts and entities occurring inside the textual parts extracted in advance.
A compressed trie [Morrison, 1968] is used for extracting candidates from the particular search queries.
[Cheng *et al.*, 2007] deals with query construction and ranking of entities in an entity-based search engine, though we do not want a user to learn a new syntax for search queries. [Guo *et al.*, 2009] describes a probabilistic approach for finding named entities in queries. However, we are of the opinion that within a search engine, a statistical approach is slower than using a trie. Within [Kumar and Tomkins, 2009] the behaviour of online search queries is discussed and it is shown that queries can be divided into different classes, e.g. URL-queries.

3 Entities in Queries and Documents

In order to show that an entity-based search engine adds some value compared to a traditional one we have to show that a certain percentage of queries contains an entity, in our case a person. [Kumar and Tomkins, 2009] show that 52.9% of all web queries contain a structured object, e.g. a product, a location or a person. But for our work we are temporarily only interested in the amount of queries containing a person. Since no entity-annotated corpus for search queries is, to our knowledge, freely avaible, we try to approximate this number, with different algorithms explained in the following.

3.1 Approximation of Search Queries with Entities

Named Entity Recognition usually considers the semantic structure of a text. However search queries have in most cases no semantic structure, e.g. from the query "Bill Cutting money" no conclusion can be made whether "Bill Cutting" refers to a person or not. For the approximation of search queries with entites we use an approach which disregards the semantics of the queries. It will be described in the following.
As a data base for the approximation we use the AOL Query Log[1], which includes a total of $\sim$ 36M search queries. Since the corpus contains many duplicate queries and this would influence our measurement results we removed redundant elements and created a list which contains $\sim$ 11M unique queries. Our approach to find persons in queries is a lookup in a name list. We first split queries at whitespaces into single words. After that we check if one of the words is included in the name list. At this point we ignore the capitalization of the single words and names, because web search queries mostly consist of lowercase letters. The list itself was created from the data of the 1990 U.S. Census list of surnames and first names[2]. This method shows the result that 50.64% of the queries contain a name. This outcome is based on the fact that some elements of the name list have ambiguous meanings, e.g. the list includes the last name 'in' which is one of the most common words in english. Due to this we tried to filter the list, removing all words which have ambiguous meanings. Our first approach uses WordNet (see [Miller, 1995]), a lexical database for the English language. We make the assumption that an element of the name list, which is also included in WordNet, has other meanings than just the name and remove the element from the list. Our result with this filtered list is that 13.75% of all queries contain a name. However WordNet includes also names, mostly of famous or historical persons, so our first assumption isn't quite correct and we removed words from the list which are actual names and have no other meaning. Consequently we implemented a second algorithm for filtering the name list (see algorithm 1).

Algorithm 1: filterList(names)

```
   // Generate the filtered name list
   FN from the given census name list
1  for n ∈ names do
       // Generate similary word list
       from WordNet
2      S.add(n)
3      S.add(similarWords(name, α))
       // Check the shape of each element
       of S. If it's a potetial name add
       it to C
4      for s ∈ S do
5          if hasNameShape(s) then
6              C.add(s)
       // Sum up the probability of every
       word in S and C
7      PS = sumProb(S)
8      PC = sumProb(C)
       // Check if the relation between
       PC and PS ist greater then a given
       factor β.  If yes add it to the
       filtered name list
9      if PS != 0 then
10         if PC / PS >= β then
11             FN.add(n)
```

Thereby the method "similarWords" generates a list of similar words from the original name using WordNet. Initially, all synsets (a set of synonyms) of the "name" parameter are loaded from WordNet. Every word of a single synset is checked whether the Levenshtein distance (see [Levenshtein, 1966]) is less than the passed paramter α. In this case, the word is added to a return list which forms the set S together with the original name. A small α parameter allows only words which differ in a few letters from the original word, whereas a high α parameter, e.g. ten, allows completely different words. Thus, a set C is generated, which contains all the potential names of S. A word in S is a potential name if it passes the "hasNameShape" method, which checks if the first character the given word is uppercase and the following characters are lowercase, with the return value true. Then the summed word probabilities of the sets C and S are calculated and stored in the variables PS and PC. For this the word distribution, within the two corpora Reuters TRC2 and RCV1[3], serves as a data base. If the ratio of these probabilities is greater than β, the word is added to the filtered name list. In figure 1 the number of queries that contain entities, for different values of α and β is shown.
We see that the higher the selected values of α and β

[1] http://www.gregsadetsky.com/aol-data/, accessed 2013-06-20

[2] http://www.census.gov/genealogy/www/data/1990surnames/names_files.html, accessed 2013-06-20

[3] http://about.reuters.com/researchandstandards/corpus/, accessed 2013-06-18

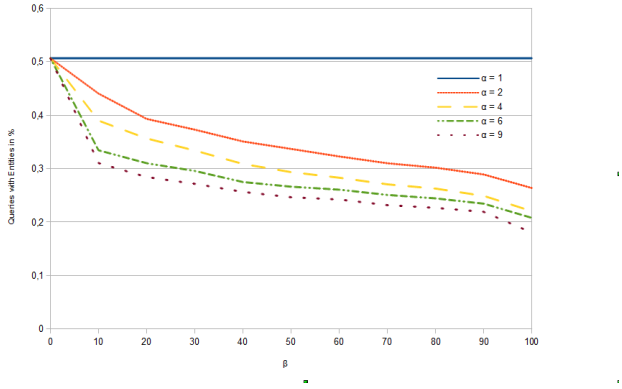


Figure 1: Queries with entities for different α and β values

are, the more names are filtered out of the list. From a certain level of the α-value the number of filtered names does not increase any longer , because every word from every synset is added to S. With a β value of 100% only names are accepted, for which the Sets C and S are identical. We cannot make definitive statements about accurate values of alpha and beta, at the moment. However, we have manually checked 2000 queries and 326 of them contain a person (16.3%). So we imply that a α value of nine and a β value of 95% brings the best results.

3.2 Distribution of Entities to Documents

To determine the distribution of entities to documents the Reuters RCV1 corpus was automatically annotated using the Stanford NER system (see [Finkel *et al.*, 2005]) which is part of the Stanford Core NLP[4]. As models we used the "english-left3words-distsim.tagger" for the part of speech tagger and "english.all.3class.distsim.crf.ser.gz" for the ner system. The corpus contains a total of 806K documents from those 491K include entities, in our case persons. Overall the corpus contains $\sim$ 486K entities. As $\frac{491000}{806000} \approx 61\%$ of the total document corpus contains person names as entities and our work focusses on crawling web sites containing news, we are optimistic that also in a realistic data environment a large percentage of the crawled documents will contain entities.

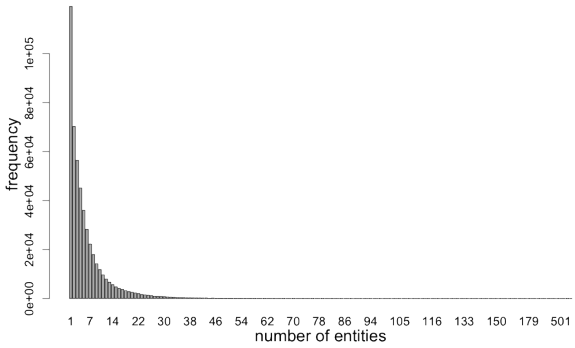


Figure 2: Distribution of Entities to Documents

[4]http://nlp.stanford.edu/software/corenlp.shtml, accessed 2013-06-20

The distribution of entities to documents of Reuters RCV1 dump can be seen in figure 2. The maximum number of entities inside documents is 1064. All documents from Reuters RCV1 that contain persons have an average size of 289.69 words and 5.73 entities. The distribution shows an extreme positive skewness which means that most of the documents contain less than or exactly 7 entities (third quartile). Therefore, we are optimistic that on the one hand building the hybrid index, described later, and (re-)building the trie is sufficiently fast.

4 Search Engine Architecture

As the main contributions of this work, entity detection and hybrid indexing, are embedded in a search engine application, we will describe shortly the process to get documents from the web, analyze the data and store them in a database for retrieval via a search engine.

Figure 3 displays an overview over the entire search engine

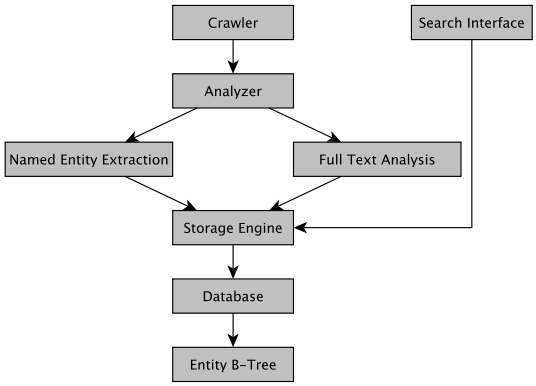


Figure 3: Overview over the Search Engine Architecture

architecture with the particular main contributions of our research work embedded inside. The particular individual parts are described in a little more detail in the following subsections.

4.1 Crawling and Analyzing

The first two parts in the document generation are crawling and analysis of the crawled documents. Therefore, two possibilities are given to retrieve documents from the web. On the one hand, there exists the possibility to crawl the web documents just by seeding a list of uniform resource identifiers (URIs) which is then input into a deep web analysis. Then, the list of URIs to be crawled from each initial URI is generated sequentially and queued to be crawled in a later step. On the other hand, there exists a module to crawl based on RSS feeds. There again, an initial seed of sites is generated (manually) whose RSS information are retrieved repeatedly based on the time restrictions given by each individual site which offers the RSS feed information. The crawler may, based on the given information, therefore download portions of information from the web which are analyzed in a further step. The boilerpipe[5] library is thereafter used to extract the boilerplate free fulltext. This approach serves to derive the "really" relevant information from the crawled sites to remove, e.g. advertisements or link lists from the web page content which is then used for

[5]https://code.google.com/p/boilerpipe/, accessed 2013-06-10

further analysis steps.
The analysis needs to take into account the particular features to be extracted from the texts. These should be prepared in order to store them in the database environment, we will describe in the following. The fulltext part is split into individual terms using normalization (e.g. character normalization and stemming) and thus pre-processed for the use of the hybrid index structure. This process is executed using the Apache Lucene[6] library. The other part, besides the fulltext to be extracted from the crawled web documents are the entities. As described in section 3, the named entities in our case with focus on person names are extracted from the documents using the Stanford NLP library modules with appropriate models to analyze the web articles.

4.2 Indexing Environment

For indexing the data crawled and analyzed in advanced, we use a relational database system. The h2 database[7] is used as the database server, in this case. This section details the setup and adaptions made for the h2 database to run properly for storage of the search engine.
The database itself was extended to allow custom index structures to be set up on tables which are loaded from external places. As the database is implemented in Java, there exists a mechanism to load the index structures from jar files and instantiate them as access methods for certain columns.
For enabling the h2 database to load external index structures from jar files, two tables are introduced which store information about the access structures:

- *INDICES*, storing the information about the index structures available for the database and
- *OPCLASSES*, storing information about so called operator classes which make the index structures, which are generalized, work for specific data types.

Table	Column	Meaning
INDICES		
	ID	Primary Key
	NAME	Name of the index structure
	FILE	File name to find the index in
OPCLASSES		
	ID	Primary Key
	NAME	Name of the operator class
	FILE	File name to find the operator class in
	INDEX	Index reference (Foreign Key)

Table 1: Table Definition of the custom tables for index structure dynamic loading

The indices table stores information about the index name and the file name which contains the index structure. The table *OPCLASSES* stores information about a certain operator class which is linked to an index structure. The operator class serves as concrete implementation of certain functionality required by a particular index structure to be able to deal with a certain kind of data stored inside the database table. An operator class for a B-Tree implementation could thus provide functions for comparing or ordering several values.
Besides the two database tables, also the SQL command for creating index structures was extended inside the h2 database so that the names of the index structure (as seen in table 1) and the name of the respective operator class may be passed to this command. The syntax of the modified

```
CREATE { [ UNIQUE ] [ HASH ] [ indexType ]
 INDEX [ [ IF NOT EXISTS ] newIndexName ]
 | PRIMARY KEY [ HASH ] } ON
 tableName ( indexColumn [,...] )
 [ USING opclassType ]
```

Figure 4: Syntax of the modified Create Index Command

command can be seen in figure 4. The fields *"indexType"* and *"USING opclassType"* are the basic extensions done to the command. The index loader then looks inside the indices table for the name specified by *"indexType"* and the associated operator class from opclasses given in *"USING opclassType"*. Therefore, for setting up an index called "bitlistbtreeindex" on a table called "documents" using the column "doc" and the associated operator class "docopclass" can be done as in the following statement:
`CREATE bitlistbtreeindex INDEX ON documents (doc) USING docopclass.`
These are the general adaptions we have done to the h2 database to enable it to extend the index structures currently available and to create new index structures based on the h2 basic definitions.

4.3 Query Interface

The query interface we present to the user does not differ from standard query interfaces from retrieval engines. Our goal is not to let the user select a certain type of entity to search for based on a pre-defined input field but to parse the query components directly from the query. Therefore, we decided only to have one field to enter the keywords in and not distinguish between fulltext keyword part and named entity (person name) part. The query interface is directly connected to the storage engine which serves the data stored inside the database using the hybrid index structure, described in section 6.

5 Entity Detection in Search Queries

Extracting the existing entities from the unstructured search queries is one of the most important aspects in this search engine. We try to achieve this without the need of specifying additional input fields. Therefore, we need methods to separate the query keywords from entity search candidates in unstructured queries.
Hypothetically, each of the used query keywords might be a candidate for an entity to be found. Initially, every sequence of query keywords may be regarded as a candidate for the entities. Therefore, a mechanism has to exist to filter candidates in order to leave the search effort low when communicating with the database. We chose a compressed trie variant stored in main memory for entity

[6] http://lucene.apache.org/core/, accessed 2013-06-19
[7] http://www.h2database.com/html/main.html, accessed 2013-06-07

candidate filtering.
This compressed trie is set up on top of the entities extracted from the already stored documents inside the existing database. This structure might also be used for auto-completion functionalities in the future. Based on the document statistics used in our test setup, we built the compressed trie to determine the resource allocation inside a real world scenario.
For approximating the performance of the compressed trie, we choose to evaluate the annotated RCV1 corpus (see 3.2).
The different candidates are generated based on the assumption that entities containing more than one word are always written in sequence. Consider, e.g., the query phrase $S = (\texttt{Barack}, \texttt{Obama}, \texttt{election})$ as input to the search. Therefore, we generate a set of a set of n-tuples $Q = \{T_1, T_2, \ldots, T_n\}$ with cardinalities $|Q| = n$ and $|T_i| = n - (i-1), 0 < i \leq n$ from the initial n-tuple of keywords $S = (w_1, w_2, \ldots, w_n)$, where w_k is the keyword at position k and n is the length of the search query. For each $T_i = \{E_{1,i}, \ldots, E_{n-(i-1),i}\} \in Q$ applies that $E_{k,i} = (w_k, w_{k+1}, \ldots, w_{k+i}), k \leq n - i$ is a tuple of length i. That means, that from an initial query, we generate a set of all candidates $C = T_1 \cup T_2 \cup \ldots \cup T_i$ with the cardinality $|C| = \sum_{x=1}^{n} x$ which are subsequently checked for being entities. From a human point of view, it is probably obvious that "Barack Obama" is the entity meant by the query issuer. The approach, described above, then generates the following set of tuples: $Q = \{T_1 = \{(\texttt{Barack}), (\texttt{Obama}), (\texttt{election})\}, T_2 = \{(\texttt{Barack Obama}), (\texttt{Obama election})\}, T_3 = \{(\texttt{Barack Obama election})\}\}$.
As there exist only $\sim$ 486K entities, generated by the approach, described above, in total (inside RCV1) and, based on our measurements, the compressed trie consists of $\sim$ 618K nodes, storing this structure in main memory results in $\sim$ 200 MB. The memory measurement is carried out in Java and thus can only serve as an approximate value. Storing this structure in main memory should not result in any problems on currently used server machines.
The check for candidates is performed using the previously generated compressed trie. Each of the generated tuples is looked up inside the trie and if it is found there, it may be considered a final candidate entity. For being able to retrieve the candidates as described before , they are also stored in the database like this while extracting the information from the documents. These candidate entities are then sent to the database, currently using an *"OR"* conjunction, and retrieve the data using the hybrid index structure, described in section 6. Additional ranking procedures might, in future, take the presence of multiple individual entities into account. The ranking procedure and final retrieval process is, however, not yet implemented and still subject of discussion.
The average query size, determined from AOL Query Log, is $\sim$ 3.014 words per query. Therefore, if we ceil the value to 4, we get a total average amount of entities to be checked first in the trie and afterwards inside the hybrid index of $\sum_{x=1}^{4} x = 10$.
We executed short performance measures on the compressed trie. The most important property, here, is the insertion time as it affects the entire process during crawling and the search effort of a trie is well known as $O(1)$ because it is only related to the length of the input. Therefore, we instantiated the compressed trie from the database to create it freshly (e.g. in case of a data loss, when it is simply contained in main memory). We used, again the prepared Reuters data and imported all known entities into the trie implementation. Importing the existing data of $\sim$ 486K entities extracted as described before, the insertion procedure took in average $\sim 7500ms$. The entire process of querying the data and inserting them into the trie took $\sim 33s$ in average. These numbers show that inserting the data into the trie on the fly while writing them to the database does not cost much as inserting one element into the trie results in $\frac{7500ms}{486000} \approx 0.015ms$. For an average document of $\sim$ 5.73 entities this makes a total average effort of $0.086ms$ for handling the trie per document. This seems to be reasonable as the remaining operations take much longer (e.g. analysis of the documents or inserting them into the database).

6 Indexing and Database Storage

Besides other tables for the search engine architecture described in section 4, there exists one table storing the documents in a denormalized form. This table stores the data to be handled by the index structure. As a full text search and a search for an entity is supposed to be done simultaneously, both parts of data have to be stored inside the table to compute a document representation from these which is thereafter indexed by the specialized index structure, described in the next subsection. Therefore, the table consists of a fulltext part where the text, pre-processed by the application using textual normalization, and an entity part storing an array of named entities associated with the particular textual document are stored. An overview over the

Table	Column	Meaning
DOCUMENTS		
	ID	Primary Key
	WORDS	TXT object (normalized words)
	ENTITIES	ARRAY of entities (varchar)
	DOC	computed column from words and entities

Table 2: Table Definition of the document table

main table to put the index on can be seen in table 2. The index itself is constructed on top of the computed column "doc" which is a composition of the two column "words" and "entities". The entities are stored inside an array of string values (varchar).
The queries are supposed to retrieve documents in which the queried entities occur and the textual part is also present. To support this type of queries efficiently, a new hybrid indexing method is introduced enabling efficient retrieval of this kind of data. The index structure used here supports the retrieval of combined data of entities and textual content.
In this case, we use a hybrid index supporting storage of the heterogeneous data types directly inside one structure. This technique can be used to support searches of the given types efficiently. Therefore, there is no need to search in two different tables or access structures and generate intermediate result sets which are intersected at the end but to directly navigate to search results fulfilling both types of search criteria (entities and keywords).

6.1 Architecture

The index structure, used here, is similar to the one described in [Göbel *et al.*, 2009]. The main changes are based on the fact that it is implemented in a real world database system and the change of the augmented structure. In our case, a base index structure whose elements will be augmented with the bitlist has to be able to handle entities (in string representation) efficiently. Therefore, as the main hybrid structure, we chose a B-Tree (or B+-Tree, more precisely) whose elements are augmented with a bitlist which represents the sets of terms valid inside the subtree pointed to by a specific B+-Tree element.

Figure 5 shows a conceptual overview of the components

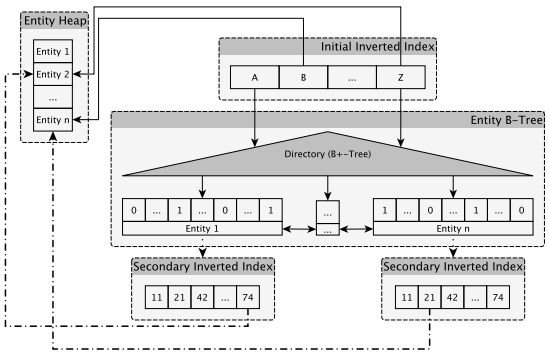


Figure 5: Conceptual Overview of the Entity B+-Tree

of the hybrid B+-Tree.

There are some additional structures used for administration of the index which are omitted in this graphic for simplicity reasons.

The central component of the index structure is the entity B-Tree. It stores combined keys of entity references and the bitlists which represent the presence or absence of terms from the initial inverted index. The initial inverted index itself stores all terms contained in the words column of the database table. If a term has a frequency higher than a pre-defined limit, it is moved into the hybrid part of the entity B-Tree and gets assigned a certain term index. This term index can then be used as identifier inside the bitlist. Primarily, the entities are not inserted into the entity B-Tree. They are stored in the so called "entity heap" where sequential comparisons of items which do not yet exceed the previously mentioned limit with respect to the absolute term frequency. If the limit is exceeded by one term, all entities referred to by each document the particular term points to are inserted into the entity B-Tree and the respective bits are set referring to the term identified by the assigned term index.

6.2 Algorithms

The insertion is done as described in algorithm 2. First, the items are added to the entity heap. After this operation, the references always point to the entity heap, which stores the document to entity assignment, and not the direct references to the documents any more. After that, the insertion operation continues to insert the full text terms into the initial inverted index (line 2). This operation also generates the list of terms and assignment to the entity heap references which exceed the artificial limit. If the references of one particular term does not exceed the artificial

Algorithm 2: addDoc(doc)

```
  // add the entities to the entity
  heap
1 ref = entityHeap.add(doc.getEntities())
  // add the terms to the initial
  inverted index using the reference
  from the entityHeap
2 overflowTerms = initialInvInd.add(doc.getTerms(), ref)
  // generate assignments of entities
  to terms and documents
3 assignments =
  generateEntityTermAssignment(overflowTerms)
4 for entry ∈ assignments do
      // insert the entity into the
      B-Tree
5     leaf = insertEntity(entry.entity)
      // add the term code to documents
      assignment at the secondary
      inverted index
6     appendCodesDocs(leaf, entry)
      // adjust the B-Tree, update the
      keys, nodes and bitlists
7     adjustTree(leaf)
```

limit, the references are stored directly inside the initial inverted index. After that, assignments are generated which point from one entity to all term indices referring to lists of entity heap references to pre-calculate the operations to be performed at the secondary inverted index. The elements of this assignment list are then distributed into the entity B-Tree where first each entity element is inserted (or already found) and then the term index to document list entries from the particular assignment are inserted at the respective secondary inverted index. After that, the "adjustTree" method of the B-Tree is executed. This method performs the standard B-Tree operations. Additionally, it is extended to perform adjustments on bitlists to update the B-Tree, correctly for being able to descend to a subtree which has a certain term code set and additionally fulfill a search condition focussed towards an entity.

Searches are executed as described in algorithm 3. It starts using a list of terms and a set of entities to be contained in the documents inside the result set. First, the initial inverted index is searched for the terms and entities. If one term does not exceed the artificial limit regarding its absolute frequency in the document collection, it is filtered sequentially for containment of the set of entities. If the frequency is higher than the artificial limit the term code is returned. If the set of term codes is empty, either nothing has been found or all terms were found inside the initial inverted index and the intersections are already calculated. Otherwise, the search is continued inside the hybrid index. In each element there, checks are performed if the element satisfies the search condition for an entity and the term index obtained from the initial inverted index, simultaneously (line 5). Afterwards, the final result set is built by searching in each secondary inverted index for the set of term codes obtained from the initial inverted index. In this step, the entities to retrieve can be simply ignored as the hybrid search part only delivers valid leaf nodes inside the entity B-Tree which already have to satisfy the entity search condition. We also know that there exist entries satisfying both search conditions as the bitlist represents exactly this behaviour.

Algorithm 3: search(terms, entities)

```
   // search references in initial
   inverted index
1  entries = initialInvInd.search(terms, entities)
2  if entries.termCodes = ∅ then
       // No term codes (term indices)
       available, so the final list is
       already built by sequential
       filtering
3      return entries.documents
4  else
       // Search for entities/bits in
       B-Tree (obtain leaf nodes)
5      leafEntries = hybridSearch(entries.termCodes,
       entities)
       // Search attached secondary
       inverted index structures for
       elements that satisfy all
       predicates
6      resultSet = searchSecondary(leafEntries,
       entries.termCodes)
7      return resultSet
```

7 Conclusions and Future Work

In this paper, we presented a search engine for person names and full texts intermixed. The approach used here, may be extended to the use of named entities, in general. The main focus was on the extraction of named entities from unstructured queries as well as database indexing. As this is project still continues, there are still open questions to be answered in future. A subject of investigation in the future will be the proper ranking which could also be integrated directly inside the retrieval process of the index structure. Another possible subject of future investigation is the real distribution of entities inside search queries as, currently, we take already present queries from AOL log. So, in future, when the system is finally running, it is probably more meaningful to investigate "real" queries put to the search engine in order to be able to analyze the real performance of our approach. We also want to compare the current trie approach for entity extraction with probabilistic methods.

References

[Bayer and McCreight, 1972] R. Bayer and E. M. McCreight. Organization and maintenance of large ordered indexes. *Acta Informatica*, 1:173 – 189, 1972. 10.1007/BF00288683.

[Chen *et al.*, 2013] Lisi Chen, Gao Cong, Christian S. Jensen, and Dingming Wu. Spatial keyword query processing: an experimental evaluation. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 217–228. VLDB Endowment, 2013.

[Cheng *et al.*, 2007] Tao Cheng, Xifeng Yan, and Kevin Chen chuan Chang. Entityrank: Searching entities directly and holistically. In *In VLDB*, pages 387–398, 2007.

[Felipe *et al.*, 2008] Ian De Felipe, Vagelis Hristidis, and Naphtali Rishe. Keyword search on spatial databases. *International Conference on Data Engineering*, 0:656–665, 2008.

[Finkel *et al.*, 2005] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Göbel and Kropf, 2010] Richard Göbel and Carsten Kropf. Towards hybrid index structures for multi-media search criteria. In *DMS'10*, pages 143–148, 2010.

[Göbel *et al.*, 2009] Richard Göbel, Andreas Henrich, Raik Niemann, and Daniel Blank. A hybrid index structure for geo-textual searches. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1625–1628, New York, NY, USA, 2009. ACM.

[Guo *et al.*, 2009] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 267–274, New York, NY, USA, 2009. ACM.

[Kumar and Tomkins, 2009] Ravi Kumar and Andrew Tomkins. A characterization of online search behavior. *IEEE DATA ENGINEERING BULLETIN*, 32(2):3–11, 2009.

[Levenshtein, 1966] VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[Miller, 1995] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[Morrison, 1968] Donald R. Morrison. Patricia-practical algorithm to retrieve information coded in alphanumeric. *J. ACM*, 15(4):514–534, 10 1968.

[Zhang *et al.*, 2009] Dongxiang Zhang, Yeow Meng Chee, Anirban Mondal, Anthony K. H. Tung, and Masaru Kitsuregawa. Keyword search in spatial databases: Towards searching by document. *Data Engineering, International Conference on*, 0:688–699, 2009.

[Zipf, 1949] George Kingsley Zipf. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley, 1949.

Detecting Documents with Complaint Character

Sebastian Ebert
University of Munich (LMU),
Munich, Germany
ebert@cis.uni-muenchen.de

Benjamin Adrian
Insiders Technologies GmbH
Kaiserslautern, Germany
B.Adrian@insiders-technologies.de

Abstract

Recognizing complaint documents as early and as fast as possible is a worthwhile goal for companies. In this paper we present an analysis showing the complexity of this practically relevant problem. Therefore, we define the task and its challenges and investigate statistical methods for automated Complaint Detection in incoming text documents. Two different approaches for handling complaint documents are presented. First, we analyze various term weightings in a standard bag-of-words approach. Second, we show the effect of feature engineering techniques known from Natural Language Processing. The results on four German and one English corpora show that already a linear classifier achieves valuable results and is competitive to more sophisticated methods in most cases.

1 Introduction

Complaints express a person's dissatisfaction and usually contain displeasure, anger, or other negative mood, since the sender is unhappy with some circumstance. Triggering events may be a company's products or services. Complaints are valuable for companies. If handled appropriately, i.e., if there is a good working management of complaints, both customer as well as the company will win satisfaction. The customer receives help and the company has a more satisfied customer. Additionally, complaints are opportunities, which can point out general problems. Fixing such issues improves the quality of products and services and reaches many customers at once.

Many of today's companies detect and handle complaints in the following way. A writing, e.g., a letter or an email, is received, scanned and forwarded to a document analysis system. Such a system extracts the text from the scanned writing and converts it into digital text. Then, information is extracted from the writing, which helps classifying it into company specific document categories, like car insurance or health insurance. Often the document is forwarded to a specific employee group based on this category. Such a group reads the text and if it is a complaint, either handles it herself/himself or forwards it to specialized complaint team. As a consequence, a complaint is handled only when an employee has recognized it. Since complaining customers are likely to cancel a company's services there is a need for prioritized handling of complaint documents. An automated Complaint Detection (CD) system is able to detect complaints even before an employee had to read a single document. This will dramatically reduce a company's reaction time.

In this paper we deal with the automatic detection of complaint documents in incoming mail. We investigate several Machine Learning (ML) methods on their suitability for this task. Using automated CD combines the benefits of complaint management, e.g., prioritized handling of complaints, with the faster approach of computer-supported detection of complaint documents.

The major challenges arising from complaints are the following:

Domain dependency Every company or even every department in a company needs to define what a complaint is. Thus, the definition can be totally different from department to department. Such differences lead to a tight domain dependency. We present a trainable method that can be adapted to different domains.

Consistent guidelines Instructing employees to recognize complaints is a difficult task, because there must be consistently and precisely formulated decision guidelines. Otherwise, one employee might say it is a complaint, another one may say it is not. Our statistical method ensures that a consistent definition of complaints is enforced and human error is eliminated as a source of inconsistencies.

Amount of documents The amount of incoming documents in a company can be higher than 1 million a day. Here, even a very low relative rate of misclassified documents leads to a high absolute number of not found complaints or writings wrongly declared as complaints. The former case vanishes the advantage of prioritized complaint handling. Furthermore, a low false negative rate is of particular importance in CD because reliable detection of the first complaint about a new problem and a quick elimination of the root cause can prevent a large number of subsequent complaints about the same problem as well as the high cost of losing dissatisfied customers and undoing damage that has already been done. A high false positive rate is undesired, because companies fear too much additional manual reclassification effort for their complaint team.

2 Related Work

Generally, detecting complaint documents is a classification task. In opposite to other classification tasks, e.g., topic classification, we have only two classes, namely complaints and non-complaints. The documents in either class do not share a certain topic. Instead, the similarity of all complaint documents is that the sender is unsatisfied with some circumstance; reasons are quite diverse. The diversity within the non-complaint documents is even larger. They can deal

with any topic, product, or service. Documents can be for example invoices, offers, or notification letters. The only thing these documents have in common is that the sender does not complain.

We belief that CD is similar to the task of Sentiment Analysis (SA). It is likely that complaint documents are written in a negative way. Much research in SA has been carried out in the movie domain. For example, Pang et al. [2002] classified the polarity (negative or positive) of movie reviews using ML algorithms, namely Naïve Bayes, Maximum Entropy, and Support Vector Machine (SVM). The authors studied the effect of term weighting schemes (binary, term frequency), bigrams, and the position of terms in a review on the polarity classification performance. In our work we carry out a more thorough research on term weighting schemes and also evaluate the use of trigrams, which allow to find longer structures. Furthermore, we look at several additional feature selection and feature extraction methods, not performed by Pang et al. [2002]. Lastly, our experiments are carried out on four German corpora and a larger movie review corpus.

The subfield of subjectivity classification deals with the distinction between subjective and objective texts [Wiebe, 2000]. Intuitively, non-complaints are always objective, like invoices, orders, etc.. However, subjective texts that are non-complaints are common, e.g., praises or accident reports in insurance companies. Such texts contain many polar words and often subjective language, but are no complaints. Moreover, somebody can complain without using subjective or polar speech. Consider the example sentence: "Why do you require 2 months for responding to my letter?" There is no explicit sentiment, i.e., a sentiment detector would probably classify it as a *neutral* sentence. Nevertheless, the sender is unhappy with the fact that nobody took care of her/his letter.

3 Term Weightings

The task of classifying a single document as being either a complaint or a non-complaint is a typical example of Text Classification (TC). In TC a given text document is assigned to one or more predefined classes [Sebastiani, 2002]. In this work, we formalize CD as a binary TC task, where the possible categories are *complaint* c_c and *non-complaint* c_n.

Documents are represented as bag-of-words: $d = \begin{bmatrix} w_1 & \dots & w_{|\mathcal{V}|} \end{bmatrix}^T$, where w_t is the weight of term t in this document and $\mathcal{V}$ is the vocabulary of all possible terms. A term weight is a numerical value that is assigned to a term. Salton and Buckley [1988] introduced a notation for term weights for their SMART retrieval system. This notation leads to a general definition for term weights: $w_{td} = f_t * f_c * f_n$, where the term weight for term t in document d consists of three factors: a *term frequency component* f_t, a *collection frequency component* f_c, and a *normalization component* f_n[1]. Table 1 lists the used components with their SMART notation and their computation.

For example, *txx* means that the number of occurrences

	notation	computation
t_t	b (binary)	$b_{td} = \begin{cases} 1 & \text{if } t \text{ occurs in } d \\ 0 & \text{otherwise} \end{cases}$
	t (term freq.)	$tf_{td} = \#(t, d)$
	l (log)	$l_{td} = \log(1 + tf_{td})$
t_c	x (none)	1.0
	f (inv. doc. freq.)[2]	$idf_t = \log \frac{N}{1+df_t}$
	Δf' (smooth. Δidf)	$\Delta idf_t = \log \frac{N_c * df_{tn} + 0.5}{df_{tc} * N_n + 0.5}$
t_n	x (none)	1.0
	c (cosine)	$c_d = \frac{1}{\sqrt{\sum_{t=1}^{\lvert\mathcal{V}\rvert} w_{td}^2}}$

Table 1: SMART notation of weighting schemes

of term t in document d, i.e., $\#(t, d)$ [Sebastiani, 2002] is taken solely as term weight. A very common term weighting in Information Retrieval (IR) is *tf-idf*, i.e., tfx, that accounts for the distribution of a given term t over the entire document corpus. df_t is the document frequency and counts in how many document the term occurs [Sebastiani, 2002]. In order to account for the document length, the term weight can be normalized by cosine normalization as presented in [Salton and Buckley, 1988]. A promising new term weighting for SA called *delta idf* (Δidf) was introduced by Martineau and Finin [2009], which instead of calculating the idf based on all documents, considers the idf values for the both classes separately and uses their difference: $\Delta idf_t = idf_{tc} - idf_{tn} = \log \frac{N_c * df_{tn}}{df_{tc} * N_n}$, where N_c and N_n represent the number of documents in the complaint and non-complaint class, respectively. df_{tc} and df_{tn} denote the document frequency of term t in the corresponding class. Paltoglou and Thelwall [2010] integrated Δidf into the SMART notation and created the so-called *smoothed* Δidf (Δf' in Table 1), which handles terms that occur in only one of the two classes.

We evaluate all combinations of f_t, f_c, and f_n, which results in a total of 18 possible weighting schemes for a single corpus.

4 Feature Engineering

A document corpus represented as bag-of-words can contain millions of terms. Many classifiers cannot handle this amount of features, because they do not scale well [Sebastiani, 2002]. Furthermore, many algorithms are prone to overfitting if there are many features. Finally, the more features there are, the longer the training (and for some algorithms also the classification) takes. Therefore, a common approach in many Natural Language Processing (NLP) tasks is to reduce the number of features. We investigate the influence of three approaches, stemming, stop-word removal and Principal Component Analysis (PCA).

To perform stemming we use Snowball, a language created for writing stemming algorithms [Porter, 2001]. The English experiments are performed with the Snowball implementation of the Porter algorithm [Porter, 1980]. The German experiments are carried out with Snowball's German stemming algorithm.

In order to see the influence of stop-word removal, we use the German and English stop-word lists provided by

[1] The SMART notation actually consists of two triples: *ddd.qqq*, where *ddd* is the document representation and *qqq* is the query representation. We have no explicit queries and thus neglect the second triple.

[2] We want to consider new words from the test set and therefore use this version of idf instead of the common form $\log \frac{N}{df_t}$ in order to avoid division by zero.

corpus	compl.	non-compl.	no. of words
liability	55	170	6,039
car	1,088	2,610	66,961
damage 1	373	989	34,674
damage 2	372	865	31,461
IMDb	1,000	1,000	38,911

Table 2: Corpora statistics

the Snowball project. The English list contains 174 stop-words. The German list comprises 231 stop-words.

Another technique for reducing the number of features we investigate is PCA. It is an unsupervised technique that calculates a transformation T that transforms the high dimensional document term matrix M into a lower dimensional space M': $M' = TM$. Since the number of dimensions $m \ll |\mathcal{V}|$ the problem of high dimensionality is tackled.

Using only single words as features, as we have done so far, has a serious drawback. It neglects the position of terms and their context entirely. A common technique to incorporate the context of words are n-grams [Manning and Schütze, 2000]. In the experiments we use bigrams ($n = 2$) and trigrams ($n = 3$).

5 Experiments

We performed all presented techniques on four German corpora and one English corpus. The four German corpora are real data from real customers[3]. They were collected in four different German insurance company departments from daily incoming mail. The departments are liability insurance (*Liability*), car insurance (*Car*), and two different departments dealing with insurances against damage (*Damage 1* and *Damage 2*). The corpora consist of incoming paper letters or faxes. Each document ran through a typical image conversion pipeline with (i) digitizing the image, (ii) cleaning it in several preprocessing steps, and (iii) running an Optical Character Recognition (OCR) to retrieve machine readable text. The preprocessing of all digital text documents consists of lowercasing and tokenization. Every document was labeled as complaint or non-complaint by an employee of the respective department. Table 2 lists the number of complaints and non-complaints in the corpora after filtering out duplicates and documents that per se can never be a complaint, e.g., invoices. Additionally, the number of distinct words is shown. The distribution of text lengths is very similar for complaints and non-complaints.

In order to measure the difficulty of this the CD task we asked two outside parties to manually label 50 randomly chosen documents from the Car corpus (25 complaints, 25 non-complaints). Both persons were asked to label each document with either complaint or non-complaint according to their own understanding of a complaint. The two raters agreed in only $32/50$ documents ($\kappa = 0.28$), which shows the complexity of this problem and the need for *consistent guidelines*.

Since we assume that CD is similar to the field of SA, we use another corpus that is well-known in this domain. This corpus called *polarity dataset v2.0*, was introduced Pang and Lee [2004]. We refer to this document collection as *IMDb*, because it comprises 2000 movie reviews that were automatically extracted from the Internet Movie Database (IMDb) and labeled as being positive or negative. The corpus statistics are listed in Table 2. For the sake of simplicity, we treat the positive class as complaint and the negative class as non-complaint, in order to have a consistent class naming.

For classification we use the SVM implemented in libSVM from Chang and Lin [2011] with a linear kernel and default parameters. To obtain the SVM performance we perform 10-fold cross validation and average the final results to an overall performance. We measure precision, recall and F_1 for the complaint class, since we want to focus on complaints.

6 Results

There are three weighting schemes that produce the highest F_1 on at least one of the corpora. Due to the large number of combinations we only report resuts for these three weighting schemes. The configurations are: *bxx*, *tΔf'c*, and *bfc*. Table 3 lists the precision, recall, and F_1 results.

The term weighting bxx has achieved the best results on 3 out of 5 corpora with a difference of up to 10 % (Damage 1) to the second best weighting, although it is the most simple feature weighting. Bfc has a very positive effect on precision compared to bxx on all corpora. Thus, if the rate of False Positives (FPs) must be kept small, it is a better term weighting than a binary representation.

In our experiments, all combinations using the new Δidf weighting have often led to lower results than bxx. Even the best combination tΔf'c has shown inferior performance.

Although there are some differences in the performances depending on the corpus, the differences in F_1 performance have not been statistically significant for $p = 0.05$[4]. We conclude that there is no benefit computing complex weightings like Δidf, because binary weights already achieve good results. Therefore, we use bxx as the baseline for further investigations.

All dimensionality results were achieved using the bxx weighting scheme. They are listed in Table 4.

Stemming and stop-word removal have led to improved performance on only one corpus each (stemming: Damage 2, stop-word: IMDb). On all other corpora, the performance has been inferior. However, the differences have not been statistically significant. We do not recommend either of the two techniques.

In an optimal case, PCA strongly reduces the number of required features and still maintains the same performance. We have chosen the number of principal components in order to keep 95 % of the data's variance. For Liability this is 164 principal components (reduction of features by 97.3 %), for Car 2,372 (96.5 %), for Damage 1 957 (97.2 %), for Damage 2 874 (97.2 %), and for IMDb 1,439 (96.3 %). This a dramatic decrease in dimensionality. As Table 4 shows, performing PCA has not lowered the performance by much. The losses in F_1 have not been significant. Thus, PCA is very well suited to reduce the feature space and therefore reduce noise.

Using bigrams has resulted in a better F_1 performance on Car and IMDb. On the other corpora, the performance declined. Using trigrams could only improve the result on

[3] Due to privacy reasons this data may not be published.

[4] We performed a Friedman test with Holm's test as post-hoc test, following Demsar [2006].

	Liability			Car			Damage 1			Damage 2			IMDb		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
bxx	.83	.69	.75	.81	.75	.78	.75	.64	.69	.87	.84	.86	.85	.85	.85
bfc	.93	.25	.40	.89	.64	.75	.92	.36	.52	.93	.77	.84	.88	.88	.88
tΔf'c	.90	.67	.77	.76	.72	.74	.79	.47	.59	.86	.84	.85	.77	.81	.79

Table 3: Term weightings results

	Liability			Car			Damage 1			Damage 2			IMDb		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
bxx baseline	.83	.69	.75	.81	.75	.78	.75	.64	.69	.87	.84	.86	.85	.85	.85
stemming	.80	.67	.73	.80	.74	.77	.75	.64	.69	.88	.85	.86	.85	.86	.85
stop-word	.94	.56	.70	.79	.71	.75	.77	.60	.68	.87	.78	.82	.87	.85	.86
PCA	.84	.67	.75	.79	.73	.76	.74	.63	.68	.85	.83	.84	.85	.84	.84
2-grams	.84	.56	.67	.84	.75	.79	.84	.58	.69	.88	.80	.84	.88	.86	.87
3-grams	.83	.36	.51	.84	.72	.77	.84	.53	.65	.89	.77	.82	.89	.86	.88

Table 4: Feature engineering results

the IMDb corpus. This finding suggests, that n-grams cannot appropriately capture the context that is necessary to classify complaints.

7 Conclusion

In this paper we have introduced the topic of CD. We have argued that complaints are very important for companies as well as for customers.

As a first step in our research, we have shown that binary term representation has delivered as good results as more sophisticated methods or even better and their computation is both, easy and fast. But, if the system's FP rate is of importance and many documents are being misclassified as complaints, bfc should be preferred, because its precision results have generally been higher. Despite these results, the independence assumption that the unigram model makes is clearly wrong and in our case seems to be unable to capture complaints entirely. But also the use of n-grams, which consider more context, has not helped. Therefore, for the classification of complaint documents we need more linguistic knowledge, e.g., in terms of word polarities or discourse structures.

Using stemming or stop-word removal has not been beneficial, they have resulted in poorer results. Instead, PCA is well suited for drastically reducing the feature space (between 96.3 % and 97.3 %), while maintaining nearly equal results to those of the baseline. We conclude that other feature selection approaches may also be helpful in finding good complaint specific features.

This work is the basis for further analysis of complaint documents. In a next step we will investigate the usage of sentiment lexicons, which allow the incorporation of word polarities in the classification task.

Acknowledgement

We like to thank Insiders Technologies GmbH for providing support for this work and Hinrich Schütze for proof reading the paper.

References

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[Demsar, 2006] Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[Manning and Schütze, 2000] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2000.

[Martineau and Finin, 2009] Justin Martineau and Tim Finin. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *Proceedings of the 3rd International Conference on Weblogs and Social Media*, 2009.

[Paltoglou and Thelwall, 2010] Georgios Paltoglou and Mike Thelwall. A Study of Information Retrieval Weighting Schemes for Sentiment Analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395, 2010.

[Pang and Lee, 2004] Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 271–278, 2004.

[Pang *et al.*, 2002] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86, 2002.

[Porter, 1980] Martin F. Porter. An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, 14(3):130–137, 1980.

[Porter, 2001] Martin F. Porter. Snowball: A language for stemming algorithms, 2001. Access date: 07/16/2012.

[Salton and Buckley, 1988] Gerard Salton and Christopher Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[Sebastiani, 2002] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[Wiebe, 2000] Janyce M. Wiebe. Learning Subjective Adjectives from Corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 735–741. AAAI Press / The MIT Press, 2000.

Report on the development of an IR system for medical image documents

Sebastian Dungs
University of Duisburg-Essen
D-47048, Duisburg, Germany
dungs@is.inf.uni-due.de

René Donner
Technical University of Vienna
A-1090, Vienna, Austria
rene.donner@meduniwien.ac.at

Abstract

A prototype of a content based information retrieval system for clinical images is presented. It is targeted at the user group of radiologists working on diagnosing new cases amongst other scenarios. While still being in an early state, the system features sophisticated image retrieval mechanisms and a comprehensive and versatile user interface. CBIR is performed using several types of visual features aggregated over supervoxels and state-of-the art indexing regimes. The user interface uses an agent based framework infrastructure which is easily extensible and suited for complex tasks with difficult information needs [Beckers *et al.*, 2012b]. The system is one main part of the EU-project *Khresmoi* finishing in August 2014.

1 Introduction

The following paper describes the current stage of development of a information retrieval system for medical experts. It is tailored at radiologists and their special demands when it comes to diagnosing diseases by looking at medical images. These images may be taken through means of CT, PET, fMRI or X-ray and are accessible within a hospital network.

Diagnosis by visual analysis requires recognition of patterns and structures in the images that may be an indication of a specific condition the patient is in. Today radiologists often rely on text books as a reference for unknown visual structures. Asking more experienced colleagues is also often the only handy option for doctors in their first years of medical practice.

The system under development—being part of the larger Khresmoi project (see Section 3)—is aiming to overcome these problems. Users will be able to perform searches based on the images of the case at hand. The system performs an image similarity search and returns images containing the same structures as the query case. Moreover, the diagnosis is returned alongside. This is believed to give the experts a starting point in their diagnosis process. The overall goal is to speed up diagnosis of tough cases with rare or unknown diseases by reducing the need to consult external resources like textbooks or human advice.

This leads to interesting research questions, like how a useful and usable user interface is going to look like or which visual characteristics best distinguish certain diseases. An important question is also how the incorporation of both visual and semantic characteristics into the employed machine learning methods allows to improve the performance over simple visual retrieval, only.

This paper is structured as follows: Related work will be discussed in the next section. Section 3 introduces the Khresmoi project. In Section 4 the system is described from the user point of view. The interaction work-flow we envision users to follow is also presented. Furthermore, Section 5 covers the retrieval algorithm used for this prototype. Section 6 sums up and Section 7 discusses our work.

2 Related Work

Clinical image data in a hospital environment is currently typically organized in a *Picture archiving and communication system* (PACS). There are various different systems available distributed by large international companies. While these systems can handle large amounts of data generated in the hospital and are directly connected to the data sources they all have a major drawback. Content based image retrieval is not supported. Our proposed system tries to overcome this deficit.

User interface related developments described in this work are mainly based on the ezDL[1] software framework (compare [Beckers *et al.*, 2012a]). The authors describe an agent based retrieval system to access heterogeneous distributed digital libraries. While the original system is not used in the medical domain, it can be easily adapted and extended by adding further data sources. It also has built in logging functionality on user interaction level which can be activated if user experiments are about to be conducted. How this system was adapted to the Khresmoi project radiology use case is described in the following section.

3 Khresmoi Project

Khresmoi[2] is a project funded by the EU and currently is its third of four years. It aims at developing a multilingual, multi-modal IR system for biomedical information. It advances the current state of the art in several domains. These include the automated extraction of information from biomedical documents, semantic search features, linking information extracted from different sources, automated analysis and indexing of numerous medical images like X-rays as well as 3-dimensional data and supports cross-language IR. Moreover, a flexible user interface framework tailored at supporting a variety of different tasks and user interaction styles is under development. It is a versatile system supporting various platforms. This includes full featured desktop clients for PC, Mac and Linux,

[1] `http://www.ezdl.de`
[2] `http://www.khresmoi.eu`

a browser based version for high flexibility and an Android app for mobile use. All clients share a common brand identity allowing for an easy transition between them.

The project has different use cases each targeted at a special user group. All user groups have different needs which are also addressed by the aforementioned variety of user interface versions, each featuring different tools suited for the most common tasks of the user group. One of these user groups in the field of medical professionals consists of radiologists. A adapted version of the desktop client interface (see Figure 1) is under development. A prototype of this interface as a well as the underlying retrieval mechanism—also referred to as *Khresmoi for radiologists*—will be explained in the following section.

4 Work-flow description

As mentioned in Section 1 one distinct use case for the Khresmoi project is the radiology department in a hospital environment. There it will be used in the process of writing diagnoses based on visual analytics of images of various modalities. Therefore, Khresmoi for radiologists will be connected to a hospital database like a PACS system. This database also provides the basis for case retrieval. The work-flow we envision is as follows:

First a radiologist selects the case in question by means of a tabular *index perspective*. It gives basic information about all cases, like imaging technique used or patient demographics. The table can be sorted and filtered to allow a quick search. Users can request additional data about any case by clicking on it. This will load the report about that case, if any exists, as well as the actual image data. Because the system will be deployed within the hospital network, the transfer time of these potentially large files is expected to remain reasonable. If a user decides to query the system with a specific case she is able to narrow down the search by marking one or more regions in the image data as relevant. How this information is used in the retrieval process is explained in Section 5. In the prototype's current state of development marking regions of interest is based on rectangular areas that can be selected in the two dimensional images.

Results will be presented in a list of result items. Every item consists of a representative preview thumbnail of the image as well as basic meta data. These include patient age and gender, the image acquisition date and a short version of the report along with pathology tags. The default sorting is relevance based, while the results can also be sorted by acquisition date in both orders. Users are able to group the result set according to case meta data values. For instance one might be interested in cases that correspond to a certain disease. All matching cases can be highlighted within the result set or it can be filtered showing only matching cases. Radiologists can retrieve the actual image data and the report for any case by clicking on a result item in the list. It will then be shown in the central part of the interface or in a separate window. All cases can be stored in a tray or basket and will on demand be saved permanently for easy access in the future.

Further information on the actual retrieval process and the novel technological aspects our system are described in the following section.

5 CBIR

The content based image retrieval part of the described system is developed and located at the Medical University Vienna, with the CIR lab. The system includes the capability for the data management and storage of very large medical volume datasets and employs state of the art computer vision techniques to analyze and index the data. The indexed data as well and the indexing service is made available to the Khresmoi framework through a private API. The data set used in the index is a collection of 3876 3D-CTs and MRs extracted from the PACS of the General Hospital Vienna / MUW. The CTs and MRs originate from all different scanners present at the department of radiology.

Processing data flow The data is transferred from the hospital's PACS to an internal data base system, after detailed anonymization of all data and meta-data. Subsequently each volume is processed, its visual features computed, and these features are added to the index. The entire data analysis is formulated as a map-reduce graph, wherein each node in the graph can store its results on disc. The inter-dependencies of the nodes are automatically exploited such that adding a new volume to the data store only triggers the computation of the minimally required subset of nodes in the graph to ensure a valid index. This approach also implicitly provides the ability to run the computations in parallel on a compute cluster, and implicit robustness to errors in the node's computations or machine failures.

Computer vision methodology The main components of the computer vision processing pipeline include the correct orientation of the volume in regards to a reference atlas, the registration to this atlas, the computation of super-voxels and several visual features per super-voxel and finally the computation of the index. The orientation of the volumes as delivered from the PACS can be arbitrary, but is defined by the volume's headers. A simple but important aligning step ensures that the volume has the same orientation as the atlas. The atlas itself consists of one whole-body CT scan. After the orientation, alignment and affine registration is performed between the volume and the atlas. This estimates the translation, rotation and scaling parameters necessary to best align the volume to the atlas. Subsequently, a non-rigid registration estimates the complex non-linear deformation necessary to obtain a correct voxel-by-voxel alignment between the volume and the atlas. The voxels of the aligned volume are divided into non-overlapping parts, i.e. so-called super-voxels. This commonly employed computer vision technique aims to extract regions which are maximally homogeneous within each region and maximally distinct between adjacent regions. The number of super-voxels is set to be three to four orders of magnitude lower than the number of voxels, greatly speeding up the feature computation and indexing, while loosing very little information which would be relevant to the task of retrieving similar regions in the volumes. We employ an adapted variant of SLIC super-voxels incorporating the monogenic signal, which allow for smoother and more regular super-voxels. For each super-voxel a set of well-established visual features is computed, namely gray-level co-occurrence matrices with Haralick-features (as for example employed in [Valentinitsch *et al.*, 2013]) and Haar-like wavelets as described in [Donner *et al.*, 2010], as well as Bags-of-Visual words of Local Binary Patters (LPB) and gray-level histograms. The framework is not limited to these visual features, and the best set of features for each retrieval scenario can be found through a cross-validation

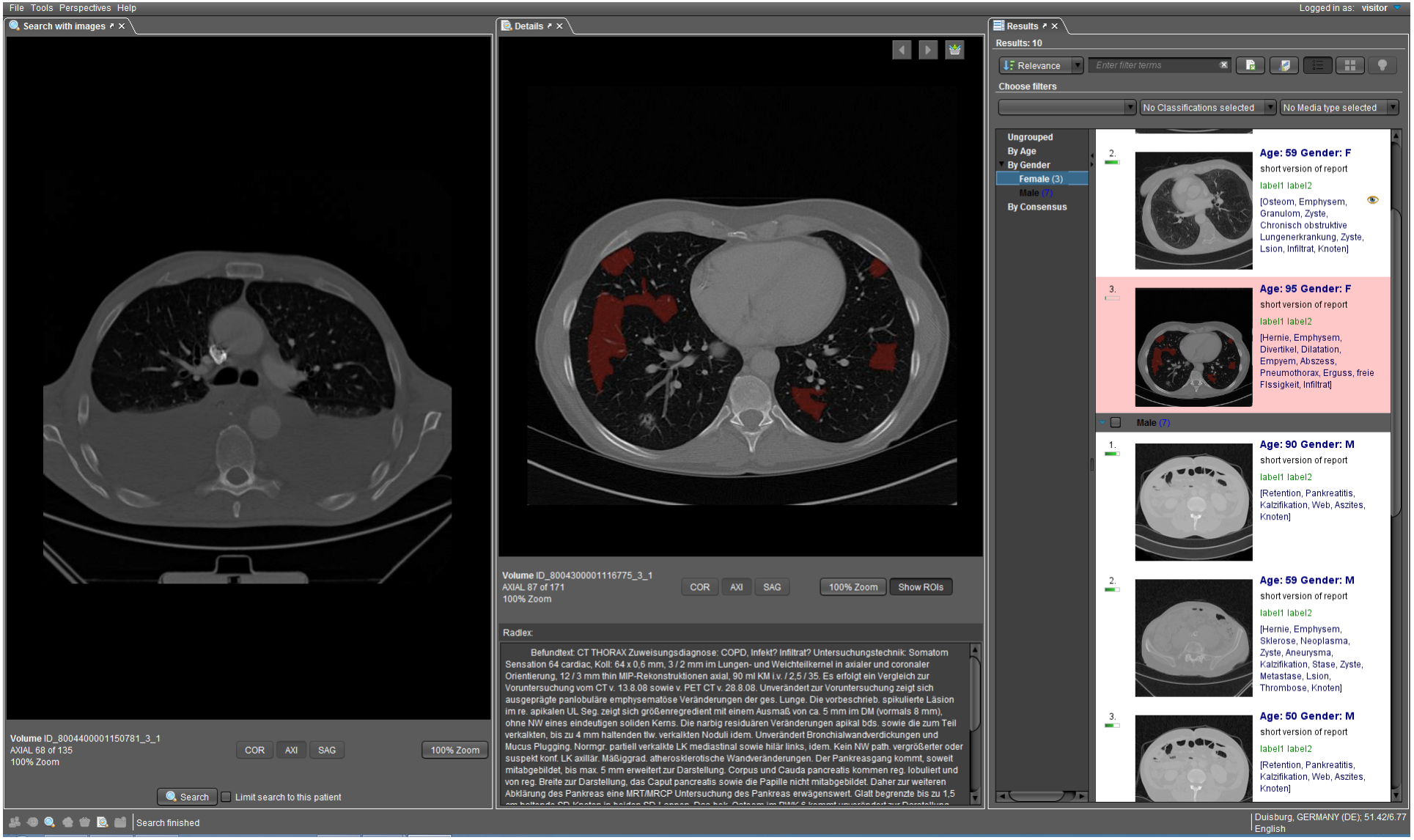


Figure 1: Screen shot of the current user interface prototype

approach. The evaluation of different visual features is the current focus of our research.

These features are concatenated, yielding a $nFeatures \times nSupervoxels$ matrix per volume. The actual indexing of these features across the data set is performed using different methods, which are currently under evaluation. ProductQuantizers [Jégou *et al.*, 2011] are used to quickly retrieve the most similar super-voxels, given a query super-voxel, in the nFeatures-dimensional feature space. The evaluation of such a system is performed on two fronts: one is concerned with measuring the effect and improvement of using this tool in clinical practice. This is mainly assessed through controlled experiments with medical experts using the system. A detailed analysis of the usage, supported by video monitoring and eye-tracking allows to measure improvements in GUI design and overall retrieval performance. Detailed interviews with the study participants are also employed in each evaluation round. The second metric is the numerical evaluation of the details of the retrieval pipeline using a large set of cases with existing diagnostic reports as ground truth. An automated semantic analysis of this corpus yields a distance metric between the cases, and the performance of the retrieval system can be evaluated against it.

6 Conclusion

We presented a system for content based image retrieval planned to be deployed in a hospital environment. Current systems like the common PACS work-stations do not support the retrieval of cases by means of image data. This leads to an inefficient and overly complex work-flow for radiologists while diagnosing cases. Our system can be used to quickly find similar cases to the one at hand without the need to rely on external sources. Therefore, data is at first anonymized and transferred to our system, where it is mapped and aligned to an atlas. By using super-voxels for computation the process is sped up significantly. The volumes are then indexed by their visuals features over super-voxels, using map-reduce graphs to structure the computational data-flow. The actual indexing can be performed using different methods. These as well as the user interface will be further evaluated in the future.

7 Future development

Also in the near future the system's prototype will be extended by additional data sources. It will include 2D image documents taken from the Pubmed[3] database as well as textual documents. This data is provided within the Khresmoi project and part of the other use case prototypes. The radiology system will benefit from this addition because users can access more information relating to the current case. Based on the image retrieval result the system will extract textual keywords which will be used for an initial query to the 2D and textual document sources. Result are presented directly in the user interface without any user interaction while making sure the work-flow is not disrupted. The user may then choose to alter the automatically generated query and reissue it while the initial 3D query and result set remain unmodified. We think this system will improve the accessibility of medical resources needed for diagnosing difficult cases as well as reduce the time between the taking of an image and the making of a diagnosis. Nevertheless, the system will have to undergo user evaluations in the future to test both the user interface and the quality of the actual retrieval. Those evaluations are foreseen to take place in the forth project year.

[3] `http://www.ncbi.nlm.nih.gov/pubmed`

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 257528 (KHRESMOI).

References

[Beckers *et al.*, 2012a] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, and Sascha Kriewel. ezdl: An interactive search and evaluation system. In *SIGIR 2012 Workshop on Open Source Information Retrieval (OSIR 2012)*, August 2012.

[Beckers *et al.*, 2012b] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, Sascha Kriewel, and Vu Tran. ezdl: A framework for interactive and professional search systems. Technical report, Universität Duisburg-Essen, Thessaloniki, Greece, 2012. Position Paper at the MUMIA Workshop about Integrating IR/MT Technologies for Professional Search.

[Donner *et al.*, 2010] René Donner, Erich Birngruber, Helmut Steiner, Horst Bischof, and Georg Langs. Localization of 3D Anatomical Structures Using Random Forests and Discrete Optimization. In *Proc. MICCAI Workshop on Medical Computer Vision*, pages 86–95, 2010.

[Jégou *et al.*, 2011] Hervé Jégou, Romain Tavenard, Matthijs Douze, and Laurent Amsaleg. Searching in one billion vectors: re-rank with source coding. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 861–864. IEEE, 2011.

[Valentinitsch *et al.*, 2013] Alexander Valentinitsch, Janina M Patsch, Andrew J Burghardt, Thomas M Link, Sharmila Majumdar, Lukas Fischer, Claudia Schuellerweidekamm, Heinrich Resch, Franz Kainberger, and Georg Langs. Computational identification and quantification of trabecular microarchitecture classes by 3-d texture analysis-based clustering. *Bone*, 54(1):133–140, 2013.

Content Based Image Retrieval using Interest Point Algorithms in Context of Scientific Cultural Image Collections of Hebraic Tombstones.

Arndt Bergner
Otto-Friedrich-University of Bamberg
96047, Bamberg, Germany
arndt1.bergner@stud.uni-bamberg.de

Abstract

The *Digital Research Infrastructure for the Arts and Humanities*-project (Dariah-DE) is dedicated to evaluate information retrieval technologies for research infrastructures of social-, human- and cultural studies like universities. One on of the main project-participants is the Salomon-Ludwig-Steinheim Institute of German-Jewish-History which documents Hebraic tombstones as a part of Jewish history and life. A query-by-example could help to improve investigations in this image-database. The *content based image retrieval* (CBIR) could be done using different features like *interest point algorithms* (IPA). These algorithms find the most stable points like corner in images and calculate a comparable representation for this point using the surrounding pixel intensities. An amount of these stable keypoints will describe the content of the image.

In this paper an example collection of Hebraic tombstone is used to evaluate IPA-detector-descriptor-pairs like SIFT-SIFT, SIFT-BRISK, SURF-SURF, SURF-SIFT, SURF-BRISK and CenSurE-SIFT. Their tolerances in the difference of object-scale, illumination and perspective angle are tested. Further user-driven test-scenarios for CBIR are used to investigate the applicability of the IPAs when similar images in context of scientific cultural research have to be retrieved.

1 Introduction

Content based image retrieval (CBIR) is one possible approach to retrieve similar images if tags, descriptions, surrounding document-text or query-terms are missing. Different requirements have emerged in the different domains for matching a query-image and retrieve relevant pictures with a similar object. Different features like those of the *interest point algorithms* (IPA) could be used to retrieve similar images. For example [Aman *et al.*, 2010] uses IPA in context of computed tomographic colonography computer-aided detection. Those algorithms like *Scale Invariant Feature Transform*-descriptor (SIFT) by Lowe can be used to describe the content of an image [Lowe, 2004]. Like Sperker and Henrich have shown these IPAs can be used in different context like car model detection [Sperker and Henrich, 2013].

The context of this work is the *Digital Research Infrastructure for the Arts and Humanities*-project (Dariah-DE) which is dedicated to strengthen the research infrastructures of for social-, human- and cultural studies like European universities institutions.[1] One goal is to evaluate the usage of CBIR for cultural databases of fields like preservation of sites of historic interest, Jewish studies, art history etc. As [Kampel *et al.*, 2009] have shown, IPAs could be used to identify historical coins. [Valle *et al.*, 2006] used them to search in databases of historical photographies. These and the previous mentioned articles lead to the question if IPAs could be used for other cultural domain in context of CBIR.

One participant of Dariah-DE is the Salomon-Ludwig-Steinheim Institute of German-Jewish-History. The goal of this institute is to preserve the historical grown graveyards of Jewish communities. Because of the holocaust a lot of these cemeteries were abandoned as the communities vanished. The institute documents these graveyards and saves images of the tombstone in a large database. These image-collections are used for research in the field of German-Jewish-History. The university of Bamberg as a participant of Dariah-DE is evaluating possible CBIR-solutions.

Those algorithms consist of a detector, which is a calculation of the most stable and unique points of an image such as corners, and a descriptor, which is a mathematical representation for those keypoints. Here the surrounding pixel values are used. For every image its descriptor represents the content and can be compared to retrieve similar images.

In this paper different detector-descriptor-pairs of IPAs are evaluated for the CBIR of scientific cultural image collection. Two main test-sets were used to investigate the performance of IPAs. The first are synthetic tests evaluating the IPAs with images containing different interference factors such as variance in illumination, scale change caused by zoom and a perspective change in the angle around a tombstone. The second set was created from image collections of the Salomon-Ludwig-Steinheim Institute of German-Jewish-History and the local Professorship for Jewish studies. It consist of different search scenarios such as the CBIR using snippets of epigraphics, ornaments and symbols as well as the search for whole similar tombstones and historical pictures from the early 1920s/1940s. The search for historical picture is done with fragments and whole tombstones.

The article is structured as follows: Section 2 will summarize some of the related evaluation work and give two examples in context of the recognition of cultural objects and the CBIR in historical image collections. Then the

[1] `https://de.dariah.eu`, last checked 31st August 2013.

evaluation-application with its functionality and the processing steps for indexing and searching are described in section 3. After that the IPAs are introduced and the evaluated algorithms are discussed in section 4. The pretest to determine applicable algorithms is discussed in 4. The test-design with the different purpose and goals of the scenarios are explained in 6 and the results are given in section 7 and 8.

2 Related work

Most of the evaluation-experiments have the goal to determine the performance of the IPAs, when 2D and 3D objects are rotated, the illumination or the scale is changed and the angle of the perspective is increased. One of the first articles for feature-based matching of images was Schmid et al. [Schmid *et al.*, 2000]. The performance of the Harris-Corner-Detector was measured via repeatability which is a highly accurate measure used in lab-environment [Harris and Stephens, 1988].

Since then Mikolajczyk et al. evaluated new IPAs like SIFT which has proven to be one of the most stable algorithms [Mikolajczyk and Schmid, 2003; 2005]. Furthermore Mikolajczyk et al. have shown the limits for IPA-detectors when changing the angle of the perspective from 30° to over 60° , which will result in less repeatability [Mikolajczyk, 2004].

Fraundorfer and Bischof made a differentiation between planar 2D- and 3D-scenes to test the algorithms [Fraundorfer and Bischof, 2005]. Here the *Maximally Stable Extremal Regions* (MSER) [Matas *et al.*, 2002] were used as detector and the repeatability was measured when the angle of the perspective was changed. The experiments showed that most of the algorithms like MSER provided less good results when 3D-scenes were used. Moreels and Perona have reported similar results when 3D-objects where perspectively transformed by 30° [Moreels and Perona, 2007]. Here the MSER-detector and the SIFT-descriptor produced only 20% stable matches.

Additional results by [Gil *et al.*, 2010] showed in the context of *Simultaneous Localization And Mapping* that SIFT and *Speeded Up Robust Feature* (SURF) [Bay *et al.*, 2008] were able to compensate worse illumination and different scale changes. In the same context *Center Surround Extrema* (CenSurE) [Agrawal *et al.*, 2008] was evaluated among others by Gauglitz et al. [Gauglitz *et al.*, 2011]. CenSurE showed stable results when zoom or illumination was changed. Again SURF- and SIFT-descriptors performed very well. Dahl et al. explained in [Dahl *et al.*, 2011] the efficient combination of a MSER-SIFT-combination but as will be shown later on MSER could not pass a standard test.

In context of CBIR of cultural pictures [Kampel *et al.*, 2009] the IPA can support the identification of unique historical coins to archive and protect them from forgery. Additional in old image collection a query with a newer picture can be used to search for historical photographies as shown by [Valle *et al.*, 2006]. But there is much more work to be done in context of *Digital Humanities*. Additionally the usage of IPAs has to be transferred to a practical level.

3 Evaluation-System

For evaluation-purpose a 32bit C++ application named PatRecEval was implemented using the functionality of OpenCV 2.4.3.[2] Most of the state-of-the-art algorithms can be found here. PatRecEval is able to index collections, save the descriptor-/keypoint-index as YAML-files and load them to enable a query with the same IPAs. A very fast implementation for the detailed view of two matched image can be used for detailed investigation. Every matched keypoint in the images of query and index is marked with a dot linking line to the correspondences.

For completeness and accuracy a bruteforce-approach with cross-validation was used to match the descriptors of the images. Every image was normalized in size for performance and equality. The matrix of the query-image is analyzed with the same detector-descriptor-pair and a direct vector-representation for the image and its keypoints is computed. To transfer matched keypoints from the query- to the indexed image a homography is used, which is determined by the *RaNdom SAmple Consensus*-algorithm (RANSAC). This normed distance from the corresponding keypoints is used to filter outlier. The images can be ranked according to the number of relevant matches (inlier) and the number of irrelevant matches (outlier). The more inlier an image has the more relevant it could be. A higher amount of outlier is assumed to decrease the relevance of an image.

4 Interest Point Algorithms

The IPAs are middle-level-feature, while color-histogram are categorized as low level feature, which can determine the most stable and unique points against changes in illumination, scale or perspective via detector-algorithm. A unique representation-matrix as a comparable numerical descriptor is produced for these keypoints. This representation can be compared via distance measure like the Euclidean distance for floating point descriptors or the Hamming distance for binary string-descriptors.

Since the applicability of the different IPA-detector-descriptor-pairs was tested, only the pairs with positive results remain (see section 5). The evaluated IPA-detectors are SIFT, SURF and CenSurE. The numerical descriptors are SIFT, SURF and the binary are *Binary Robust Invariant Scalable Keypoints* (BRISK) [Leutenegger *et al.*, 2011].

Since 2004 SIFT is one of most efficient, state-of-the-art IPAs. Lowe describes in [Lowe, 2004] that the image is transferred into scale-space and the local extrema are found via a *Difference of Gaussian*-function, known as DoG. The detected points are only accepted if they, compared to all of its pixel-neighbors in different scales, differ in their intensities. Unstable edges or points prone to contrast-changes will be filtered via the Harris-Corner-function, the determinant and the ratio of the smallest and the biggest eigenwert. The descriptor is built using the gradient strength and orientation. Around the point 4x4 subregions with 8x45° orientations form a 128-dimensional descriptor.

SURF takes the ideas of SIFT and improves them by approximating the *Laplacian of Gaussian* (LoG) with linear box-filters and integral images. With a *Determinant-of-Hessian* the local extrema are extracted. A 64-dimensional descriptor is calculated using the filter-responses of Haar-wavelets regarding different sizes and the orientation of the intensities in the subregions around the keypoint.

CenSurE approximates the LoG with a octagonal bi-level-filter and the difference of octagons of an inner and an outer region of the filter. The image is transferred into scale-space via Gauss and seven filter-scales are applied to

[2] `http://opencv.org`, last checked 31st September 2013.

the picture. After a non-maximal suppression, only those minimal and maximal extrema are accepted which pass an adapted Harris-Corner-Response-function composed of curvature and trace [Agrawal *et al.*, 2008].

A BRISK descriptor contains a string of binary values which are determined by intensity-tests. Around the keypoint a pattern of Gaussian convolved regions is applied. Two subsets of short- and long-distance-pairs are build considering distance-restrictions. The long-distance-pairs are used to determine the gradient-orientation and the pattern is rotated according to this. The tests for the short-distance-pairs are used to construct the descriptor.

5 Pretest

OpenCV provides most of the state-of-the-art IPAs for feature-detection and -description. To select applicable detector-descriptor-pairs they have to pass a standard-test. For this test an image collection of CD-covers from the Stanford university was used with the default OpenCV-configuration of the IPAs [Begen *et al.*, 2011]. Exceptional parameter adjustments were made for MSER (max. are-size 650px), FAST (edge-threshold of 28) and BRISK (edge-filtering-threshold via FAST is set to 5). Four images of one CD-cover are contained in the collection and at least three of them have to be found at the first ranks which mean a precision@4 of 75%. This test does not consider the specific mannerisms of the tombstone-images but if a IPA fails at this task, it cannot be used for more domain-specific images.

After these results a picture of a perspective transformed tombstone was evaluated with the IPA-pairs to check the results. The passed IPA-pairs are summarized in table 5. MSER, *Features from Accelerated Segment Test* (FAST) [Rosten and Drummond, 2006], *oriented FAST* and *oriented BRIEF* (ORB) [Rublee *et al.*, 2011], *Fast Retina Keypoint* (FREAK) [Alahi *et al.*, 2012] and *Binary Robust Independent Elementary Features* (BRIEF) [Calonder *et al.*, 2010] failed the standard-test and are not further discussed.

6 Test design

After this the detector-descriptor-pairs were calibrated for the given image-collection of Hebraic tombstones. The parameter of synthetic tests were used to assess the performance of the algorithms when different interference factors would occur:

- Illumination: The deviation of intensity from the auto-adjusted setting of the camera from $[-2, -1, +1, +2]$(-2 means a underexposure and +2 an overexposure).
- Zoom: The focal distance in a range of $[18mm, 25mm, 31mm, 43mm, 49mm, 55mm]$ from a default of 37mm.
- Perspective: The angle measured with a protractor from 0° to 80° in 10°-steps.

For the change of angle perspective the rate of irrelevant matches (RIM), which are none-object-correspondences, and for all three types of synthetic tests the false-positive-rate (FPR) of the matches were manually counted and calculated. For the last two tests the irrelevant background was cut. The goal of these three tests was to get the overall limits of the IPAs in case of interferences factors which could occur in the field.

The third test contains different scientific search-scenarios which were discussed with the Professorship for Jewish studies and one member of the chair of art history of the university of Bamberg. These users wanted to find similar tombstones, search in historic image-collections and retrieve tombstones with epigraphics, symbols or ornaments. For the last three scenarios snippets were cut from the images and used as query. For the other scenarios complete images were used. Every scenario had at least 3 pictures. Every set had 4 query-images. Altogether the collection for this first explorative evaluation has a size of 125 pictures from the Salomon-Ludwig-Steinheim Institute of German-Jewish-History and the local Professorship. The creation of a bigger collection was not possible due to high effort in finding similar images and time limitations. 19,2% of them were never used and were kept as noise. The following performance indicators were used in descending order of importance to give a qualitative evaluation of the IPAs:

1. Overall performance: *Normalized Discounted Cumulative Gain* (NDCG) considering the rank of relevant matches [Järvelin and Kekäläinen, 2002].
2. Detailed performance: Inspection of the first ten images / the first occuring relevant match. The following questions were important: Where are the keypoints? How much keypoints have been found using the specific descriptor? How are the keypoints spread in the indexed image?
3. Additional Indictator: The distribution of relevant matches in the ranking.

7 Experimental Results for Synthetic Tests

As was shown in the related work of section 2 the perspective transformation will result in stable results until 30° (see table 2). After this point the FPR as well as the RIM are rising. Irrelevant matches (Ir) occur on several parts of the images like moss on the tombstone, background vegetation like trees or graveyard walls. Until 60° the results show worse performance and with an angle of 80° no relevant matches are found. This leads to the result that a possible limit for perspective change is 30°. After this point the results for the use of IPAs become unstable. Some of the algorithm like SIFT and SURF are having trouble dealing with regions with high intensity variation caused by moos. The algorithms found a lot of keypoints which affected FPR and RIM. One example is shown in table 2 for SIFT-BRISK displaying the limit of 30°. After this point the FPR rises as well as the RIM. Note that this detector-descriptor-pair finds less keypoints, the RIM and FPR are directly affected if a correspondence is irrelevant/false.

Angle	*In*	*Out*	Ir	RIM	False	FPR
10°	191	264	30	15,71%	1	0,62%
20°	124	300	8	6,45%	1	0,86%
30°	116	305	12	10,34%	1	0,96%
40°	30	305	12	40,00%	1	5,56%
50°	57	304	57	100,00%	0	100,00%
60°	19	372	18	94,74%	1	100,00%
70°	17	317	11	64,71%	6	100,00%
80°	18	367	12	66,67%	6	100,00%

Table 2: Evaluation data of the FPR and RIM for SIFT-BRISK when a change in perspective occurs.

The test in scale change caused by zoom showed that the images should not differ to greatly in the focal distances. Only the range of 31mm until 43mm from a point of 37mm

Detector\Descriptor	SIFT	SURF	BRISK	FREAK	BRIEF	ORB
SIFT	✓	**X**	✓	**X**	**X**	**O**
SURF	✓	✓	✓	**X**	**O**	**X**
MSER	**X**	**X**	**X**	**X**	**O**	**X**
FAST	**X**	**O**	**X**	**O**	**O**	**O**
CenSurE	✓	**X**	**X**	**X**	**O**	**X**
ORB	**O**	**X**	**O**	**X**	**O**	**X**
Keys:	**O** := not tested		**X** := failed		✓ := passed	

Table 1: Results of the standard-test using the Stanford image collections of CD-cover [Begen *et al.*, 2011].

caused a low FPR. Altogether combinations like CenSurE-SIFT did not create enough matches on the tombstones.

When dealing with a change in illumination-intensities the IPAs cannot handle underexposure. The darker the image gets the more equal the regions of intensity-values become until their difference it too low. This leads to less extrema, corners and stable regions. The FPR was rising as well as the poor distribution on the tombstone. In contrast the overexposure can be compensated. The count of extrema is rising when the image gets brighter which means lots of keypoints. The FPRs is low and a good distribution of relevant points on the tombstones exists. The results can be displayed in the table 3 for SIFT-SIFT. Here a change from the auto-detected illumination-norm of +2 creates a high FPR while in contrast a value of -2 give only 0,92%.

	In	*Outl*	False	FPR
+2	18	127	16	88,89%
+1	39	133	1	2,56%
-1	228	60	2	0,88%
-2	218	70	2	0,92%

Table 3: Evaluation data of the FPR for SIFT-SIFT when the illumination from the norm of the auto-detected illumination is changed.

8 Experimental Results for Scientific Search-Scenarios

As mentioned before the following scenarios were discussed with the users. The collection is composed of different subsets representing the scenarios. These sets are differing in their size but have a minimum of three pictures which could be found as a relevant match. Every image in the collection was normalized in size to equalize the advantages of bigger images where lots of keypoints could be found. Altogether 23 subsets exist with four query-images except the historical search scenarios which have only one. The evaluation results are summarized and example tables and pictures are only given for for the scenarios of a floral ornament, fragments or historical picture and similar tombstones.

8.1 Snippet Queries

The performance of IPAs for the specific scenario are summarized in this section. The IPAs cannot be used to describe the epigraphics on the the surface of the tombstones. The NDCG values are very low because the textures between the Hebraic letters interfere greatly. Even great results show no reliable performance of the IPA as the descriptors of the snippet are matched with background elements like an ivy.

The subset of the floral ornaments is one of the largest and contains almost identical designs. The best detector-descriptor-combinations like SURF-SIFT, SIFT-SIFT and SURF-SURF always found at least two relevant images in the first ten ranks. But in the overall performance they show low NDCG values, wrong correspondences when the images were directly evaluated. The distribution of relevant images in the ranking is very high. A little example is given in the table 4 for CenSurE-SIFT. Even almost ideal rankings are mostly not caused by correct correspondences. The same behavior occurs using the subset of shell ornaments. Here even good NDCG-values don't indicate good performance cause relevant images are not found by similar descriptors.

	5-1					
n	1	2	3	4	5	6
IDCG	1,00	2,00	2,63	3,13	3,56	3,95
n	2	34	44	63	112	113
NDCG	0,50	0,26	0,30	0,33	0,37	0,40
	5-4					
n	1	2	3	8	10	42
NDCG	1,00	1,00	1,00	0,64	0,70	0,74

Table 4: Evaluation data for CBIR of similar floral ornaments using CenSurE-SIFT (rank n and NDCG). The table displays good performance in the first query and worse performance at the fourth query of the fifth subset.

Same results were shown when searching for similar symbols on the upper part of the tombstone. The subsets contained a hexagram (corners only and less texture), praying hand and a Levites can with lib. Both of the last two symbols have more textures and corners to be represented. In this scenario even high NDCG values had to be checked in detail. In most cases wrong correspondences were found between the indexed and the query image. Less keypoints were found on the symbol itself so it was not described by the IPAs. The distribution of the images in the ranking was very high which leads to the conclusion that IPAs cannot describe symbols, ornaments or epigraphics when using snippet images for queries.

8.2 Fragments / Historical Photography

The subset of historical photographies are from 1942-1954 and 1912. These microfilms are in bad shape suffering from overexposure or scratches. Additionally their were scanned with a small size of 740px x 1024px and always have a different perspective in comparison to the newer images from 2004. If an image is in good condition the performance of CenSurE-SIFT is outstanding. SURF-SIFT works well too but with lots of false correspondences. If the image condition is very bad the performance of all IPAs drops drasti-

cally. The NDCG values for the ranking of CenSurE-SIFT in table 5 show worse performance.

	1			**2**		
n	1	2	3	1	2	3
IDCG	1	2	2,63	1	2	2,63
n	60	78	80	75	77	81
NDCG	0,06	0,13	0,19	0,06	0,12	0,18
	3			**4**		
n	41	44	83	1	44	62
NDCG	0,07	0,14	0,2	0,38	0,45	0,51
	5			**6**		
n	1	2	3	6	24	63
NDCG	1	1	1	0,15	0,23	0,29
	7			**8**		
n	58	68	100	1	2	3
NDCG	0,07	0,13	0,19	1	1	1

Table 5: Evaluation data for CenSurE-SIFT showing the ranking and the NDCG values. Eight queries were tested.

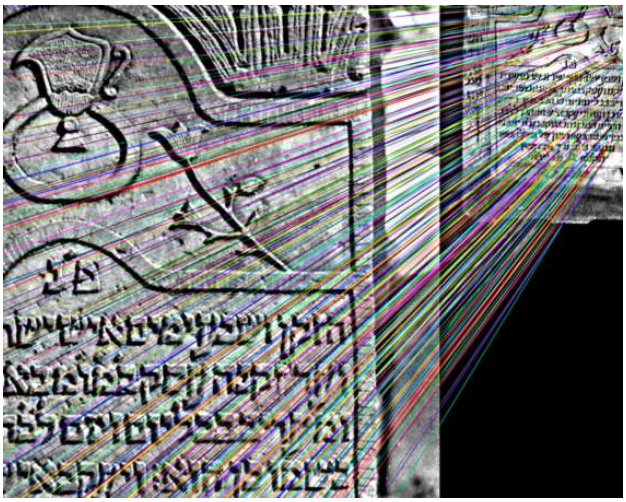


Figure 1: Working example at rank 1. All relevant images were found using CenSurE-SIFT mostly caused by surface and epigraphcs.

If a microfilm image is in good condition and the tombstone is plainly shown without interfering background-element it can be found. But most of the cases show that keypoints exist on the plain surface of the tombstone, not on other details like the ornaments (see image 1). So the use of IPAs is restricted to collections in good conditions which cannot always be the case. The overall performance is not reliable for CBIR-purpose.

8.3 Similar Tombstones

The subsets for similar tombstones vary greatly in their details like ornaments etc. A less shaped tombstone means less keypoints because of the homogenous surface. It is questionable if the remaining details like unique ornaments can be described by enough keypoints/descriptors. The object in the images have to be highly textured. If the indexed image is very similar, preferably identical in design, detector-descriptor pairs like CenSurE-SIFT and SURF-SIFT could be used to retrieve the relevant images as been shown in the example image 2. The table 6 shows good performance of both IPAs.

	2-3					**Pair**
n	1	2	3	4	5	
IDCG	1,00	2,00	2,63	3,13	3,56	
n	1	2	4	12	120	SU-SI
NDCG	1,00	1,00	0,70	0,78	0,82	
n	7	10	34	120	122	Ce-SI
NDCG	0,10	0,19	0,24	0,28	0,32	0,28

Table 6: Evaluation data for CBIR of similar tombstones using SURF-SIFT and CenSurE-SIFT (rank n and NDCG). The table displays good performance at the third query of the second subset .

Figure 2: Relevant image at rank 2 found by SURF-SIFT show good performance. Ornaments are mostly correct described.

Figure 3: Relevant image with wrong correspondences at rank 9 found by CenSurE-SIFT.

But as objects in the image-collection are unique in design and even if the picture in the subset look very similar, the performance is unstable for CBIR. As an example the table 7 shows worse performance of both IPAs in contrast. The NDCG drops as the relevant images have higher ranks. As figure 3 shows, even a relevant image found in the first ten ranks has wrong correspondences resulting in a low rank. Also the descriptors can represent irrelevant content like in figure 5. Additionally the details of a image will not be described as figure 4 show.

As a conclusion the performance of the IPAs are to unstable to use them for CBIR. A tombstone has to be too identical and even if it was found as a similar image, irrelevant parts interfere or details like ornaments are not described.

	3-4					Pair
n	1	2	3	4	5	
IDCG	1,00	2,00	2,63	3,13	3,56	
n	12	14	20	116	123	SU-SI
NDCG	0,08	0,15	0,22	0,26	0,30	
n	1	2	3	13	120	Ce-SI
NDCG	1,00	1,00	1,00	0,82	0,86	

Table 7: Evaluation data for CBIR of similar tombstones using SURF-SIFT and CenSurE-SIFT (rank n and NDCG). The table displays worse performance at the fourth query of the third subset.

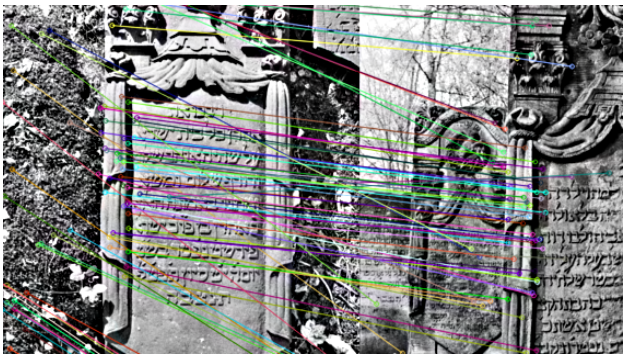


Figure 4: Relevant image at rank 1 found by CenSurE-SIFT. Details of tombstone are not described, surface intensities are too strong this could cause instable performance.

Figure 5: Relevant image rank 6 found by SURF-SIFT. Most of the matches are coming from the soil surface.

9 Conclusion

This article has used a scientific cultural collection of Hebraic tombstones to evaluate the performance of IPA for CBIR. The OpenCV 2.4.3 implementations of the state-of-the-art algorithms were evaluated: SIFT-SIFT, SIFT-BRISK, SURF-SIFT, SURF-SURF, SURF-BRISK and CenSurE-SIFT. The tolerances for these IPAs were determined in the field. They showed stable performance with a change in the view-angle up to 30°, overexposure or a change in scale caused by zoom. The difference of the focal length of two images showed that the IPA could handle up to 6mm with very stable results.

The test scenarios for CBIR showed that snippet-images dangerously decreases the amount of possible keypoints. Additional it is not guaranteed that keypoints are found on the specific parts of the image which has to be described. In some cases the descriptors are compared with irrelevant image-parts and returned as a better match than similar details on the tombstone.

If the collection contains older pictures like microfilm images the condition of these pictures is a crucial factor for CBIR. As shown in the synthetic tests for illumination change, overexposure can be handled, but if the perspective differs greatly, the image is scratched, some parts are suffering from overexposure the IPA cannot be used for CBIR.

If whole images with similar tombstones should be found, the objects have to be too equal or have to contain very similar, detailed and distinct attributes to describe the content. The overall performance for CBIR of similar tombstone is too unstable as IPA can be used to find images effectively.

Even though some algorithms show outstanding performance. The combination of CenSurE-detector and SIFT-descriptor showed good NDCG-values, enough keypoints on the tombstone and a good distribution of the relevant images in the ranking. Other algorithms like SIFT-SIFT, SURF-SIFT or SURF-SURF had the advantage to create lots of keypoints which meant a higher probability that images could be found.

Because only the implementations of IPAs contained in the OpenCV-distribution were used, there are more algorithms like affine-SIFT to be evaluated [Morel and Yu, 2009]. Another approach could be to create hybrid descriptors using the Daisy-descriptor with most common ones like SIFT. The *Local Energy based Shape Histogram* (LESH) could be used to describe the shape and to filter irrelevant outlier. Additionally other algorithms than RANSAC like [Moisan and Stival, 2004] could be used to determine correspondences.

References

[Agrawal *et al.*, 2008] M. Agrawal, K. Konolige, and M. R. Blas. CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. In *Computer Vision – ECCV 2008*, volume 5305 of *Lecture Notes in Computer Science*, pages 102–115. Springer, 2008.

[Alahi *et al.*, 2012] A. Alahi, R. Ortiz, and P. Vandergheynst. FREAK: Fast Retina Keypoint. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517. IEEE, 2012.

[Aman *et al.*, 2010] J.M. Aman, J. Yao, and R.M. Summers. Content-based image retrieval on CT colonography using rotation and scale invariant features and bag-of-words model. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1357–1360. IEEE, 2010.

[Bay *et al.*, 2008] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[Begen *et al.*, 2011] A.C. Begen, K. Mayer-Patel, V.R. Chandrasekhar, J. Bach, B. Girod, D.M. Chen, S.S. Tsai, N. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, and R. Grzeszczuk. The stanford mobile visual search data set. In *Proceedings of the second annual ACM conference on Multimedia systems - MMSys '11*, pages 117–122. ACM Press, 2011.

[Calonder *et al.*, 2010] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: binary robust independent elementary features. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 778–792. Springer-Verlag, 2010.

[Dahl *et al.*, 2011] A.L. Dahl, H. Aanæs, and K.S. Pedersen. Finding the Best Feature Detector-Descriptor Combination. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 318–325. IEEE, 2011.

[Fraundorfer and Bischof, 2005] F. Fraundorfer and H. Bischof. A novel performance evaluation method of local detectors on non-planar scenes. In *In Workshop Proc. Empirical Evaluation Methods in Computer Vision (CVPR)*, 2005.

[Gauglitz *et al.*, 2011] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking. *International Journal of Computer Vision*, 94(3):335–360, 2011.

[Gil *et al.*, 2010] A. Gil, O.M. Mozos, M. Ballesta, and O. Reinoso. A comparative evaluation of interest point detectors and local descriptors for visual SLAM. *Machine Vision and Applications*, 21(6):905–920, 2010.

[Harris and Stephens, 1988] C. Harris and M. Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[Järvelin and Kekäläinen, 2002] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[Kampel *et al.*, 2009] M. Kampel, R. Huber-Mörk, and M. Zaharieva. Image-Based Retrieval and Identification of Ancient Coins. *IEEE Intelligent Systems*, 24(2):26–34, 2009.

[Leutenegger *et al.*, 2011] S. Leutenegger, M. Chli, and R.Y. Siegwart. BRISK: Binary Robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, pages 2548–2555. IEEE, 2011.

[Lowe, 2004] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Matas *et al.*, 2002] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *Proceedings of the British Machine Vision Conference 2002, BMVC 2002, Cardiff, UK, 2-5 September 2002*. British Machine Vision Association, 2002.

[Mikolajczyk and Schmid, 2003] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, pages II–257–II–263. IEEE, 2003.

[Mikolajczyk and Schmid, 2005] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[Mikolajczyk, 2004] K. Mikolajczyk. Scale & Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[Moisan and Stival, 2004] L. Moisan and B.r Stival. A Probabilistic Criterion to Detect Rigid Point Matches Between Two Images and Estimate the Fundamental Matrix. *International Journal of Computer Vision*, 57(3):201–218, 2004.

[Moreels and Perona, 2007] P. Moreels and P. Perona. Evaluation of Features Detectors and Descriptors based on 3D Objects. *International Journal of Computer Vision*, 73(3):263–284, 2007.

[Morel and Yu, 2009] J.-M. Morel and G Yu. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[Rosten and Drummond, 2006] E. Rosten and T. Drummond. Machine Learning for High-Speed Corner Detection. In *Computer Vision – ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 430–443. Springer Berlin Heidelberg, 2006.

[Rublee *et al.*, 2011] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571. IEEE, 2011.

[Schmid *et al.*, 2000] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of Interest Point Detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[Sperker and Henrich, 2013] H.-C. Sperker and A. Henrich. Feature-based Object Recognition a case study for car model detection. In *11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 127–130. IEEE, 2013.

[Valle *et al.*, 2006] E. Valle, M. Cord, and S. Philipp-Foliguet. Content-Based Retrieval of Images for Cultural Institutions Using Local Descriptors. In *Geometric Modeling and Imaging–New Trends (GMAI'06)*, pages 177–182. IEEE, 2006.

In search of Honduras – Case report of developing local search for a developing country

Dirk Ahlers
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
dirk.ahlers@idi.ntnu.no

Abstract

This paper reports on the efforts to establish a research project on a geospatial search engine for the Latinamerican country of Honduras as well as establishing an encompassing research group on information retrieval during the author's stay there at a local university. Honduras is an interesting example of the challenges for information and knowledge management in developing countries as it combines many of the issues that might be encountered. These include low Web coverage in a low-resource country, cross-language information retrieval, and generally, work in challenging circumstances. The specific focus on geospatial information uncovers further issues that need to be addressed, such as informal addressing systems, broad or incorrect location references, or insufficient ground truth in databases. The research stay yielded valuable experiences, even if the tangible results of the project stay behind the original goals.[1]

1 Introduction

Location is as an important organizing principle for many Web search tasks. In most industrialized nations the search for locations features prominently within search engines and users are used to seamlessly working local search with a multitude of correct results. It works this well because there is both good data available and tailored technology to make use of it.

But in many developing countries, the situation is gravely different. Local search may not be as accessible, important places are missing, or the information density is rather low, only mentioning a name without more in-depth information. Errors or inaccuracies may further complicate the situation, if information is even available in the first place. The basic usefulness of mobile phones and their applications in developing countries have already been shown[Duncombe and Boateng, 2009; Donner, 2008; Hagan *et al.*, 2012; Frias-Martinez *et al.*, 2012; Frohlich *et al.*, 2012], but search in general has received less attention [Chen *et al.*, 2010; 2009; Kothari *et al.*, 2009]. In this paper, we want to examine the special case of geospatial search in developing regions and we explore it by means of the Latinamerican country of Honduras. The basis question then is, how can one make local search (or even Web search in general) work in a low-resource country with only very little Web coverage where few people even have Internet access? The available potential has to be assessed and a roadmap drawn for the realization.

An overview of the project's challenges of a country-specific search engine has previously been published, discussing the challenges at the beginning of the project [Ahlers, 2011]. We elaborate on these challenges with some added hindsight of research performed in the meantime in the form of this consolidated report.

1.1 Social background

Some socio-economic statistics abridged from [Ahlers and Henze, 2012] should help to better understand the expected background and population. Honduras is a developing country which is classified as a lower-middle-income economy[2], ranks 121st in the Human Development Index worldwide, and is the sixth-poorest country in Latin America[3] with 23% of the population below the poverty line[4] and 60% of the population below the national poverty line[5].

Despite these numbers, Internet use is rising fast, with 11% of Internet users in the population in 2010 up from 1% in 2000[6]. Very little reliably data is available on mobile phone or mobile Internet use in Honduras. Even a project partner, a local telecommunications company, could not readily provide such information. While computer ownership is at only 2.5%, peoples' mobile subscriptions are much more promising. From 3% in 2000, they have surpassed 100% in 2008 and were at 125% in 2010[7]. This oversaturation can be explained by the practice of having mobile phones for different providers to take advantage of lower calling cost. More useful is the estimate of people actually owning a mobile phone at about 75% in 2010[8]. Informal estimates for the share of smartphones with Internet

[1] The research described here was carried out at UNITEC – Universidad Tecnológica Centroamericana, Tegucigalpa, Honduras.

This paper was written during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme which has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 246016.

[2] http://data.worldbank.org/country/honduras

[3] http://hdrstats.undp.org/en/countries/profiles/HND.html

[4] http://www.unicef.org/infobycountry/honduras_statistics.html

[5] http://data.un.org/Data.aspx?d=MDG&f=seriesRowID\%3A581

[6] http://devdata.worldbank.org/ict/hnd_ict.pdf

[7] http://data.worldbank.org/indicator/IT.CEL.SETS.P2/countries/HN?display=graph

[8] http://www.latinobarometro.org/

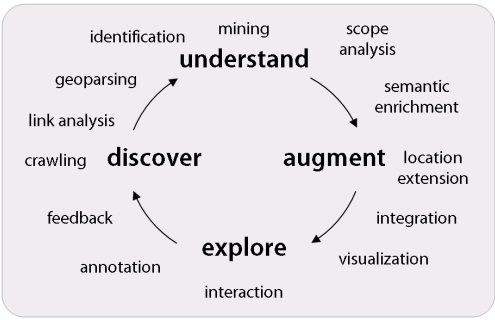


Figure 1: Process life cycle of geospatial Web information (from [Boll and Ahlers, 2008])

access are around 10%, which hints that a lot of Internet use happens on mobile devices.

While the very unevenly distributed use of communication technology denotes a strong digital divide in Honduras, mobile phone and Internet use is rising. This is reflected in the strategies of mobile phone providers, who aim to cater to the bimodal use by offering initial smartphone solutions, but chiefly providing texting and USSD (menu-based dialogs) services. Overall, this means that Honduras is a latecomer but is quickly catching up, which makes local search a viable option for future applications. With increasing use of online services, this might also reach larger parts of the population [Chong and Micco, 2002].

This might also help answer the question of whether such a service is really needed and whether there are not more pressing needs. Currently a large part of the population cannot afford expensive services, lives mostly by subsistence farming and might only buy at local small shops in their neighborhood. These people mostly will not need these services. But with a slowly growing middle class and more people expecting local search to work, many users would benefit. Thus, while local search might be regarded as a luxury, especially in view of a large poor population, the adoption of services for the population capable of affording the necessary technology to participate, can still be a worthwhile undertaking and may, by uncovering and presenting available information, improve the general data situation and might lead to other undertakings that might benefit a larger amount of the population.

2 Development process for geospatial search

Many research projects have been undertaken to extract the location information for a special vertical geospatial Web search, e.g., [Ahlers and Boll, 2007; Purves *et al.*, 2007; Markowetz *et al.*, 2005; Borges *et al.*, 2003]. Also commercial services such as Google Maps, Google Earth, Bing Maps, or Yahoo! Maps are building location-based search applications and creating indexes of geospatial information [Ahlers, 2012a]. However, no tailored geospatial search for Honduras exists and the existing services showed some shortcomings. The initial project idea therefore was to work towards a geospatial database of Honduran locations, places, points of interest and Web pages in close cooperation with a local mobile telecom provider [Ahlers, 2011].

The main challenges concern the requirements and data situation; and the analyzing, extracting and indexing of location data. We base this on the general process life cycle of finding and using geospatial Web information presented in [Boll and Ahlers, 2008]. This includes processes of discovery, understanding, augmentation, and exploration as shown in Fig. 1. Extending this data-centric viewpoint towards the process for a whole search engine, we arrive at a development process for a geospatial search engine [Ahlers, 2013c]:

General situation assessment and understanding: The first step in creating a new geospatial search engine is to get an overview of what applications and services may be interesting and relevant. Specific undertakings range from market analysis, viability analysis, and data source investigations over requirements engineering to data analysis and user studies, regarding search and/or mobile applications. These can provide initial insight into the needs and gaps in the currently available systems and would also provide requirements for subsequent steps.

- **Market:** Quite an obvious step, a market analysis should find out if the intended solution is actually new and needed or if other participants can provide some help.
- **Data:** The data situation is the most important, as it plays an important role in the feasibility analysis. This includes how much data is available, what its characteristics and quality is, which additional sources are available.
- **Users:** As a search engine is an offer towards users, their requirements and situation have to be taken into account. Also, the target group needs to be identified or selected first. Informal interviews, user studies, or usage observation can provide valuable insight.

Building a knowledge base: To aid in the extraction of geospatial data, a bootstrapping of known geographical placenames is normally used. Such gazetteer data can either be directly available or needs t be collected and combined from multiple sources.

Discovery and analysis of data sources: To get a good overview, many data sources have to be explored to understand the type of information they offer and estimate the amount of data available.

Extraction and analysis of data: Specific extraction methods have to be developed. These especially concern geoparsing, i.e., the identification and extraction of location references, and geocoding, i.e., the grounding of location references to geographic coordinates.

Web crawling: For crawling a country-specific Web, the characteristics and boundaries of the country in the Web have to be defined to setup parameters for the crawl.

Source integration: The different identified sources need to be integrated. Local search is not just a document search but is also an entity search in the sense that it models georeferenced documents as well as the actual georeferenced entities described in the documents. For this cross-correlation and entity resolution across all results is performed.

Building interfaces: Finally, based upon the available data and potential augmentations from additional analysis steps, interfaces for users to actually access the index are to be developed.

Then we further have cross-sectional issues such as problems encountered along the way in all the different steps,

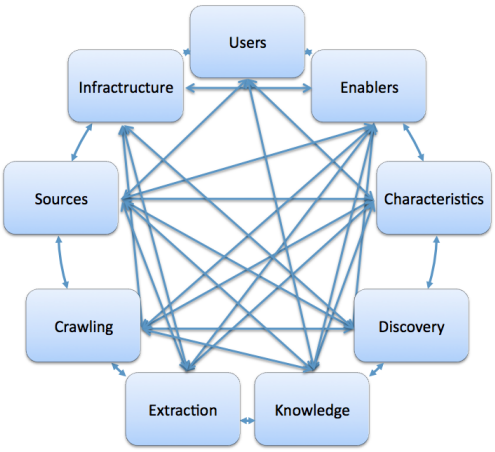


Figure 2: Components in the development of geospatial search (from [Ahlers, 2013c])

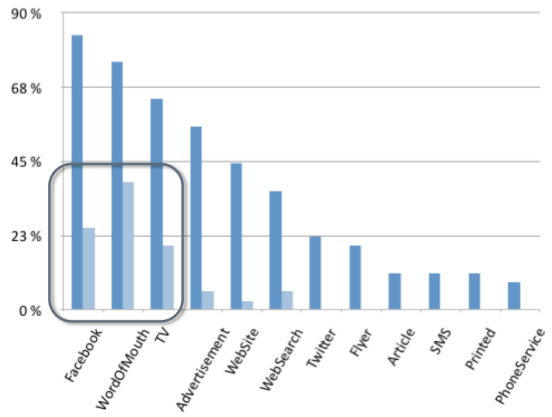


Figure 3: Information seeking behaviour: used sources. Answers to the question "How do you find out about local things?" Dark blue: multiple choice answers, light blue: only primary choice. (from [Ahlers and Henze, 2012])

which have to be solved and integrated. A major point here, especially for a researcher who is a foreigner in the country, is to be aware of own biases and preconceived notions or assumptions. To this end, one should try to talk to as many people as possible and try to collect as much information as possible, especially if it is contradictory. Actually, this is a good hint that something is not as easy as it might seem. As in good journalism, it is better to ask at least two sources so that assumptions cannot go wrong that easily. This classification and more formal description is a work in progress [Ahlers, 2013c] which is visualized with its strong interdependencies in Fig. 2.

3 Challenges and realization

We will shed some light onto this process by examining it from the viewpoint of geospatial search in Honduras. We discuss challenges identified in [Ahlers, 2011] and organize them along the development process. We also expand the presentation with research results gathered in the meantime.

3.1 Search engine market

The big search engines provide map data, sometimes at very good quality and also provide some map-based local search. However, compared to other regions of the world, there is very little local information available and its depth is very low, often offering nothing more than a name and a rough location. Instead of waiting for other players to take up the market, the current situation offers a unique potential to build a Honduran geospatial search engine. While the Web coverage is still low and the address scheme makes exact location extremely difficult, Web usage is rising, potential data sources exist, and people begin using location-based services which creates sufficient demand and support. Honduras has no own search engine, and people regularly use the big international search engines, mainly Google. There exists a local domain, google.hn, which seems to give a slight preference to pages about Honduras as part of the location customization[9]. Local search [Ahlers, 2012a] has initial data, but is far from a comprehensive coverage – which is in part the topic of this research. Similar research projects have been described, for example, for Chile [Mendoza *et al.*, 2009], Portugal [Gomes *et al.*, 2008], Brazil [Borges *et al.*, 2003], or Germany [Markowetz *et al.*, 2005; Ahlers and Boll, 2007].

[9] http://support.google.com/websearch/bin/answer.py?hl=en&answer=179386

3.2 Users

The most comprehensive overview on mobile phone use in developing countries was a literature study [Donner, 2008] reviewing about 200 studies. An ethnographic study of 26 participants by means of interview and shadowing was done to examine the use of mobile phones to maintain a social network for migrant workers in cities in China [Lang *et al.*, 2010]. It was noted that social interactions happen throughout the day, with little distinction between work and spare time, for non-factory workers. The use of mobile phones can also be understood as a method of empowerment in developing nations. [Blumenstock and Eagle, 2010] analyzed patterns of mobile phone use in Rwanda with a joint approach of using demographic surveys and call detail records analysis of a mobile operator, additionally discussing other surveys done in the developing world.

For insights focused on Honduras, we did a user study on the use of local search and local information seeking behavior in general [Ahlers and Henze, 2012]. Among other things, we found that the preferred modes and sources of information search are word-of-mouth or existing knowledge about locations, combined with a knowledgeable social circle as seen in Fig. 3. This social aspect may also explain why the most used online source is Facebook, followed with a distance by search engines. Overall, local search is not that prevalent and social aspects are very strong. Due to rather little search happening on the move, there is less of the usually associated here-and-now mentality [Ahlers and Henze, 2012]. One implication of these findings is that the search engine most probably would have to follow a hybrid approach to access a variety of data sources, also including the social networks, or even employing crowdsourcing to establish both relevant information and trust.

Privacy and security in geolocalization

In Honduras, privacy concerns are connected to bigger security concerns. Due to a high level of targeted criminality, many people prefer to keep their personal information, especially their location, very private [Ahlers and

Figure 4: Entity example with street-level address from different sources, counter-clockwise: Facebook, yellow pages, foursquare, OpenStreetMap (from [Ahlers, 2012b])

Henze, 2012]. Yet, many entries in location-sharing services explicitly concern people's own houses ("Mi casa", "My house"). In these cases, the functionality seems to override security concerns. Additionally, in part due to the security situation, people will not walk and rather take cars, taxis, or buses and would only get their phone out in safe places, but not just on the street. This influences usage, which is not as spontaneous as in other countries and happens less on the move. Therefore, developed services need to ensure safe handling of people's location data and also consider a less spontaneous mobile use.

3.3 Country-specific characteristics

A very challenging characteristic of Honduran location references is that exact locations in the form of addresses with house numbers in a formal, high-granularity addressing scheme are usually not given. This seriously impedes a high-granularity approach that would try to map information to individual buildings. There are some areas or smaller cities where a rectangular street grid exists, which usually also allows for a better addressing scheme. However, in most regions, location references are given by city name, city district and sometimes the street name. Various other forms of descriptions have evolved that allow finding a certain building. Often these are given additional directional information such as nearby landmarks or well-known buildings. Sometimes a description is accompanied by a sketched map, a so-called croquis to help with orientation. The usually encountered low-granularity location references – in common Web pages as well as in databases – pose a particular problem to geoparsing, the extraction of location references from general text. The example in Fig. 4 shows varying descriptions and given locations for a place.

3.4 Data

The official language of Honduras is Spanish, and normally, the articles from this language would be expected to sufficiently cover the country. However, at the Caribbean coast and Bay islands of the country, English is a recognized regional language and is more frequently spoken, mainly by the Garifuna population. Additionally, this is the main area for foreign tourism and much information of a tourism nature is more comprehensively available in English than in Spanish. Therefore, a search engine that should cover the whole of Honduras needs to use cross-language retrieval techniques.

Furthermore, there is an interesting anomaly for Honduras in that much information about the country exists in English instead of Spanish. For example, while examining Wikipedia, we noted that there exist more English

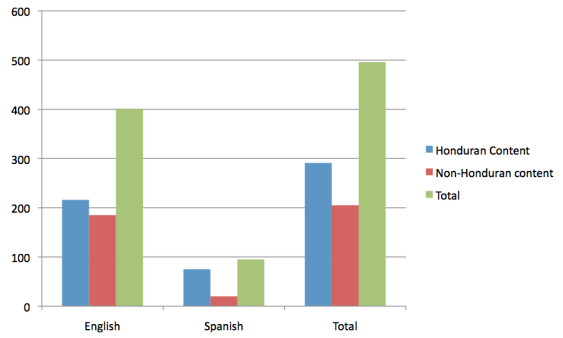


Figure 5: Provenance of pages with topic Honduras in DMOZ [Ahlers *et al.*, 2012]

geotagged articles than Spanish ones. This is shown as an example in Section 4.

In trying to define the Honduran Web, we took two separate approaches as discussed in [Ahlers *et al.*, 2012]. First, we had a look at DMOZ, and second, we built a list of all Honduran domains under a .hn ccTLD. The DMOZ Open Directory has relatively little coverage for Honduras. In the English hierarchy, it contains 421 entries, with 10 from .hn (2.5%) and 411 others; for Spanish, there are only 96 entries, but 46 are from .hn (48%), only 50 from others. The English part contains mainly travel sites and general descriptions while the Spanish contains actual local pages. Fig. 5 shows the provenance of the DMOZ data for the English and Spanish categories, classified by being from inside or outside the country.

We were able to obtain a list of registered domains through inofficial channels. Of the 5780 registered domains, we were able to only find 893, meaning that probably a lot of them are not used. Trying to confirm a suspicion of a majority of Web hosting happening outside the country, we used a commercial database to estimate the hosting locations of the reachable domains. The country assignment is shown in Fig. 6 on a logarithmic scale. A large amount is actually assigned in Honduras, but the majority in the US, with other American countries following behind. A deeper inspection of the US hosts revealed these to be mostly in southern countries associated with a large Latinamerican population, while the other countries are often related to the owners or investors of businesses. However, we also found 25% of governmental domains hosted in the US, confirming the suspicion of a 'digital divide' [Nakahira *et al.*, 2006].

Due to unknown deterrents, possibly high server costs or similar, a very high number of Honduran businesses not only host outside the country, but also do not use the .hn domain and instead opt for a generic .com domain. Further complications arise from the fact that many businesses forego an own Web presence and instead create a Facebook page. This makes it more difficult to gather all relevant domains for Honduras and actually include all relevant information. As a first rough estimate on the sites that are available, all Honduran DMOZ links are correlated with the known domains registered for .hn. This gives a number from a conservative 6200 to a probably heavily overestimated 225,000 domains overall, with its geometric mean below 40,000 domains. This means that .hn domains only represent an estimate of between 5% to 61% of all relevant domains for the country.

We are looking into ways to reliably identify .com do-

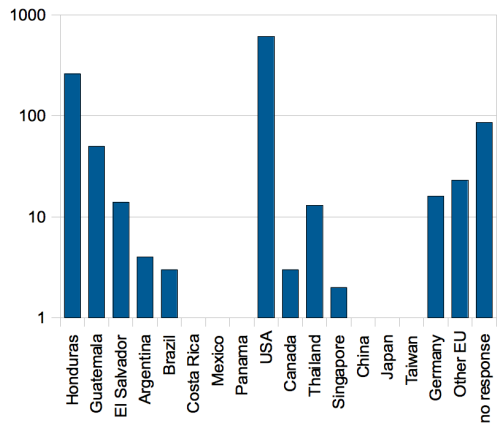


Figure 6: Distribution of hosting countries for .hn domains [Ahlers *et al.*, 2012]

mains carrying Honduran content, using a mixture of location analysis, language, and interlinking.

3.5 Building a gazetteer

For a small country with very limited Web coverage, the geotagged Wikipedia articles can provide some reliable initial knowledge (cf. Section 4). We injected the Wikipedia data into a larger gazetteer that we are generating for Honduras. For the gazetteer, we use initial data from geonames.org to serve as a bootstrapping of the search engine's knowledge about places and placenames.

3.6 Source integration

Since the Web shows only very low coverage for Honduras, we aim to additionally include specific structured datasources into the search engine index. In view of the user behavior, we also aim to include social networks, which carry a lot of location-relevant information. This mandates that the search engine follows a hybrid approach of both Web search with georeferencing of documents and additionally database access and merging for specific data sources. However, even for only Web-based location search, cross-correlation and entity resolution is needed to identify identical entities (cf. [Ahlers, 2013a; Sehgal *et al.*, 2006]). The multiple sources also may provide a remedy for low-granularity locations, as they may be combined in some cases to improve the accuracy of geocoding [Ahlers, 2012b] (cf. Fig. 4).

4 Wikipedia cross-language linkage

Wikipedia includes a vast amount of articles about places, many of which include a geographic coordinate that locates the content in the real world. This makes it a good starting point to bootstrap out knowledge about a country. Cross-language links are possible between articles describing the same place in multiple languages. Ideally, any article about a place of interest would include its correct coordinate and links to other Wikipedia language versions. This poses the question of how we can identify geotagged Wikipedia articles that describe the same place across different languages and what are useful similarity measures? This section is based on [Ahlers, 2013b].

The official language of Honduras is Spanish, and normally, the articles from this language would be expected to sufficiently cover the country. However, much information about the country comes from outside. We therefore also looked into English articles to see if these would increase the coverage. This prompted the discovery of an interesting anomaly: Honduras has more English geotagged articles than Spanish ones.[10] Our aim is to merge both language versions and identify identical places on the article level.

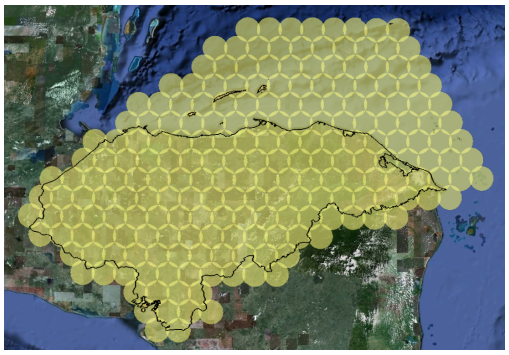


Figure 7: Hexagonal grid arrangement of packed circles for API queries

Table 1: Examples of sibling articles

Spanish title	English title
Tegugigalpa	Tegucigalpa
Aeropuerto Internacional Toncontín	Toncontín International Airport
Santa Bárbara (Santa Bárbara)	Santa Bárbara, Honduras
El Paraíso	El Paraíso Department
Departamento de Copán	Copán Department
Virginia (Honduras)	Virginia, Lempira
Parque nacional Pico Bonito	Pico Bonito National Park

The approach we are following is called, varyingly, record linkage, entity fusion, entity resolution, or duplicate detection. [Sehgal *et al.*, 2006] gives an overview on geospatial entity resolution. [Overell and Rüger, 2006] use Wikipedia to ground and disambiguate place names. Merging geonames data to Wikipedia, [Hoffart *et al.*, 2012] use a simple approach that if more than one entity exists in geonames with the same name, the closest entity within a distance of 5km is chosen. [Liu and Yoshioka, 2011] add a translation approach to improve the title matching. This work is closest to ours, however, is still lacking in a graded consideration of both textual and positional similarity.

We use the APIs of geonames and wikilocation to retrieve articles, because we do not want to retrieve the whole Wikipedia dump for the relevant languages. Since the services put a limit of 20km in the radius for simple circle search, we have to use a covering of queries for the region, which is shown in Fig. 7.

4.1 Wikipedia Language Fusion

We define the data fusion method in terms of finding language *siblings*. We combine text- and entity-based merging methods with geographic conflation techniques. For each article, we select and rank candidate siblings in the respective other language. The merging is based on the title and the location as shown in the examples in Table 1. The geographic feature type is rarely present, so it can only used as second-level evidence. For a comparison of two potential siblings, there are four cases to consider, 1. Names

[10] Such anomalies exist in many countries: `http://www.zerogeography.net/2012/10/dominant-wikipedia-language-by-country.html`

and coordinates match, 2. Names match, coordinates do not match, 3. Names do not match, coordinates match, 4. Names do not match, coordinates do not match. The first case is obviously trivial. All other cases are modeled by similarity measures based on non-exact matching.

Coordinates can vary due to different interpretations of the center of an area or variations in user-generated coordinates, especially for larger entities [Ahlers and Boll, 2009]. We limit the amount of candidate siblings we have to examine by cutting off the location similarity with a perimeter of 10km around an article's *location*, inside of which all candidates are examined.

For all candidates' *titles* within the radius, three cases would constitute a match, 1) titles match exactly, 2) titles match with small variations, 3) title can be translated and transposed to match. We define a title translation distance TTD as an editing distance similarity measure based on partial translations and permutations. The first case is easy, the second case only needs to account for spelling variations, which we do with a Levenshtein editing distance adapted with a weight relative to the term length and with a reduced penalty for accents and tildes. Interestingly, most proper nouns are identical or very similar in both languages and can be well accounted for with the adapted Levenshtein distance. However, common nouns have to be translated and the order of terms within a placename also be changed. The translation table was filled mostly with relevant geographical feature types, taken from geonames[11] (e.g., airports, islands, mountains, stadiums, cities, parks, etc.). Heuristics were generated about some conventions that we observed for both languages. For example, for municipality and department names, *Santa Bárbara (Honduras)*$_{ES}$ puts the higher-level administrative body, in this case the country name, in brackets, while *Santa Bárbara Department*$_{EN}$ uses the administrative type without a hint towards the country. This is helpful as often, departments and capital cities have the exact same coordinates.

To cover permutations, we employ a list of transposition heuristics as part of the translation. The inverted-first-pair translation swaps the first two terms: *Congreso Nacional de Honduras*$_{ES}$ $\rightarrow$ *National Congress of Honduras*$_{EN}$. The inverse order translation swaps first and last terms: *Río coco*$_{ES}$ $\rightarrow$ *Coco river*$_{EN}$; and the inverted-first-pair-moved translation inverts the order of the first two words and moves them to the end: *Parque nacional Pico Bonito*$_{ES}$ $\rightarrow$ *Pico Bonito National Park*$_{EN}$. We generate all potential variations of the title, including translations, and chose the variation with the minimum TTD and the smallest location distance as a sibling.

4.2 Evaluation

Honduras had 342 Spanish and 405 English articles, an 18% English overrepresentation. We use the Wikipedia language interlinks as a ground truth for the evaluation. For all articles, the Wikipedia page and its interlinks were manually examined to determine siblings.

The algorithm resulted in 317 article pairs, 25 only Spanish articles, and 88 only in English (Fig. 8). Of these, 99.4% are correct pairs. The articles without siblings are 84% correct, with 16% false negatives. Only two pairs were false positives. The first wrongly identifies *Comayaguëla*$_{ES}$ and *Comayagua*$_{EN}$ because they have both

[11] http://www.geonames.org/export/codes.html

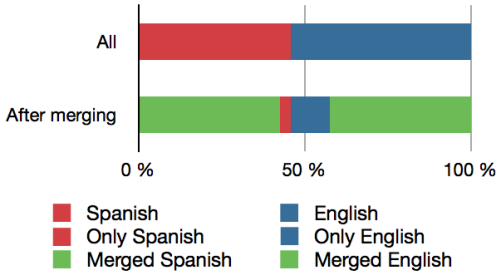


Figure 8: Results of merging Wikipedia articles

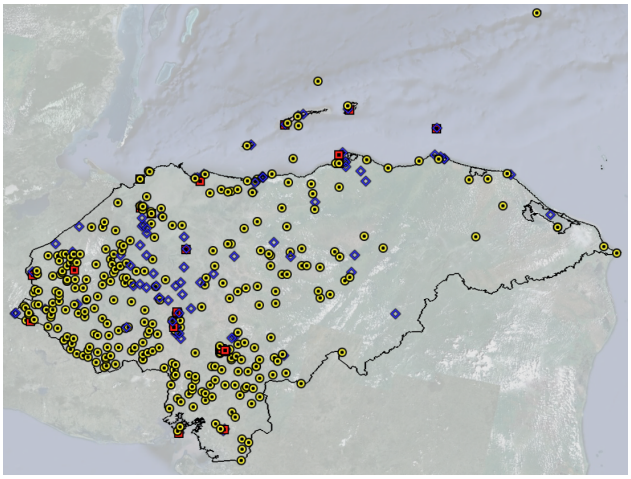


Figure 9: Mapping of geotagged Wikipedia articles, yellow ⊙: merged articles, red ⊡: Spanish, blue ◇: English

the exact same coordinates, even if the cities are about 80km apart. In this case the error lies with the incorrect coordinate in the article. The second assigns the department *Comayagua*$_{ES}$ to the city *Comayagua*$_{EN}$, which surprisingly is also wrong in the interlinks. This induces a subsequent error in the false negatives: *Comayagua (ciudad)*$_{ES}$ and *Comayagua Department*$_{EN}$ each had no siblings, but should have been matched to the previous pair. The other false negatives concern mostly slight mismatches paired with distanced coordinates, but also some more debatable ones, such as *Roatán (municipio)*$_{ES}$ and *Coxen Hole*$_{EN}$. When mapping articles as shown in Fig. 9, we see no language dominating certain regions but both languages distributed rather similarly.

The approach is to be extended by using the learned characteristics in an entity fusion approach for gazetteer data as well, which will make stronger use of the feature type. This is expected to help in cleaning up and linking geonames data to other sources.

5 Conclusion and sustainability

Overall, Honduras provides an ideal ground for research due to its numerous challenges that will require the combination of many different fields of search engine technology and geographic information retrieval. Furthermore, due to the small size of the country, even a research prototype can be expected to cover a huge fraction of the Honduran Web, thus building up a comprehensive and usable index.

Even with the complications described above such as low coverage, ambiguous or insufficient location references, non-local Web hosting, etc., there exists sufficient data to develop a prototypical search engine, starting with

some "easier" aspects of the data. A more organizational issue was that doing research in the country was very different as there was not a strong background of research or even development present. In some cases, resources had to be procured in a very backhanded way. For example, there did not seem to be an official way to get certain numbers, but a student knew someone who might have access to certain data. There were little official or formal ways of interaction and in terms of doing research, Honduras very clearly shows the signs of a developing country. A rather sad aspect of the project is that, even while there was enthusiastic support of the general idea, there was not enough motivation to continue the project during a funding issue or even to properly recover it afterwards so that no substantial sustainability could be reached until this point for the full project, but the partial solutions described here are still useful for future projects. However, the whole project was a great experience and due to some students moving into related industry jobs working on similar ideas, some knowledge will remain in the country and be developed further.

Overall, the project offered very good potential for research. It also drives one to challenge certain assumptions, as many factors have to be established here that can be taken for granted in other regions. We hope that the results will be used further in the country and we also hope that may be applicable to other regions as well. Furthermore, the work carries a large potential for follow-up research, as many interesting questions are still open.

Acknowledgment

We thank our students Isaac Martinez and Jose Matute of UNITEC – Universidad Tecnológica Centroamericana, Tegucigalpa, Honduras, who worked within the research project and provided initial implementations.

References

[Ahlers and Boll, 2007] Dirk Ahlers and Susanne Boll. Location-based Web search. In Arno Scharl and Klaus Tochterman, editors, *The Geospatial Web. How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society*. Springer, London, 2007.

[Ahlers and Boll, 2009] Dirk Ahlers and Susanne Boll. On the Accuracy of Online Geocoders. In *Geoinformatik 2009, Osnabrück*, volume 35 of *ifgiprints*, pages 85–91, Münster, 2009.

[Ahlers and Henze, 2012] Dirk Ahlers and Niels Henze. ¿Donde está? – Surveying Local Search in Honduras. In *MWB2012 – Workshop on Mobility and Web Behavior at MobileHCI2012*, 2012.

[Ahlers *et al.*, 2012] Dirk Ahlers, Jose Matute, Isaac Martinez, and Chandan Kumar. Mapping the Web resources of a developing country. In *GI Zeitgeist 2012, Proceedings of the Young Researchers Forum on Geographic Information Science*, volume 44 of *IfGi prints*, pages 117–122. AKA, 2012.

[Ahlers, 2011] Dirk Ahlers. Towards Geospatial Search for Honduras. In *LACNEM 2011*, San José, Costa Rica, 2011.

[Ahlers, 2012a] Dirk Ahlers. Local Web Search Examined. In Dirk Lewandowski, editor, *Web Search Engine Research*, volume 4 of *Library and Information Science*, pages 47–78. Emerald, 2012.

[Ahlers, 2012b] Dirk Ahlers. Multi-source conflating index construction for local search in a low-coverage country. In *LA-WEB 2012 (8th Latin American Web Congress)*, pages 28–31, 2012.

[Ahlers, 2013a] Dirk Ahlers. Business Entity Retrieval and Data Provision for Yellow Pages by Local Search. In *Integrating IR technologies for Professional Search Workshop @ ECIR2013*, 2013.

[Ahlers, 2013b] Dirk Ahlers. Lo mejor de dos idiomas – Cross-lingual linkage of geotagged Wikipedia articles. In *ECIR2013*, 2013.

[Ahlers, 2013c] Dirk Ahlers. Towards a development process for geospatial information retrieval and search. WWW '13, pages 143–144, Geneva, Switzerland, 2013. IW3C.

[Blumenstock and Eagle, 2010] J. E. Blumenstock and N. Eagle. Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda. In *4th International Conference on Information and Comunications Technologies and Development (ICTD)*, 2010.

[Boll and Ahlers, 2008] Susanne Boll and Dirk Ahlers. A Web more Geospatial: Insights into the Location Inside. In David De Roure and Wendy Hall, editors, *Workshop on Understanding Web Evolution: A Prerequisite for Web Science (WebEvolve2008) held at WWW08*, volume 300, pages 40–43, Beijing, China, 2008. Web Science Research Initiative.

[Borges *et al.*, 2003] Karla A. V. Borges, Alberto H. F. Laender, Claudia B. Medeiros, Altigran S. Da Silva, and Jr. Clodoveu A. Davis. The Web as a Data Source for Spatial Databases. In *Anais do V Brazilian Symposium on Geoinformatics*, 2003.

[Chen *et al.*, 2009] Jay Chen, Lakshminarayanan Subramanian, and Jinyang Li. RuralCafe: Web Search in the Rural Developing World. In *18th International World Wide Web Conference*, WWW '09, pages 411–420, New York, NY, USA, 2009. ACM.

[Chen *et al.*, 2010] Jay Chen, Lakshmi Subramanian, and Eric Brewer. SMS-Based Web Search for Low-end Mobile Devices. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, MobiCom '10, pages 125–136, New York, NY, USA, 2010. ACM.

[Chong and Micco, 2002] Alberto Chong and Alejandro Micco. The internet and the ability to innovate in latin america. RES Working Papers 4291, Inter-American Development Bank, Research Department, January 2002.

[Donner, 2008] Jonathan Donner. Research approaches to mobile use in the developing world: A review of the literature. *The Information Society*, 24:140–159, May 2008.

[Duncombe and Boateng, 2009] Richard Duncombe and Richard Boateng. Mobile phones and financial services in developing countries: a review of concepts, methods, issues, evidence and future research directions. *Third World Quarterly*, 30(7):1237–1258, 2009.

[Frias-Martinez *et al.*, 2012] Vanessa Frias-Martinez, Jesus Virseda, and Aldo Gomero. Educamovil: a mobile learning tool for low-income schools. MobileHCI '12, pages 169–172, New York, NY, USA, 2012. ACM.

[Frohlich *et al.*, 2012] David Frohlich, Simon Robinson, Kristen Eglinton, Matt Jones, and Elina Vartiainen. Creative cameraphone use in rural developing regions. MobileHCI '12, pages 181–190, New York, NY, USA, 2012. ACM.

[Gomes *et al.*, 2008] Daniel Gomes, André Nogueira, João Miranda, and Miguel Costa. Introducing the Portuguese web archive initiative. In *Proceedings of the 8th International Web Archiving Workshop*, Aaarhus, Denmark, September 2008.

[Hagan *et al.*, 2012] Margaret Hagan, Nan Zhang, and Joseph 'Jofish' Kaye. Safe mathare: a mobile system for women's safe commutes in the slums. MobileHCI '12, pages 47–52, New York, NY, USA, 2012. ACM.

[Hoffart *et al.*, 2012] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal, Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources*, 2012.

[Kothari *et al.*, 2009] Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarthy, and L. Venkata Subramaniam. SMS based Interface for FAQ Retrieval. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 852–860, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Lang *et al.*, 2010] Xueming Lang, Elisa Oreglia, and Suzanne Thomas. Social practices and mobile phone use of young migrant workers. In *Proceedings of the 12th international conference on Human Computer Interaction with mobile devices and services*, MobileHCI '10, pages 59–62, New York, NY, USA, 2010. ACM.

[Liu and Yoshioka, 2011] Yiqi Liu and Masaharu Yoshioka. Construction of large geographical database by merging Wikipedia's Geo-entities and GeoNames. Technical Report SIG-SWO-A1102-03, 2011.

[Markowetz *et al.*, 2005] Alexander Markowetz, Yen-Yu Chen, Torsten Suel, Xiaohui Long, and Bernhard Seeger. Design and Implementation of a Geographic Search Engine. In AnHai Doan, Frank Neven, Robert McCann, and Geert Jan Bex, editors, *WebDB 2005*, pages 19–24, Baltimore, Maryland, USA, 2005.

[Mendoza *et al.*, 2009] Marcelo Mendoza, Hipolito Guerrero, and Julio Farias. Inquiro.CL: a New Search Engine in Chile. In *WWW '09: 18th International World Wide Web Conference (WWW in Ibero-America track)*, WWW '09. ACM, 2009.

[Nakahira *et al.*, 2006] Katsuko T. Nakahira, Tetsuya Hoshino, and Yoshiki Mikami. Geographic locations of web servers under african domains. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 989–990, New York, NY, USA, 2006. ACM.

[Overell and Rüger, 2006] Simon E. Overell and Stefan M. Rüger. Identifying and grounding descriptions of places. In Ross Purves and Chris Jones, editors, *Proceedings of the 3rd ACM Workshop on Geographic Information Retrieval, GIR 2006*, Seattle, WA, USA, 2006. Department of Geography, University of Zurich.

[Purves *et al.*, 2007] Ross S. Purves, Paul Clough, Christopher B. Jones, Avi Arampatzis, Benedicte Bucher, D. Finch, Gaihua Fu, Hideo Joho, Awase Khirni Syed, Subodh Vaid, and Bisheng Yang. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science*, 21(7):717–745, 2007.

[Sehgal *et al.*, 2006] Vivek Sehgal, Lise Getoor, and Peter D Viechnicki. Entity Resolution in Geospatial Data Integration. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, GIS '06, pages 83–90, New York, NY, USA, 2006. ACM.

Konzeption und Implementierung einer Android-App für das *ezDL*-System

Stefan Muno, Thomas Beckers, Sascha Kriewel
Universität Duisburg-Essen
Fachgebiet Informationssysteme
47048 Duisburg
{muno,tbeckers,sk}@is.inf.uni-due.de

Abstract

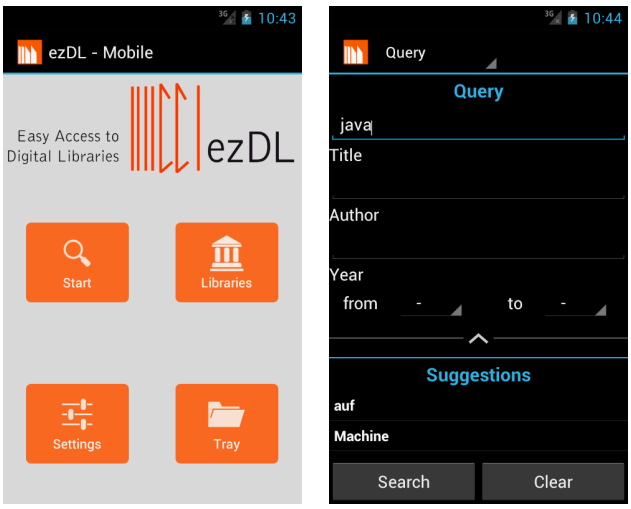
ezDL ist ein interaktives Suchwerkzeug und Framework für interaktive IR-Systeme. Für die Nutzung auf mobilen Geräten ist das bestehende Desktop-Frontend, sowohl aus technischen als auch aus konzeptionellen Gründen (z.B. Bildschirmgröße, Eingabemethoden), nicht geeignet. Deshalb wurde nun eine App für das Android-Betriebssystem entwickelt. Diese App wird außerdem für das Projekt Khresmoi angepasst, um die Suche nach verlässlichen medizinischen Informationen zu ermöglichen.

1 Einführung und *ezDL*

ezDL[1] [2] ist ein quelloffenes, interaktives Suchwerkzeug und Framework für interaktive IR-Systeme. Neben dem Einbinden unterschiedlicher Datenquellen bietet *ezDL* dem Suchenden umfangreiche Funktionen für den Suchprozess an. Zu den Funktionen zählen unter anderem das Vorschlagen weiterer Suchbegriffe, die Extraktion häufiger Terme, Erscheinungsjahre und Autoren aus einer Ergebnisliste und das Speichern von gefundenen Dokumenten, Termen und Autoren in einer Ablage. *ezDL* basiert auf einer serviceorientierten Architektur, die aus einem Backend und verschiedenen Frontends besteht. Dem existierenden Desktop-Frontend wurde nun eine Anwendung für mobile Endgeräte (App) zur Seite gestellt, da diese besondere (technische) Anforderungen an die Benutzerschnittstelle stellen, die durch eine klassische Desktop-Benutzerschnittstelle nicht erfüllt werden können [4].

2 Entwurf und Implementierung

Um die Funktionalität des existierenden Desktop-Frontends auf ein mobiles Endgerät zu portieren, wurde zunächst ein UI-Konzept für Smartphones und Tablets entwickelt. Die Schwerpunkte des Konzepts umfassen die platzsparende Darstellung von Inhalten, die Bedienung über einen Touchscreen, die Darstellung des Inhalts in horizontaler und vertikaler Ausrichtung des Endgeräts und die Verwendung der mit neueren Android-Versionen eingeführten Action Bar (siehe z.B. Abbildung 2) als Hauptbedienelement der App. Abbildung 1 zeigt den Startbildschirm (a) und die erweiterte Suchansicht (b) mit Suchtermvorschlägen auf einem Smartphone. Wenn die App auf einem Tablet ausgeführt wird, wechselt diese in eine Ansicht mit zwei Fenstern, die miteinander interagieren können. Abbildung 2 zeigt die Ergebnisliste



(a) Der Startbildschirm (b) Die Suchanfrage

Abbildung 1: Der Startbildschirm mit Verknüpfungen zu den wichtigsten Funktionen und die Suchanfrage mit Termvorschlägen

und die Detailansicht des angewählten Elements der App auf einem 10-Zoll-Tablet, welches im Querformat gehalten wird. Suchergebnisse können aus der Ergebnisliste in eine Ablage gelegt oder exportiert werden. Außerdem ist es möglich, z.B. häufige Terme oder Autoren aus den Suchergebnissen oder der Ablage zu extrahieren. Die App ist über den Google Play Store erhältlich[2].

Während zwischen der Desktop-Anwendung und dem *ezDL*-Backend eine konstante Verbindung existiert, über welche die beiden Komponenten mit Hilfe von Nachrichten in Gestalt serialisierter Java-Objekte kommunizieren, ist dies für den Datenaustausch mit einer mobilen Benutzerschnittstelle nicht praktikabel. Die potentielle Instabilität der Verbindung und die Inkompatibilität zwischen Java und Android machten die Entwicklung eines neuen Protokolls erforderlich, das Anfragen im JSON-Format über HTTP entgegennimmt und diese an das *ezDL*-Backend weiterleitet. Die Antworten des *ezDL*-Backends werden im JSON-Format aufbereitet und an die App zurückgeschickt.

3 Evaluierung

Zur Prüfung der Gebrauchstauglichkeit der entwickelten App wurde eine Evaluierung mit 8 Probanden durchgeführt (siehe [4] für eine detaillierte Darstellung). Zunächst

[1] http://ezdl.de

[2] http://goo.gl/ObdDc

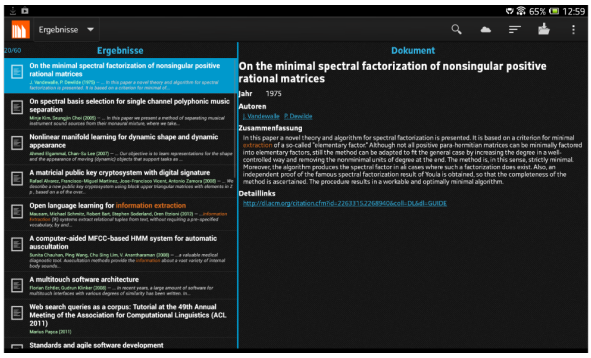
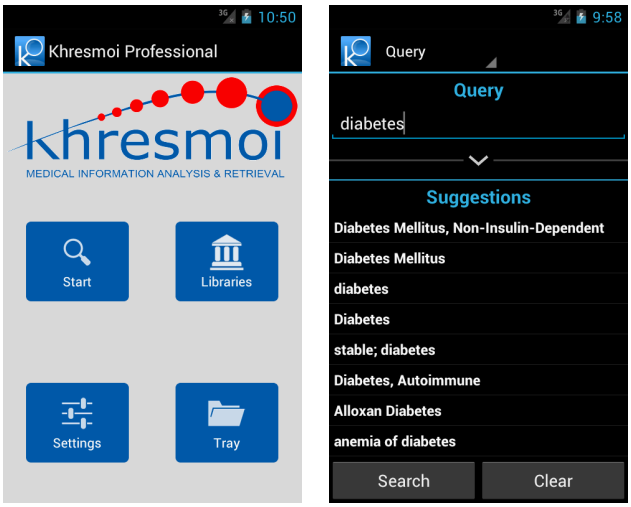


Abbildung 2: Die Ergebnisliste und die Detailansicht auf einem 10-Zoll-Tablet (Sony Xperia Tablet Z). Am oberen Bildschirmrand ist die Action Bar zu sehen



(a) Startbildschirm

(b) Medizinische Terme als Vorschläge

Abbildung 3: Die angepasste App für medizinische Informationssuche

wurden per Fragebogen die demographischen Daten und Vorkenntnisse erhoben. Das Durchschnittsalter der Probanden war 28 (Standardabweichung: 1,51). Bis auf einen Probanden hatte alle Erfahrung im Umgang mit Smartphones. Nur zwei Probanden hatten keine Vorkenntnisse im Umgang mit Android. Insgesamt vier Probanden hatten bereits Erfahrungen mit dem Desktop-Frontend von *ezDL* gesammelt.

Im Anschluss haben die Probanden mehrere Aufgaben mit Hilfe der App gelöst. Die Aufgaben führten den Anwender durch alle Funktionen der App und waren zunehmend schwieriger zu lösen. Zur Einschätzung der Gebrauchstauglichkeit durch die Probanden wurde der SUS-Fragebogen [3] verwendet, da dieser sich bereits in zahlreichen Studien bewährt hat.

Die Auswertung des Fragebogens ergab mit 72,5 von 100 Punkten eine eher durchschnittliche Bewertung der Gebrauchstauglichkeit. Die Probanden, die mit dem Desktop-Frontend vertraut waren, bewerteten die App hierbei allerdings deutlich besser (durchschnittlich 83 Punkte, Standardabweichung: 9) als die übrigen Probanden (durchschnittlich 62 Punkte, Standardabweichung: 12,5). Ein Hauptkritikpunkt der Probanden, die *ezDL* noch nicht kannten, war ein hoher Lernaufwand zum Verwenden der App. Da *ezDL* ein komplexes Suchsystem ist, kann man einen gewissen Lernaufwand zur Verwendung des Systems nicht vermeiden. Die Ergebnisse des SUS-Fragebogens entsprachen daher in etwa den Erwartungen.

Aus Beobachtungen der Probanden während des Versuchs ging hervor, dass diese zu Anfang teilweise Schwierigkeiten mit der Bedienung der App hatten. Das ab Android-Version 3 eingeführte Konzept der Action Bar zur Navigation und zur Bereitstellung von Aktionen in einer Android-App war einigen Probanden nicht bekannt und wirkte sich daher negativ auf die Gebrauchstauglichkeit aus.

4 Einsatz für die medizinische Informationssuche

Die entwickelte App wird im Projekt Khresmoi[3] unter dem Namen *Khresmoi Professional* für die Suche nach verlässlichen medizinischen Informationen eingesetzt. Für den Einsatz von Apps in der Praxis ist die Vertrauenswürdigkeit der gefundenen Informationen wichtig. Die Zielgruppe dabei sind Gesundheitsexperten. Nach einer Studie [1] verfügten im Jahre 2012 in den großen europäischen Ländern über 80% der Gesundheitsexperten ein Smartphone und ca. 47% ein Tablet-Computer. Der Anteil der beruflichen Nutzung von Smartphones betrug 36%. Um die App für die Suche nach medizinischen Informationen nutzen zu können, werden diverse Anpassungen durchgeführt. Neben der Anbindung an das auf *ezDL*-basierende Khresmoi-Backend ist es außerdem notwendig, *i*) andere Objekttypen zu unterstützen *ii*) die entsprechenden Dienste für medizinische Termvorschläge einzubinden (siehe Abbildung 3(b)) und *iii*) zusätzliche Funktionen speziell für die Suche nach medizinischen Informationen bereitzustellen. Darüber hinaus ist die Benutzerschnittstelle an das Design von Khresmoi angepasst worden (siehe Abbildung 3(a)).

[3] http://www.khresmoi.eu/

5 Fazit

Der Zugriff auf das *ezDL*-System ist nun auch über mobile Endgeräte möglich. Eine speziell angepasste Version der App wird zur Suche nach medizinischen Informationen eingesetzt.

Danksagung

Die Forschungen, die zu diesen Ergebnissen geführt haben, wurden teilweise durch das siebte Forschungsrahmenprogramm der Europäischen Union (FP7/2007-2013) mit der Finanzhilfevereinbarung Nr. 257528 (KHRESMOI) finanziert.

Referenzen

[1] Healthcare professional use of mobile devices: A comparative study between Europe and US in 2010 and 2012. Studie, EPG Health Media, 2012.

[2] T. Beckers, S. Dungs, N. Fuhr, M. Jordan, and S. Kriewel. ezDL: An interactive search and evaluation system. In *SIGIR 2012 Workshop on Open Source Information Retrieval (OSIR 2012)*, August 2012.

[3] J. Brooke. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, pages 189–194, 1996.

[4] S. Muno. Entwicklung eines Android-Frontends für ezDL. Diplomarbeit, Universität Duisburg-Essen, 2013.

Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2013)

Topics of Interest

We invite submissions on all aspects of data mining, knowledge discovery and machine learning. Beside new work, we also invite preliminary results and resubmissions of articles published or disseminated at international conferences or journals no earlier than 2013.
Topics of interest include but not limited to:

- Methods of supervised, semi-supervised and unsupervised learning
- Distributed data mining and ubiquitous knowledge discovery
- Data Stream Mining
- Temporal, Spatial & Spatiotemporal Data Mining
- Multiobjective learning
- Rule-based learning
- Network and graph mining
- Text Mining
- Web Mining
- Mining for recommendation engines
- Visual Analytics
- Data Mining in bioinformatics
- Data Mining for business applications
- Data mining in resource constrained environments

Program Chairs

- Martin Atzmüller
- Ute Schmid
- Thomas Seidl

Program Committee

- Bettina Berendt, KU Leuven
- Johannes Fuernkranz,TU Darmstadt
- Stephan Günnemann, Carnegie Mellon University
- Alexander Hinneburg, Universität Halle

- Andreas Hotho, Universität Würzburg
- Eyke Hüllermeier, Universität Marburg
- Christian Igel, University of Kopenhagen
- Kristian Kersting, Universität Bonn
- Ralf Klinkenberg, rapid-i
- Stefan Kramer, Universität Mainz
- Peer Kröger, LMU München
- Daniel Kudenko, University of York
- Ulrike von Luxburg, Universität Hamburg
- Emmanuel Müller, Karlsruher Institut für Technologie
- Lars Schmidt-Thieme, Universität Hildesheim
- Myra Spiliopoulou, Universität Magdeburg
- Gerd Stumme, Universität Kassel

Adaptive Speed Tests*

Daniel Bengs and **Ulf Brefeld**†
German Institute for Educational Research
Frankfurt am Main, Germany
{bengs,brefeld}@dipf.de

Abstract

The assessment of a person's traits such as ability is a fundamental problem in human sciences. We focus on assessments of traits that can be measured by determining the shortest time limit allowing a testee to solve simple repetitive tasks, so-called speed tests. Existing approaches for adjusting the time limit are either intrinsically non-adaptive or lack theoretical foundation. By contrast, we propose a mathematically sound framework in which latent competency skills are represented by belief distributions on compact intervals. The algorithm iteratively computes a new difficulty setting, such that the amount of belief that can be updated after feedback has been received is maximized. We provide theoretical analyses and show empirically that our method performs equally well or better than state of the art baselines in a near-realistic scenario.

1 Introduction

The assessment of a person's traits such as ability is a fundamental problem in the human sciences. Perhaps the most prominent example is the Programme for International Student Assessment (PISA) launched by the Organisation for Economic Cooperation and Development (OECD) in 1997. Traditionally, assessments have been conducted with printed forms that had to be filled in by the testees (paper and pencil tests). Nowadays, computers and handhelds become more and more popular as platforms for conducting studies in social sciences; electronic devices not only facilitate data acquisition and processing, but also allow for real-time adaptivity and personalization.

Psychological testing differentiates between two types of tests, namely *power* and *speeded* tests [Furr and Bacharach, 2007]. The former uses items with a wide range of difficulty levels, so that testees will almost surely be unable to solve all items, even when they are given unlimited time. On the contrary, speeded tests deploy homogeneous items that are easy to solve. The difficulty in speeded tests is realized by narrow time intervals in which the response has to be given. In adaptive speed tests, the latent competency parameter $\hat{\theta}$ encodes for instance reaction time, concentration, or awareness of the testee. An example of such a test is the Frankfurt Adaptive Concentration Test II (FACT-II) [Goldhammer and Moosbrugger, 2007] where a simple multiplicative update of the estimate $\hat{\theta}$ is applied for the adaptation process.

In this paper, we present a novel framework for learning competency parameters in speeded tests. The formal problem setting resembles a game played in rounds. In each round, the goal is to gain as much information as possible on the difficulty setting θ corresponding to the testee's competency. The uncertainty of an estimate $\hat{\theta}$ is represented by a belief distribution over a compact interval. At round t, a new estimate $\hat{\theta}_t$ is drawn, such that $\hat{\theta}_t$ divides the belief mass in two equally sized halves. The testee solves the item which realizes a difficulty level of $\hat{\theta}_t$. The agent observes the response ρ_t. We differentiate three cases: (i) if $\hat{\theta}_t < \theta_t$, the difficulty induced by $\hat{\theta}_t$ was *too easy* for the testee and $\rho_t = 1$, (ii) in case $\hat{\theta}_t > \theta_t$, the setting as *too difficult* and $\rho_t = -1$, and (iii) $\theta_t = \hat{\theta}_t$ which corresponds to a *just right* setting and response $\rho_t = 0$. A similar scenario for discrete variables has been studied by Missura and Gärtner [2011] in the context of computer games.

Before we continue with the presentation of our method, note that the problem setting does not match traditional approaches, including standard supervised (e.g., binary classification) and unsupervised (e.g., density estimation) settings, as the feedback needs to be viewed a directional and not a point-wise one and we cannot make assumption on the testee or stationarity of the observations due to learning effects and tiredness. Thus, the directional feedback is used to update exactly half of the belief mass for maximal information gain. The rationale behind this update strategy is the following: once we observe that $\hat{\theta}$ is *too difficult*, it is highly probable that all difficulty levels $\tilde{\theta} > \hat{\theta}$ are also *too difficult*. A similar argument holds vice versa for *too easy*. The directional feedback is therefore used as a nominal reward that triggers the update process. We present results on the step size of the proposed algorithm and show that it performs equally well or better than state of the art baselines in a near-realistic scenario modelling testee behaviour.

The remainder is organized as follows. Section 2 reviews related work. We present our main contributions, the learning agent and a theoretical analysis in Sections 3 and 4, respectively. Section 5 reports on simulation studies and Section 6 concludes.

2 Related Work

Motivated by applications in computer games as well as teaching systems, Missura and Gärtner [2011] considered the problem of dynamic difficulty adjustment. They formalized the problem setting as a game between a master and a player played in rounds $t = 1, 2, \cdots$, where the mas-

*This paper is a short version of [Bengs and Brefeld, 2013].
†UB is also affiliated with Technische Universität Darmstadt.

ter predicts the difficulty setting for the next round. After the player has finished his turn, the master receives feedback and updates the belief on the difficulty settings and predicts the setting for the next round. The authors introduce the Partial Ordered Set Master (POSM) algorithm that represents the set of admissible difficulty settings as a finite discrete set $\mathcal{K}$ endowed with a partial ordering $\prec$. We will show later that the POSM algorithm for the case of a totally ordered set of difficulty settings is contained as a special case within our framework.

Csáji and Weyer [2011] investigate the problem of estimation in the presence of noise using a binary sensor with adjustable threshold. Their approach estimates a constant $\theta^* \in \mathbb{R}$ that is disturbed by additive, i.i.d. noise. The threshold θ_t is assumed to be adjustable based on all previous observations and threshold values. Under mild assumptions on the distribution of the noise, they derive a strongly consistent estimator for θ^* based on stochastic approximation. In contrast to them , we do not make any assumptions on the distribution of the value to be estimated or on its stationarity.

In the field of psychometrics, only a few adaptive speed tests have been designed. For the assessment of concentration ability, Goldhammer and Moosbrugger [2007] suggested the Frankfurt Adaptive Concentration Test II (FACT-II). As FACT-II conceptualizes concentration as the ability to respond to stimuli in the presence of distractors, testees are shown a set of items comprising of target and non-target items. They are instructed to hit one button, if a target item is present, and another button, if no target item is among the items shown. After each round t, exposure time is adjusted until a liminal exposure time is reached that just allows the testee to solve the task. Starting with a fixed initial exposure time θ_1, updating is performed multiplicatively depending on whether a response is given in time or not.

3 A Learning Agent for Parameter Estimation in Speeded Tests

We cast the problem of learning competency parameters in speeded tests as a game between an agent $\mathcal{A}$ and a testee $\mathcal{T}$ played in rounds $t = 1, 2, \ldots$ on a continuous interval of difficulty settings $\Theta = [a, b]$. Θ is governed by a total order relation $>$ induced by the real numbers corresponding to the *more-difficult-than relation*. We assume that at each round, there is a *just right* setting $\theta_t \in \Theta$ for the testee $\mathcal{T}$. At round t, (i) the agent chooses a setting $\hat{\theta}_t \in \Theta$ based on the current belief, (ii) the testee responds, and (iii) the agent observes directional feedback of the form $\rho_t \in \{-1, 0, +1\}$ subject to the following rule:

$$\rho_t = \begin{cases} +1 & \text{if } \hat{\theta}_t < \theta_t, \text{ too easy} \\ 0 & \text{if } \hat{\theta}_t = \theta_t, \text{ just right} \\ -1 & \text{if } \hat{\theta}_t > \theta_t, \text{ too difficult} \end{cases}$$

Note that the *just right* setting remains hidden to the agent at all times.

In the course of the game, the agent is choosing actions $\hat{\theta}_t$ from the space of possible actions Θ that lead to a reward signal ρ_t depending on the state of the environment θ_t. The goal of the agent is to reach the rewarding state of having selected the *just right* setting by avoiding the punishing signals associated with *too difficult* or *too easy* settings.

The general idea of our approach is the following: We use a function $w_t : [a, b] \to (0, \infty)$ to model the agent's belief at time t about the optimal action based on the experience gathered at time-steps $1, \ldots, t-1$. Suppose that the agent selects a setting $\hat{\theta}_t$ and receives feedback $\rho_t = +1$ (*too easy*). Because of the transitivity of the ordering of difficulty settings, the agent not only learns about $\hat{\theta}_t$ as an isolated point, but also learns that all settings $\tilde{\theta}$ which are easier than $\hat{\theta}_t$, i.e., $\tilde{\theta} < \hat{\theta}_t$, would also have been *too easy* and the agent updates the belief on the whole interval $[a, \hat{\theta}_t]$. The mass of belief that can be updated is then given by

$$A_t(\hat{\theta}_t) := \int_a^{\hat{\theta}_t} w_t(x)dx.$$

Similarly, if $\rho_t = -1$, the belief in the interval $[\hat{\theta}_t, b]$ can be updated according to

$$B_t(\hat{\theta}_t) := \int_{\hat{\theta}_t}^{b} w_t(x)dx.$$

If $\rho_t = 0$, there is no reason to update belief, because current knowledge has led to a correct prediction. We devise the following strategy for predicting $\hat{\theta}_t$ and updating belief: The difficulty setting $\hat{\theta}_t$ for the upcoming round is selected in order to allow to update as much belief as possible after feedback has been obtained. That is, we select $\hat{\theta}_t$ so that

$$\hat{\theta}_t = \operatorname{argmax}_{\tilde{\theta} \in [a,b]} \min \left\{ A_t(\tilde{\theta}), B_t(\tilde{\theta}) \right\}. \qquad (1)$$

It can easily be seen that this amounts to selecting $\hat{\theta}_t$ such that

$$A_t(\hat{\theta}_t) = \frac{1}{2} \int_a^b w_t(x)dx.$$

Equivalently, $\hat{\theta}_t$ can be characterized by $A_t(\hat{\theta}_t) = B_t(\hat{\theta}_t)$. Because w_t is non-negative by assumption, the mapping $\hat{\theta}_t \mapsto A_t(\hat{\theta}_t)$ strictly increasing and thus bijective, so $\hat{\theta}_t$ is uniquely determined if only $\int_a^b w_t(x)dx \neq 0$. In order to derive an algorithm from this framework, we need to specify the space of belief functions $\mathcal{W}$ and the belief updating rule

$$\mathcal{W} \times \{-1, 0, 1\} \to \mathcal{W}, \quad (w_t, \rho_t) \mapsto w_{t+1}.$$

The next section introduces strategies to learn the agent.

3.1 Interval Subdivision Agent

While there is no restriction on the space of belief functions arising from the general framework, we choose to use the space of non-negative step functions on $[a, b]$ for $\mathcal{W}$ and an exponential updating rule based on interval subdivision. That is, we divide the interval containing the actual prediction $\hat{\theta}_t$ at $\hat{\theta}_t$ and update the belief values to the left or right of $\hat{\theta}_t$ depending on the feedback ρ_t by multiplying with a parameter $\beta \in (0, 1)$. Formally, denoting by χ_M the characteristic or indicator function of a set $M \subset \mathbb{R}$, we write w_t as a sum

$$w_t = \sum_{i=1}^{N_t} y_i^{(t)} \chi_{I_i^{(t)}}$$

for some $N_t \in \mathbb{N}$, where $y_i^{(t)} \geq 0$ is the value w_t takes on the i^{th} interval given by

$$I_i^{(t)} = [x_{i-1}^{(t)}, x_i^{(t)})$$

for $i = 1 \cdots, N_t - 1$ and $I_{N_t}^{(t)} = [x_{N_t-1}, x_{N_t}]$. The interval endpoints are defined by a partition

$$a = x_0^{(t)} < x_1^{(t)} < x_2^{(t)} < \cdots < x_{N_t}^{(t)} = b$$

of $[a, b]$. Denoting the index of the interval containing $\hat{\theta}_t$ by i_t^*, we update

$$w_{t+1} = \sum_{i=1}^{i_t^*-1} \beta y_i \chi_{I_i^{(t)}} + \beta y_{i_t^*} \chi_{[x_{i_t^*-1}, \hat{\theta}_t)} + y_{i_t^*} \chi_{[\hat{\theta}_t, x_{i_t^*})} + \sum_{i=i_t^*+1}^{N_t} y_i \chi_{I_i^{(t)}},$$

in case $\rho_t = 1$ and analogously for $\rho = -1$,

$$w_{t+1} = \sum_{i=1}^{i_t^*-1} y_i \chi_{I_i} + y_{i_t^*} \chi_{[x_{i_t^*-1}, \hat{\theta}_t)} + \beta y_{i_t^*} \chi_{[\hat{\theta}_t, x_{i_t^*})} + \sum_{i=i_t^*+1}^{N_t} \beta y_i \chi_{I_i}.$$

Finally, if $\rho_t = 0$ no update is necessary and $w_{t+1} = w_t$. The belief function can be stored and updated efficiently by storing the endpoints $x_1^{(t)}, \cdots, x_{N_t-1}^{(t)}$ and function values $y_1^{(t)}, \cdots, y_N^{(t)}$. Also, our particular choice of $\mathcal{W}$ makes the computation of $\hat{\theta}$ simple and inexpensive: As w is a step function, its integral over θ is given by

$$\int_a^b w_t(x)dx = \sum_{i=1}^{N_t-1} y_i \left(x_{i+1} - x_i\right)$$

The initial belief function w_1 can be tailored to incorporate prior knowledge about where to expect θ_1. In the absence of prior knowledge on the distribution of θ, $w_1 \equiv 1$ serves as a possible initialization.

3.2 Limited-memory Interval Subdivision Agent

The memory usage of the internal subdivision agent (ISA) at time t is in $\mathcal{O}(t)$. Indeed, if w_0 is represented by N interval-value pairs, each step adds at most one node in the belief function. A limit on the amount of memory consumed by ISA can be imposed by limiting interval subdivision. Thus, the limited-memory ISA (LISA) only subdivides intervals when subdivision results in intervals of width greater than a given parameter $\epsilon > 0$.

4 Theoretical Analysis

In this section we present a theoretical analysis of the ISA algorithm. We are interested in characterizing convergence properties of ISA under different assumptions. The simplest assumption that can be made about the *just right* setting is that it remains constant at all times. That is, $\theta_t \equiv c$ for $c \in [a, b]$ and all $t \in \mathbb{N}$. We now present a bound on the step size between successive predictions by ISA. The bound follows directly from Lemma 1.[1]

Lemma 1. *Let $f : [a, b] \to (0, \infty)$ be bounded and integrable on $[a, b]$. Let $\beta \in (0, 1)$. Let $\theta_1, \theta_2 \in [a, b]$ be numbers such that $\int_a^{\theta_1} f(x)dx = \frac{1}{2}\int_a^b f(x)dx$ and $\int_a^{\theta_2} \hat{f}(x)dx = \frac{1}{2}\int_a^b \hat{f}(x)dx$, where*

$$\hat{f}(x) = \begin{cases} \beta f(x) & \text{if } a \le x \le \theta_1 \\ f(x) & \text{if } \theta_1 < x \le b \end{cases}.$$

Then $\theta_1 < \theta_2$ and

$$\frac{1-\beta}{4M}\int_a^b f(x)dx \le \theta_2 - \theta_1 \le \frac{1-\beta}{4m}\int_a^b f(x)dx. \quad (2)$$

where $M := \max_{x\in[a,b]} f(x)$ and $m := \min_{x\in[a,b]} f(x)$.

Lemma 1 says that if the difficulty level $\hat{\theta}_t$ estimated by ISA is *too easy* ($\rho_t = 1$), the new estimate will be greater than its predecessor, that is $\hat{\theta}_{t+1} > \hat{\theta}_t$ holds. Analogously the case $\rho_t = -1$ implies $\hat{\theta}_{t+1} < \hat{\theta}_t$. We use the inequality to derive a bound on the step size of ISA in the following Theorem 1.

Theorem 1. *Let $\left(\hat{\theta}_t\right)_{t=1}^N$ be a sequence of estimations generated by ISA with parameter β. Then for $t = 1, \ldots, N-1$ it holds that*

$$\frac{1-\beta}{4M_t}\int_a^b w_t(x)dx \le \left|\hat{\theta}_{t+1} - \hat{\theta}_t\right| \le \frac{1-\beta}{4m_t}\int_a^b w_t(x)dx,$$

where

$$M_t := \max_{x\in[a,b]} w_t(x)$$

and

$$m_t := \min_{x\in[a,b]} w_t(x).$$

Theorem 1 bounds the minimal and maximal difference between successive estimates by ISA. Note that the bounds are invariant under rescaling of the belief function, but depend on the parameter β that controls learning rate: If β is small, new experience is given more weight and the lower bound on step size is greater than its analogue for $\beta \approx 1$ which gives less weight to new information.

We now investigate the relation between LISA and POSM for a completely ordered set which we denote by $\Theta' = \{1, \cdots, N\}$ for some $N \in \mathbb{N}$, endowed with the natural ordering. The following proposition holds:

Proposition 1. *Let $N \in \mathbb{N}$, $\Theta' = 1, \ldots, N$ endowed with the natural ordering be the set of difficulty levels for POSM and let $[a, b] = [0, N]$. Let $\beta \in (0, 1)$, $\epsilon < 1$. Define the initial belief function w_0 for LISA by $x_i = i$ for $i = 0, \cdots, N$ and $y_i = 1$ for $j = 1, \cdots, N$. Denote by $\mathrm{ind}(x)$ the function mapping $x \in [a, b]$ to Θ' such that $x \in [x_{\mathrm{ind}(x)-1}, x_{\mathrm{ind}(x)})$. Then, given a sequence of feedback $(\rho_t)_{t\in\mathbb{N}}$, the estimates $(\tilde{k}_t)$ produced by POSM coincide with $(\mathrm{ind}(\hat{\theta}_t))_{t=1}^N$.*

The result stated in Proposition 1 explains to some extent why ISA and LISA expose a behaviour qualitatively similar to that of POSM in the setting of our experiments. As we show in the next section, the LISA and ISA algorithms are able to exploit the continuous setting, outperforming POSM by a significant margin.

5 Empirical Results

For our experiments, we simulate near-realistic scenarios to create settings that reflect behaviour observed in adaptive psychological speed tests or computer games. We compare the empirical performance of ISA and LISA to state-of-the-art baselines POSM and the algorithm used by FACT-II .

Throughout all our experiments, we use $\Theta = [0, 1]$. Note that this does not limit generality, as every compact interval can be rescaled and shifted to match Θ. To allow for a fair comparison, the set of difficulty settings for POSM consists of N equidistantly sampled points in Θ, where N is

[1] Detailed proofs are presented in [Bengs and Brefeld, 2013].

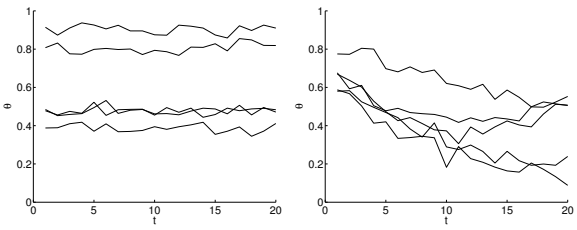


Figure 1: Randomly parametrized functions modelling θ in absence (left) and presence of drift (right). In both scenarios white noise is added.

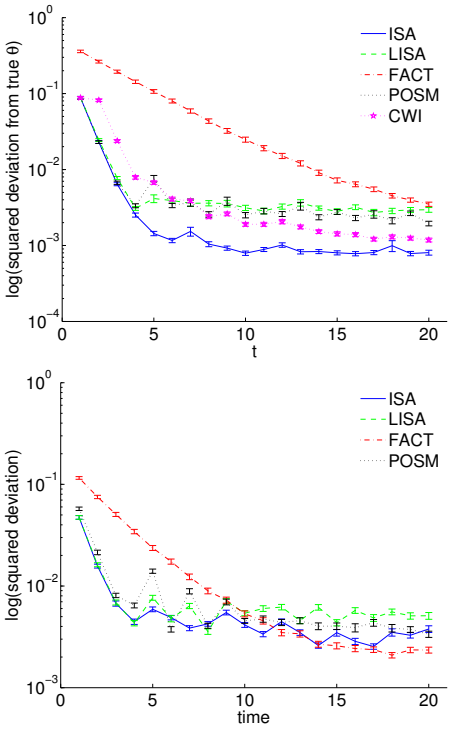


Figure 2: Squared deviations from true θ for the constant (top) and the drift (bottom) setting.

the number of time steps used. This choice guarantees that the number of subdivisions made by ISA and LISA is less than or equal to the number of settings available to POSM. Thus, all approaches have access to the same amount of resources. We use optimal parameters for ISA, LISA and POSM chosen by model selection.

We consider two distinct settings: In the first setting, the true parameter θ remains constant and is sampled from a uniform distribution. For the constant setting, we also include Csáji-Weyer-Iteration (CWI) [Csáji and Weyer, 2011] as an additional baseline. In the second setting, we simulate learning and tiredness effects. The true parameter θ thus underlies drifts and the resulting distribution is not stationary. Additionally, observations are disturbed by additive noise originating from a Gaussian distribution. Figure 1 shows sample observations for the two settings. In both settings, we conduct 500 repetitions with randomly drawn sequences θ_t and report on averaged deviations and standard errors.

Figure 2 (top) shows the results for the constant setting. All algorithms need some time to adapt to the noisy θ_t. The three learning algorithms and CWI, however, approach the true θ significantly faster than FACT. CWI and

Table 1: Sum of squared deviations from true θ, average over 500 runs.

	ISA	LISA	POSM	FACT	CWI
const.	3.3842	4.3905	4.2441	34.8336	5.9575
drift	3.4027	4.0825	4.4171	9.4808	–

ISA approximate the true θ more closely with ISA realizing quicker convergence and smaller error. The squared error is smallest for ISA, followed by the almost equally performing LISA and POSM. FACT is outperformed by all four competitors by a large margin (see also Table 1).

Figure 2 (bottom) summarizes the results for the drift setting. ISA performs best, followed by LISA and POSM. Again FACT is outperformed significantly by the others. The squared errors are similar or smaller for all algorithms than they are in absence of drift (see Table 1), showing that all algorithms can deal with drift well. The performance of FACT even proves significantly better than in the setting without drift. This effect can be explained by the fact that the model of drift employed here favors evolutions of θ starting in the upper range of Θ. Note that FACT always initializes θ_0 with the highest possible value which highly affects its performance in the first iterations. The other algorithms thus benefit in the beginning from intitalizing θ with the mean of the search space. However, different choices are possible.

6 Conclusion

We have introduced a mathematically sound learning framework for parameter adaptation in speeded tests. Our approach does not make any assumptions on the distribution of the true parameter and is therefore deployable in settings characterized by parameter drift and additive noise. Empirically, we have shown that the algorithm performs equally or better than state of the art baselines in different scenarios modelling testee behaviour under different assumptions.

References

[1] D. Bengs and U. Brefeld. A Learning Agent for Parameter Adaptation in Speeded Tests. In *Proceedings of the ECML/PKDD Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Reward*, 2013.

[2] B. C. Csáji and Weyer. System identification with binary observations by stochastic approximation and active learning. In *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*:3634–3639 , 2011.

[3] R. M. Furr and V. R. Bacharach. *Psychometrics: An Introduction*, SAGE Publications, Thousand Oaks, California, 2007.

[4] O. Missura and T. Gärtner. Predicting Dynamic Difficulty. In *Advances in Neural Information Processing Systems 24*: 2007–2015, 2011.

[5] H. Moosbrugger and F. Goldhammer. *FAKT-II Frankfurter Adaptiver Konzentrationsleistungs-Test II*, Huber, Bern, 2007.

Using a Classifier Pool in Accuracy Based Tracking of Recurring Concepts in Data Stream Classification[†]

Mohammad Javad Hosseini, Zahra Ahmadi, Hamid Beigy
Sharif University of Technology
Tehran, Iran
{mjhosseini, z_ahmadi, beigy}@ce.sharif.edu

Abstract

Data streams have some unique properties which make them applicable in precise modeling of many real data mining applications. The most challenging property of data streams is the occurrence of "concept drift". Recurring concepts is a type of concept drift which can be seen in most of real world problems. Detecting recurring concepts makes it possible to exploit previous knowledge obtained in the learning process. This leads to quick adaptation of the learner whenever a concept reappears. In this paper, we propose a learning algorithm called Pool and Accuracy based Stream Classification (PASC), which takes the advantage of maintaining a pool of classifiers to track recurring concepts. Each classifier is used to describe an existing concept. Two methods are presented for classification task: active classifier and weighted classifiers methods. For the updating of the pool we use two methods: Bayesian and Heuristic. Experimental results on real and artificial datasets show the effectiveness of weighted classifiers method while dealing with sudden concept drifting datasets. In addition, the proposed updating methods outperform the existing algorithms in datasets with arbitrary attributes.

1 Introduction

As the data available on the web increases, processing the large volume of data and extracting knowledge from them is needed. These data are changing and they cannot be saved and processed wholly in the same way as classical data mining assumes. So, presenting new algorithms which could learn and classify using this continuous and unlimited stream of data is a challenging problem. Data streams have some properties [Tsymbal, 2004]:

- They could not be saved completely and so a forgetting mechanism is needed to forget ineffective data.
- The processing of data should be done online and the algorithm complexity should be simple.
- Most of the time, feature (or class) distribution is changed over the time. This is known as concept drift. If the drift takes effect in the target function, it is named real concept drift.

The concept drift could be sudden, gradual, incremental or recurring [Zliobaite, 2010a]. When the underlying distribution of data changes suddenly at time t_k, sudden drift occurs. Gradual drift happens when in a period of time, the data is drawn from two distributions and over time, the probability of the old distribution decreases and the probability of the new distribution increases. Incremental drift can be thought of as a generalized version of gradual drift. Here in the drift period, there could be more than two distributions to draw data from. However the difference between the distributions should be small. The other type of drift is recurring concept, where previously seen concepts reappear after some time. One important challenge in learning from data streams in the presence of concept drift is distinguishing the drift from the noise. It is important to note that I.I.D (Independent Identically Distribution) condition is not valid in the streams in which concept drift occurs, but it is rational to think that small size batches of data satisfy the I.I.D condition.

There have been extensive studies on sudden and gradual concept drift detection and learning [Baena-García et al., 2006; Gama and Castillo, 2006; Helmbold and Long, 1994; Klinkenberg and Joachims, 2000; Klinkenberg, 2004; Kolter and Maloof, 2007; Kuh et al., 1991; Gao et al., 2008; Nishida, 2008; Bifet et al., 2010a; Bifet et al., 2010b; Kuncheva and Zliobaite, 2009; Garnett, 2010; Ikonomovska et al., 2010; Scholz and Klinkenberg, 2006; Zliobaite, 2010b]. Early systems in data stream support recurring concepts [Schlimmer and Granger, 1986; Widmer and Kubat, 1993; Widmer and Kubat, 1996], however, they are mostly considered recently [Lazarescu, 2005; Gama and Kosina, 2009; Katakis et al., 2009; Morshedlou and Barforoush, 2009; Gomes et al., 2010], and identified as a challenging problem in data streams.

In this paper we propose a learning algorithm which tries to improve classifying concept drifting data streams by exploiting the existence of recurring concepts. This is done by maintaining a pool of classifiers which is updated continuously while processing consecutive batches of data (same as previous approaches, e.g. [Gomes et al., 2011; Katakis et al., 2009; Ramamurthy and Bhatnagar, 2007]). Each classifier of this pool is used to describe one of the existing concepts. When a new batch of data is received, first it is classified and after receiving the true labels of instances, it is used to update an existing classifier in the pool or add a new classifier to it. Deciding which classifier should be updated or whether a new one is needed is done by some examinations on the new batch of data and the pool. Classification of the instances is done by using the classifiers in the pool in an effective and adaptive way. This algorithm is similar to the one used in [Katakis et al.,

[†] This paper is the resubmission of a paper with the same topic published in Evolving Systems 4(1): 43-60 (2013). (an earlier version was published ICDM Workshops 2011).

2009], but there are major changes in the steps of the algorithm. In fact, our contribution is to propose a new method to classify instances called weighted classifiers method. The other novel part of the paper is the presentation of new methods to update the pool using Bayesian formulation and a heuristic method. Finally the presented methods are compared with the existing ones.

The results show the effectiveness of our algorithm in terms of accuracy and time especially in data streams of sudden drifts. In addition, it is tried to solve some parameter setting problems that exist in some of the previous methods.

The structure of this paper is as follows: in the next section the related works of recurring concepts is discussed. In section 3 the proposed algorithm is presented. Section 4 evaluates the proposed algorithm and compares the experimental results to some previous methods. Section 5 concludes the paper and discusses some future developments which can be done.

2 Related Work

Concept drift learning of data streams has been studied extensively in recent decades. As discussed previously, drifts can be of different types. Most of studies are done on the learning of sudden and gradual drifts. But one possible drift is the change of the current concept to one of the previously seen concepts. As in data streams the learner forgets some unused concepts passing the time, if instances from a previously seen concept is presented to the learner, it may classify them incorrectly. So the learner may be fallen into the trap. Recurring concepts detection and learning is a hard and challenging problem which has been studied in recent years [Lazarescu, 2005; Gama and Kosina, 2009; Katakis et al., 2009; Morshedlou and Barforoush, 2009; Gomes et al., 2011]. All of presented methods try to extract concept from received instances and maintain the concept in a pool of concepts. Every time a new instance arrives, the similarity to available concepts is measured and a model is selected or created. The rest of this section reviews the researches done in the area of recurring concepts in data streams.

The first algorithm supporting recurring concepts consists of an ensemble of classifiers [Ramamurthy and Bhatnagar, 2007]. Each classifier is built on a data chunk and none of the classifiers are deleted. Then while choosing classifiers for the ensemble, the algorithm selects only pertinent classifiers and so it supports the recurring concepts.

Reference [Katakis et al., 2009] presents a framework for the problem of recurring concepts. It extracts a conceptual vector from the arrived batch of data using a transformation function. We name the instances of a labeled batch as

$$\mathrm{B_L} = \{\boldsymbol{x}_{L(k)}, \boldsymbol{x}_{L(k+1)}, \dots, \boldsymbol{x}_{L(k+b-1)}\}, \quad (1)$$

where $\boldsymbol{x}_{L(k+i)}$ is the $(i+1)^{th}$ instance of the labeled batch of data. A conceptual vector $\boldsymbol{Z} = (z_1, z_2, \dots, z_n)$ is extracted from the batch where z_i is a conceptual feature and is calculated from

$$z_i = \begin{cases} \{P(f_i = v|c_j): i = 1..n, j = 1..m, v \in V_i\} & \text{if } f_i \text{ is nominal} \\ \{\mu_{i,j}, \sigma_{i,j} : j = 1..m\} & \text{if } f_i \text{ is numeric} \end{cases}, \quad (2)$$

where f_i is the i^{th} feature, V_i is the set of possible values of a nominal feature, $\mu_{i,j}$ and $\sigma_{i,j}$ are the mean and standard deviation of the j^{th} class of feature i. Then by using a clustering algorithm on the available concepts, the algorithm detects the recurring concepts. For each concept in the pool, the algorithm preserves a classifier which will be updated through the time. Clustering is done on the conceptual vectors and using the Euclidean distance as the similarity (difference) measure. If the similarity of a new conceptual vector is more than a threshold, an available concept and its classifier will be updated otherwise a new cluster and classifier will be created. One major problem of this framework is how to determine the threshold. The threshold value is a problem specific parameter and should be regularized by try and error.

Mean and standard deviation is used for the presentation of models in [Morshedlou and Barforoush, 2009] too. This approach uses a proactive behavior versus drifts: by knowing the current concept, it calculates the probability of next concept. If the probability is more than a threshold, the concept will be added to the buffer. If the algorithm detects a drift and decides to behave proactively, it selects a concept from the buffer. If the concept matches the batch, it will be updated. If the concept does not match the data and the algorithm behaves proactively, the next concept will be selected else if the reactive behavior is selected, a new classifier will be trained on the batch. [Morshedlou and Barforoush, 2009] uses a heuristic approach to select proactive or reactive action. Here a threshold parameter should be selected as well as doing some computations to select the suitable behavior each time which is a time consuming action.

The other approach uses meta-learners which can detect the reoccurrence of concepts and activate the previous classifiers using proactive behaviors [Gama and Kosina, 2009]. The meta-learner learns the space where the base learner does well. When the algorithm enters the warning phase of drift, meta-learners determine the performance of their corresponding base learners. If the performance is more than a threshold, the algorithm will use the base learner to classify next instances. Here all base learners and their corresponding meta-learners (referees) are maintained in the pool.

Another idea used in this domain is the use of context space model to extract concept from learning model [Gomes et al., 2010]. A context space is a *N*-tuple of the form $R = (a_1^R, a_2^R, \dots, a_N^R)$, where a_i^R determines the acceptable regions of feature a_i. Each classifier has a context space description and all of them will be saved in a repository. To select the appropriate model, the algorithm uses their corresponding contexts.

3 Proposed Learning Algorithm

Our goal is to propose a new method named *Pool and accuracy based Stream Classification (PASC)*. The idea followed in this method is similar to the method proposed in [Katakis et al., 2009]. We maintain a pool of classifiers which contains a number of classifiers each describing a particular concept which is being updated through the time. After receiving a batch of data, first we predict the labels of its instances and then receive the true labels. Then we can use the instances and their labels to update a classifier in the pool or create a new classifier on this batch of data and add it to the pool, if necessary. The classifiers added to the pool cannot grow arbitrarily the maximum number of classifiers in the pool cannot exceed a predefined limit which is a parameter of our algorithm.

To update or create a classifier in the pool, first of all the most relevant concept to the batch of labeled data is selected. If the similarity is more than a predefined threshold or the pool is full, we update the most relevant classifier with the newly arrived labeled batch. Otherwise we construct a new classifier on it. The classifiers used in our method can be any kind of updateable classifiers.

In the rest of this section, we seek how to classify the batches of data and update the pool. As mentioned above, after receiving each batch of data, the classification is done and after receiving their labels, we update the pool. In the proposed method, iteratively after receiving the t^{th} batch of unlabeled data $B_t = (x_{t,1},x_{t,2},\ldots,x_{t,k})$ such that $x_{t,i}$ is the i^{th} data of the t^{th} batch, and its labels $L_t=(l_{t,1},l_{t,2},\ldots,l_{t,k})$ such that $l_{t,i}$ is the label of $x_{t,i}$, we follow the general framework shown in Procedure 1.

	Input: an infinite stream of batches of instances B_t. After classification of each instance $B_{t,i}$, its label is revealed to the algorithm. **Output:** Predicted labels of instances $B_{t,i}$.
1	Pool = Ø; // the pool of classifiers
2	C = make_classifier(B_1,L_1);
3	RDC = new classifier(); //only used in Bayesian
4	//method
5	ac = 1; // active classifier
6	W_1=1;
7	Pool = Pool U {C};
8	X_1=sum_data(B_1);
9	RDC.update(X_1,1); //1 is the label of X_1
10	**for** j=2 to infinity **do**
11	Classify B_t.
12	Update Pool with B_t and L_t;
13	determine active classifier (classifier weights);
14	**end for**

Procedure 1. The main framework of PASC.

In line 2, *C* is the first classifier which will be added to the pool and W_1 (in line 6) is its weight. *RDC* is a classifier and *ac* contains the active classifier which will be used in the rest of the procedure. In line 8, X_1 is an instance constructed from B_1. This procedure contains three main phases which can be seen in lines 11 to 13.

In the following subsections, we consider the details of the parameters discussed above and the three phases of the algorithm.

3.1 Phase 1: Classifying the Batch

In this phase, after receiving a batch of unlabeled data B_t, we classify the batch using the classifiers in the pool. This task can be done in two ways. The first is similar to the method used in [Katakis et al., 2009] and the second tries to classify the batch using the weights assigned to the classifiers.

Classifying the Batch According to the Active Classifier

method is used in [Katakis et al., 2009] to classify instances using the classifiers in the pool. The classifier selected to classify the batch is named active classifier. This classifier is defined according to the last iteration. If in the last iteration, a new classifier was added to the pool, it would be the active classifier. Otherwise, the classifier that has the most relevance to the batch would be the active classifier. The pseudocode of this method is shown in Procedure 2. In line 2, *ac* is the active classifier and *pl* stores the predicted labels of the instances.

1	**for** i=1 to k **do**
2	pl[$B_{t,i}$] = Pool[ac].classify($B_{t,i}$);
3	**end for**

Procedure 2. Classify batch according to active classifier.

Classifying the Batch According to the classifiers' weights.

The first way of classifying a batch uses the active classifier that is appropriate for the last batch of data. However, when a sudden concept drift occurs, the method's performance decreases significantly, because the appropriate classifier for the last batch is not appropriate for the current batch anymore. We suggest using the classifiers in the pool in an adaptive way. A positive weight is assigned to each classifier in the beginning of processing the batch according to the performance of the classifier on the previous batch and when we want to classify an instance, we use the classifier with the highest weight. When the true label is revealed to the algorithm, the classifiers' weights can be updated. Updating the weights is done according to the following rule:

$$w'(j) = w(j) * \beta^{M(j,i)}, \quad (3)$$

where $w(j)$ is the current weight of j^{th} classifier and $w'(j)$ is its new weight and β is a parameter in [0,1). If the j^{th} classifier classifies the i^{th} instance correctly, $M(j,i)$ will be 0, otherwise it is 1. Equation (1) is inspired from [Freund and Schapire, 1996] which models the online prediction problem with a two-player repeated game. The first player is the learner and the second is the environment. The leaner can choose a mixed strategy *P* that determines how to classify the instances determined by the mixed Strategy *Q* of the environment. The mixed strategy *P*, determines the weight of each concept to be used in the weighted majority method of classifying instances. The mixed strategy *Q* determines how to present instances to the learner. The game is as follows: First, the learner chooses mixed strategy *P* that determines how it would classify the instances, and then the environment chooses mixed strategy *Q* that determines how the instances are presented to the algorithm. In the next step, learner can observe the loss of using these strategies and so it can change its mixed strategy in the next iteration by updating the weights. It has been shown that for sufficient number of instances, the error of ensemble with the weights determined by (3) is sufficiently close to the best classifier's error [Freund and Schapire, 1996]. So if the size of the batch is large enough, the performance of our ensemble classifier on the current batch is close to the performance of the best classifier in the pool. But this size should not be so large that it violates the I.I.D condition in the batch or makes difficulty in storing data in the memory.

Although using this method is guaranteed to work well, we slightly modify the method to improve its efficiency. First, Instead of using weighted majority to classify an instance we use only the classifier with the highest weight. Second, Instead of applying the updating rule for every instance, we use it for a subsample of the batch that has the size equal to square root size of the batch.

The initial values of weights are 1 and after processing each batch, the weights are set according to the rule discussed in phase 3. The pseudocode of this method is

shown in Procedure 3. In line 1, S_t is a subsample of the batch B_t and m is its size which is set to the square root size of the batch. After classifying each instance in line 4, if the instance is a member of the subsample, classifiers' weights will be updated.

```
1   S_t=sub_sample(B_t,m);
2   /* makes a sub_sample of size m*/
3   for i=1 to k do
4       pl[B_t,i] = classifyw(Pool,W,B_t,i);
5       /*Uses the most weighted classifier*/
6       if S_t does not contains B_t,i
7           continue;
8       end if
9       for j=1 to size(Pool) do
10          W_j=W_j* Pool[j].error(B_t,i,L_t,i);
11      end for
12  end for
```

Procedure 3. Classify batch according to classifier weights.

3.2 Phase 2: Updating the Classifiers' Pool

After receiving L_t, the true labels of B_t, a classifier in the pool will be updated incrementally or a new classifier will be created on the batch. If we assume the size of the batch is small enough, it will be relevant to only one of the available concepts, because the concepts in the pool represent different hypotheses. So the relevant concept should be updated using the current batch of data. So we need to find the concept which describes B_t and L_t with the highest probability and also find a measure of its correspondence to the batch. In the following two subsections, two alternatives of performing this task are discussed. The first is a straightforward method and uses Bayes' theorem to find the probabilities. The second is a heuristic method which is more efficient than the first.

Bayesian method for Updating the Classifiers' Pool

In this method, we estimate the relevance probability of each available concept to B_t and L_t. As previously mentioned, in the environments subject to concept drift, the I.I.D condition does not hold. But we can assume that this condition holds for a batch of data that is sufficiently small. So the probability that B_t and L_t correspond to concept h_i can be formulated as:

$$P(h_i|B_t, L_t) = \frac{P(B_t, L_t|h_i) * P(h_i)}{P(B_t, L_t)}, \tag{4}$$

where the right side of the equation follows from Bayes' theorem. Thus the best concept to describe B_t and L_t is:

$$\begin{aligned} &argmax_i P(h_i|B_t, L_t) \\ &= argmax_i\ P(B_t, L_t|h_i) *\ P(h_i). \end{aligned} \tag{5}$$

Equation (5) uses the fact that the best concept does not relate to the probability of B_t and L_t. As the environment is non-stationary and we cannot have any assumption about the concepts, we consider $P(h_i)$ which is the prior probability of the i^{th} concept to be identical for all concepts. So equation (5) becomes:

$$\begin{aligned} &argmax_i P(h_i|B_t, L_t) \\ &= argmax_i\ P(B_t, L_t|h_i) \\ &= argmax_i P(B_t|h_i) * P(L_t|B_t, h_i). \end{aligned} \tag{6}$$

Hence we should estimate $P(L_t|B_t,h_i)$ and $P(B_t|h_i)$. The former is the conditional probability that the labels of the instances $(x_{t,1},x_{t,2},\ldots,x_{t,k})$ be $(l_{t,1},l_{t,2},\ldots,l_{t,k})$ given that the instances and their labels are described by the i^{th} concept and the latter is the probability that the batch is produced in an environment described by the i^{th} concept.

According to I.I.D condition in a batch, we have:

$$P(L_t|B_t, h_i) = \prod_{j=1}^{j=k} P\left(l_{t,j}\middle|x_{t,j}, h_i\right). \tag{7}$$

Notice that $P(l_{t,j}|x_{t,j},h_i)$ can be estimated using the posterior probability calculated by the i^{th} classifier. To estimate $P(B_t|h_i)$, according to I.I.D we have:

$$P(B_t|h_i) = \prod_{j=1}^{j=k} P\left(x_{t,j}\middle|h_i\right). \tag{8}$$

There is a straightforward way to determine $P(x_{t,j}|h_i)$ by using a classifier which we call *raw data classifier*. The input of this classifier is the unlabeled instances $x_{t,j}$ and its output is the probability of the instances to belong to the concepts. So to train the raw data classifier, first the concept which describes B_t and L_t best, is determined. Then all instances in the batch and the concept index (or its id) as the class label are given to the classifier to be updated. To determine the relevant concept of the batch, we can give all of the batch instances to the classifier. But this will take much time to find $P(B_t|h_i)$ and therefore we use an alternate way: instead of using all instances in the batch we make an instance X_t for the batch B_t and use it to train raw data classifier (RDC). X_t has the same number of features as the original instances and its i^{th} feature is simply the sum of all the i^{th} features of the instances in the batch.

After receiving unlabeled batch B_t, X_t is built and the probability of each of its instances to belong to any of the concepts in the pool is estimated by the probability of X_t to belong to the concept which can be calculated by RDC. Then the best concept matching B_t and L_t is determined (it may be a new concept added to the pool) and X_t and the best concept index are given to RDC to be updated. So $P(B_t|h_i)$ can be estimated as:

$$P(B_t|h_i) = {p_i}^k, \tag{9}$$

where p_i is the probability of belonging X_t to i^{th} concept which is calculated by RDC. Therefore, to determine the best concept describing B_t and L_t we can use:

$$\begin{aligned} &argmax_i P(h_i|B_t, L_t) \\ &= argmax_i\ {p_i}^k * \prod_{j=1}^{j=k} P\left(l_{t,j}\middle|x_{t,j}, h_i\right). \end{aligned} \tag{10}$$

To prevent underflow of the products we use (11) Instead of (10) to find the best concept:

$$\begin{aligned} &argmax_i P(h_i|B_t, L_t) \\ &= argmax_i\ k * log\ p_i + \sum_{j=1}^{j=k} log\ P\left(l_{t,j}\middle|x_{t,j}, h_i\right). \end{aligned} \tag{11}$$

If the pool is not full and the result of the expression computed in (11) is less than a parameter θ_1, a new classifier will be added.

Using this method, we must find the posterior probability of k instances for finding the best concept and this will take much time. To resolve this problem, relying on the fact that the instances in the batch are I.I.D, only a sub-

sample of the square root size of the batch is used to estimate the best concept. The pseudocode of this method is shown in Procedure 4. In line 2, S_t contains a subsample of the batch B_t and m is its size which is set to the square root size of the batch. SL_t stores the labels of S_t. Lines 5 to 7 find the best describing classifier of the batch according to Bayesian method. The variable *bestC* refers to the best classifier and *maxA* indicates the result of the expression computed in (11) for *bestC*.

```
1   X_t = sum_data (B_t);
2   S_t = sub_sample (B_t,m);
3   SL_t = sub_sample (L_t,m);
4   /* stores the labels of the S_t*/
5   (maxA , bestC) = (max,argmax)_{j:1..size(Pool)}
6   (m* log (RDC.prob(x_t,j)) +
7   Σ_{i=1:m} log (Pool[j].prob(S_i,SL_i)) );
8   if (maxA>θ_1 or size(Pool)>maxC)
9       Pool[bestC].update(B_t,L_t);
10  else
11      C = make_classifier(B_t,L_t);
12      Pool = Pool U {C};
13      bestC = size(Pool);
14  end if
15  RDC.update(x_t,bestC);
```

Procedure 4. Bayesian method for updating classifiers' pool.

Heuristic method for Updating the Classifiers' Pool

To find the best concept describing B_t and L_t, the accuracy of all classifiers on B_t will be measured. If the pool is full and a new classifier cannot be added, the best classifier is updated with B_t and L_t. But if the pool is not full and the accuracy of the best classifier for this batch of data is more than a parameter θ_2, then the best existing classifier is updated by B_t and L_t. Otherwise if the accuracy of classifier is less than θ_2, a new classifier is created and trained on this batch. The reason of using this approach is that the more the accuracy of a classifier on the current batch is, the more relevance it may have to the batch. Therefore, the concept this classifier describes can be refined or extended using the current batch of data. The pseudocode of this method is shown in Procedure 5. Lines 4 and 5 find the best classifier describing the batch according to heuristic method. The variable *bestC* refers to the best classifier and *maxA* indicates the accuracy of that classifier on the current batch.

```
1   S_t = sub_sample (B_t,m);
2   SL_t = sub_sample (L_t,m);
3   /* stores the labels of the S_t*/
4   (maxA , bestC) = (max,argmax)_{j:1..size(Pool)}
5   (pool[j].accuracy(S_t,SL_t));
6   if (maxA>theta or size(Pool)>maxC)
7       Pool[bestC].update(B_t,L_t);
8   else
9       C = make_classifier(B_t,L_t);
10      Pool = Pool U {C};
11      bestC = size(Pool);
12  end if
```

Procedure 5. Heuristic method to update classifiers' pool.

3.3 Phase 3: determining the active classfier (or classfier weiths)

After phases 1 and 2 are done, some final operations should be done before moving to the next iteration. If phase 1 is done according to the active classifier, the active classifier should be set. Active classifier is the one that has been updated with the current batch of data, i.e. the bestC parameter of our algorithm.

If phase 1 is done in the second way, the weights should be initialized for the next iteration. The weights of the classifiers in the pool are set so that in the next iteration, the performance of the method will be high. Each classifier is tested on a subsample of the square root size of the batch and its weight is set by:

$$w_0(i) = \beta^{(2^{A(i)})}, \tag{12}$$

Where *A(i)* is the accuracy of the i^{th} classifier. A classifier which classifies the current batch poorly, will have a less initial weight. Some kind of locality assumption is used in (12) for setting the initial weights which does not work properly when a sudden concept drift occurs. Phase 1 tries to handle this problem by updating the weights while processing the batch. The pseudocode of this method is shown in Procedure 6.

```
1   S_t = sub_sample (B_t,m);
2   SL_t = sub_sample (L_t,m);
3   for j=1 to size(Pool) do
4       c_error= Pool[j].error(S_t,SL_t);
5       W_j= beta^(2^c_error);
6   end for
```

Procedure 6. Determine classifier weights.

4 Experimental Results

In this section, we first introduce the data sets containing recurring concepts which are used in the experiments. Then we discuss the parameter tuning of our method and compare it to the parameters of CCP framework. In the last subsection the proposed methods are compared with each other and the CCP framework, one of the most promising frameworks developed in the tracking of recurring concepts. The experiments show the effectiveness of our method.

4.1 Data sets

Three real datasets and one artificial dataset are chosen for the experiments given in this section. The artificial dataset is moving hyperplanes and contains sudden concept drift. Real datasets are emailing list [Katakis et al., 2009], spam filtering and sensor data. Emailing list and spam filtering are high dimensional datasets and sensor data is a very large real dataset. Emailing list and hyperplane datasets contain sudden concept drift and spam filtering and sensor data contain gradual drift.

Emailing List Dataset

The emailing list (elist) dataset which is used in [Katakis et al., 2009] contains a stream of emails about different topics shown to the user one after another and are labeled as interesting or junk. To construct this dataset, the data in usenet posts [Frank, 2010] which exists in 20 newsgroups collection is used and three topics are selected. The user is interested in one or two topics in each concept and so he/she labels the emails according to his/her interest. The interests of the user can be changed in time and so this dataset simulates recurring concepts and concept drift (Table 1). The dataset contains 1500 instances with 913 attributes and is divided into 5 time periods with equal number of instances .

Spam Filtering Dataset

This dataset is obtained from Spam Assassin[‡] collection and contains email messages. The dataset consists of 9324 instances with 500 attributes and represents gradual concept drift .

Table 1. Emailing List Dataset (elist) [Katakis et al. 2009]

	1-300	300-600	600-900	900-1200	1200-1500
Medicine	+	-	+	-	+
Space	-	+	-	+	-
Baseball	-	+	-	+	-

Hyperplane Dataset

This dataset simulates the problem of predicting class of a rotating hyper plane. In an *n*-dimensional space, a hyper plane decision surface is the equation $g(\vec{x}) = \vec{w}.\vec{x} = 0$ where $\vec{w}$ determines the orientation of the surface and $\vec{x}$ is an instance in the space. If $g(\vec{x}) > 0$, $\vec{x}$'s label is 1, otherwise it is 0. To simulate concept drift, the orientation of the hyper plane is changed over time. Our dataset has 8000 instances with 30 real attributes. There is a concept drift after each 2000 instances. There are only two concepts which reappear after the first 4000 instances. This dataset shows the problem of sudden concept drift and recurring concepts.

Sensor dataset

Sensor dataset is a real dataset which consists of the information collected from 54 sensors deployed in Intel Berkeley Research laboratory in a two-month period [Zhu, 2010]. The class label is the sensor ID, so there are 54 classes, 5 attributes and 2,219,803 instances. The type and place of concept drift is not specified in the dataset but it is obvious that there are some drifts. For example, lighting or the temperature of some specific sensors during the working hours is much stronger than nights or weekends

4.2 Parameter Tuning

One of the advantages of the proposed method is that its parameters can be tuned in a much simpler way compared to the CCP framework method and small changes of parameter values, do not lead to major variations in performance. On the other hand, the CCP framework method has a θ parameter which is somehow similar to our θ_1 and θ_2 parameters. If this parameter is set wrongly in CCP framework method, the accuracy of the classification will decrease significantly. For example, θ should be 4 for elist and 2.5 for spam filtering dataset. If we set θ to 2.5 instead of 4 for elist dataset, its accuracy will be 55% rather than 77%.

If weighted classification method is used in phase 1, a parameter β is required to update the weights which is by definition in [0,1). The more sudden the concept drift is, the smaller the parameter should be. We have set this parameter to 0.1 for all datasets. Another parameter is the maximum classifier number (*maxC*) which is set to 10 and implies that we expect to have at most 10 different concepts. In addition, we have a parameter θ_1 in the heuristic method which is a threshold for the accuracy of the best classifier. So the more the *maxC* parameter is and the less sudden the concept drift is, the higher θ_1 should be. We have set this parameter to 0.95 for all datasets which means that only when a classifier has the accuracy more than 0.95 on a batch, it will describe the concept of the batch correctly. For the other parameter, θ_2, in the Bayesian method, we have set it to $2 * \log(0.75) * m$, according to its definition. This is because we believe if each of the *2m* probabilities of (11) is at least 0.75, then the concept can be relevant to the batch and its labels. The batch size is set to 50 for elist and spam filtering datasets and 500 for hyperplane dataset.

As a result, parameter tuning for our method is simpler than CCP framework method and the same parameters work well for all datasets with different natures we have chosen. The only parameter that does not have the same value for all datasets in our experiments is the batch size. This problem also exists in the CCP Framework method and must be resolved according to the properties of the dataset.

The reason behind our claim that our parameter setting is simple is that most of these parameters can be expressed as some property of the datasets, but setting the parameters correctly needs some knowledge about the dataset.

4.3 Results and Discussion

We compared our method with the CCP Framework method [Katakis et al., 2009] in terms of accuracy, precision, recall and running time. We have discussed how to tune our method's parameters in the previous subsection. The results of our experiments on elist, spam filtering, hyperplane and sensor datasets are shown in tables 2, 3, 4 and 5, respectively.

comparison of methods' accuracies, precisions and recalls

The results for elist and hyperplane datasets that simulate sudden concept drift are much better when using the weighted classifiers method rather than active classifier method. The difference of about 8% in the accuracies can be seen. We have tested the weighted classifiers method in conjunction with the CCP framework method and the same result can be seen in terms of increase in the accuracy. This is reasonable, because when a sudden concept drift occurs, the active classifier which is appropriate for the last batch works poorly in classifying the current batch. When the weighted classifiers method is used, after receiving the first few instances of the batch, the classifier' weights are adapted so that the concept drift is taken into account and the classification task will have a higher accuracy.

As a comparison, our weighted classifiers method outperforms the CCP framework method for sudden concept drift and has similar results for gradual concept drift. Our batch assignment methods (Bayesian and heuristic) have results similar to the CCP framework method without having parameter setting problems discussed previously.

In sensor dataset, CCP and Bayesian batch assignment methods have lower performances (between 9% and 15% of accuracy) than Heuristic method. This means that CCP framework and Bayesian method have some problems in determining the true concept of a batch in sensor dataset. One problem with CCP framework method is that it uses the Euclidean distance as the measure of similarity of a batch to a concept. ConDis, the distance measure used in CCP, is dependent on the magnitude of the attribute values and an attribute with large values can reduce the effects of the other attributes in the distance calculation. The problem of Bayesian method could be possibly the I.I.D as-

[‡] The Apache SpamAssasin Project - http://spamassassin.apache.org/

sumptions made in it. However, Bayesian method still outperforms than CCP framework method (about 3%).

Table 2. Results of all methods on elist dataset.

Batch assignment Method	Classification Method	Acc.	P	R	F-measure	Time
CCP	Active	0.77	0.73	0.81	0.77	1004
	Weighted	0.82	0.79	0.83	0.81	1274
Heuristic	Active	0.75	0.71	0.77	0.74	1816
	Weighted	0.82	0.8	0.83	0.81	1843
Bayesian	Active	0.75	0.71	0.8	0.75	2089
	Weighted	0.82	0.8	0.84	0.82	2462

Table 3. Results of all methods on spam filtering dataset.

Batch assignment Method	classification method	Acc.	P	R	F-measure	Time
CCP	Active	0.91	0.91	0.84	0.94	2217
	Weighted	0.89	0.92	0.87	0.93	2820
Heuristic	Active	0.89	0.91	0.84	0.93	3942
	Weighted	0.89	0.92	0.89	0.93	4112
Bayes	Active	0.89	0.9	0.86	0.93	4537
	Weighted	0.88	0.91	0.91	0.92	5405

Table 4. Results of all methods on Hyperplane dataset.

Batch assignment Method	classification method	Acc.	P	R	F-measure	Time
CCP	Active	0.76	0.72	0.81	0.78	868
	Weighted	0.83	0.81	0.83	0.84	947
Heuristic	Active	0.76	0.73	0.77	0.78	974
	Weighted	0.84	0.81	0.83	0.85	970
Bayes	Active	0.78	0.75	0.8	0.8	876
	Weighted	0.86	0.83	0.84	0.87	899

Table 5. Results of all methods on sensor dataset.

batch assignment	classification method	Accuracy	Time
CCP	Active	0.71	370560
	Weighted	0.71	813398
Heuristic	Active	0.87	929289
	Weighted	0.86	846226
Bayes	Active	0.74	883682
	Weighted	0.74	1299652

Comparision of methods' run times

The run time of each method is shown in the last column of the result tables (Table 2-5). The most time consuming part of these methods is the time spent calling the training and test methods of the classifiers. In the CCP framework method additional time is spent on the construction of the conceptual vectors and the clustering task. In all methods, each instance of the batch is used once to update a classifier in the pool. The difference is in the number of times an instance is classified or its posterior probability distribution is measured by the classifiers. Simply, assume that T_0 is the time taken to classify an instance and T_1 is the time taken to find the posterior probabilities for it. In the classification task, each data is classified only once in all batch assignment methods and so the only major differences are in updating the classifiers' weights and in phase 2 where the updating of the classifiers' pool is done. Suppose that the subsample size of the batch used in both the heuristic and the Bayesian methods is m. In the heuristic method, each of the m instances is classified once using all of the classifiers in the pool and in the Bayesian method, the posterior probabilities of each of the m instances are measured by each of the classifiers. In the Bayesian method, one posterior probabilities estimation and one update by the raw data classifier is also required for each batch but this can be ignored. So the time required in the heuristic method is at most $m * maxC * T_0$ and in the Bayesian method is at most $m * maxC * T_1$. T_1 is greater or equal to T_0 according to their definitions. So in general, we expect using the Bayesian method is more time consuming rather than the heuristic method, because the maximum time computed for Bayesian method is greater. This can be seen in tables 2 and 3, but not in the last dataset, because in this problem setting only two classifiers are added to the pool for the Bayesian method (among 10 possible classifiers).

In addition, we use a subsample of the batch to update the weights in the weighted classifiers method. Each of the instances in this subsample is classified by each of the classifiers in the pool to find the classifiers' errors. So if we use the same subsample of the batch for both updating the classifiers and their weights, we will obtain a time saving when using Heuristic and weighted classifiers methods. Therefore for each of batch assignment methods, using weighted classifiers method will consume more time than using active classifier. This can be seen in tables 2 to 4 for our three datasets, except in the Heuristic method because of the time saving we mentioned.

At last, Bayesian method takes the most time among all batch assignment methods while Heuristic and CCP methods take almost the same time using active classifier and Heuristic method is better when using weighted classifiers.

5 Conclusion and Future Works

We have proposed a method with some variations for streaming data classification in the presence of concept drift and recurring concepts. The general framework used in this paper maintains a pool of classifiers and updates them according to consecutive batches of data. The classifiers in the pool are used to classify new batches of data. The most similar method to our method is the CCP framework. Our method improves the accuracy while its parameter tuning is simpler.

Some future research works related to this study might include the followings. First, managing the classifiers in the pool can be done more complexly. For example, classifiers can be merged or removed to handle more complicated situations. Second, parameters of the algorithm are

dependent on the datasets. If they can be set dynamically according to the datasets, the algorithm will work properly for all datasets. Third, the algorithm should be run on more real datasets in order to achieve more reliable results.

References

[Baena-García et al., 2006] Manuel Baena-García, José del Campo-Ávila, Raul Fidalgo, Albert Bifet, Ricard Gavaldà and Rafael Morales-Bueno, Early Drift Detection Method, in ECML PKDD Workshop on Knowledge Discovery from Data Streams. 2006.

[Bifet, 2009] Albert Bifet, Adaptive Learning and Mining for Data Streams and Frequent Patterns, in Departament de Llenguatges i Sistemes Informatics. 2009, Universitat Politecnica de Catalunya.

[Bifet et al., 2010a] Albert Bifet, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Accurate ensembles for data streams: Combining restricted Hoeffding trees using stacking. in 2nd Asian Conference on Machine Learning. 2010. Tokyo, Japan: JMLR.

[Bifet et al., 2010b] Albert Bifet, A., Geoff Holmes, Richard Kirkby, Bernhard Pfahringer, MOA: Massive Online Analysis. Journal of Machine Learning Research. 2010, **99**: pp. 1601-1604.

[Frank, 2010] Frank, A., UCI Machine Learning Repository. 2010. accessed on May 2011; Available from: http://archive.ics.uci.edu/ml.

[Freund and Schapire, 1996] Yoav Freund, and Robert .E. Schapire. Game theory, on-line prediction and boosting. in Proceedings of the ninth annual conference on Computational learning theory. 1996: ACM.

[Gama and Castillo, 2006] Joa Gama and Gladys Castillo, Learning with local drift detection, in Advanced Data Mining and Applications, Proceedings, X. Li, O.R. Zaiane, and Z.H. Li, Editors. 2006, pp. 42-55.

[Gama and Kosina, 2009] Joa Gama and Petro Kosina, Tracking Recurring Concepts with Meta-learners, in Proceedings of the 14th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence. 2009.

[Gao et al., 2008] Jing Gao, Bolin Ding, Wei Fan, Jiawei Han, Classifying Data Streams with Skewed Class Distributions and Concept Drifts. IEEE Internet Computing, 2008. **12**(6): pp. 37-49.

[Garnett, 2010] Roman Garnett, Learning from Data Streams with Concept Drift, in Department of Engineering Science. 2010, University of Oxford. pp. 163.

[Gomes et al., 2011] Joao B. Gomes, Ernestina Menasalvas, and Pedro A.C. Sousa, Learning recurring concepts from data streams with a context-aware ensemble, in Proceedings of the 2011 ACM Symposium on Applied Computing, 2011, pp. 994-999.

[Helmbold and Long, 1994] David P. Helmbold and P. M. Long, Tracking Drifting Concepts by Minimizing Disagreements. Machine Learning, 1994, 14, pp.27-45.

[Ikonomovska et al., 2010] Elena Ikonomovska, Joa Gama, and S. Deroski, Learning model trees from evolving data streams. Data Mining and Knowledge Discovery, 2010. **23**(1): pp. 128-168.

[Katakis et al., 2009] Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas, Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems, 2009. **22**(3): pp. 371-391.

[Klinkenberg and Joachims, 2000] Ralf Klinkenberg and Thorsten Joachims, Detecting Concept Drift with Support Vector Machines. In the Proceedings of the Seventeenth International Conference on Machine Learning (ICML), 2000, pp.487-494.

[Klinkenberg, 2004] Ralf Klinkenberg. Learning Drifting Concepts:Example Selection vs. Example Weighting. Intelligent Data Analysis, Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, 2004, **8**(3), pp.281-300.

[Kolter and Maloof, 2007] J. Zico Kolter and Marcus A. Maloof, Dynamic weighted majority: An ensemble method for drifting concepts. Journal of Machine Learning Research, 2007. **8**: pp. 2755-2790.

[Kuh et al., 1991] A. Kuh, T. Petsche and H. Rivest, Learning Time-Varying Concepts. In Advances in Neural Information Processing Systems (NIPS), 1991, pp.183-189.

[Kuncheva and Zliobaite, 2009] Ludmila I. Kuncheva and Indre Zliobaite, On the window size for classification in changing environments. Intell. Data Anal., 2009. **13**(6): pp. 861-872.

[Lazarescu, 2005] Mihai M. Lazarescu, A Multi-Resolution Learning Approach to Tracking Concept Drift and Recurrent Concepts, in 5th IAPR Workshop on Pattern Recognition in Information Systems (PRIS). 2005: Miami, USA. pp. 52-61.

[Morshedlou and Barforoush, 2009] Hossein Morshedlou, and Ahmad A. Barforoush, A New History Based Method to Handle the Recurring Concept Shifts in Data Streams. World Academy of Science, Engineering and Technology, 2009. **58**: pp. 917-922.

[Nishida, 2008] Kyosuke Nishida, Learning and Detecting Concept Drift, in Information Science and Technology. 2008, Hokkaido University: Hokkaido.

[Ramamurthy and Bhatnagar, 2007] Sasthakumar Ramamurthy and Raj Bhatnagar. Tracking Recurrent Concept Drift in Streaming Data Using Ensemble Classifiers. in Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA '07). 2007. pp. 404-409.

[Scholz and Klinkenberg, 2006] Martin Scholz and Ralf Klinkenberg, Boosting Classifiers for Drifting Concepts. Intelligent Data Analysis, Special Issue on Knowledge Discovery from Data Streams, 2007, **11**(1), pp.3-28.

[Schlimmer and Granger, 1986]Jeffrey C Schlimmer and Richard H Granger Jr, Incremental learning from noisy data. Machine learning,1986, 1(3), pp. 317-354.

[Widmer and Kubat, 1993] Gerhard Widmer and Miroslav Kubat, Effective learning in dynamic environments by explicit context tracking. in Proceedings of the European Conference on Machine Learning, 1993, pp.227-243.

[Widmer and Kubat, 1996] Gerhard Widmer and Miroslav Kubat, Learning in the Presence of Concept Drift. Machine Learning, 1996, 23, pp.69-101.

[Witten et al., 2005] Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical machine learning tools and techniques. 2005: Morgan Kaufmann.

[Tsymbal, 2004] Alexey Tsymbal, The Problem of Concept Drift: Definitions and Related Work. 2004.

[Zliobaite, 2010a] Indre Zliobaite, Learning under Concept Drift: an Overview. 2010.

[Zliobaite, 2010b] Indre Zliobaite, Adaptive Training Set Formation. 2010, Vilnius University.

[Zhu, 2010]Xingquan Zhu, *Stream Data Mining repository*.2010. Accessed on Jan 2012; Available from: http://www.cse.fau.edu/~xqzhu/stream.html

Cross Domain Active Learning

Christian Poelitz
TU Dortmund University, Dortmund 44227, Germany
christian.poelitz@cs.uni-dortmund.de

Abstract

In this paper, we propose a solution to reduce the labeling costs by applying domain adaption methods coupled with active learning to reduce the number labels needed to train a classifier. We assume to have only one task but different domains in the sense that we have texts that come from different distributions. Our approach uses multi domain learning together with active learning to find a minimum number of texts to label from as few domains as possible to train a classifier with a certain confidence in its predictions.

1 Introduction

A large cost factor in computer linguistic rises from the labeling of texts. For example, we want to investigate the hypothesis that certain statements occur always in positive or negative context in a large set of texts. A usual approach in computer linguistic would be to go through the texts and label parts of it as positive or negative to use them as examples for a classifier. This can be quite expensive with respect to the texts and the task.

The problem gets an additional twist, in case we have a large corpus or many corpora of texts from different domains. To investigate the texts, we would start to label the texts from a certain domain. Now, it is easier to keep labeling only texts from this domain instead of switching to another.

To reduce labeling costs, we propose to use active learning techniques to support us on what to label. We base our decision on what to label on a trained classifier and the confidence of the classification of unlabeled texts. First, we need to know when we can stop labeling in the current domain. This will happen when further labels will not increase the quality of the classifier anymore. Next, we want to find out if we actually need labels from the other domains when there are not further texts in the current domain or the quality of the classifier saturates. In case we have large confidence in predicting the texts in all domains we do not need further labels. This means, we train a classifier on only some domains and expect a generalization on all domains. Unfortunately, this is usually only possible under strong assumptions on the distributions of the texts in the different domains. By distributions of the texts we mean the probability distribution of a stochastic process that generates the text.

Further, we assume a nonlinear cost model with respect to the number of labels. We expect that at the beginning the labeling of the texts demands largest effort but decreases with more and more labels. This is intuitively clear since we will grow accustom to the texts. Further, when we start labeling texts from an other domain, we must consider that the effort will be again higher at the beginning and smaller after a while.

The paper is organized as the following. First, we explain how we can statistically model the distribution of texts in the different domains and what classifier we use in our training. Then, we describe how we use domain adaptation and active learning to train a classifier for all domains. Finally, we report results on our propose method on a benchmark data set.

2 Related Work

We leverage methods from active learning, domain adaptation and multi domain learning.

Active learning tries to direct the labeling process considering intermediate results. A classifier that is trained on a small amount of labeled texts is used to estimate which further texts should be labeled to increase the quality of the classifier when trained also on these labeled texts. As candidates for further labeling we use the texts that are classified with least confidence. This strategy is called

uncertainty sampling (LC94). There are different sampling strategies in the literature. A general overview is given by (Set09).

We assume that the texts have different distributions in the domains but the labels have the same distribution given a text. In this case, instance weights can be used. In (JZ07), a classifier is trained on examples with labels and weights for each example. The weights are chosen such that the mass distribution of the examples from one domain adapts to the mass distributions of an other domain. By this, they train a classifier using examples and labels from one domain that generalizes to an other domain. An other approach is to model the commonalities of different domains as proposed by (BMP06) or (DM06) for instance.

In multi domain learning, a classifier is learned over several domains. A classifier shall be generated that performs best over all domains while using only a small amount of training data that comes from very few domains. In best case we need only to training the classifier on a single domain. An overview on existing multi learning methods is given by (JCDR12).

3 Statistical modeling

Since we pose assumptions on the distributions of texts, we need to model these distributions based on text examples from the domains. There are many approaches to model the probability distribution of texts. Here, we use the language model (PC98).

The probability of a word w or a sequence of words (a text) can be estimated by the frequency of the occurrences of the word. Formally, we note $p(w_1 \cdots w_n)$ as the probability of the event to see (or to read) the sequence $w_1 \cdots w_n$ in the domain. Further, $p(w_n|w_1 \cdots w_{n-1})$ is the probability of seeing word w_n, after we have already seen the words $w_1 \cdots w_{n-1}$. By assuming independence of words that are farther away from each other than a given context size, we can estimate the probability by frequencies easily. Using a context of only one word we also speak of a bigram model in contrast to a unigram model when we assume all words are independent, hence: $p(w_1 \cdots w_n) = \prod_{i=1}^{n} p(w_i)$. This naturally generalizes to ngram models, when we consider a context of $n-1$ words.

The concrete probabilities for unigrams can be estimated for a given domain by the maximum likelihood estimate of the Multinomial distribution, hence $p(w) = \frac{N_w}{N}$ for N_w number of occurrences of word w among the N words in the domain.

4 Classifier

As classifier we use support vector machines that have proven to be efficient in text classification, see (Joa02) for example. Given a set of texts with labels, we find a separating hyperplane in a Reproducing Kernel Hilbert space. In this paper we use the bag of word representation. Each text is mapped to a large vector (a word vector) such that each component tells how many times a certain word occurs in the text.

During SVM training we minimize a regularized loss, formally $min_f \frac{1}{N} \sum_{i=1}^{N} [(1 - y_i \cdot f(x_i))_+] + \lambda \cdot ||f||$ using the hinge loss $()_+$, y_i the labels and x_i the texts. We use an adaptation that integrates weights on the texts. This means we solve the following minimization problem: $min_f \frac{1}{N} \sum_{i=1}^{N} \beta_i \cdot [(1 - y_i \cdot f(x_i))_+] + \lambda \cdot ||f||$ See (LLW02) for further details.

In order to retrieve confidence in the prediction of our classifier we use the approach by (Pla99) to derive posterior probabilities using the outcome of an SVM. The probability of a prediction given an example (here a text) is modeled as sigmoid function: $P(y = 1|f(x)) = \frac{1}{1+exp(A \cdot f(x)+B)}$. The parameters A and B are estimated using the labeled texts. Using this method, we get the confidence of the prediction of a text as.

5 Domain Adaptation

Assuming that the texts are differently distributed in different domains we use the SVM with weighted examples as described above. The weights are estimated based on the difference of the distributions of texts using importance sampling based on language models.

5.1 Importance sampling

If P_s and P_t are the text distributions from domain s and domain t with the same support, we can estimate the expected loss under the domain t using texts from domain s, using importance sampling. In importance sampling we sample from P_s but weight the examples by $\beta(x)$ such that $\beta(x) \cdot x$ has approximately the distribution P_t. For further reading we refer to (OZ00). We integrate these weights into the risk minimization framework for

the SVM using the hinge loss L. This results to the following:

$$E(L(x,y,f)) = \int L(x,y,f) \cdot P_t(x,y) \cdot dx$$

$$= \int \frac{P_t(x,y)}{P_s(x,y)} \cdot L(x,y,f) \cdot P_s(x,y) \cdot dx$$

$$\simeq \frac{1}{N} \sum_{i=1}^{N} \frac{P_t(x,y)}{P_s(x,y)} \cdot L(x,y,f)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \beta_i \cdot L(x,y,f)$$

In this paper we concentrate on covariate shifts. This means, we expect that conditional probabilities of the labels, given an observation, are the same over two different domains. This means, $P_s(y|x) = P_t(y|x)$. Hence, we can write $\frac{P_t(x,y)}{P_s(x,y)} = \frac{P_t(x) \cdot P_t(y|x)}{P_s(x) \cdot P_s(y|x)} = \frac{P_t(x)}{P_s(x)}$.

5.2 Multi Domain Classifier

We want to train a classifier that can be applied on different domains but the training is only done on texts from a single domain or a small amount of domains. Using the language model we estimate the probability distributions of the texts from each domain i, noted as $P_i(x)$. Further, we define an ensemble of classifiers $f_i(x)$. Each classifier f_i is trained with respect to the distribution of domain i using importance sampling on an other domain. Given a trained set of classifiers f_i we perform the prediction on a given text - from any domain - as: $F(x) = f_{i(x)}(x)$ with $i(x) = argmax_i\{P_i(x)\}$.

6 Active Learning across different domains

In this section we describe how we use active learning and domain adaptation in order to reduce the labeling effort over different domains in a classification task. We generally assume that the distribution of the texts differ among different domains. Formally this means $P_i(x) \neq P_j(x)$, for two different domains i and j and a text x. Further, we assume that the distributions of the labels for a given example are the same among the domains, hence $P_k(y = i|x) = P_l(y = i|y)$.

The goal is to train the classifier only on few domains and examples but apply it to all domains. To achieve this goal, we use an active learning technique to ask for labels in a certain domain such that the number of overall labels are minimized while maximizing the expected quality over all domains. Therefore, we train an SVM with probabilistic outputs to estimate the confidence in the predictions. This means, for each domain i we train an SVM on the texts from a single domain j, but weight them as described above by $\frac{P_i(x)}{P_j(x)}$. These classifiers f_i are then combined to the multi domain classifier F. The multi domain classifier is applied to all unlabeled examples from all domains. In case all the resulting predictions have at least a certain level of confidence we can stop here and use F as final classifier. When there are still predictions with less confidence we need further labels.

There are two possibilities to continue. First, we can ask for more labels from the current domain. Second, we ask for labels from any different domain. We propose to suggest to switch to a different domain only when there are no further unlabeled examples in the current domain. Then, the next domain can be any domain that still contains unlabeled examples. This is a valid approach since we expect that the texts samples in the different domains are independent identical distributed. Then, we only need to start asking for labels from a next domain in case we have no further examples in the current domain. In general, with enough examples in one domain we expect the same quality of the classifier when we train only on this domain as when we train on all domains. This holds because we expect only a covariance shift.

When we continue - maybe with a new current domain - we apply each classifier f_i on all unlabeled examples from the current domain, but each time we weight the examples before applying the classifier. By this we adapt the mass distribution to the corresponding domain. Then, the examples that have the least confident predictions among all classifiers should be used. Hence, among the least confident predictions of the unlabeled data we sample k examples and ask for their labels. Afterwards, we train the classifiers again using also the newly labeled samples, build the multi domain classifier and test if we have enough confident predictions now. If we still have some examples with low confidence, we simply perform the steps again.

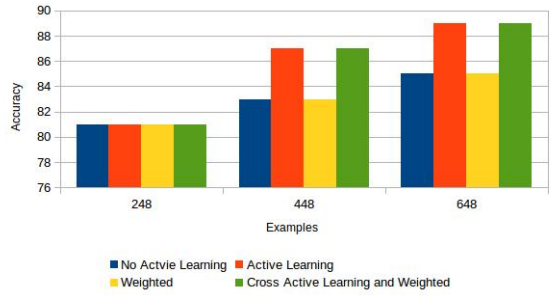


Figure 1: Results on Reuters People.

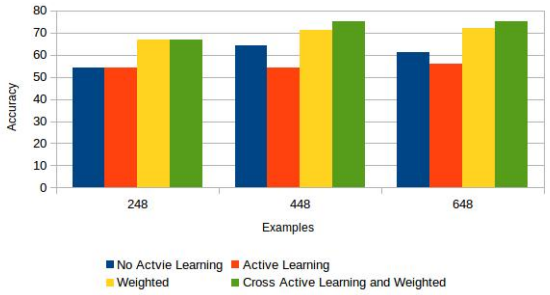


Figure 2: Results on Reuters Organizations.

7 Experiments

We test our proposed method on a standard benchmark data set that is commonly used in NLP. We use the Reuters-21578[1] data set with the topics people and organizations. For both domains we estimate a language model to model $P_i(d)$ the probability that document d was generated in the domain i for $i = s$ the source domain containing texts talking about people and $i = t$ the target domain containing texts talking about organizations. We use the weighted SVM as classifier and weighted texts from the source domain for training. The weights are chosen with respect to the probability of the texts on target domain and source domain as explained above.

We split the data from the source domain into 3 parts each having $1/5, 1/5, 3/5$ of the original data. One split is used for the first training without active learning. One split is used for testing the classifier and the final split is used for active learning as described above. We use always batches of 200 examples for which we ask labels. In iteration k these 200 are the examples that are classified with the least confidence by the classifier trained with all labeled examples so far.

To investigate the behavior of our proposed model we conducted several experiments. First, we tested how good we perform on the source domain when we use an active learning approach. Then, we investigate how our trained classifier performs on the target domain. We are specially interested in how much benefit we get from weightening the examples. Finally, we test our proposed active learning strategy across the two domains.

Figure 1 and Figure 2 shows the accuracy on the source domain respectively target domain for different experiments. For the first batch, we cannot perform an active learning strategy since we need a trained classifier first. Only after we already have trained a classifier we can perform an active learning strategy. The first two bars show the results on the accuracy when we perform an active learning strategy and when not. We get faster a better accuracy on the source domain when we actively ask for the next labels. Next, we test the classifier on the target domain. The first two bars show the accuracy when the classifier is only trained on the source domain without weights. We see that the accuracy is low and more train data increases the quality only slightly. Furthermore, we see that the active learning strategy, which considers only the source domain here, performs worse on the target domain. Next, we investigate the domain adaption by weighting the texts with respect to the language models. The third bar shows that the accuracy on the target domain increases when we use importance sampling. Finally, we test our active learning strategy across the two domains. The last bar shows that when we actively ask for labels with respect to both domains, we get the best overall accuracy.

8 Conclusion and Future Work

We explained an approach to perform active learning across different domains. We used importance sampling and statistical language models to adapt an SVM trained on a certain source domain to a different target domain. Our proposed active learning strategy that considers both domains shows good results on a benchmark data set. In the future we want to investigate how other probability models for the texts can be used. Further, we plan to extend our approach to multi task learning.

[1] http://www.daviddlewis.com/resources/testcollections/

References

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Hal Daumé, III and Daniel Marcu. Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1):101–126, May 2006.

Mahesh Joshi, William W. Cohen, Mark Dredze, and Carolyn P. Rosé. Multi-domain learning: when do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1302–1312, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

Jing Jiang and Chengxiang Zhai. Instance weighting for domain adaptation in nlp. In *In ACL 2007*, pages 264–271, 2007.

David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1994.

Yi Lin, Yoonkyung Lee, and Grace Wahba. Support vector machines for classification in nonstandard situations. *Mach. Learn.*, 46(1-3):191–202, March 2002.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):pp. 135–143, 2000.

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

Burr Settles. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

Smoothed Discretization for Simplified Cutpoints

Georg Dietrich, Florian Lemmerich, and Frank Puppe
University of Würzburg
{dietrich, lemmerich, puppe}@informatik.uni-wuerzburg.de

Abstract

This paper describes work in progress. Discretization is one of the most common pre-processing steps in data mining and machine learning. We propose a novel approach to obtain simpler discretization cutpoints, which are easier to capture for human users, e.g., as they require less non-zero digits. For that purpose, a post-processing step is performed after applying an arbitrary conventional discretization method. It trades-off the necessary modifications in comparison to the original discretization scheme with the reduction in complexity of the cutpoints. Experiments with classification tasks show, that this leads to considerably simpler cutpoints with only marginal influence on the algorithmic performance, i.e., the prediction accuracy.

1 Introduction

The following paper reports preliminary results of ongoing research. Many machine learning and data mining algorithms, e.g., rule learners or decision-tree algorithms, can be applied automatically, but aim at models, which allow for introspection by the user. Other approaches, such as subgroup discovery, are not intended for automatic application at all, but provide patterns, which are directly interpreted by human experts. Both categories of algorithms require simple input data to build understandable models.

Discretization is a key pre-processing technique. It transforms numeric attributes into nominal ones in order to apply algorithms, which allow only for nominal attributes as inputs. Over the last decades a large number of sophisticated discretization methods have been proposed [7; 4]. Until now, research on these methods has focussed almost exclusively on the predictive power of the thresholds, but mostly ignored the resulting complexity of the discretization thresholds. This leads to discretization intervals that are inconvenient for humans, e.g., $income = [38952.4; 60427.2[$. Findings for such boundaries are not only unintuitive, but also potentially less useful in the application domain, as they are difficult to compare with previous knowledge. Additionally, such discretization bounds are potentially subject to *over-fitting* on the training data.

In this paper, we present a novel meta-method for discretization that aims at obtaining discretization thresholds, which are more intuitive for human users. For example, a very similar, but much simpler discretization interval for the above interval could be $income = [40000; 60000[$. Our approach obtains such simpler intervals by post-processing the resulting cutpoints of an arbitrary discretization method. In doing so, we combine the advantages of sophisticated discretization algorithms with intuitive discretization thresholds. The extent of the modification is traded-off against the complexity reduction of the results. Although our approach is applicable as a pre-processing method for arbitrary data mining tasks, the evaluation focuses in this work-in-progress on the classification tasks, since their results can be easily compared.

The rest of this paper is structured as follows: Section 2 introduces notations and discusses some related work. Next, Section 3 presents our novel approach of smoothed discretization bounds. First experimental results are provided in Section 4. The paper concludes with pointers to future work in Section 5.

2 Background and Related Work

In this paper, a dataset $\mathcal{D} = (\mathcal{I}, \mathcal{A})$ is formally defined as an ordered pair of a set of *instances* $\mathcal{I} = i_1, i_2, \ldots, i_y$ and a set of attributes $\mathcal{A} = A_1, A_2, \ldots, A_z, C$. Each attribute $A \in \mathcal{A} : I \rightarrow dom(A)$ is a function that indicates a characteristic of an instance by mapping it to a value in its range. Consequently, $A_m(i)$ denotes the value of the attribute A_m for the instance i. In our setting, there is one class attribute in each dataset A_C, which is to be predicted by a classification algorithm. We assume the class attribute to be *nominal* and all other attributes $A_1, \ldots, A_i$ to be *numeric*, i.e., $dom(A_i) = \mathbb{R}$.

Many data mining algorithms are not directly suited for numeric attributes, but require nominal attributes as input data. Therefore, *discretization algorithms* are used in a pre-processing step to transform a numeric attribute A into a new nominal attribute A'. These methods split the range of a numeric attribute into $n + 1$ disjunct intervals defined by a set of cutpoints $cp_1, \ldots, cp_n$: $\mathbb{R} =]-\infty; cp_1],]cp_1; cp_2], \ldots,]cp_{n-1}; cp_n],]cp_n; +\infty[$. The new attribute has one value for each of these intervals. Instance values are mapped accordingly:

$$A'(i) = \begin{cases} 0, & \text{if } A(i) \leq cp_1 \\ k & \text{if } cp_k < A(i) \leq cp_{k+1}, k = 1, \ldots n \\ n & \text{if } A(i) > cp_{k+1} \end{cases}$$

The cutpoints for different attributes are determined independent from each other by using a discretization method. For this task, a large amount of discretization methods have been proposed in literature, see [7; 4] for two recent overviews. The most popular methods include Equal-frequency discretization, top-down entropy-based discretization [3] and bottom-up discretization based on chi-values [6; 8]. Discretization methods, which lead to easy-to-read intervals, have received only little attention so far. An exception to this is the *intuitive partitioning* proposed by Han and Kamber [5] that discretizes an attribute into "natural" segments: In a top-down approach, the range of the attribute A is split into three, four, or five subintervals depending on the difference in the most significant digit in the attribute range. In contrast to this technique, our novel method joins the power of supervised discretization

algorithms with the goal of easy-to-read cutpoints. It can be combined with arbitrary discretization methods.

3 Smoothed Discretization

In the next section, we present our novel approach for discretization. The main idea is as following: First, any traditional discretization algorithm is run. The resulting set of cutpoints $cp_1, \ldots, cp_n$ is used as the input for our technique. A new discretization scheme is obtained by modifying the cutpoints cp_i one-by-one. For each cutpoint, an alternative new cutpoint is determined. The selection of the new cutpoints follows two criteria: 1. The replacement cutpoint should be "natural", i.e., less complex and easier-to-read. 2. The replacement cutpoint should be as close to the original cutpoint as possible. In the following sections, we present novel measures to quantify these criteria as well as a simple scoring function, which allows to trade-off between them. Furthermore, we outline a simple algorithm that allows to identify the best alternative cutpoint. It generates a number of candidate cutpoints, which are scored by the presented measure.

3.1 Complexity

The perceived complexity of a number differs from user to user. Due to this inherent subjectiveness quantifying its complexity is a difficult issue. One can consider several different intuitions to measure the complexity of a number, which are plausible for most users: First, short numbers are easier to comprehend than longer numbers: As a consequence the number 624 should receive a lower complexity score than a number like 7245. Second, one benchmark could be, how difficult it is to remember a number. Therefore, 1.000 would have a lower complexity score than 8103. A potential third intuition is, that numbers should receive a lower complexity score, if they are used more often by humans.

Next, we present one simple method to capture the complexity of a number. We are fully aware that this is definitely not the only solution for this problem and one can think of several variations of this measure. Our measure is based on the decimal representation of a number x. The scoring is based on the number of digits $k(x)$, which are required to write x, excluding trailing zeros. The count $k(x)$ is increased by one, if it contains a decimal point. Then, the complexity for x is defined as:

$$complexity(x) = \begin{cases} 0, & \text{if } x = 0 \\ 1 & \text{if } x = 10^n, n \in \mathbb{N} \\ 1 + k(x) & \text{else} \end{cases}$$

The following table shows some examples for this complexity measure.

x	$complexity(x)$	x	$complexity(x)$
0	0	400	2
1	1	725	4
100	1	-725	4
4	2	7.25	5

This basic measure could be improved in a variety of directions: One may argue that a number ending with the digit 5 is simpler than other numbers. E.g., 95 can be considered as a simpler, more intuitive bound than 93. Another issue is, if a decimal really increases the complexity, i.e., if 0.4 is a more complex number than 4. Although these considerations could be incorporated in more sophisticated variations in future approaches, we focus in this paper on the complexity measure presented above for the sake of simplicity and transparency.

3.2 Modification measures

Additionally, our approach requires a measure that compares, how strongly the original discretization scheme is modified, if a candidate cutpoint cp'_k is used instead of the respective original cutpoint cp_k. To quantify this amount of modification we propose two measures.

Distance-based deviation measures

The distance-based measure describes the difference in the range of the discretized attribute. It is computed as the percentage of the interval between the original cutpoint cp_k and the candidate cutpoint cp'_k in relation to the distance between the current cutpoint cp_k and the neighboring cutpoint in the original discretization scheme. For candidates smaller than the original cutpoint, the neighboring cutpoint is given by the next lowest cutpoint cp_{k-1}, otherwise it is the next highest cutpoint cp_{k+1}. For the special cases, that the current cutpoint is the first one ($k = 1$) or the last one ($k = n$), the instances in the dataset with the lowest, respectively highest, attribute values are used as neighboring cutpoints. Formally it is computed as (ignoring the special cases):

$$mod^{dist}(cp'_k, cp_k) = \begin{cases} 0 & \text{if } cp'_k = cp_k \\ \frac{cp_k - cp'_k}{cp_k - cp_{k-1}} & \text{if } cp'_k < cp_k \\ \frac{cp'_k - cp_k}{cp_{k+1} - cp_k} & \text{if } cp'_n > cp_k \end{cases}$$

Instance-based deviation measures

The distance-based deviation measure just considers the difference between the original cutpoint and the candidate cutpoint, independent of additional information contained in the dataset. The second approach, the instance-based deviation, additionally takes the values of each instance i for the attribute A, which is discretized, into account. It measures the percentage of the instances in the interval, which are relocated to another interval, if the original cutpoint cp_k is exchanged with the candidate cp'_k:

$$mod^{inst}(cp'_k, cp_k) = \begin{cases} 0 & \text{if } cp'_k = cp_k \\ \frac{|\{i | cp'_k \leq A(i) < cp_k\}|}{|\{i | cp_{k-1} \leq A(i) < cp_k\}|} & \text{if } cp'_k < cp_k \\ \frac{|\{i | cp_k \leq A(i) < cp'_k\}|}{|\{i | cp_k \leq A(i) < cp_{k+1}\}|} & \text{if } cp'_k > cp_k \end{cases}$$

3.3 Smoothed cutpoint selection

Cutpoint smoothing is a trade-off between reducing the complexity of a cutpoint and modifying the intervals generated by the original discretization method. For that purpose, we propose the following family of functions that balances between these two goals using the complexity and modification measures presented above. The candidate with the lowest score according to this measures is considered the best cutpoint.

$$score(cp'_k) = complexity(cp'_k) + \frac{1}{\alpha} \cdot mod(cp'_k, cp_k)$$

Here, α is a user chosen parameter. For high values of α less complex cutpoints are preferred, even if they strongly modify the original solution. In contrast, lower values of α emphasize the similarity to the original discretization scheme, even if the resulting cutpoints are only slightly less complex than the original ones. α can be interpreted, which ratio of an interval the cutpoint can be moved to reduce the complexity by one point. E.g., if $\alpha = 0.05$ the algorithm will shift the cutpoint by up to 5% of the adjacent interval (based on the pure difference or the number of contained instances), if this decreases the complexity by one.

3.4 Computation of smoothed cutpoints

The computation of the best smoothed cutpoint is straight forward: First, candidates are generated in two directions. For that purpose the cutpoint is iteratively rounded up with decreasing precision. This is repeated, until either zero or

the middle of the adjacent interval is reached. This prevents, that two different original cutpoints are smoothed to identical values. Candidate cutpoints smaller than the original cutpoint are obtained analogously by rounding down. Additionally, the original cutpoint is also considered as a candidate. Then, every candidate is evaluated by the score-function with user-chosen parametrization. The best (lowest) scoring candidate then replaces the original cutpoint in the smoothed discretization scheme.

3.5 Example

We demonstrate our approach in a small example: Initially the user chooses a parameter α for the scoring function and one of the two proposed modification measures. We assume an α value of 0.01 and the distance-based deviation measure in this example. To discretize an attribute A with our approach, first a traditional discretization algorithm, e.g., frequency-based discretization, is executed. We assume, this method resulted in the 3 cutpoints $cp_1 = -724$, $cp_2 = 692$, and $cp_3 = 1525$. For each of these cutpoints, our approach determines a smoothed cutpoint cp_i^*, which should be easier-to-read. In this example we focus on the cutpoint cp_2. For this cutpoint, first candidates for alternative cutpoints are determined by rounding up and down. This results in the candidates 700 and 1000 for rounding up and 690, 600, and 0 for rounding down. Additionally, the original cutpoint 692 is considered as a candidate. For each of these six cutpoint candidates the score is determined as described in Section 3.3. For example, the score for the candidate 700 is determined as follows: The complexity of the candidate is $complexity(700) = 2$. Its distance is computed as $\frac{700-692}{1525-692} \approx 0.0096$. The score for this candidate is $score(700) = 2 + \frac{1}{0.01} \cdot 0.0096 = 2.96$. Analogously, the score for the cutpoint 1000, which has a complexity of 1 is determined as $1 + \frac{1}{0.01} \cdot \frac{1000-692}{1525-692} \approx 37.97$. As another example, the original cutpoint has a complexity of 4 and a distance of 0 and thus a score of $score(692) = 4$. As it turns out after computing the scores for all six cutpoint candidates, 700 has the lowest (best) score and thus is used as a replacement for the original cutpoint in the novel discretization scheme.

4 Evaluation

To evaluate the effectiveness of our novel approach, we performed an experimental study on a classification task, using the well-known decision tree algorithm *C4.5* [9]. We used 12 data sets from the UCI Machine Learning Database Repository [2] and from the KEEL data set repository [1]. Except for the class attributes, these data sets consists of numerical attributes only.

We applied a standard 10-fold cross-validation procedure. For each training data set the discretization cutpoints were determined for the three popular discretization methods *equal-frequency discretization*, *entropy-based discretization*, and *Chi2 discretization*. Afterwards, the introduced smoothing techniques were performed with distance-based and instance-based modification measures and with different settings for the parameter α in the scoring function. A very low α value (here $\alpha = 10^{-7}$) in combination with an instance-based distance measure means, that the cutpoint is replaced with the alternative cutpoint with the lowest complexity, which implies no reallocation of any instance to another discretization interval. For the basic discretization and for the classification algorithm implementations from the KEEL software suite were used with default parametrization. In particular, the equal-frequency discretization performed a split into 10 intervals. For the resulting discretized data a classifier was learned on the training data and the accuracy was measured in the test data. Summarized results, which are averaged over all datasets, are shown in Tables 1a, 1b and 1c. Exemplary detailed result for all datasets using entropy-based discretization are denoted in Tables 2a and 2b. These tables show the predictive accuracy of the classifier with the discretized attributes as input as well as the averaged complexity score of the smoothed discretized cutpoints. All base discretizers lead to a high complexity of the used cutpoints: the cutpoints are overall hard-to-read for humans. For all discretizers, smoothing these cutpoints even with only low values of α leads to a drastic decrease of the complexity. The complexity of cutpoints is further reduced for increased parameter values of α. These adaptations influence the accuracy of the classifiers only marginally, i.e., the improved classification accuracy of entropy-based discretization is maintained even for substantially simplified cutpoints. This may hint at possible overfitting of the discretization algorithms. Only for the highest settings of α ($\alpha \geq 0.1$) a slight decrease of the accuracy can be observed for the supervised discretization algorithms. The reduction of the accuracy is smaller for instance-based smoothing methods, while similar complexity reductions are achieved. Therefore, this variation is to be preferred based on the current results. These experiments overall demonstrate the effectiveness of our novel approach, as it succeeds in decreasing the complexity of the used cutpoints with only marginal influence on the main algorithm, which uses the discretization intervals.

	distance		instance	
α	comp	acc	comp	acc
0.0	9.293	0.812		
10^{-7}			4.766	0.811
0.01	4.668	0.810	4.350	0.807
0.05	3.810	0.812	3.697	0.813
0.1	3.240	0.811	3.337	0.817
0.3	2.672	0.800	2.665	0.816
0.5	2.509	0.799	2.619	0.813

(a) Entropy-based discretization

	distance		instance	
α	comp	acc	comp	acc
0.0	9.102	0.781		
10^{-7}			4.879	0.780
0.01	4.818	0.784	4.520	0.780
0.05	4.002	0.784	4.017	0.785
0.1	3.488	0.774	3.645	0.788
0.3	2.942	0.768	2.887	0.780
0.5	2.746	0.771	2.841	0.779

(b) Chi2 discretization

	distance		instance	
α	comp	acc	comp	acc
0.0	9.042	0.784		
10^{-7}			4.981	0.784
0.01	5.121	0.784	4.730	0.785
0.05	4.301	0.780	4.266	0.782
0.1	3.912	0.782	3.931	0.780
0.3	3.277	0.780	3.372	0.774
0.5	2.992	0.785	3.298	0.780

(c) Equal-frequency discretization

Table 1: Results for different values for the parameter α in the scoring function and both modification measures. Each table refers to a different discretization technique. For each setting, the prediction accuracy of the classification algorithm and the cutpoint complexity are denoted averaged over all data sets.

α	0.0		0.01		0.05		0.1		0.3		0.5	
	comp	acc	comp	acc	comp	acc	comp	acc	comp	acc	comp	acc
appendicitis	8.729	0.834	5.184	0.852	4.936	0.852	4.436	0.832	3.607	0.868	3.547	0.878
banana	9.331	0.748	4.423	0.752	3.335	0.745	3.207	0.745	2.296	0.745	2.296	0.745
glass	11.973	0.758	4.658	0.735	3.789	0.731	3.173	0.747	2.879	0.700	2.844	0.700
movement	10.978	0.606	5.356	0.567	4.847	0.589	4.534	0.589	4.035	0.586	3.425	0.569
pageblocks	6.359	0.968	4.205	0.966	3.483	0.965	3.071	0.965	2.452	0.963	2.298	0.965
phoneme	9.030	0.812	4.909	0.814	3.771	0.817	3.203	0.813	2.539	0.806	2.353	0.804
segment	11.255	0.939	4.454	0.937	3.424	0.944	3.116	0.938	2.679	0.936	2.556	0.941
sonar	7.794	0.764	5.407	0.759	4.882	0.744	4.533	0.759	3.926	0.700	3.753	0.701
spambase	8.121	0.927	4.202	0.924	2.806	0.920	2.410	0.918	2.139	0.918	2.044	0.920
titanic	11.333	0.771	4.333	0.771	4.000	0.771	1.667	0.771	1.667	0.771	1.667	0.771
vowel	8.814	0.719	4.601	0.736	3.047	0.720	2.600	0.731	1.549	0.687	1.305	0.669
wine	7.801	0.898	4.286	0.904	3.401	0.944	2.935	0.922	2.302	0.915	2.014	0.921

(a) Distance-based modification

α	10^{-7}		0.01		0.05		0.1		0.3		0.5	
	comp	acc	comp	acc	comp	acc	comp	acc	comp	acc	comp	acc
appendicitis	5.250	0.834	5.250	0.834	4.749	0.842	4.447	0.860	3.933	0.859	3.933	0.859
banana	5.186	0.749	4.215	0.751	3.363	0.747	2.662	0.746	2.315	0.746	2.216	0.741
glass	4.814	0.753	4.755	0.753	4.147	0.765	3.840	0.754	2.775	0.756	2.618	0.721
movement	5.572	0.603	5.384	0.564	4.896	0.589	4.606	0.569	4.047	0.614	4.003	0.631
pageblocks	4.243	0.968	3.891	0.968	3.550	0.966	3.259	0.966	2.821	0.965	2.750	0.965
phoneme	5.773	0.813	4.955	0.815	4.041	0.823	3.611	0.829	2.769	0.825	2.723	0.825
segment	5.152	0.939	4.194	0.938	3.372	0.944	3.086	0.942	2.605	0.936	2.558	0.936
sonar	5.718	0.759	5.396	0.735	4.943	0.750	4.633	0.783	4.285	0.778	4.267	0.768
spambase	4.851	0.927	4.450	0.929	3.930	0.928	3.549	0.925	1.238	0.896	1.165	0.890
titanic	0.667	0.771	0.667	0.771	0.667	0.771	0.667	0.771	0.667	0.771	0.667	0.771
vowel	5.461	0.720	4.778	0.725	3.091	0.718	2.549	0.738	2.028	0.718	2.028	0.718
wine	4.507	0.898	4.267	0.898	3.614	0.915	3.137	0.921	2.494	0.933	2.494	0.933

(b) Instance-based modification

Table 2: Detailed results for different values of the parameter α in the scoring function. Each table refers to a different modification measure. For each setting and each data set, the average prediction accuracy of the classification algorithm and the average cutpoint complexity are denoted. As basis entropy-based discretization was used.

5 Conclusions

In this paper we proposed a novel approach on discretization, which aims at cutpoints, which are easy-to-read for human users, e.g., as they require less non-zero digits. For that purpose, a post-processing step is performed after applying an arbitrary conventional discretization method. It trades-off the necessary modifications in comparison to the original discretization scheme with the reduction in complexity. In that direction novel functions for measuring the complexity of a number and for measuring the difference between the original cutpoints and candidates for alternatives have been discussed. Experiments with classification tasks showed, that our approach leads to considerably simpler cutpoints, while the algorithmic performance, i.e., the prediction accuracy, is only marginally influenced.

Since this paper presents work in progress, we plan to extend it in several areas: The proposed function to measure the complexity of a number is currently very simple and could be replaced by a more sophisticated one. This, however, will require an extensive user study to evaluate human perception. Furthermore, we will extend the performed experiments to descriptive data mining tasks such as subgroup discovery. Since the results of these tasks are directly interpreted by domain experts, natural, easy-to-read intervals are especially useful in these areas.

References

[1] Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garcia, S.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Multiple-Valued Logic and Soft Computing 17(2-3), 255–287 (2011)

[2] Blake, C., Merz, C.J.: {UCI} Repository of machine learning databases (1998)

[3] Fayyad, U.M., Irani, K.: Multi-interval discretization of continuous-valued attributes for classification learning (1993)

[4] García, S., Luengo, J., Saez, J., Lopez, V., Herrera, F.: A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering 25(4), 734–750 (2013)

[5] Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques. Morgan kaufmann (2006)

[6] Kerber, R.: Chimerge: Discretization of numeric attributes. In: Proceedings of the tenth national conference on Artificial intelligence. pp. 123–128. AAAI Press (1992)

[7] Kotsiantis, S., Kanellopoulos, D.: Discretization techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering 32(1), 47–58 (2006)

[8] Liu, H., Setiono, R.: Chi2: feature selection and discretization of numeric attributes. Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence pp. 388–391 (1995)

[9] Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)

Evolution of Contacts and Communities in Networks of Face-to-Face Proximity

Extended Abstract*

Mark Kibanov Martin Atzmueller Christoph Scholz Gerd Stumme
Knowledge and Data Engineering Group, University of Kassel
Wilhelmshöher Allee 73, D-34121 Kassel, Germany
Email: {kibanov,atzmueller,scholz,stumme}@cs.uni-kassel.de

Abstract

Communities are a central aspect in the formation of social interaction networks. In this paper, we analyze the evolution of communities in networks of face-to-face proximity. As our application context, we consider four scientific conferences. We compare the basic properties of the contact graphs to describe the properties of the contact networks and analyze the resulting community structure using state-of-the-art automic community detection algorithms. Specifically, we analyze the evolution of contacts and communities over time to consider the stability of the respective communities. In addition, we assess different factors which have an influence on the quality of community prediction. Overall, we provide first important insights into the evolution of contacts and communities in face-to-face contact networks.

1 Introduction

In this paper, we consider the evolution of both contacts and communities at academic conferences. Specifically, we consider the LWA 2010, LWA 2011, LWA 2012 and Hypertext 2011 conferences, where the CONFERATOR[1] system [1] was applied. Using RFID technology, it allows us to collect face-to-face contact data [3], which we can utilize for analyzing contacts and communities.

Our contribution is summarized as follows:

1. We analyze if the structure of the contact graphs is similar for different conferences.
2. We investigate the progress of face-to-face contacts during the respective conferences.
3. We consider automatically detected communities, and analyze the quality of the used algorithms.
4. Finally, we analyze how communities develop over time during a conference and whether detected communities stay stable and thus predictable.

To the best of the authors' knowledge, this is the first time, that these research questions have been addressed in the context of human face-to-face contact networks.

*This extended abstract summarizes the paper [4]: Mark Kibanov, Martin Atzmueller, Christoph Scholz, and Gerd Stumme. On the Evolution of Contacts and Communities in Networks of Face-to-Face Proximity. Proc. IEEE CPSCom 2013, IEEE Computer Society, Boston, MA, USA, 2013

[1] http://www.conferator.org

2 Analysis

In the following, we first briefly describe the utilized dataset, before we summarize the evolution of contacts and communities. For a detailed discussion, we refer to [4].

2.1 Datasets

At the LWA 2010, 2011, 2012 and Hypertext 2011 conferences we asked each participant to wear proximity tags, so they could use the CONFERATOR [1] system. These tags can detect close-range face-to-face proximity (1-1.5 meters) of the participants wearing them [3] - 77 (LWA 2010), 69 (Hypertext), and 42 (LWA 2011 and LWA 2012) for the respective conferences.

2.2 Evolution of Contacts

In summary, the number of edges in contact graph grows nearly linearly during all three LWA conferences. The number of new contacts at the beginning and at the end of these conferences can be explained by the small number of participants who come early or stay longer. An interesting fact for the Hypertext conference is a slow growth of contacts during the second part of the conference. This "tail" is much longer compared to the end of the LWA conferences. We assume that the Hypertext conference has a different "social profile", so the participants are more focused on "socializing" during the first day.

Another important observation shows that graphs with "long" talks ($\geq$ 180 seconds) have almost half of the number of edges of the graphs with all conversations, but their total length is equal to 80% – 90% of the whole length of the whole graph.

2.3 Evolution of Communities

For analyzing the stability of community structure we define a *c-pair* (*Community-pair*) as follows: If two nodes u and v belong to the same community, then $cp = (u, v)$ is a c-pair. CP denotes the set of all possible *c-pairs*. The more c-pairs stay over time, the more stable is a community structure.

To estimate and compare the stability of communities during different conferences, we applied a "simple" predictor $P : I \times J \rightarrow CP$, where $I \subseteq \mathbb{N}, J \subseteq \mathbb{N}$. This predictor assumes that all the c-pairs that were built during (a) reference day(s) in I will be also formed during the subsequent day(s) in J. In the case where I and J contain only single elements, we will drop the set notation for simplicity. Let CP_i be the set of c-pairs of day i: $CP_i = \{(u, v) \mid u, v \in C_j \subseteq V_i\}$, where V_i is the set of the nodes of the contact graph of the day i. We applied

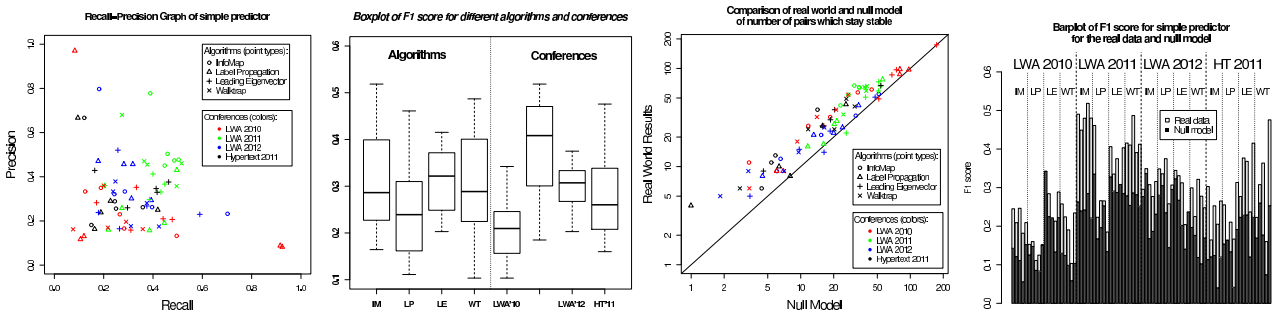


Figure 1: (a) Recall-Precision Graph of the "simple" predictor of the considered algorithms (marked by point types) and conferences (marked by colors). (b) Boxplots of the F1 score of different algorithms (on the left side) and conferences (on the right side). The abbreviations mean: IM – InfoMap, LP – Label Propagation, LE – Leading Eigenvector, WT – Walktrap. (c) Comparison of the real community stability with the null model of the considered algorithms (marked by the different point types) and conferences (marked by different colors): The x-axis contains the respective null model values, the y-axis contains the respective real values. Both axes are scaled logarithmically. (d) Barplots of the F1 scores of the predictions compared to the respective null model for different conferences and different algorithms shown on the top.

the predictor five times for each algorithm and each conference. For computing the 'correct' predictions, we consider the intersection with a subsequent day, and the respective c-pairs. The more c-pairs are predicted correctly, the more stable is the computed community structure.

Figure 1 shows the respective recall and precision values. The larger the value of precision, the more c-pairs from the "training"-day tend to appear also during the "result"-day. The larger the value of recall, the less new c-pairs tend to appear during the "result" day. The type of the point defines the applied algorithm and the color of the point defines the conference. Red circles, for example, show recall and precision of predictions made by the InfoMap algorithm for the LWA 2010 conference. The LWA 2011 data (green points) tend to show a better performance compared to the other conferences and thus we assume the community structure during LWA 2011 is more stable. Similarly, the communities of LWA 2012 are also rather well "predictable". A potential explanation is given by the significant community structure of the four special interest groups constituting the LWA conferences, see [2]. Summarizing both precision and recall, the F1 scores for each applied algorithm and each conference are shown in Figure 1.

The choice of the community detection algorithm did not have a big impact on the performance of our "simple" algorithm and thus on the obtained communities. On the other hand, the choice of the event has a crucial influence on the stability of the communities: The F1 scores confirm the stability of community structure computed for the LWA 2011 conference (green points). The stability of the community structures detected for the LWA 2012 conference show the smallest deviation (The F1 score lies between 0.2 and 0.4).

As another interesting observation, the active communication does not make communities stable – even vice versa. Comparing the LWA 2011 and LWA 2012 conferences with the similar number of participants, we see that the LWA 2012 communications were less active than those at the LWA 2011 in terms of graph density and the total length of communication; overall, we observe more stable communities during LWA 2011. We observed the same phenomenon considering LWA 2010 and HT 2011 – two conferences with the same number of participants but very different dynamics of face-to-face communications. On hypothesis for explaining the negative correlation of community stability and communication is the following: The participants stick to the known persons and tend to have less contacts with new persons which implies both lack of new contacts and stability of the existing communities over the whole conference.

So far, our proposed measures compare the overall stability of communities of different conferences. However, in order to clarify that these stabilities are significant and not accidental, we apply a null model NM computed using the following formula: $NM = 2 \times \frac{CP_t}{n \times (n-1)} \times CP_{t+1}$, where CP_i is the number of c-pairs at day i, and n is the number of nodes in the considered graph. As shown in Figure 1, the majority of points lies above the null model line which means the stability of communities is not a random phenomenon. Some of the results obtained using the $LeadingEigenvector$ algorithm lie below the null model line, while some of the $LabelPropagation$ measurements are just placed on the line. These findings would seem to show some randomness of the stability of community structures computed with these algorithms. In order to characterize the stability further, we compare the F1 score of the real data and the null model (see Figure 1). On average the real world F1 score is 1.65 times larger than the obtained null model F1 score. This shows, that persons tend to stay in the same communities over one conference; the choice of algorithm also does not affect this.

References

[1] M. Atzmueller, D. Benz, S. Doerfel, A. Hotho, R. Jäschke, B. E. Macek, F. Mitzlaff, C. Scholz, and G. Stumme. Enhancing Social Interactions at Conferences. *it*, 53(3):101–107, 2011.

[2] M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme. Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles. In *Modeling and Mining Ubiquitous Social Media*, volume 7472 of *LNAI*. Springer Verlag, Heidelberg, Germany, 2012.

[3] A. Barrat, C. Cattuto, V. Colizza, J.-F. Pinton, W. V. den Broeck, and A. Vespignani. High Resolution Dynamical Mapping of Social Interactions with Active RFID. *PLoS ONE*, 5(7), 2010.

[4] M. Kibanov, M. Atzmueller, C. Scholz, and G. Stumme. On the Evolution of Contacts and Communities in Networks of Face-to-Face Proximity. In *Proc. IEEE CPSCom 2013*, Boston, MA, USA, 2013.

On the Semantics of User Interaction in Social Media

(Extended Abstract*)

Folke Mitzlaff[1], Martin Atzmueller[1], Gerd Stumme[1] and Andreas Hotho[2]

[1]Knowledge and Data Engineering Group, University of Kassel
[2]Data Mining and Information Retrieval Group, University of Würzburg

Abstract

In ubiquitous and social web applications, there are different user traces, for example, produced explicitly by "tweeting" via twitter or implicitly, when the corresponding activities are logged within the application's internal databases and log files. Each set of user interactions can then be mapped to a network, with links between users according to their observed interactions.
In this paper, we analyze correlations between different interaction networks. We collect for every user certain external properties which are independent of the given network structure. Based on these properties, we then calculate semantically grounded reference relations among users and present a framework for capturing semantics of user relations. The experiments are performed using different interaction networks from the twitter, flickr and BibSonomy systems.

1 Introduction

By interacting with social and ubiquitous systems, the user is leaving traces within the different databases and log files, e. g., by updating the current status via twitter or chatting with social acquaintances via facebook. Ultimately, each type of such traces gives rise to a corresponding network of user relatedness, where users are connected if they interacted either explicitly (e. g., by establishing a "friendship" link within in an online social network) or implicitly (e. g., by visiting a user's profile page). We consider a link within such a network as evidence for user relatedness and call it accordingly *evidence network* or *interaction network*. These interaction networks are of large interest for many applications, such as recommending contacts in online social networks or for identifying groups of related users [8]. Nevertheless, it is not clear, whether every such interaction network captures meaningful notions of relatedness and what the semantics of different aggregation levels really are. As multifaceted as humans are, as many reasons for individuals being related exists. Ultimately, it is therefore not possible to judge whether an interaction network is "meaningful" or not. Nevertheless, certain networks are more probable than others and give rise to more traceable notions of relatedness.

*This extended abstract summarizes the paper [9]: Folke Mitzlaff, Martin Atzmueller, Gerd Stumme, and Andreas Hotho. Semantics of User Interaction in Social Media. In Gourab Ghoshal, Julia Poncela-Casasnovas, and Robert Tolksdorf (Eds.), Complex Networks IV, Springer Verlag, Heidelberg, Germany, 2013.

2 Experiments and Results

This paper summarizes work presented in [9], focussing on an experimental methodology for assessing the semantics of evidence networks and similarity metrics therein. The methodology is applied to a broad range of evidence networks. The obtained results thus yield a *semantic grounding* of evidence networks and similarity metrics, which are merely based on structural properties of the networks. Furthermore, we consider both established reference sources such as tagging data, as well as geographical locational data as a proxy for semantic relatednesss.

Evidence Networks in BibSonomy Beside explicit relations among users, i. e., the "*friends*" in BibSonomy, different relations are established implicitly by user interactions, e. g., when user u looks at user v's resources. In particular, we considered the directed *Friend-Graph*, containing an edge (u, v) iff user u has added user v as a friend, the directed *Copy-Graph* which contains an edge (u, v) with weight $c \in \mathbb{N}$, iff user u has copied c resources, i. e., a publication reference from user v and the directed *Visit-Graph*, containing an edge (u, v) with label $c \in \mathbb{N}$ iff user u has navigated c times to the user page of user v.

Evidence Networks in twitter Each user publishes short text messages ("*tweets*") which may contain freely chosen *hashtags*, i. e., distinguished words being used for marking keywords or topics. Furthermore, users may "cite" each other by "retweeting": A user u retweets user v's content, if u publishes a text message containing "RT @v:" followed by (an excerpt of) v's corresponding tweet. Users may also explicitly follow other user's tweets by establishing a corresponding friendship-like link. For analysis, we considered the directed *Follower-Graph*, containing an edge (u, v) iff user u follows the tweets of user v and the *ReTweet-Graph*, containing an edge (u, v) with label $c \in \mathbb{N}$ iff user u cited (or "retweeted") exactly c of user v's tweets.

Evidence Networks in flickr In flickr, users mainly upload images and assign arbitrary tags but also interact, e. g., by establishing contacts or commenting on other users images. For our analysis we extracted the directed *Contact-Graph*, containing an edge (u, v) iff user u added user v to its personal contact list, the directed *Favorite-Graph*, containing an edge (u, v) with label $c \in \mathbb{N}$ iff user u added exactly c of v's images to its personal list of favorite images as well as the directed *Comment-Graph*, containing edge (u, v) with label $c \in \mathbb{N}$ iff user u posted exactly c comments on v's images.

Table 1: High level statistics for all networks with density d, the number of strongly connected components #scc and the size of the largest strongly connected component SCC.

	$\lvert V_i\rvert$	$\lvert E_i\rvert$	d	#scc	SCC
Copy	$1,427$	$4,144$	$2 \cdot 10^{-3}$	$1,108$	309
Visit	$3,381$	$8,214$	10^{-3}	$2,599$	717
Friend	700	$1,012$	$2 \cdot 10^{-3}$	515	17
ReTweet	$826,104$	$2,286,416$	$3,4 \cdot 10^{-6}$	$699,067$	$123,055$
Follower	$1,486,403$	$72,590,619$	$3,3 \cdot 10^{-5}$	$198,883$	$1,284,201$
Comment	$525,902$	$3,817,626$	$1,4 \cdot 10^{-5}$	$472,232$	$53,359$
Favorite	$1,381,812$	$20,206,779$	$1,1 \cdot 10^{-5}$	$1,305,350$	$76,423$
Contact	$5,542,705$	$119,061,843$	$3,9 \cdot 10^{-6}$	$4,820,219$	$722,327$

General Structural Properties Table 1 summarizes major graph level statistics for the considered networks which range in size from thousands of edges (e. g., the Friend-Graph) to more than one hundred million edges (flickr's Contact-Graph). All networks obtained from BibSonomy are complete and therefore not biased by a previous crawling process. In return, effects induced by limited network sizes have to be considered.

3 Analysis of Network Semantics

In the following, we tackle the problem of assessing the "meaning" of relations among pairs of vertices within such a network. This analysis then gives insights into the question, whether and to which extent the networks give rise to a common notion of *semantic relatedness* among the contained vertices. For this, we apply an experimental methodology, which was previously used for assessing semantical relationships within co-occurrence networks [10]. The basic idea is simple: We consider well founded notions of relatedness, which are naturally induced by external properties of the corresponding vertex sets, as, e. g., similarity of the applied tag assignments in BibSonomy or geographical distance between users in twitter. We then compute for each pair of vertices within a network these "semantic" similarity metrics and correlate them with different measures of structural similarity in the considered network.

3.1 Vertex Similarities

Below, we apply two well-established similarity functions in corresponding unweighted variants, namely the cosine similarity COS and the Jaccard Index JC as well as the corresponding weighted variants $\widetilde{COS}$ and $\widetilde{JC}$, following the presentation in [2]. Additionally we apply a modification of the *preferential PageRank* which we adopted from our previous work on folksonomies [3]: For a column stochastic adjacency matrix A and damping factor α, the *global* PageRank vector $\vec{w}$ with uniform *preference vector* $\vec{p}$ is given as the fixpoint of $\vec{w} = \alpha A\vec{w} + (1-\alpha)\vec{p}$. In case of the *preferential PageRank* for a given node i, only the corresponding component of the preference vector is set. For vertices x, y we set accordingly $PPR(x,y) := \vec{w}_{(x)}[y]$, that is, we compute the preferential PageRank vector $\vec{w}_{(x)}$ for node x and take its y'th component. We calculate the adopted preferential PageRank score by subtracting the global PageRank score PR from the preferential PageRank score in order to reduce frequency effects and set

$$PPR\text{+}(x,y) := PPR(x,y) - PR(x,y).$$

3.2 Semantic Reference Relations

For assessing the semantic similarity of two nodes within a network, we consider the similarity of users based on the applied tags or hashtags, respectively, and the geographical distance of users in twitter and flickr.

Tag Similarity In the context of social tagging systems like BibSonomy, the cosine similarity is often used for measuring semantic relatedness (see, e. g., [1]).

We compute the cosine similarity in the vector space $\mathbb{R}^T$, where, for user u, the entries of the vector $(u_1, \ldots, u_T) \in \mathbb{R}^T$ are defined by $u_t := w(u,t)$ for tags t where $w(u,t)$ is the number of times user u has used tag t to tag one of her resources (in case of BibSonomy and flickr) or the number of times user u has used hash tag t in one of her tweets.

Geographical Distance In twitter and flickr, users may provide an arbitrary text for describing his or her location. Accordingly, these location strings may either denote a place by its geographic coordinates, a semi structured place name (e. g., "San Francisco, US"), a colloquial place name (e. g., "Motor City" for Detroit) or just a fantasy name. Also the inherent ambiguity of place names (consider, e. g., "Springfield, US") renders the task of *exactly* determining the place of a user impossible. Nevertheless, by applying best matching approaches, we assume that geographic locations can be determined up to a given uncertainty and that significant tendencies can be observed by averaging over many observations.

We used Yahoo!'s Placemaker™ API for matching user provided location strings to geographic locations with automatic place disambiguation. In case of flickr, we obtained geographic locations for $320,849$ users and in case of twitter for $294,668$ users. Geographical distance of users is then simply given by the distance of the centroids for the correspondingly matched places.

3.3 Grounding of Shortest Path Distance

For analyzing the interdependence of *semantic* and *structural* similarity between users, we firstly consider a very basic measure of structural relatedness between two nodes in a network, namely their respective shortest path distance. We ask, whether users which are direct neighbors in an evidence network tend to be more similar than distant users. That is, for every shortest path distance d and every pair of nodes u, v with a shortest path distance d, we calculated the average corresponding similarity scores $COS(u,v)$, $JC(u,v)$, $PPR(u,v)$ with variants and geographic distance. To rule out statistical effects, we repeated for each network G the same calculations on shuffled null model graphs.

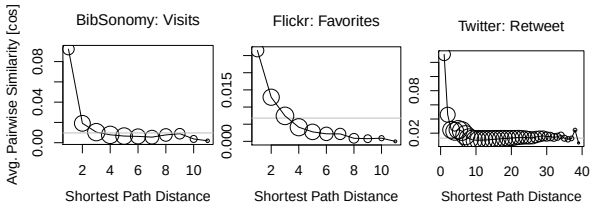


Figure 1: Average pairwise cosine similarity based on the users' tag assignments relative to the shortest path distance in the respective networks where the global average is depicted in gray and the point size scales logarithmically with the number of pairs.

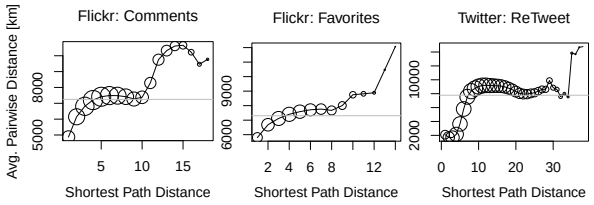


Figure 2: Shortest path distance vs. average pairwise geographic distance in flickr. The global average is depicted in gray and the point size scales logarithmically with the number of pairs.

Semantic Similarity Figure 1 shows the resulting plots for each considered network separately. Though the obtained average similarity scores vary greatly in magnitude for different networks (e. g., a maximum of 0.22 for the Friend-Graph in BibSonomy compared to a maximum of 0.1 for the Visit-Graph), they also share a common pattern: Direct neighbors are in average significantly more similar than distant pairs of users. And with a distance of two to three, users tend to be less similar than in average (in case of the ReTweet graph, users are more similar than in average up to a distance of eight). For the Visit-Graph, the Comment-Graph, the Follower-Graph and the ReTweet graph, the average similarity scores approach the global average similarity again. For distances around a network's diameter, the number of observations is too small, resulting in less pronounced tendencies for very distant nodes.

Geographic Distance For average geographic distances of users in flickr and twitter, we repeated the same calculations, as depicted in Figure 2. Firstly, we note the overall tendency, that direct neighbors tend to be located more closely than distant pairs of users within a network. Additionally, the average geographic distance of users then approaches the global average, and increases again after a certain plateau. As for the ReTweet-Graph, the average geographic distance remains at the global average level, once reached at a shortest path distance of ten.

Discussion It is worth emphasizing, that in all considered evidence networks, the relative position of users already gives rise to a semantically grounded notion of relatedness, even in case of implicit networks, which are merely aggregated from usage logs as, e. g., the Visit-Graph. But one has to keep in mind that all observed tendencies are the result of averaging over a very large number of observations (e. g., $34,282,803,978$ pairs of nodes at distance four in the Follower-Graph). Therefore, we cannot deduce geographic proximity from topological proximity for a given pair of users, as even direct neighbors in the Follower-Graph are in average located $4,000$ kilometers apart from each other. But the proposed analysis aims at revealing semantic tendencies within a network and for comparing different networks (e. g., the Retweet-Graph better captures geographic proximity of direct neighbors in the graph).

3.4 Grounding of Structural Similarity

We now turn our focus towards different measures of structural similarity for nodes within a given network. There is a broad literature on such similarity metrics for various applications, such as link prediction [7] and distributional semantics [4; 10]. We thus extend the question under consideration in Section 3.3, and ask, which measure of structural similarity best captures a given semantically grounded notion of relatedness among users. In the scope of the present work, we consider the cosine similarity and Jaccard index, which are based only on the direct neighborhood of a node as well as the (adjusted) preferential PageRank similarity which is based on the whole graph structure (refer to Section 3.1 for details).

Ultimately, we want to visualize correlations among structural similarity in a network and semantic similarity, based on external properties of nodes within it. We consider, again, semantical similarity based on users' tag assignments in BibSonomy, flickr and hash tag usage in twitter as well as geographic distance of users in flickr and twitter. In detail: For a given network $G = (V, E)$ and structural similarity metric S, we calculate for every pair of vertices $u, v \in V$ their structural similarity $S(u, v)$ in G as well as their semantic similarity and geographic distance. For visualizing correlations, we create plots with structural similarity at the x-axis and semantic similarity at the y-axis. As plotting the raw data points is computationally infeasible (in case of the Contact-Graph $30,721,580,000,000$ data points), we binned the x-axis and calculated average semantical similarity scores per bin. As the distribution of structural similarity scores is highly skewed towards lower similarity scores (most pairs of nodes have very low similarity scores), we applied logarithmic binning, that is, for a structural similarity score $x \in [0, 1]$ we determined the corresponding bin via $\lfloor \log(x \cdot b^N) \rfloor$ for given number of bins N and suitable base b. Pragmatically, we determined the base relative to the machine's floating point precision ϵ resulting in $b := \epsilon^{\frac{-1}{N}}$.

Semantic Similarity Figure 3 shows the obtained results for each considered network separately. We firstly note, that the cosine similarity metric and the Jaccard index are highly correlated. Secondly, the adjusted preferential PageRank similarity consistently outperforms the other similarity metrics with respect to magnitude and monotonicity (except for BibSonomy's Friend-Graph and flickr's Contact-Graph).

Geographic Distance As for geographic distances, Figure 4 shows the observed correlations for structural similarity in the different evidence networks and the corresponding average pairwise distance. In all but flickr's Favorite-Graph, for both local neighborhood based similarity metrics COS and JC, the average distance first decreases, but then increases again. This behavior is most pronounced in twitter's ReTweet-Graph. In the Favorite-Graph, both COS and JC monotonically decrease with increasing similarity score. On the other hand, the average distance decreases monotonically with increasing preferential PageRank score PPR consistently in all considered networks, ex-

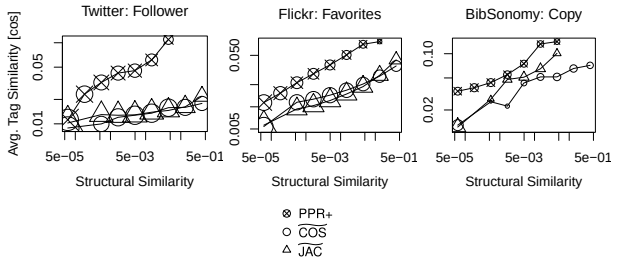


Figure 3: Average pairwise semantic similarity based on tags users assigned to resources in BibSonomy and flickr or hash tag usage in twitter, relative to different structural similarity scores in the corresponding networks. The point size scales logarithmically with the number of pairs.

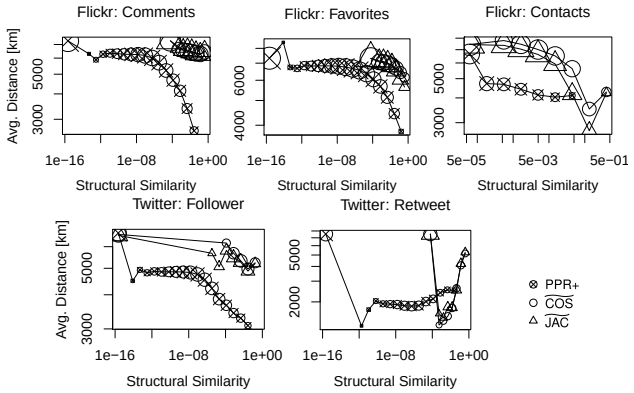


Figure 4: Average pairwise distance relative to different structural similarity scores in the corresponding networks. The point size scales logarithmically with the number of pairs.

cept the ReTweet-Graph, where the average distance stays at a level of around 2.000 kilometers for similarity scores > 0. Generally (except for the ReTweet-Graph), it yields average distance values which are magnitudes below those obtained via the local similarity metrics.

Discussion Again, the obtained results only point at tendencies of the considered similarity metrics in capturing geographic proximity by means of structural similarity. Nevertheless, the adjusted preferential PageRank similarity consistently outperforms the other considered metrics. We therefore conclude that from all considered similarity metrics, the adjusted preferential PageRank similarity best captures the notion of geographic proximity. This is especially of interest, as the geographic proximity is a prior for many properties users may have in common, such as, e. g., language, cultural background or habits. twitter's ReTweet-Graph seems to encompass the strongest geographic binding, as indicated in the relative low average distance for direct neighbors (cf. Figure 2 and the overall low average distance for higher preferential PageRank similarity scores (cf. Figure 4). Of course, other established similarity metrics (e. g., [6; 5; 4]) can be applied as well and are the subject of future considerations.

4 Conclusion & Future Work

With the present work, we introduced an experimental framework for assessing the semantics of social networks. The proposed methodology has a broad range of applications, such as *user recommendation* or *community mining* tasks, as it allows semantically grounded pre processing of given networks (e. g., merging different small networks, scaling edge weights, selecting certain groups of users or directedness of networks). The conducted experiments give insights into the semantics of evidence networks from flickr, twitter and BibSonomy and well known similarity metrics.

Ultimately, the proposed experimental setup allows to formulate the assessment of semantic user relatedness as a regression task, which will be subject to future work.

References

[1] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *The Semantic Web – Proc. ISWC 2008*, volume 5318 of *LNAI*, pages 615–631, Heidelberg, 2008. Springer.

[2] H. de Sá and R. Prudencio. Supervised link prediction in weighted networks. In *Neural Networks (IJCNN), The 2011 Int. Joint Conference on*, pages 2281–2288. IEEE, 2011.

[3] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information Retrieval in Folksonomies: Search and Ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426, Heidelberg, June 2006. Springer.

[4] A. Islam and D. Inkpen. Second Order Co-Occurrence PMI for Determining the Semantic Similarity of Words. In *Proc. of the Int. Conference on Language Resources and Evaluation (LREC 2006)*, pages 1033–1038, 2006.

[5] G. Jeh and J. Widom. SimRank: A Measure of Structural-Context Similarity. In *Proc. of the eighth ACM SIGKDD int. conference on Knowledge discovery and data mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM.

[6] E. A. Leicht, P. Holme, and M. E. J. Newman. Vertex Similarity in Networks, 2005. cite arxiv:physics/0510143.

[7] D. Liben-Nowell and J. Kleinberg. The Link-Prediction Problem for Social Networks. *J. of the American society for inf. science and technology*, 58(7):1019–1031, 2007.

[8] F. Mitzlaff, M. Atzmueller, D. Benz, A. Hotho, and G. Stumme. Community Assessment using Evidence Networks. In *Analysis of Social Media and Ubiquitous Data*, volume 6904 of *LNAI*, 2011.

[9] F. Mitzlaff, M. Atzmueller, G. Stumme, and A. Hotho. Semantics of User Interaction in Social Media. In G. Ghoshal, J. Poncela-Casasnovas, and R. Tolksdorf, editors, *Complex Networks IV*, volume 476 of *Studies in Computational Intelligence*. 2013.

[10] F. Mitzlaff and G. Stumme. Relatedness of Given Names. *Human Journal*, 1(4):205–217, 2012.

Learning Shortest Paths in Word Graphs*

Emmanouil Tzouridis and **Ulf Brefeld**†
Technische Universität Darmstadt,
Hochschulstr. 10, 64289 Darmstadt, Germany
{tzouridis,brefeld}@kma.informatik.tu-darmstadt.de

Abstract

In this paper we briefly sketch our work on text summarisation using compression graphs. The task is described as follows: Given a set of related sentences describing the same event, we aim at generating a single sentence that is simply structured, easily understandable, and minimal in terms of the number of words/tokens. Traditionally, sentence compression deals with finding the shortest path in word graphs in an unsupervised setting. The major drawback of this approach is the use of manually crafted heuristics for edge weights. By contrast, we cast sentence compression as a structured prediction problem. Edges of the compression graph are represented by features drawn from adjacent nodes so that corresponding weights are learned by a generalised linear model. Decoding is performed in polynomial time by a generalised shortest path algorithm using loss augmented inference. We report on preliminary results on artificial and real world data.

1 Introduction

In this paper we study the intelligent summarisation of related sentences to quickly serve information needs of users. Given a collection of sentences dealing with the same real-world event, we aim at generating a single sentence that is (i) a summarisation of the input sentences, (ii) simply structured and easily understandable, and (iii) minimal in terms of the number of words/tokens. The input sentences are represented as a word graph [Filippova, 2010], where words are identified with nodes and directed edges connect adjacent words in at least one sentence, and the output summary is thus a path in the graph fulfilling conditions (i-iii). In this paper, we cast sentence compression as learning a mapping from word graphs to their shortest paths. Edges of the graphs are labeled with costs and the shortest path realises the lowest possible costs from a start to an end node.

Learning mappings between arbitrary structured and interdependent input and output spaces challenges the standard model of learning a mapping from independently drawn instances to a small set of labels. For capturing the involved dependencies it is helpful to represent inputs $\mathbf{x} \in \mathcal{X}$ and outputs $\mathbf{y} \in \mathcal{Y}$ in a joint feature representation. The standard approach to learn to predict structured outputs seeks the most likely output structure given an input. By contrast, we aim at finding the shortest and thus minimal path that leads from a start to an end node of the compression graph. Therefore the task is rephrased as finding a function $f : \mathcal{X} \times \mathcal{Y} \to \Re$ such that

$$\hat{\mathbf{y}} = \operatorname*{argmin}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \tag{1}$$

is the desired output for any input $\mathbf{x}$ [Tsochantaridis *et al.*, 2005; Taskar *et al.*, 2004]. The function f is a linear model in a joint space $\Phi(\mathbf{x}, \mathbf{y})$ of input and output variables and the computation of the argmin is performed by an appropriate decoding strategy such as a shortest-path algorithm.

The remainder is organised as follows. Section 2 introduces preliminaries. Our main contribution is presented in 3. We briefly discuss empirical results in Section 4 and Section 5 concludes.

2 Preliminaries

2.1 Related Work

Barzilay and Lee [Barzilay and Lee, 2003] study sentence compression using dependency trees. Aligned trees are represented by a lattice from which a compression sentence is extracted by an entropy-based criterion over all possible traversals of the lattice. Wan et al. at [Wan *et al.*, 2007] use a language model in combination with maximum spanning trees to rank candidate aggregations that satisfy grammatical constrains.

While the previous approaches to multi-sentence compression are based on syntactic parsing of the sentences, word graph approaches have been proposed, that do not make use of dependency trees or other linguistic concepts. Filippova [Filippova, 2010] casts the problem as finding the shortest path in directed word graphs, where each node is a unique word and directed edges represent the word ordering in the original sentences. The costs of these edges are given by a heuristic that is based on word frequencies. Recently, Boudin and Morin [Boudin and Morin, 2013] propose a re-ranking scheme to identify summarising sentences that contain many keyphrases. The underlying idea is that representative key phrases for a given topic give rise to more informative aggregations.

2.2 Word Graphs and Shortest Paths

Word graphs intend to build a non-redundant representation for possibly redundant sequences by merging identical observations. From a collection of related sentences we iteratively construct a word graph by adding sentences one-by-one as follows: We begin with an empty graph and add the first sentence, where every word in the sentence becomes a

*This paper is a short version of [Tzouridis and Brefeld, 2013].

†UB is also affiliated with the German Institute for Educational Research (DIPF), Frankfurt am Main, Germany.

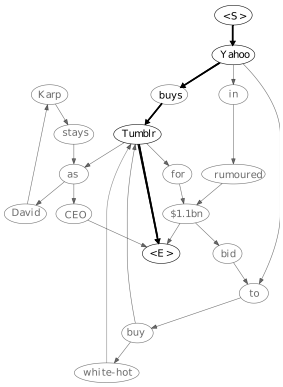


Figure 1: The word graph constructed from the sentences: "Yahoo in rumoured $1.1bn bid to buy white-hot Tumblr", "Yahoo buys Tumblr as David Karp stays as CEO", "Yahoo to buy Tumblr for $1.1bn". The corresponding shortest path is highlighted.

node and a directed edge connects nodes of adjacent words. Words from the next sentences are incorporated by creating a new node for the word or by mapping the word to the corresponding already existing node. A directed edge is inserted to connect the word to its predecessor. We continue until all sentences are incorporated.

Auxiliary words indicating the start (e.g, x_s) and the end (e.g., x_e) of the sentence are added to the sentences. The sketched procedure merges identical words but preserves the structure of the sentences along the contained paths and the original sentences can often be reconstructed from the compressed representation. Fig. 1 shows related sentences and the corresponding word graph.

The described construction gives us a directed weighted graph $\mathbf{x} = (N, E)$, where N is the set of nodes and E the set of edges. As every word graph $\mathbf{x}$ also defines the sets N and E, we will use $N(\mathbf{x})$ and $E(\mathbf{x})$ in the remainder to denote the set of nodes and edges of graph $\mathbf{x}$, respectively. Every edge $(x_i, x_j) \in E(\mathbf{x})$ is assigned a positive weight given by a cost function $cost : (x_i, x_j) \mapsto \Re^+$. A path $\mathbf{y}$ in the graph $\mathbf{x}$ is a sequence of connected nodes of $\mathbf{x}$ and the cost of such a path is given by the sum of the edge costs for every edge that is on the path. Given the word graph $\mathbf{x}$, the shortest path problem is finding the path in $\mathbf{x}$ from x_s to x_e with the lowest costs,

$$\operatorname*{argmin}_{\mathbf{y}} \sum_{(x_i,x_y)\in N(\mathbf{x})} y_{ij} cost(x_i, x_{i+1}) \text{ s.t. } \mathbf{y} \in path(x_s, x_e).$$

There exist many algorithms for computing shortest paths efficiently [Bellman, 1958; Ford, 1956; Dijkstra, 1959]. Usually, these methods are based on relaxation integer programming, where an approximation of the exact quantity is iteratively updated until it converges to the correct solution. Figure 1 shows an example that visualises the shortest path for a compression graph.

3 Learning the Shortest Path

3.1 Representation

To learn the shortest path, we need to draw features from adjacent nodes in the word graph to learn the score of the connecting edge. Let x_i and x_j be connected nodes of the compression graph $\mathbf{x}$, that is $x_i, x_j \in N(\mathbf{x})$ and $(x_i, x_j) \in E(\mathbf{x})$. We represent the edge between x_i and x_j by a feature vector $\phi(x_i, x_j)$. A path in the graph is represented as an $n \times n$ binary matrix $\mathbf{y}$ with $n = |N(\mathbf{x})|$ and elements $\{\mathbf{y}_{ij}\}$ given by $y_{ij} = [[(x_i, x_j) \in path]]$ where $[[z]]$ is the indicator function returning one if z is true and zero otherwise. The cost of using the edge (x_i, x_j) in a path is given by a linear combination of those features which is parameterised by $\mathbf{w}$,

$$cost(x_i, x_j) = \mathbf{w}^\top \phi(x_i, x_j).$$

Replacing the constant costs by the parameterised ones, we arrive at the following objective function (ignoring the constraints for a moment) that can be rewritten as a generalised linear model.

$$\sum_{(x_i,x_j)\in E(\mathbf{x})} y_{ij}\, \mathbf{w}^\top \phi(x_i, x_j) = \mathbf{w}^\top \Phi(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y})$$

Given a word graph $\mathbf{x}$, the shortest path $\hat{\mathbf{y}}$ for a fixed parameter vector $\mathbf{w}$ can now be computed by

$$\hat{\mathbf{y}} = \operatorname*{argmin}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}),$$

where f is exactly the objective of the shortest path algorithm and the argmin consequently computed by an appropriate solver, such as Yen's algorithm [Yen, 1971].

3.2 Learning Shortest Paths with SVMs

In our setting, word graphs $\mathbf{x} \in \mathcal{X}$ and the best summarising sentence $\mathbf{y} \in \mathcal{Y}$ are represented jointly by a feature map $\Phi(\mathbf{x}, \mathbf{y})$ that allows to capture multiple-way dependencies between inputs and outputs. We apply a generalised linear model $f(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top \Phi(\mathbf{x}, \mathbf{y})$ to decode the shortest path

$$\hat{\mathbf{y}} = \operatorname*{argmin}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}),$$

where the quality of f is measured by the Hamming loss

$$\Delta_H(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2} \sum_{(x_i,x_j)\in E(\mathbf{x})} [[y_{ij} \neq \hat{y}_{ij}]]$$

that details the differences between the true $\mathbf{y}$ and the prediction $\hat{\mathbf{y}}$, where $[[\cdot]]$ is again the indicator function from Section 3.1. Using the loss Δ_H, structural support vector machines [Tsochantaridis *et al.*, 2005] minimise the regularised empirical risk

$$\hat{R}[f] = \|f\|^2 + \sum_{i=1}^{m} \Delta_H \left(\mathbf{y}, \operatorname*{argmin}_{\bar{\mathbf{y}}} f(\mathbf{x}, \bar{\mathbf{y}}) \right).$$

It is often beneficial to rescale the induced margin by the loss to implement the intuition that the confidence of rejecting a mistaken output is proportional to its error. Combining everything, we arrive at the following optimisation problem

$$\begin{aligned} \min_{f,\boldsymbol{\xi}} \quad & \|f\|^2 + \sum_{i=1}^{m} \xi_i \\ s.t. \quad & \forall i \forall \bar{\mathbf{y}} \neq \mathbf{y}_i : f(\mathbf{x}_i, \bar{\mathbf{y}}) - f(\mathbf{x}_i, \mathbf{y}_i) \geq \Delta_H(\mathbf{y}_i, \bar{\mathbf{y}}) - \xi_i \\ & \forall i : \xi_i \geq 0 \end{aligned}$$

which can be solved in polynomial time by cutting planes. The idea behind cutting planes is to instantiate only a minimal subset of the exponentially many constrains. This is achieved by decoding for every training sample the shortest path using our current model, if this is not the correct path, then it is added to the constrains and the model is updated. If the decoded path is the correct one, we need to

decode the second best path to verify wether the associated margin constraint is fulfilled; if not, the pair is added to the constraints and the model is updated accordingly. Luckily, we do not need to rely on an expensive two-best shortest path algorithm but can compute the most strongly violated constraint directly via the cost function

$$Q(\bar{\mathbf{y}}) = \Delta_H(\mathbf{y}_i, \bar{\mathbf{y}}) - \mathbf{w}^\top \Phi(\mathbf{x}_i, \bar{\mathbf{y}}) + \mathbf{w}^\top \Phi(\mathbf{x}_i, \mathbf{y}_i) \quad (2)$$

that has to be maximised wrt $\mathbf{y}$. The following proposition shows that we can equivalently solve a shortest path problem for finding the maximiser of Q.

Proposition 1 (Loss augmented inference for shortest path problems)**.** *The maximum* $\mathbf{y}^*$ *of* Q *in Equation* (2) *can be equivalently computed by minimising a shortest path problem with cost* $(x_i, x_j) = y_{ij} + \mathbf{w}^\top \phi(x_i, x_j)$.

Proof. Omitted for lack of space. □

Given a parameter vector $\mathbf{w}$ and start and end nodes x_s and x_e, respectively, the optimisation of Q can be performed with the following linear program.

$$\begin{aligned}
\min_{\bar{\mathbf{y}}} \quad & \sum_{ij} \left(y_{ij} + \mathbf{w}^\top \phi(x_i, x_j)\right) \bar{y}_{ij} \\
s.t. \quad & \forall k \in N(\mathbf{x})/\{s,t\} : \sum_j \bar{y}_{kj} - \sum_i \bar{y}_{ik} = 0 \\
& \sum_j \bar{y}_{sj} - \sum_i \bar{y}_{is} = 1 \\
& \sum_i \bar{y}_{ie} - \sum_j \bar{y}_{ej} = 1 \\
& \forall (i,j) : \; y_{ij} \leq x_{(i,j)} \quad \wedge \quad \forall (i,j) : \mathbf{y}_{ij} \in \{0,1\}
\end{aligned}$$

The first constraint guarantees that every inner node of the path must have as many incoming as outgoing edges, the second line of constraints guarantees the path to start in x_s and, analogously, the third line ensures that it terminates in x_e. The last line of constraints forces the edges of the path $\bar{\mathbf{y}}$ to move along existing paths of $\mathbf{x}$.

4 Empirical Results

In this section, we empirically compare learning shortest paths to traditional unsupervised approaches. To this end, we also deploy a structural perceptron [Collins and Duffy, 2002; Altun *et al.*, 2003] as a special-case of the presented large-margin approach.

4.1 Artificial Data

We generate artificial graphs with $|N| \in \{10, 20, 30, 40\}$ nodes as follows. For every node in a graph, we sample the number of outgoing edges uniformly in the interval $[\frac{|N|}{2}, |N|]$, and for every edge, a receiving node is sampled uniformly from the remaining nodes. To annotate the optimal path we first draw its length uniformly in the interval $[\frac{|N|}{2}, |N|]$ and randomly select the respective number of nodes from N, while enforcing that every edge in the path is included in the graph as well.

To ensure that the optimal path is actually the one with lowest costs, edge features are sampled from a one-dimensional Gaussian mixture distribution, where the generating component is chosen according to whether the respective edge lies on the the shortest path or not. That is, we introduce two Gaussian components $G_{1,2}$, so that costs for edges lying on the shortest path are drawn from $G_1(\mu_1, \sigma_1^2)$ while costs for all other edges are sampled from $G_2(\mu_2, \sigma_2^2)$.

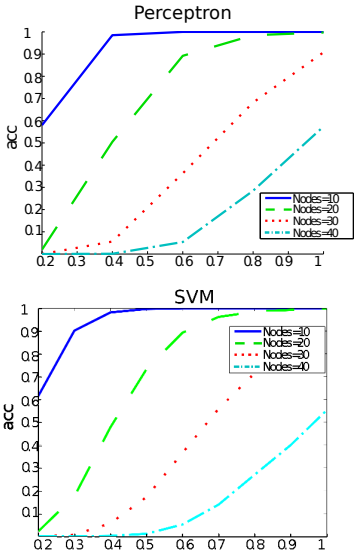


Figure 2: Performance on artificial data for perceptrons (top) and SVMs (bottom).

The difficulty of the experimental setup is controlled by a parameter α that measures the distance of the two means, i.e., $\alpha = |\mu_1 - \mu_2|$. We sample the means from the following normal distributions $\mu_1 \sim G(-\frac{\alpha}{2}, 0.1)$ and $\mu_2 \sim G(\frac{\alpha}{2}, 0.1)$. We use $\sigma_1 = \sigma_2 = 0.01$ and report on averages over 100 repetitions. The results for perceptrons and SVMs are shown in Figure 2. The distance α is depicted on the x-axis. The y-axis shows the top-one accuracy. The performance of both algorithms highly depends on the distance of the cost-generating components and the size of the graph. Both algorithms perform similarly.

4.2 News Headlines

The real-world data originates from titles of crawled news articles from several web sites on different days. We use categories *Technology, Sports, Business* and *General*. Related sets with more than 4 news headlines are manually identified and grouped together, and word graphs are built according to the procedure described in Section 2.2. Ground truth is annotated manually by selecting the best sentence among the 20 shortest paths computed by Yen's algorithm [Yen, 1971] using frequencies as edge weights. This process leaves us with 87 training examples.

We intend to learn the costs for the edges that give rise to the optimal compression of the training graphs and compare our algorithms to the unsupervised approach presented in [Filippova, 2010] that uses $(\#(x_1) + \#(x_2))/\#(x_1, x_2)$ as edge weights. We devise two different sets of features. The first feature representation consists of only two features that are inspired by the heuristic. That is, for an edge (x_i, x_j), we use

$$\phi_1(x_i, x_j) = \left(\frac{\#(x_1)}{\#(x_1, x_2)}, \frac{\#(x_2)}{\#(x_1, x_2)} \right)^\top,$$

where $\#$ denotes the frequency of nodes and edges, respectively. The second feature representation which is again inspired by [Filippova, 2010] and uses the ingredients of the heuristic instead of precomputing the surrogates to have the

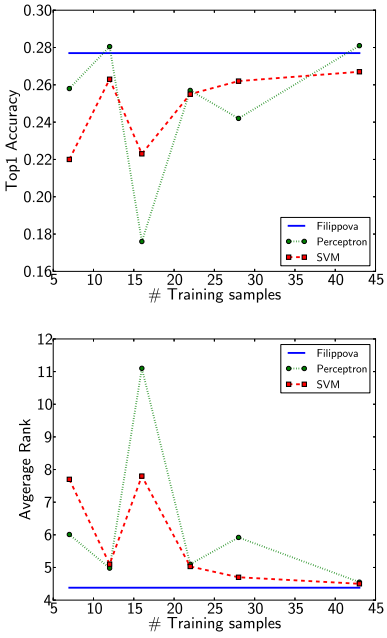


Figure 3: Results on news headlines: Accuracies and average ranks for perceptrons and SVMs.

Table 1: Leave-one-out results for news headlines using feature representation ϕ_2.

	avg. acc.	avg. rank
Filipova	0.277	4.378
Perceptron	0.115	7.58
SVM	0.252	6.942

algorithm pick the best combination,

$$\phi_2(x_i, x_j) = (\#(x_1), \#(x_2), \#(x_1, x_2), \log(\#(x_1)), \log(\#(x_2)), \log(\#(x_1, x_2)))^\top.$$

Figure 3 shows average accuracy (top) and average rank (bottom) for perceptrons and SVMs, respectively, for different training set sizes, depicted on the x-axis. Every curve is the result of a cross-validation that uses all available data. Thus, the rightmost points are generated by a 2-fold cross validation while the leftmost points result from using 11-folds. Due to the small training sets, interpreting the figures is difficult. The unsupervised baseline outperforms the learning methods although there are indications that more training data could lead to better performances of perceptrons and SVMs. The first feature representation shows better performances than the second. However, these conjectures need to be verified by an experiment on a larger scale.

Using only the second feature representation, Table 1 shows average accuracies and average ranks for a leave-one-out setup to increase the sizes of the training sets. The results are promising and not too far from the baseline, however, as before, the evaluation needs to be based on larger sample sizes to allow for interpretations.

5 Conclusion

In this paper, we proposed to learn shortest paths in compression graphs for summarising related sentences. We addressed the previously unsupervised problem in a supervised context and devised structural support vector machines that effectively learn the edge weights of compression graphs, so that a shortest path algorithm decodes the best possible summarisation. We showed that the most strongly violated constrains can be computed directly by loss-augmented inference and rendered the use of two-best algorithms unnecessary. Empirically, we presented preliminary results on artificial and real world data sets. Due to small sample sizes, conclusions cannot be drawn yet, although the results indicate that learning shortest paths could be an alternative to heuristic and unsupervised approaches. Future work will address this question in greater detail.

References

[Altun *et al.*, 2003] Y. Altun, M. Johnson, T. Hofmann. Investigating loss functions and optimization methods for discriminative learning of label sequences, Proc. EMNLP, 2003.

[Barzilay and Lee, 2003] R. Barzilay, L. Lee, Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment, in Proc. of NAACL-HLT, 2003.

[Bellman, 1958] R. Bellman (1958). "On a routing problem". Quarterly of Applied Mathematics 16: 8790. MR 0102435.

[Brefeld, 2007] U. Brefeld. Cost-based Ranking in Input Output Spaces. Proceedings of the Workshop on Learning from Non-vectorial Data, 2007.

[Boudin and Morin, 2013] F. Boudin and E. Morin. Keyphrase Extraction for N-best Reranking in Multi-Sentence Compression, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013.

[Collins and Duffy, 2002] M. Collins and N. Duffy. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron, ACM, 2002

[Dijkstra, 1959] E. W. Dijkstra (1959). "A note on two problems in connexion with graphs". Numerische Mathematik 1: 269271. doi:10.1007/BF01386390.

[Filippova, 2010] K. Filippova. Multi-sentence compression: Finding shortest paths in word graphs, COLING, 2010

[Ford, 1956] J. Ford, R. Lester (August 14, 1956). Network Flow Theory. Paper P-923. Santa Monica, California: RAND Corporation.

[Taskar *et al.*, 2004] B. Taskar and D. Klein and M. Collins and D. Koller and C. Manning. Max-margin parsing, Proc. EMNLP, 2004.

[Tsochantaridis *et al.*, 2005] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, Large Margin Methods for Structured and Interdependent Output Variables, Journal of Machine Learning Research, 6 (Sep):1453-1484, 2005

[Tzouridis and Brefeld, 2013] E. Tzouridis, U. Brefeld. Learning Shortest Paths for Text Summarization. Proceedings of the ECML/PKDD Workshop on Mining Ubiquitous and Social Environments, 2013.

[Wan *et al.*, 2007] S. Wan, R. Dale, M. Dras, C. Paris. Global revision in summarisation : generating novel sentences with Prim's algorithm, Conference of the Pacific Association for Computational Linguistics, 2007.

[Yen, 1971] J. Y. Yen, Finding the k Shortest Loopless Paths in a Network. Management Science 17 (11): 712716, 1971

Erkennung von Sequenzen mimischer Schmerzausdrücke durch genetische Programmierung*

Christoph Stocker[1], **Michael Siebers**[2] **und Ute Schmid**[3]
Kognitive Systeme, Fakultät Wirtschaftsinformatik und Angewandte Informatik
Otto-Friedrich-Universität Bamberg
[1]c.stocker.mail@googlemail.com, [2]michael.siebers@uni-bamberg.de, [3]ute.schmid@uni-bamberg.de

Abstract

Aus empirisch erhobenen Videos von Gesichtern, die Schmerzmimik zeigen, werden über Abfolgen von Action Units (AUs) reguläre Ausdrücke generalisiert, um für Schmerz charakteristische Sequenzen von Bewegungen im Gesicht zu identifizieren. Für die Generierung regulärer Ausdrücke über AU-Sequenzen wurde ein genetischer Algorithmus implementiert. Als Datenstruktur werden Syntaxbäume regulärer Ausdrücke verwendet. Die Individuen sind Grammatiken, die die in den Trainingsdaten gegebenen Sequenzen erkennen können. Die resultierende Präzision der Grammatiken bewegt sich zwischen ca. 80–100%. Problem bei der Optimierung des Verfahrens ist die Justierung der zahlreichen Mutationsparameter. Zudem erschwert das Fehlen von negativen Trainingsdaten eine Evaluation der gelernten Ausdrücke.

1 Einleitung

Angesichts einer kontinuierlich alternden Gesellschaft wird die Betreuung pflegebedürftiger Menschen immer wichtiger. Die Entwicklung von computergestützten Systemen, die das Pflegepersonal bei dieser Aufgabe unterstützen, ist daher ein interessantes Forschungsgebiet. Ein möglicher Teilaspekt dabei ist die Kommunikation mit Menschen, die aufgrund einer Erkrankung nicht oder nur erschwert zu sprachlichen Äußerungen oder Gestik in der Lage sind. In vielen Fällen sind jedoch reaktive Bewegungen, wie sie etwa von Schmerzreizen ausgelöst werden, voll funktionsfähig. Es können daher Systeme entwickelt werden, die diese Bewegungen erkennen, um die nonverbale Kommunikation mit dem Patienten zu ermöglichen.

Verfahren zur Analyse von Gesichtsausdrücken lassen sich in solche unterscheiden, die direkt Klassifikatoren über metrische Bildmerkmale generieren und solche, die in einem Zwischenschritt sogenannte *Action Units* (AUs) in Bilder identifizieren [Fasel and Luettin, 2003]. Die Nutzung von AUs basiert auf den psychologischen Arbeiten von Ekman und Friesen [Ekman and Friesen, 1978], die gezeigt haben, dass sich Emotionen kulturunabhängig und stabil durch die Bewegung von Teilmengen der 43 Gesichtsmuskeln charakterisieren lassen. Das sogenannte *Facial Action Coding System* (FACS) beschreibt die Menge der möglichen AUs und die Zuordnung von Teilmengen dieser AUs zu Emotionen. Das FACS wurde auch für andere, nicht-emotionale mimische Reaktionen – beispielsweise für Schmerz – erweitert [Lucey *et al.*, 2011]. Üblicherweise erfolgt die Analyse von mimischen Ausdrücken über die *Menge* der in einem Zeitintervall aufgetretenen AUs. Es besteht jedoch die Möglichkeit, dass die *sequentielle Abfolge* von AUs zusätzlich diagnostisch relevante Information enthalten kann [Schmid *et al.*, 2012].

Im Folgenden werden zunächst die Datengrundlage sowie Vorarbeiten zum Lernen von charakteristischen AU-Sequenzen dargestellt. Danach wird der genetische Algorithmus zum Lernen regulärer Ausdrücke beschrieben. Es folgt eine Evaluation des Algorithmus' an zwei Datensätzen. Schließlich werden die Ergebnisse bewertet und ein Ausblick auf weitere Arbeitsschritte gegeben.

2 Lernen von Grammatiken aus Sequenzen von Action Units

Im Rahmen eines psychologischen Experiments [Kunz *et al.*, 2007] wurden Patienten mit einer dementiellen Erkrankung ($n = 42$), ältere gesunde Personen ($n = 54$) sowie Studierende ($n = 28$) kurzen Episoden von Druckschmerz ausgesetzt. Dabei wurden die mimischen Bewegungen im Gesicht aufgezeichnet. Die Videos wurden von einer ausgebildeten FACS-Coderin mit AUs annotiert. Da bei mimischen Reaktionen die onset-Zeiten von AUs klar identifizierbar sind, die offset-Zeiten jedoch nicht, da mimische Reaktionen üblicherweise langsam „zerfallen", wurden aus den FACS-Kodierungen Sequenzen von AUs bezüglich der onset-Zeiten extrahiert. Zudem wurden nur diejenigen AUs in die Sequenz aufgenommen, bei denen die Intensität der verabreichten Schmerzreize – gemessen auf einer Skala von 1 bis 5 – mindestens 4 betrug, da bei schwächeren Reizen häufig keine mimische Reaktion gezeigt wurde. Haben zwei oder mehr AUs die gleiche onset-Zeit wurden diese AU-compounds in die Sequenz übernommen. Das den Sequenzen zugrundeliegene Alphabet besteht aus insgesamt 76 AUs und AU-compounds.

In einer Vorgängerarbeit [Schmid *et al.*, 2012] wird die Datengrundlage detailliert beschrieben. In dieser Arbeit wurde ein Verfahren zum Lernen regulärer Grammatiken (*Alignment Based Learning* ABL [van Zaanen, 2002]) zur Generalisierung einer Schmerzgrammatik verwendet. Dabei wurden bereits vielversprechende Ergebnisse erzielt. Allerdings waren die resultierenden Grammatiken zur Charakterisierung von Schmerzmimik sehr komplex. Nun sollte geprüft werden, ob mittels genetischer Algorithmen, bei denen die Komplexität der Grammatik als Kriterium in die Fitness-Funktion aufgenommen wird, kompaktere Beschreibungen gelernt werden können.

*Ergebnisse eines studentischen Projekts im Master-Studiengang *Computing in the Humanities* an der Universität Bamberg

3 Konzeption eines genetischen Algorithmus

3.1 Datenstrukturen

Ein regulärer Ausdruck ist eine intensionale Beschreibung von Mengen von Zeichenketten. Der Ausdruck selbst ist dabei eine syntaktische Beschreibung, welche die in der Menge enthaltenen Sequenzen erfüllen müssen. Durch reguläre Ausdrücke lassen sich reguläre Mengen beschreiben, d.h. Mengen, die durch die Operationen der Vereinigung, Konkatenation und Sternbildung entstehen. Im Gegensatz zu einer Darstellung durch Produktionsregeln können allerdings nur reguläre Grammatiken dargestellt werden. Dieser Nachteil wurde angesichts der entfallenden Konsistenzprüfung in Kauf genommen.

Um reguläre Ausdrücke adäquat darstellen zu können, wurde entschieden, diese als Syntaxbäume zu implementieren, wie dies im Bereich der genetischen Programmierung üblich ist [Mitchell, 1997, S. 262]. Jeder Knoten im Baum entspricht dabei einem Operator eines regulären Ausdrucks. Aus Sicht der genetischen Programmierung entspricht jeder Operator einer Funktion sowie jede AU einem Terminal. Es wurden folgende Operatoren implementiert: Verknüpfung ('Und-Knoten'), Alternation ('Oder-Knoten') sowie die Quantoren '*','+' und '?'.

3.2 Algorithmus

Die Konzeption des genetischen Algorithmus orientiert sich an den in diesem Bereich üblichen Konzepten der Selektion bezüglich einer Fitness-Funktion [Mitchell, 1997, S. 251].

Eine Gruppe von Individuen verbessert sich im Laufe mehrerer Lebenszyklen, indem starke Individuen ihr Erbgut durch Fortpflanzung weitergeben, während schwache Individuen im Lauf der Zeit verschwinden. Die Fortpflanzung besteht aus Generierung neuer Individuen, welche eine zufällige Kombination des Erbguts der Eltern enthalten. Zusätzliche spontane Mutationen verhindern dabei, dass das Verfahren in lokalen Maxima zum Erliegen kommt. Unter einem Individuum wird hierbei im Sinne dieser Arbeit eine einzelne Grammatik verstanden. Das Erbgut eines Individuums sind demzufolge die Regeln, aus denen die Grammatik besteht.

In einem Selektionsschritt wird bestimmt, welche Individuen zur Fortpflanzung geeignet sind. Dies wird entschieden, indem für jedes Individum geprüft wird, wie gut es die gestellte Aufgabe erfüllt. Im Falle dieser Arbeit bedeutet dies, dass eine Grammatik dann geeignet ist, wenn sie möglichst viele Sequenzen effektiv erkennen kann.

Die wichtigsten Schritte des Verfahrens sind demnach die Selektion, die Mutation sowie das Crossover. Zudem ist die Umsetzung der Fitnessfunktion ein weiteres Schlüsselelement.

Selektion

Für die Selektion wurde eine probabilistische Variante verwendet. Die Selektion eines Individuums h_i erfolgt nach einer Gewichtung gemäß der Fitness, bezogen auf die aufsummierte Fitness der gesamten Population:

$$Pr(h_i) = \frac{Fitness(h_i)}{\sum_{j=1}^{p} Fitness(h_j)}$$

Fitness

Die Fitnessfunktion bewertet die Effizienz eines Individuums. Es ist daher naheliegend die Präzision einer Grammatik als Maß der Fitness in Betracht zu ziehen. Ein Individuum ist folglich umso fitter, je mehr Sequenzen der Trainingsdaten es erkennen kann. Aus der Umsetzung der Grammatiken sowie einigen technischen Aspekten ergeben sich allerdings noch weitere Kriterien welche für die Bewertung einer Grammatik in Betracht gezogen werden können.

Zum einen erhöhen große Grammatiken signifikant die Laufzeit des Algorithmus. Der Grund hierfür ist neben der steigenden Anzahl an Knoten im Allgemeinen die steigende Komplexität, welche sich durch Oder-Knoten sowie Quantoren ergibt. Enthält eine Grammatik viele Oder-Knoten, so müssen beim Matching der regulären Ausdrücke für jede Veroderung alternative Möglichkeiten in Betracht gezogen werden. Dies kann durchaus dazu führen, dass die Laufzeit des Algorithmus nicht mehr im akzeptablen Bereich liegt. Demzufolge ist es sinnvoll, Grammatiken niedriger zu bewerten, welche viele Oder-Knoten enthalten.

Zum anderen lässt sich über die Fitnessfunktion das Generalisierungsverhalten des Algorithmus steuern. Grundsätzlich lässt sich festhalten, dass durch die Verwendung von regulären Ausdrücken als Repräsentationsgrundlage sowie durch das Fehlen negativer Trainingsbeispiele im Prinzip jede Sequenz durch einen einfachen regulären Ausdruck erkennen lässt:

$$(au1|au2|au3|au4|\ldots|auN)^*$$

Dieser Ausdruck liefert zwar für jede beliebige Menge von Sequenzen eine Präzision von 100%, ist jedoch nicht sehr aussagekräftig, da er Sequenzen, welche Schmerz anzeigen, nicht von anderen Sequenzen unterscheiden kann. Eine Klassifikation ist somit nicht mehr möglich.

Aufgrund des Mangels an negativen Trainingsdaten lässt sich dieses Phänomen nicht vollständig verhindern. Allerdings kann das Generalisierungsverhalten des Algorithmus auf ein Mindestmaß eingeschränkt werden, indem extrem generalisierende Ausdrücke vermieden werden. Der oben genannte Ausdruck lässt sich beispielsweise vermeiden, indem Grammatiken niedrig bewertet werden, welche viele Veroderungen enthalten. Hierbei ist jedoch anzumerken, dass dies das Voranschreiten der Evolution negativ beeinflusst, da potenziell wertvolle Oder-Knoten ebenfalls seltener auftreten.

Des Weiteren macht es gemäß des Sparsamkeitsprinzips durchaus Sinn, die Anzahl der Knoten eines Syntaxbaums ebenfalls in die Bewertung einfließen zu lassen. Von zwei Grammatiken, welche die gleiche Präzision erzielen, wird demzufolge diejenige bevorzugt, welche weniger Knoten enthält und damit tendenziell eher spezifischer ist.

In die Implementierung der Fitnessfunktion wurden alle genannten Aspekte miteinbezogen. Die endgültige Fitness einer Grammatik h besteht somit aus der gewichteten Summe aller Teilaspekte. Hierbei bezeichnet a_i einen Teilaspekt, sowie w_i das dazugehörige Gewicht:

$$Fitness(h) = \frac{\sum a_i w_i}{\sum w_i}$$

Die Präzision einer Grammatik wird direkt anhand der Trainingsdaten berechnet. Der Aspekt der vorhandenen Oder-Knoten kann einfach quantifiziert werden, indem man die Anzahl der Oder-Knoten einer Grammatik der Gesamtanzahl an Knoten gegenüberstellt, für den Aspekt der verwendeten Quantoren wurde ebenso verfahren. Die Länge einer Grammatik berechnet sich aus $\frac{1}{1+n}$, wobei n die Gesamtanzahl der Knoten einer Grammatik bezeichnet. Wie genau die Einstellung der Gewichte vorzunehmen ist,

hängt von der konkreten Implementierung sowie deren Einsatzgebiet ab.

Mutation
Eine Mutation wird für jede in der Population vorhandene Grammatik mit einer vorgegebenen Wahrscheinlichkeit angewendet. Da eine Grammatik als ein Syntaxbaum repräsentiert wird, ist die naheliegendste Form der Mutation, einen beliebigen Knoten im Syntaxbaum durch einen neuen zu ersetzen.

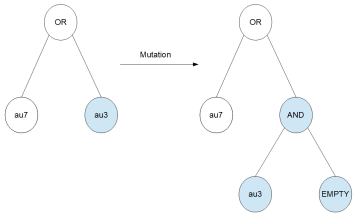


Abbildung 1: Beispiel für eine Mutation

Ein Beispiel für dieses Vorgehen ist in Abbildung 1 dargestellt. Die Action Unit *au3* wird durch einen Und-Knoten ersetzt. Hier zeigt sich bereits das breite Spektrum an Entscheidungen auf, welche bei der Implementierung der Mutation auftreten. Wird wie im genannten Beispiel eine Action Unit durch einen Und-Knoten ersetzt, so muss entschieden werden, welche Kindknoten im Und-Knoten eingefügt werden. Eine mögliche Option wäre es, die alte Action Unit sowie einen leeren Knoten einzufügen. Die resultierende Grammatik hätte demnach die selbe Aussagekraft wie die ursprüngliche, hätte jedoch durch den zusätzlichen Leerknoten weiteres Potenzial durch zukünftige Mutation eine höhere Präzision zu erreichen. Eine andere Option wäre es zufällig generierte Knoten als Kindknoten einzusetzen. Dies bewirkt einen größeren Sprung im Suchraum, erzeugt jedoch mit einer größeren Wahrscheinlichkeit ineffiziente Grammatiken.

Für jede implementierte Mutationsvariante muss festgelegt werden, mit welcher Wahrscheinlichkeit diese auftritt. Es hat sich erwiesen, dass Mutationen, welche neue Teilbäume erzeugen, nur sehr selten vorgenommen werden sollten (Größenordnung 1-2%). Werden zu häufig neue Teilbäume erstellt, so entstehen sehr schnell breite Syntaxbäume, was die Laufzeit signifikant erhöht. Insgesamt sind alle festgelegten Mutationswahrscheinlichlichkeiten als Parameter zu betrachten, da sie die Wirkungsweise des Algorithmus sowie die Laufzeit beeinflussen und keine allgemein gültige optimale Konfiguration möglich ist.

Crossover
Im Gegensatz zur Mutation gestaltet sich das Crossover bei der genutzten Repräsentation als Syntaxbäume relativ einfach. Ein Crossover wird durchgeführt, indem von zwei Grammatiken jeweils zufallsbasiert Teilbäume ausgetauscht werden. Welche Grammatiken durch Crossover verändert werden sollen wird zufallsbasiert, jedoch gewichtet nach der jeweiligen Fitness entschieden. Die dazu verwendete Berechnungsvorschrift entspricht der Formel aus Abschnitt 3.2.

4 Evaluation

Die Vollständigkeit des Algorithmus wird anhand der Präzision gemessen. Unter Präzision ist hierbei der prozentuale Anteil an Trainingsbeispielen zu verstehen, der von einer Grammatik erkannt werden kann.

	Datensatz 1	Datensatz 2
Anzahl AUs	30	76
Anzahl Sequenzen	59	347
Längste Sequenz	17	17
Kürzeste Sequenz	1	1
Mittlere Sequenzlänge	3.59	4.03

Tabelle 1: Eckdaten der Datensätze

Im Rahmen dieser Arbeit wurde mit zwei Datensätzen gearbeitet, die unterschiedliche Schwierigkeitsgrade aufgrund ihrer Komplexität aufweisen. Datensatz 2 umfasst FACS-kodierte Sequenzen aus AUs und AU-compounds aller im Experiment von Kunz et. al. betrachteten Personengruppen [Kunz *et al.*, 2007] – also von älternen Menschen mit dementieller Erkrankung, älteren gesunden Probanden und studentischen Untersuchungsteilnehmern. Datensatz 1 enthält nur die Sequenzen der Studentinnen, da diese Personengruppe die klarsten mimischen Reaktionen zeigte und die Sequenzen weniger variationsreich sind als bei den anderen Probandengruppen.

Die Eckdaten der beiden Datensätze sind in Tabelle 1 aufgelistet. Der Hauptunterschied der Datensätze besteht in der Anzahl der Sequenzen, sowie in der Anzahl der Action Units, die verwendet werden. Datensatz 1 ist aufgrund dieser Einteilung als der „einfache“ Datensatz zu verstehen, sowie Datensatz 2 als der „komplexe“ Datensatz.

Es ist anzumerken, dass dies nur eine grobe Unterteilung darstellt, die nicht effektiv auf die Komplexität der Daten hinweisen kann. Da es sich um eine rein statistische Einteilung handelt wird darin beispielsweise nicht berücksichtigt, ob die enthaltenen Sequenzen nur mit kontextsensitiven Regeln zu erkennen sind. Einteilungen nach Komplexität dieser Art müssten für genauere Untersuchungen berücksichtigt werden.

Um die Vollständigkeit des Algorithmus auszuwerten wurden auf jeden Datensatz 10 Iterationen ausgeführt. Im Allgemeinen lässt sich feststellen, dass die hier vorgestellte Umsetzung die beiden Datensätze recht gut erkennen kann. Für Datensatz 1 wurde der Zielwert von 95% Präzision bei jedem Durchgang erreicht. Hierbei waren, bis auf eine Ausnahme, nie mehr als 100 Iterationen notwendig.

Die Ergebnisse für Datensatz 2 sind ebenfalls als erfolgreich zu erachten. Die Präzision erreicht den erstrebten Zielwert von 95% in 5 der 10 durchgeführten Testläufe. Die restlichen Resultate bewegen sich in einem Bereich von ca. 78% bis 90%. Um diese Ergebnisse zu erreichen, waren bei Datensatz 2 jedoch deutlich mehr Iterationen erforderlich. In 5 von 10 Fällen erreichte der Algorithmus das festgelegte Maximum von 200 Iterationen. Bei diesen 5 Testläufen handelt es sich um diejenigen, die nicht den gewünschten Grenzwert erreichten. Für die erfolgreichen Durchläufe waren jedoch ebenfalls stets deutlich mehr als 100 Iterationen notwendig. Die dafür notwendige Laufzeit ist ebenfalls deutlich höher als bei Datensatz 1.

Um das generelle Verhalten der Umsetzung bezüglich Laufzeit und Präzision evaluieren zu können, wurde in jeder Testreihe die dabei erreichte Präzision festgehalten. Abbildung 2 und 3 stellen die durchschnittlich erreichte Präzision pro Iteration dar. Die Standardabweichung ist dabei als Y-Balken mit angegeben. Bei Betrachtung dieser beiden Abbildungen ist zu erkennen, dass unabhängig vom Schwierigkeitsgrad der Daten ca. die erste Hälfte der Iterationen eine stetige Verbesserung der Präzision erzielt,

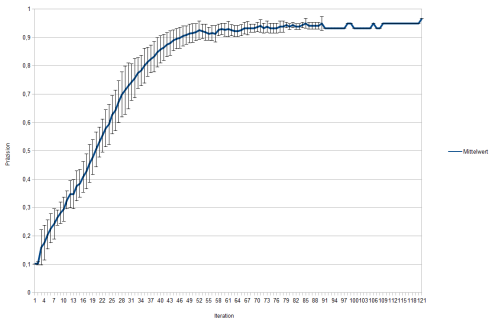


Abbildung 2: Mittlere Präzision, Datensatz 1

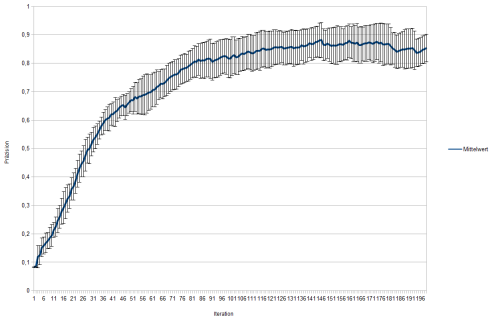


Abbildung 3: Mittlere Präzision, Datensatz 2

während die zweite Hälfte zur „Perfektion“ der gefundenen Grammatik dient. Betrachtet man die Standardabweichungen, so stellt man fest, dass die Testreihen nicht gleichmäßig verlaufen. Dies lässt sich durch lokale Maxima erklären, welche bei Datensatz 2 häufiger auftreten.

Eine Auswertung der Korrektheit gestaltet sich aufgrund des Mangels an Negativbeispielen sehr schwierig. Kontrollversuche wie etwa die Auswertung der Erkennungsrate von Zufallssequenzen sind nur bedingt sinnvoll, da für generierte Sequenzen nicht entschieden werden kann, ob diese Schmerz repräsentieren. Es lässt sich jedoch festhalten, dass durchaus eine Generalisierung der Trainingsdaten stattfindet, da durch die Verwendung von Quantoren und Alternation eine infinite Menge unterschiedlicher Sequenzen erkannt werden kann.

5 Ausblick

In dieser Arbeit wurde ein genetischer Algorithmus zur Induktion regulärer Ausdrücke vorgestellt, der in der Lage ist Sequenzen von Action Units zu erkennen, welche Schmerzausdrücke in der Mimik eines Probanden darstellen. Hierbei wurde ein genetischer Algorithmus implementiert. Grundsätzlich konnte der Klassifikator erfolgreich umgesetzt werden. Allerdings gibt es durchaus Möglichkeiten, den in dieser Arbeit vorgestellten Ansatz noch zu verbessern.

Beispielsweise könnten komplexere Operatoren wie etwa Rückwärtsreferenzen implementiert werden, mit welchen mehr als nur reguläre Sprachen erzeugt werden können. Dies würde demnach die Ausdrucksmächtigkeit der Grammatiken erhöhen.

Zudem enthält die vorgestellte Umsetzung eine sehr große Anzahl an Parametern. Eine optimale Einstellung dieser Parameter konnten in Rahmen dieser Arbeit nicht ergründet werden.

Zur Validierung der Hypothese, dass die Abfolge von Action Units zusätzliche Information gegenüber des bloßen Vorhandenseins der Action Units gibt, um den mentalen Zustand zu erschließen, der einem mimischen Ausdruck zugrunde liegt, wurde ein erstes psychologisches Experiment durchgeführt. Hier sollte bei Gesichtsavataren beurteilt werden, ob diese Schmerz oder Ekel zeigen [Siebers *et al.*, submitted]. Es zeigt sich, dass die Erkennungsraten bei Abfolgen höher liegen als bei gleichzeitiger Umsetzung der Action Units. Darüber hinaus beurteilen Probanden die AU-Sequenzen als natürlicher. Entsprechend kann die Identifikation von Grammatiken über AUs sinnvoll verwendet werden, um natürlichere Animationen für die Interaktion mit Avataren und humanoiden Robotern zu realisieren.

Literatur

[Ekman and Friesen, 1978] Paul Ekman and Wallace V. Friesen. *The Facial Action Encoding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[Fasel and Luettin, 2003] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.

[Javier *et al.*, 2012] Cano Fco. Javier, Macias Fernando, and Milovec Martina. Learning grammars with swarm genetic approach. Master Project Cognitive Systems, SS 2012, 2012.

[Kunz *et al.*, 2007] Miriam Kunz, Siegfried Scharmann, Uli Hemmeter, Karsten Schepelmann, and Stefan Lautenbacher. The facial expression of pain in patients with dementia. *Pain*, 133:221–228, 2007.

[Lautenbacher *et al.*, 2007] S. Lautenbacher, M. Kunz, V. Mylius, S. Scharmann, U. Hemmeter, and K. Schepelmann. Mehrdimensionale schmerzmessung bei demenzpatienten. *Der Schmerz*, 21:529–538, 2007.

[Lucey *et al.*, 2011] Patrick Lucey, Jeffrey F. Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 41(3):664–674, 2011.

[Mitchell, 1997] T.M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[Schmid *et al.*, 2012] Ute Schmid, Michael Siebers, Dominik Seuss, Miriam Kunz, and Stefan Lautenbacher. Applying grammar inference to identify generalized patterns of facial expressions of pain. *Proceedings of the 11th International Conference on Grammatical Inference*, 21:1–6, 2012.

[Siebers *et al.*, submitted] Michael Siebers, Tamara Engelbrecht, and Ute Schmid. On the relevance of sequence information for decoding facial expressions of pain and disgust – An avatar study. In Dirk Reichardt, editor, *Proceedings of the 7th Workshop on Emotion and Computing - Current Research and Future Impact (KI 2013)*, submitted.

[van Zaanen, 2002] Menno van Zaanen. Bootstrapping structure into language: Alignment-based learning. *CoRR*, cs.LG/0205025, 2002.

CMI: An Information-Theoretic Contrast Measure for Enhancing Subspace Cluster and Outlier Detection

Hoang Vu Nguyen[◦] **Emmanuel Müller**[◦•] **Jilles Vreeken**[•] **Fabian Keller**[◦] **Klemens Böhm**[◦]

[◦] Karlsruhe Institute of Technology, Germany
{hoang.nguyen, emmanuel.mueller, fabian.keller, klemens.boehm}@kit.edu

[•] University of Antwerp, Belgium
{emmanuel.mueller, jilles.vreeken}@ua.ac.be

Abstract

In many real world applications data is collected in multi-dimensional spaces, with the knowledge hidden in subspaces (i.e., subsets of the dimensions). It is an open research issue to select meaningful subspaces without any prior knowledge about such hidden patterns. Standard approaches, such as pairwise correlation measures, or statistical approaches based on entropy, do not solve this problem; due to their restrictive pairwise analysis and loss of information in discretization they are bound to miss subspaces with potential clusters and outliers.

In this work[1], we focus on finding subspaces with strong mutual dependency in the selected dimension set. Chosen subspaces should provide a high discrepancy between clusters and outliers and enhance detection of these patterns. To measure this, we propose a novel contrast score that quantifies mutual correlations in subspaces by considering their cumulative distributions—without having to discretize the data. In our experiments, we show that these high contrast subspaces provide enhanced quality in cluster and outlier detection for both synthetic and real world data.

1 Introduction

Clustering and outlier detection are two key data mining tasks. They are widely used, such as in bioinformatics, for detecting functionally dependent genes, in marketing, for customer segmentation, in health surveillance, for anomaly detection, and so on. For these techniques to work well, some kind of dependency between the objects in a given data space is required, i.e., high similarity between clustered objects and high deviation between outliers and the residual data distribution.

Obviously, detecting clusters and outliers in uniformly random distributed spaces, e.g., considering a data space with independent dimensions, does not make sense at all. With more and more dimensions such effects tend to hinder data mining tasks, which is widely known as the "curse of dimensionality" [Beyer *et al.*, 1999]. Not just a fringe theoretical case, we observe this effect in practice, for example, in gene expression analysis where each gene is described with very many expression values under different medical treatments. In general, we observe a loss of contrast between clusters and outliers in the full space (all given dimensions) of the data, while the meaningful knowledge is hidden in subspaces (i.e., subsets of the dimensions) that show a high dependency between the selected dimensions.

Recently, more attention has been placed on *subspace clustering* [Agrawal *et al.*, 1998; Aggarwal *et al.*, 1999; Sequeira and Zaki, 2004; Müller *et al.*, 2009a] and *subspace outlier detection* [Aggarwal and Yu, 2001; Kriegel *et al.*, 2009; Müller *et al.*, 2011]. Both of these paradigms detect a set of relevant dimensions for each individual cluster or oultier. Hence, they are able to detect meaningful patterns even if only few dimensions are relevant for the individual pattern. However, they all face a common problem in the selection of subspaces. Each of the techniques re-invents a very specific subspace selection scheme according to the underlying cluster or outlier model. Only few techniques have focused on general solutions to the problem of *subspace search* designed for clustering [Cheng *et al.*, 1999; Baumgartner *et al.*, 2004] or outlier mining [Ye *et al.*, 2009; Keller *et al.*, 2012]. In this work, we follow this general idea of subspace search. We aim at a further generalization for the selection of relevant subspaces.

More specifically, we aim at selecting *high contrast subspaces* that potentially provide high contrast between clustered and outlying objects. Due to its generality this problem statement poses several open questions. First, it is unclear how to measure the contrast of a given set of dimensions. Solutions based on correlation analysis and entropy measures seem promising but show major drawbacks w.r.t. pairwise analysis, discretization, and the empty space problem, as we will explain later. Second, one requires robust statistics to capture the mutual dependence of dimensions. Existing solutions performing a pairwise analysis miss important higher-order dependencies that can only be identified when multiple dimensions are considered together. Finally, a subspace selection has to be performed in an efficient manner in order to scale with the increasing number of dimensions, i.e., an exponential search space.

We tackle all three of these challenges by our contrast measure. It is independent of any cluster or outlier model and purely based on the statistical dependence of data observed in a multi-dimensional subspace. Furthermore, it is directly applicable to continuous data and does not fall prey to the information loss by previous discretization techniques. It is designed to capture mutual dependencies, and thus, quantifies the subspace deviation from the condition of uncorrelated and independent dimensions: "The larger the deviation from the mutual independence assumption, the higher the contrast of a subspace." Hence, we instantiate our measure based on the analysis of cumulative dis-

[1] This work has been published in Proc. SIAM International Conference on Data Mining (SDM 2013) [Nguyen *et al.*, 2013]

tributions in different subspaces. Cumulative distributions have the advantage that they can be computed directly on empirical data. Furthermore, we propose a scalable processing scheme to select high contrast subspaces. Due to the exponential search space we rely on an approximative solution based on beam search.

Overall, our contributions are as follows: (a) a set of abstract quality criteria for subspace search based on contrast analysis, (b) our multi-variate contrast measure based on cumulative distributions for continuous data, (c) a scalable subspace search method applying our contrast measure for subspace selection, and (d) quality enhancement for both subspace clustering and subspace outlier mining as a result of high contrast.

2 Related Work

Pairwise measures and space transformations. First, we discuss approaches that assess dependencies between dimensions. Spearman correlation and modern variants [Reshef *et al.*, 2011] are aimed at pairwise correlations. However, higher order interactions (i.e., mutual dependence) among several dimensions can be missed. Similarly, dimensionality reduction techniques [Lee and Verleysen, 2007], including PCA, are not aware of locally clustered projections; they only measure the (non-)linear dependence between dimensions, meaning that they consider one (global) projection, and may hence miss interesting local projections containing subspace clusters and outliers. Our method, on the other hand, is not limited to a pairwise assessment and provides multiple projections for clustering and outlier mining. It can cope with mutual dependencies in arbitrary subspace projections.

Feature selection. Next, we consider methods for unsupervised feature selection. Recent methods [Dy and Brodley, 2004; Law *et al.*, 2004] perform iteratively a partitioning and feature selection. They first partition the data (e.g., by EM clustering), and then they evaluate feature subsets based on the obtained clusters. Another approach [Roth and Lange, 2003] aims at different feature subsets for different clusters. However, it focuses on disjoint clusters and does neither allow overlapping clusters nor outliers. Our method is more general and is aware of outliers and overlap of clusters. In general, feature selection differs from our approach in major aspects. Current feature selection methods are specifically bound to clustering. In contrast, our method is more general and suitable for both cluster and outlier mining in multiple subspaces. Most approaches [Dy and Brodley, 2004; Law *et al.*, 2004] select a single projection of the data space, which uncovers some certain cluster structure in the data. These methods are limited to one subspace, while we mine multiple possibly overlapping subspaces. Yet keeping only one subspace may miss local projections containing different subspace clusters [Müller *et al.*, 2009b].

Subspace search. We now discuss methods for selecting relevant subspaces. They avoid the limitations of the above paradigms, and focus on multiple projections with arbitrary dimensionality. Existing methods, however, rely on discretization of continuous dimensions [Cheng *et al.*, 1999; Ye *et al.*, 2009] or only work with binary data [Zhang *et al.*, 2008] and/or discrete data [Chanda *et al.*, 2010].

ENCLUS [Cheng *et al.*, 1999] and PODM [Ye *et al.*, 2009] detect subspaces with low entropy and high interest, discretizing continuous dimensions into equi-width bins in order to compute the entropy measure. By requiring discretization, these methods have unintuitive parameters, and are hence inherently susceptible to knowledge loss and to the curse of dimensionality. To some extent, these limitations have been tackled by HiCS [Keller *et al.*, 2012], which works directly on continuous data. It quantifies the differences between the marginal and conditional distribution in a random dimension of the considered subspace; by its random nature it may hence miss relevant subspaces. Further, it is exposed to the curse of dimensionality w.r.t. conditional distributions in high dimensional spaces.

Our method, on the other hand, can reliably score contrast, regardless of subspace dimensionality. Furthermore, for each subspace we aim to find that permutation of dimensions that yields optimal contrast.

Cluster and outlier detection in subspaces. Specific methods for clustering and outlier detection have been proposed. However, they do not provide a general notion of subspace selection. They select subspaces very specifically to the underlying cluster [Agrawal *et al.*, 1998; Aggarwal *et al.*, 1999; Sequeira and Zaki, 2004; Müller *et al.*, 2009a] or outlier [Aggarwal and Yu, 2001; Kriegel *et al.*, 2009; Müller *et al.*, 2011] definitions. In contrast to all these solutions, our goal is to design a contrast measure that is applicable to subspace selection for different mining paradigms. We show its instantiations to clustering and outlier detection and evaluate its quality.

3 Basic Notions for Contrast Assessment

Given a database DB of size N and dimensionality D, we want to measure the contrast of any lower dimensional subspace S with dimensionality $1 \leq d \leq D$. Our assessment is based on the full space of all dimensions given by $F = \{X_1, \ldots, X_D\}$. Each dimension i is associated with a random variable X_i that has a continuous value domain $dom(X_i) = \mathbb{R}$. We use the notion of density distribution $p_{X_i}(x_i)$ for the projected database on dimension i. We write $p_{X_i}(x_i)$ as $p(x_i)$ when the context is clear. Any non-empty subset $S \in \mathcal{P}(F)$ is called a subspace of DB. The dimensionality of S is denoted as $dim(S)$. W.l.o.g., $\{X_1, \ldots, X_d\}$ is used as representative for any d-dimensional subspace S in our analysis.

3.1 Contrast Assessment.

As our general notion of a contrast measure we have the following formalization:

Definition 1 ***Contrast Measure of Subspaces**:*

$$C : \mathcal{P}(F) \setminus \{\emptyset\} \to \mathbb{R}$$

In general, the contrast score $C(S)$ quantifies the difference of S w.r.t. the baseline of d independent and randomly distributed dimensions. In the following we provide different instantiation of this contrast measure and discuss formal properties of the instantiations. Let us first formalize the independence baseline. For d random variables $X_1, \ldots, X_d$, there are two types of independence we are interested in.

Definition 2 ***Mutual Independence**:*
$X_1, \ldots, X_d$ are mutually independent iff

$$p(x_1, \ldots, x_d) = p(x_1) \cdots p(x_d)$$

Definition 3 ***m-wise Independence**:*
$X_1, \ldots, X_d$ are m-wise independent with $m \leq d$ iff any subset $\{X_{i_1}, \ldots, X_{i_m}\} \subseteq \{X_1, \ldots, X_d\}$ is mutually independent.

Please note that pairwise independence is modeled as a special case of m-wise independence when $m = 2$. However, pairwise analysis misses important higher-order dependencies that can only be identified when multiple dimensions are considered altogether. Therefore, we focus on higher-order dependencies and their contrast assessment. A subspace is referred to as uncorrelated if its dimensions are mutually independent. Our goal is to design a contrast measure C that quantifies as closely as possible the deviation of subspaces from uncorrelated ones. In other words, for a d-dimensional subspace S with dimensions $\{X_1, \ldots, X_d\}$, its contrast depends on the difference between two functions: $p(x_1, \ldots, x_d)$ and $p(x_1) \cdots p(x_d)$.

$$C(S) \sim \mathit{diff}(p(x_1, \ldots, x_d), p(x_1) \cdots p(x_d))$$

Contrast of one-dimensional subspaces is undefined. Thus, we restrict the contrast measure C to two- or higher-dimensional subspaces. In the following, we propose three properties for a meaningful contrast assessment based on the idea "deviating from uncorrelated subspaces":

Property 1 (Discriminative contrast scores): For subspaces S_1 and S_2 such that $dim(S_1) = dim(S_2)$, if S_1 is more correlated than S_2 then $C(S_1) > C(S_2)$.

Property 2 (Zero contrast score): $C(S) = 0$ if and only if the dimensions of S are mutually independent.

Property 3 (Awareness of m-wise independence): If the dimensions of S are m-wise independent but not mutually independent then $C(S)$ is small but not zero. This is because m-wise independence does not guarantee mutual independence.

Furthermore, C should be directly applicable to continuous data, i.e., we do not require discretization to obtain the probability mass functions. Since discretization causes knowledge loss, this property is mandatory.

3.2 Discussion of Properties.

Looking at existing techniques, ENCLUS [Cheng *et al.*, 1999] instantiates the *diff* function by the well-known total correlation $\sum_{i=1}^{d} H(X_i) - H(X_1, \ldots, X_d)$ where $X_1, \ldots, X_d$ are *discretized* versions of the original dimensions. PODM [Ye *et al.*, 2009] also discretizes data and instantiates the *diff* function as $\sum \frac{1}{p(x_1,\ldots,x_d)}$ where $p(x_1, \ldots, x_d) \neq 0$. The instantiation of HiCS [Keller *et al.*, 2012] is done by averaging over multiple random runs of the form $\mathit{diff}(p(x_i), p(x_i|\{x_1, \ldots, x_d\} \setminus \{x_i\}))$ where X_i is picked randomly.

None of these techniques fulfills all properties mentioned. Considering Property 1, the measure of ENCLUS is unreliable because of the knowledge loss caused by data discretization. Further, the use of the joint probability mass function $p(x_1, \cdots, x_d)$ also is problematic. In particular, $H(X_1, \ldots, X_d) = -\sum p(x_1, \ldots, x_d) \log p(x_1, \ldots, x_d)$ with $p(x_1, \ldots, x_d)$ measured by the relative number of points in the respective hypercube. For increasing d, most hypercubes are empty and the non-empty ones most likely contain only one data point each [Aggarwal and Yu, 2001; Lee and Verleysen, 2007]. Taking into account that $\lim_{x \to 0} x \log x = 0$, $H(X_1, \ldots, X_d)$ approaches $-\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N$. Hence, when d is large enough and all X_i have a similar distribution (e.g., uniformly dense), any d-dimensional subspaces S_1 and S_2 have very similar contrast: $C(S_1) \approx C(S_2)$. In other words, the measure of ENCLUS produces indifferent contrasts for high dimensional subspaces. Thus, it fails to satisfy Property 1, i.e., the most basic property. PODM relying on data discretization and the joint probability $p(x_1, \ldots, x_d)$ suffers the same issue. As for HiCS, the random choice of X_i causes potential loss of contrast as some attribute may not be tested against the remaining ones. In addition, HiCS uses conditional probability distributions with $(d-1)$ conditions and exposes itself to the same problem of empty space.

Considering Properties 2 and 3, since ENCLUS works with discretized data that causes loss of knowledge, it only satisfies these properties with a proper grid resolution. Such a resolution is data-dependent. PODM misses both Properties 2 and 3 since its measure just relies on the joint probability, i.e., it does not measure dependency. A zero contrast assigned by HiCS does not imply uncorrelated spaces since there is no guarantee that all dimensions are assessed against the others at least once. Thus, HiCS does not meet Property 2. Furthermore, HiCS does not aim at m-wise independence and thus does not address Property 3.

4 Methodology

In order to address all three properties, we first introduce a novel notion of mutual information, called Cumulative Mutual Information (CMI), which is instantiated based on a new notion of entropy, called Cumulative Entropy (CE). We then verify that CMI addresses Properties 1 to 3. Since CMI is dependent on the order of subspace dimensions, we then devise an approach to select a dimension permutation that approximates the optimal CMI value for a given subspace. Due to space limitation, all proofs for the following theorems will be provided as an extended version of this paper.

4.1 Cumulative mutual information.

Given continuous random variables $X_1, \ldots, X_d$, their cumulative mutual information $CMI(X_1, \ldots, X_d)$ is defined as:

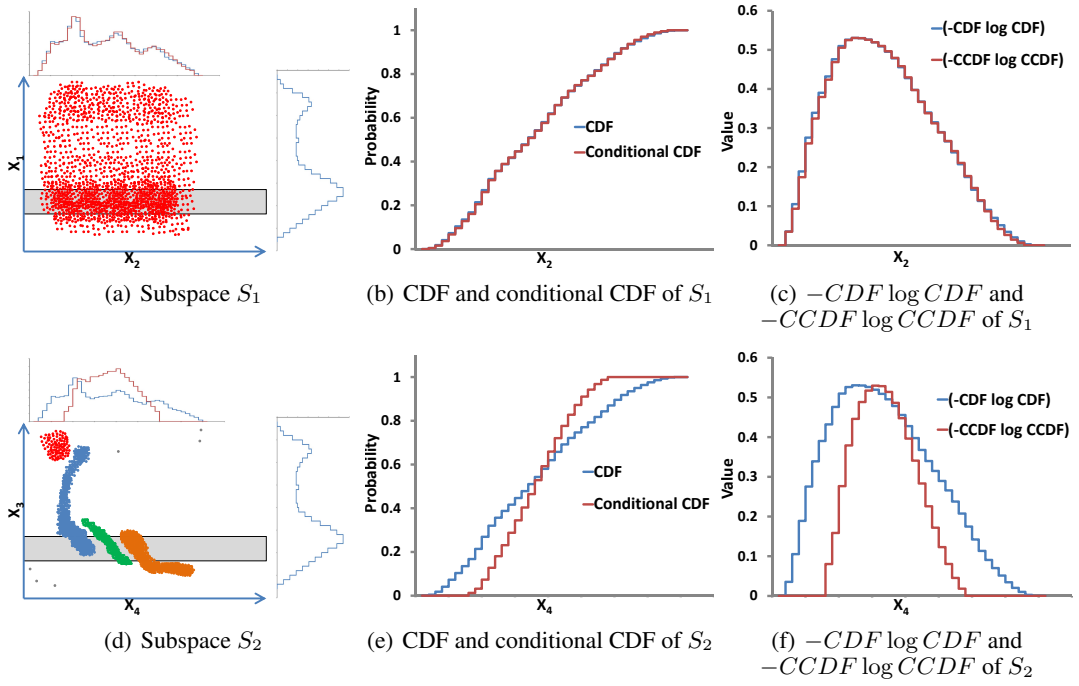
$$\sum_{i=2}^{d} \mathit{diff}(p(x_i), p(x_i|x_1, \ldots, x_{i-1}))$$

Intuitively, $CMI(X_1, \ldots, X_d)$ measures the mutual information of $X_1, \ldots, X_d$ by aggregating the difference between $p(x_i)$ and $p(x_i|x_1, \ldots, x_{i-1})$ for $2 \leq i \leq d$. Loosely speaking, it is the sum of the contrasts of subspaces $(X_1, X_2), \ldots, (X_1, \ldots, X_i), \ldots, (X_1, \ldots, X_d)$ if we consider $\mathit{diff}(p(x_i), p(x_i|x_1, \ldots, x_{i-1}))$ to be the contrast of the subspace $(X_1, \ldots, X_i)$. The reason for using lower-dimensional subspace projections is to avoid the empty space phenomenon. Since probability functions are not available at hand and can only be roughly estimated, e.g. by discretization, we aim at a direct solution using cumulative distributions as difference of two functions $\mathit{diff}(p(x_i), p(x_i|x_1, \ldots, x_{i-1}))$. We instantiate CMI by means of CE and conditional CE that are based on cumulative distributions. We demonstrate in Section 5 how these allow efficient contrast calculation without discretizing data. Their definitions are given below:

Definition 4 *The cumulative entropy for a continuous random variable X, denoted $h_{CE}(X)$, is defined as:*

$$h_{CE}(X) = -\int_{dom(X)} P(X \leq x) \log P(X \leq x) dx$$

Our notion of cumulative entropy is based on [Crescenzo and Longobardi, 2009]. However, it is more general since



(a) Subspace S_1 (b) CDF and conditional CDF of S_1 (c) $-CDF \log CDF$ and $-CCDF \log CCDF$ of S_1

(d) Subspace S_2 (e) CDF and conditional CDF of S_2 (f) $-CDF \log CDF$ and $-CCDF \log CCDF$ of S_2

Figure 1: Example of low and high contrast subspaces with different CMIs

it is not restricted to non-negative random variables. Furthermore, we extend the notion of CE to conditional cumulative entropy and prove that it maintains some important properties of traditional conditional entropy as follows:

Definition 5 *The conditional CE of any continuous random variable X knowing that some random vector $V \in \mathbb{R}^B$ (with B being a positive integer) takes the value v is defined as:*

$$h_{CE}(X|v) = -\int_{dom(X)} P(X \leq x|v) \log P(X \leq x|v) dx$$

The CE of X conditioned by V is:

$$E_V[h_{CE}(X|V)] = \int_{dom(V)} h_{CE}(X|v)p(v)dv$$

Just like the usual conditional entropy, we denote $E_V[h_{CE}(X|V)]$ as $h_{CE}(X|V)$ for notational convenience. The conditional CE has two important properties given by the following theorems:

Theorem 1 *$E_V[h_{CE}(X|V)] \geq 0$ with equality iff there exists a function $f : dom(V) \to dom(X)$ such that $X = f(V)$.*

Theorem 2 *$E_V[h_{CE}(X|V)] \leq h_{CE}(X)$ with equality iff X is independent of V.*

Under CE, $diff(p(x), p(x|\ldots))$ is set to $h_{CE}(X) - h_{CE}(X|\ldots)$. Therefore, $CMI(X_1, \ldots, X_d)$ becomes:

$$\sum_{i=2}^{d} h_{CE}(X_i) - \sum_{i=2}^{d} h_{CE}(X_i|X_1, \ldots, X_{i-1})$$

where $h_{CE}(X_i|X_1, \ldots, X_{i-1})$ is $h_{CE}(X_i|V)$ with $V = (X_1, \ldots, X_{i-1})$ being a random vector in $dom(X_1) \times \cdots \times dom(X_{i-1})$.

Regarding the three properties, similar to traditional mutual information, the more correlated $X_1, \ldots, X_d$ are, the smaller the conditional CEs are, i.e., the larger is CMI. Thus CMI is able to capture subspace correlation (Property 1). To illustrate this property, we use the toy example in Figure 1. It depicts the scatter plots, CDF plots, and plots of the function $-P(X \leq x) \log P(X \leq x)$, namely $-CDF \log CDF$, of two subspaces S_1 and S_2 (CCDF means conditional CDF). The blue lines stand for the marginal distribution of the corresponding dimension. The red lines feature the conditional distribution of one dimension obtained by selecting a range of the remaining dimension (gray strips). One can see that S_2 has higher contrast than S_1 and hence, $CMI(X_3, X_4)_{\text{selected range}} = 4.344 > CMI(X_1, X_2)_{\text{selected range}} = 0.113$. Further, even when high-order conditional CEs may be impacted by the curse of dimensionality, CMI still yields distinguishable contrast for high dimensional subspaces due to its member low-order conditional CEs. If $X_1, \ldots, X_d$ are m-wise independent, then $CMI(X_1, \ldots, X_d)$ is low as $h_{CE}(X_i) - h_{CE}(X_i|\ldots)$ vanishes for $i \leq m$ (Property 3). However, we have proved that $CMI = 0$ iff $X_1, \ldots, X_d$ are mutually independent (Property 2).

Theorem 3 *$CMI(X_1, \ldots, X_d) \geq 0$ with equality iff $X_1, \ldots, X_d$ are mutually independent.*

4.2 Choice of permutation.

CMI can be used as our contrast measure. However, CMI changes with dimension permutations. In order to make our contrast measure permutation-independent we investigate a heuristic search of the maximal contrast.

Our goal is to find a permutation that maximizes the contrast of a given subspace $S = \{X_1, \ldots, X_d\}$. Since CMI is permutation variant, there are $d!$ possible cases in total. Together with the exponential number of subspaces, a brute-force approach is impractical. We therefore apply a heuristic to obtain a permutation that approximates the optimal one. In particular, we first pick a pair of dimensions X_a and X_b ($1 \leq a \neq b \leq d$) such that $h_{CE}(X_b) - h_{CE}(X_b|X_a)$ is maximal among the possible pairs. We then continue selecting the next dimension X_c

($c \neq a$ and $c \neq b$) such that $h_{CE}(X_c) - h_{CE}(X_c|X_a, X_b)$ is maximal among the remaining dimensions. Likewise, at each step, assuming $I = \{X_{p_1}, \ldots, X_{p_k}\}$ is the set of dimensions already picked and $R = \{X_{r_1}, \ldots, X_{r_{d-k}}\}$ is the set of remaining ones, we select the dimension $X_{r_i} \in R$ such that $h_{CE}(X_{r_i}) - h_{CE}(X_{r_i}|I)$ is maximal. The process goes on until no dimension is left. Denoting the permutation obtained by our strategy as π_{opt}, the contrast of S is defined as $CMI(\pi_{opt}(X_1, \ldots, X_d))$.

5 Algorithmic Approach

For a D-dimensional data set, there are $2^D - 1$ candidate subspaces to examine. The exponential number of subspaces makes a brute-force search impractical. A scalable subspace exploration framework is required. Moreover, the contrast measure must also permit efficient computation. In this section, we first introduce an approximate yet scalable levelwise subspace search framework. We then proceed to discuss how to compute our measure efficiently.

5.1 Scalable subspace exploration.

Our aim is to mine high contrast subspaces upon which subspace clustering and outlier detection techniques are applied. To tackle the exponential search space, we target at a processing scheme that trades off accuracy for efficiency. More specifically, we rely on the intuition that a high contrast high dimensional subspace likely has its high contrast reflected in its lower-dimensional projections. In the field of subspace clustering, there is an analogous observation: Subspace clusters tend to have their data points clustered in all of their lower-dimensional projections [Agrawal *et al.*, 1998; Müller *et al.*, 2009b]. One can then apply a levelwise scheme to mine subspaces of contrast larger than a pre-specified value. However, to facilitate parameterization of our method, we avoid imposing direct thresholds on contrast scores produced by CMI.

Instead, we design a beam search strategy to obtain efficiency. Starting with two-dimensional subspaces, in each step we use top M subspaces of high contrast to generate new candidates in a levelwise manner. A newly generated candidate is only considered if all of its child subspaces have high contrast. First, this permits tractable time complexity. Second, interaction among different subspace dimensionality is taken into account and selected subspaces are ensured to have high contrast. Third, we avoid redundancy, if $T \subseteq S$ and S has higher contrast than T. In this case, T is excluded from the final result.

5.2 Efficient contrast computation.

To compute CMI, we need to compute CE and conditional CE.

Let $X_1 \leq \ldots \leq X_n$ be i.i.d. random samples of the continuous random variable X. Then $h_{CE}(X)$ can be calculated as follows:

$$h_{CE}(X) = -\sum_{i=1}^{n-1} (X_{i+1} - X_i) \frac{i}{n} \log \frac{i}{n}$$

In contrast to this straightforward computation, it is not as simple to calculate the conditional CE in an accurate and efficient way. In the following, we first point out that due to limited data, sticking to the exact formula of conditional CE may lead to inaccurate results. We then propose a strategy to resolve this while ensuring that data discretization is not required.

First, w.l.o.g., consider the space $[-1/2, 1/2]^d$ containing N limited data points. The d dimensions are $X_1, \ldots, X_d$. Our goal is to compute $h_{CE}(X_1|X_2, \ldots, X_d)$ using limited available data. From Definition 5: $h_{CE}(X_1|X_2, \ldots, X_d) = \int_{-1/2}^{1/2} \cdots \int_{-1/2}^{1/2} h(X_1|x_2, \ldots, x_d) p(x_2, \ldots, x_d) dx_2 \cdots dx_d$. Further:

$$h_{CE}(X_1|x_2, \ldots, x_d) = \lim_{\varepsilon \to 0^+} h_{CE}(X_1 | x_2 - \varepsilon \leq X_2 \leq x_2 + \varepsilon, \ldots, x_d - \varepsilon \leq X_d \leq x_d + \varepsilon)$$

Taking into account that the total number of data points N is limited, the expected number of points contained in the hypercube $[x_2 - \varepsilon, x_2 + \varepsilon] \times \cdots \times [x_d - \varepsilon, x_d + \varepsilon]$, which is $N(2\varepsilon)^{d-1}$, approaches 0 as $\varepsilon \to 0^+$. For high dimensional spaces, the problem is exacerbated as one faces the empty space phenomenon. With empty hypercubes (or even hypercubes of one data point), $h_{CE}(X_1|x_2, \ldots, x_d)$ vanishes. Hence, $h_{CE}(X_1|X_2, \ldots, X_d)$ becomes 0. We thus encounter a paradox: For commonly used density estimation, e.g. by hypercubes [Cheng *et al.*, 1999], one might end up with an inaccurate result with the exact formula of conditional CE. To alleviate this problem, we must ensure to have enough points for meaningful calculation. Therefore, we propose data summarization by clustering.

Clustering summarizes the data by means of clusters. Since the number of clusters is generally much less than the original data size, we may have more data points in each cluster. Hence, the issue of limited data is mitigated. Assuming that a clustering algorithm is used on DB projected to $\{X_2, \ldots, X_d\}$ resulting in k clusters $\{C_1, \ldots, C_k\}$ (the support of C_i is $|C_i|$), we propose to estimate $h_{CE}(X_1|X_2, \ldots, X_d)$ by:

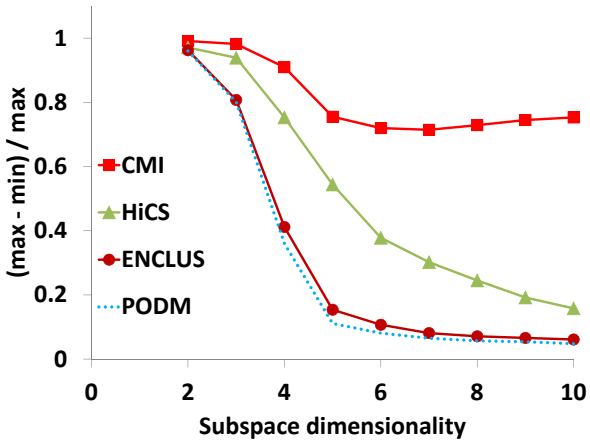
$$\sum_{i=1}^{k} \frac{|C_i|}{N} h_{CE}(X_1|C_i)$$

If k is kept small enough, we will have enough points for a meaningful computation of $h_{CE}(X_1|C_i)$ regardless of the dimensionality d. As our cluster-based approach does not rely on any specific cluster notion, it can be instantiated by any method. To ensure efficient computation of the contrast measure, we use the one-pass k-means clustering strategy introduced in [Ordonez and Omiecinski, 2004]. We obtain k clusters summarizing the data. For the parameter k, if it is set too high, we may end up with high runtime and not enough data in each cluster for a reliable estimation of conditional CE. If it is instead set to 1, i.e., no clustering at all, $h_{CE}(X_1|\cdots)$ becomes $h_{CE}(X_1)$, i.e., there is a loss of information. In all of our experiments, we set $k = 10$. Using clustering, one can verify that the conditional CE is less than or equal to its respective unconditional one.

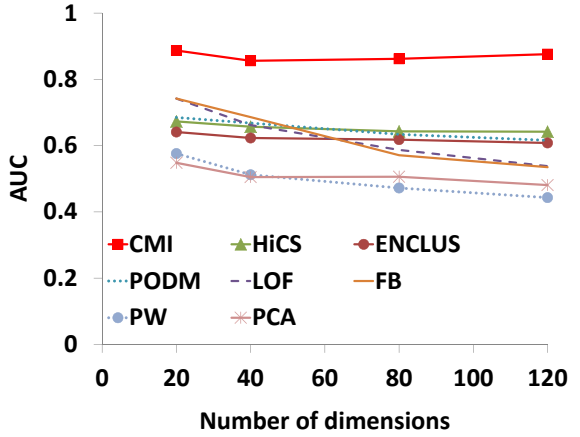
6 Experiments

We compare CMI, to three subspace search methods: ENCLUS [Cheng *et al.*, 1999], HiCS [Keller *et al.*, 2012], and PODM [Ye *et al.*, 2009]. As further baselines we include random selection (FB) [Lazarevic and Kumar, 2005], PCA [Lee and Verleysen, 2007], and pairwise correlation (PW) [Reshef *et al.*, 2011]. For CMI we use $M = 400$ and $k = 10$, unless stated otherwise. In order to assist comparability and future research in this area, we provide our algorithm, all datasets, parameters and further material on our website.[2]

[2] http://www.ipd.kit.edu/~muellere/CMI/



(a) Contrast score vs. dimensionality



(b) AUC vs. dimensionality

Figure 2: Subspace quality w.r.t. dimensionality

We evaluate how mining of high contrast subspaces improves the result quality of outlier detection and clustering techniques. Therefore, LOF [Breunig *et al.*, 2000] and DBSCAN [Ester *et al.*, 1996], two well-established methods, are used on top of the tested approaches. For fair comparison, we use the same parameter settings for both LOF and DBSCAN.

To ensure succinct sets of subspaces that allow for post-analysis, only the best 100 subspaces of each technique are utilized for clustering and outlier detection. Outlier detection results are assessed by the Area Under the ROC Curve (AUC) as in [Lazarevic and Kumar, 2005; Keller *et al.*, 2012; Müller *et al.*, 2011]. Clustering results are evaluated by means of F1, Accuracy, and E4SC as in [Müller *et al.*, 2009b; Günnemann *et al.*, 2011].

6.1 Impact of dimensionality.

To illustrate that *CMI* is robust w.r.t. increasing dimensionality of subspaces, we evaluate it on a synthetic data set of 20 dimensions and 5120 data points, generated according to [Müller *et al.*, 2011]. Please note that in this experiment, we perform an exhaustive search without any pruning. Because of the large total number of subspaces ($2^{20} - 1$), we only experiment up to $d = 10$ to avoid excessive runtime. We record $\frac{\max A_d - \min A_d}{\max A_d}$ where A_d is the set of contrast scores of all d-dimensional subspaces. For $2 \leq d \leq 10$, $\min A_d \approx 0$ (as there are uncorrelated d-dimensional subspaces) and $\max A_d \neq 0$ (as there are correlated d-dimensional subspaces with clusters and outliers). Hence, ideally $\frac{\max A_d - \min A_d}{\max A_d} = 1$ for $2 \leq d \leq 10$. The results, plotted in Figure 2(a), show that HiCS, ENCLUS, and PODM do not scale well with higher dimensionality. In contrast, *CMI* is more robust to dimensionality and yields discriminative contrast scores even for high dimensional subspaces.

	CMI	HiCS	Enclus	Podm	DBScan	FB
20 dimensions						
F1	**0.96**	**0.96**	0.72	0.75	0.65	0.67
Acc.	**0.98**	0.96	0.75	0.82	0.67	0.68
E4SC	**0.92**	0.75	0.42	0.36	0.19	0.27
40 dimensions						
F1	**0.93**	0.88	0.65	0.72	0.54	0.61
Acc.	**0.93**	0.74	0.68	0.76	0.61	0.66
E4SC	**0.89**	0.73	0.27	0.34	0.21	0.23
80 dimensions						
F1	**0.94**	0.83	0.62	0.68	0.57	0.61
Acc.	**0.95**	0.74	0.66	0.81	0.62	0.69
E4SC	**0.86**	0.57	0.22	0.34	0.24	0.25
120 dimensions						
F1	**0.94**	0.86	0.52	0.61	0.55	0.63
Acc.	**0.94**	0.72	0.68	0.71	0.58	0.62
E4SC	**0.87**	0.64	0.18	0.23	0.21	0.19

Table 1: Clustering results on synthetic data sets

6.2 Synthetic data: cluster and outlier mining.

Based on the method described in [Müller *et al.*, 2011], we generate synthetic data sets with 5120 data points and 20, 40, 80, and 120 dimensions. Each data set contains subspace clusters embedded in randomly chosen 2-6 dimensional subspaces and 120 outliers deviating from these clusters.

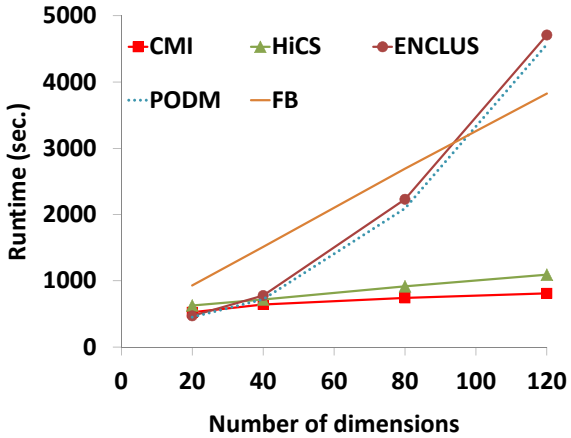
Quality for outlier mining. The quality of subspaces is evaluated by inspecting how the selected subspaces enhance outlier detection compared to LOF in the full space. The results are shown as Figure 2(b). Overall, *CMI* outperforms the competing techniques and is stable with increasing dimensionality. The performance of LOF degrades with increasing dimensionality of data. Similarly, FB [Lazarevic and Kumar, 2005] is affected by random choice of low contrast projections. The pairwise method PW [Reshef *et al.*, 2011] and PCA show worst performance, due to their inability to measure contrast in multi-dimensional subspaces. As subsequent evaluation confirmed this trend, we exclude PW and PCA in the experiments below.

Quality for clustering. Here, subspace quality is assessed by clustering results. DBSCAN is used as the baseline method. Furthermore, for all methods tested, we reduced redundancy in clustering output [Assent *et al.*, 2007]. The results in Table 1 show that *CMI* achieves best quality and best scalability for increasing dimensionality. High E4SC values of *CMI* indicate that it performs well in selecting subspaces containing clusters and outliers.

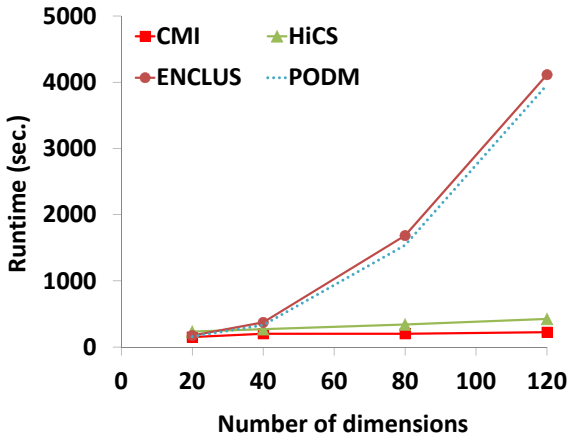
Runtime vs. Dimensionality. Besides accuracy, we are also interested in scalability w.r.t. runtime. In this experiment, previous synthetic data sets are reused. Since the tendency of all methods is similar in both outlier detection and clustering, we only present the runtime for outlier detection. We display in Figure 3(a) the total time for completing the task, i.e., time for mining subspaces (cf., Figure 3(b)) and time for outliers mining. We can see that *CMI*

scales better than our competitors.

Although *FB* does not spend much time for mining high contrast subspaces, it clearly suffers from high overall runtimes. This is due to high dimensional subspaces that have low contrast, and hence, induce the costly detection of many false alarms. ENCLUS and PODM also scale badly as their contrast measures are inefficient in terms of time complexity. Since *CMI* prunes low contrast subspaces better than HiCS, it can avoid exploring many high dimensional subspaces. In conclusion, *CMI* is faster than all tested approaches and yields higher accuracy.



(a) total runtime



(b) subspace search only

Figure 3: Runtime w.r.t. dimensionality

6.3 Evaluation on real world data.

All real world databases used in our experiments are from the UCI Machine Learning Repository [Frank and Asuncion, 2010] and have been used as benchmarks in recent publications [Lazarevic and Kumar, 2005; Müller *et al.*, 2009a; Müller *et al.*, 2009b; Keller *et al.*, 2012].

Quality for outlier mining. We evaluate the performance of all subspace search methods with outlier detection on real world data. We perform experiments on 9 benchmark datasets, using the minority class as ground truth for the outlier evaluation. In some of these data sets (e.g., *Pendigits*) all classes have identical support and we down-sample one class to 10% of its original size, which is a commonly used procedure in outlier evaluation [Lazarevic and Kumar, 2005; Keller *et al.*, 2012; Müller *et al.*, 2011]. The results in Table 2 show that *CMI* achieves the best AUC in all data sets. In addition, we show the runtimes in Table 3. Overall, our method provides the best quality enhancement for LOF.

Dataset	CMI	HiCS	Enclus	Podm	LOF	FB
Thyroid	**0.96**	0.95	0.94	0.91	0.86	0.93
WBCD	**0.95**	0.94	0.94	0.87	0.87	0.87
Diabetes	**0.73**	0.72	0.71	0.69	0.71	0.72
Glass	**0.82**	0.80	0.80	0.78	0.77	0.78
Ion	**0.83**	0.82	0.82	0.78	0.78	0.79
Pendigits	**0.98**	0.95	0.94	0.86	0.94	0.93
Segment	**0.94**	0.84	0.88	0.89	0.76	0.86
Lympho	**0.95**	0.86	0.67	0.67	**0.95**	**0.95**
Madelon	**0.60**	0.59	0.51	0.56	0.59	0.59

Table 2: Outlier mining: AUC on real world data

Dataset	CMI	HiCS	Enclus	Podm	FB
Thyroid	**17.33**	27.54	49.32	48.11	53.60
WBCD	**16.42**	17.11	33.63	34.55	24.49
Diabetes	**1.74**	1.80	4.74	4.63	5.56
Glass	**0.24**	**0.24**	0.27	0.26	0.27
Ion	**6.01**	6.19	7.31	7.19	8.07
Pendigits	**1368.23**	1616.96	2153.09	2094.36	1854.56
Segment	**101.23**	107.99	225.46	218.34	150.80
Lympho	**4.10**	6.08	6.37	5.79	5.31
Madelon	**23.45**	25.82	315.22	304.57	232.48

Table 3: Runtime (in seconds) for outlier detection

Quality for clustering. As we show in Table 4, *CMI* provides also the best quality improvement w.r.t. clustering. It outperforms traditional full space DBSCAN and existing subspace search methods that fail to identify clusters due to scattered subspace projections. In contrast to the competing approaches, we achieve a clear quality enhancement for both subspace clustering and subspace outlier detection.

	CMI	HiCS	Enclus	Podm	DBScan	FB
Wisconsin Breast Cancer						
F1	**0.79**	0.75	0.44	0.40	0.73	0.60
Acc.	**0.77**	0.72	0.69	0.67	0.71	0.69
E4SC	**0.76**	0.70	0.53	0.49	0.67	0.59
Shape						
F1	**0.82**	0.77	0.76	0.74	0.55	0.76
Acc.	**0.84**	0.78	0.66	0.69	0.34	0.41
E4SC	**0.71**	0.64	0.58	0.63	0.38	0.44
Pendigits						
F1	**0.73**	0.55	0.50	0.51	0.52	0.63
Acc.	**0.81**	0.75	0.66	0.64	0.68	0.77
E4SC	**0.68**	0.54	0.56	0.55	0.52	0.53
Diabetes						
F1	**0.71**	0.53	0.25	0.15	0.52	0.58
Acc.	**0.76**	0.66	0.67	0.63	0.68	0.70
E4SC	**0.65**	0.34	0.11	0.07	0.52	0.52
Glass						
F1	**0.59**	0.37	0.26	0.29	0.32	0.42
Acc.	**0.68**	0.54	0.52	0.55	0.32	0.44
E4SC	**0.52**	0.40	0.35	0.38	0.24	0.28

Table 4: Clustering: Quality on real world data

7 Conclusions

We proposed CMI, a new contrast measure for multi-dimensional data. It is based on cumulative entropy of subspaces and does not require data discretization. Furthermore, it is not restricted to pairwise analysis, captures mutual dependency among dimensions, and scales well with increasing subspace dimensionality. Overall, it is more accurate and more efficient than previous subspace search methods. Experiments on various real world databases show that CMI provides improvement for both cluster and outlier detection.

Acknowledgments

This work is supported by the German Research Foundation (DFG) within GRK 1194, by the YIG program of KIT as part of the German Excellence Initiative. And by two Post-Doctoral Fellowships of the Research Foundation – Flanders (FWO).

References

[Aggarwal and Yu, 2001] Charu C. Aggarwal and Philip S. Yu. Outlier detection for high dimensional data. In *SIGMOD*, 2001.

[Aggarwal *et al.*, 1999] Charu C. Aggarwal, Cecilia Magdalena Procopiuc, Joel L. Wolf, Philip S. Yu, and Jong Soo Park. Fast algorithms for projected clustering. In *SIGMOD*, 1999.

[Agrawal *et al.*, 1998] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD*, 1998.

[Assent *et al.*, 2007] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. DUSC: Dimensionality unbiased subspace clustering. In *ICDM*, 2007.

[Baumgartner *et al.*, 2004] Christian Baumgartner, Claudia Plant, Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. Subspace selection for clustering high-dimensional data. In *ICDM*, 2004.

[Beyer *et al.*, 1999] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *ICDT*, 1999.

[Breunig *et al.*, 2000] Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander Raymond T. Ng. LOF: Identifying density-based local outliers. In *SIGMOD*, 2000.

[Chanda *et al.*, 2010] Pritam Chanda, Jianmei Yang, Aidong Zhang, and Murali Ramanathan. On mining statistically significant attribute association information. In *SDM*, 2010.

[Cheng *et al.*, 1999] Chun Hung Cheng, Ada Wai-Chee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In *KDD*, 1999.

[Crescenzo and Longobardi, 2009] Antonio Di Crescenzo and Maria Longobardi. On cumulative entropies. *J. Statist. Plann. Inference*, 139, 2009.

[Dy and Brodley, 2004] Jennifer G. Dy and Carla E. Brodley. Feature selection for unsupervised learning. *JMLR*, 5, 2004.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[Frank and Asuncion, 2010] A. Frank and A. Asuncion. UCI machine learning repository [http://archive.ics.uci.edu/ml], 2010.

[Günnemann *et al.*, 2011] Stephan Günnemann, Ines Färber, Emmanuel Müller, Ira Assent, and Thomas Seidl. External evaluation measures for subspace clustering. In *CIKM*, 2011.

[Keller *et al.*, 2012] Fabian Keller, Emmanuel Müller, and Klemens Böhm. HiCS: High contrast subspaces for density-based outlier ranking. In *ICDE*, 2012.

[Kriegel *et al.*, 2009] Hans-Peter Kriegel, Erich Schubert, Arthur Zimek, and Peer Kröger. Outlier detection in axis-parallel subspaces of high dimensional data. In *PAKDD*, 2009.

[Law *et al.*, 2004] Martin H. C. Law, Mário A. T. Figueiredo, and Anil K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9), 2004.

[Lazarevic and Kumar, 2005] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *KDD*, 2005.

[Lee and Verleysen, 2007] J Lee and M Verleysen. *Nonlinear Dimensionality Reduction*. Springer, New York, 2007.

[Müller *et al.*, 2009a] Emmanuel Müller, Ira Assent, Stephan Günnemann, Ralph Krieger, and Thomas Seidl. Relevant subspace clustering: Mining the most interesting non-redundant concepts in high dimensional data. In *ICDM*, 2009.

[Müller *et al.*, 2009b] Emmanuel Müller, Stephan Günnemann, Ira Assent, and Thomas Seidl. Evaluating clustering in subspace projections of high dimensional data. *PVLDB*, 2(1), 2009.

[Müller *et al.*, 2011] Emmanuel Müller, Matthias Schiffer, and Thomas Seidl. Statistical selection of relevant subspace projections for outlier ranking. In *ICDE*, 2011.

[Nguyen *et al.*, 2013] Hoang Vu Nguyen, Emmanuel Müller, Jilles Vreeken, Fabian Keller, and Klemens Böhm. CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In *SDM*, 2013.

[Ordonez and Omiecinski, 2004] Carlos Ordonez and Edward Omiecinski. Efficient disk-based K-means clustering for relational databases. *IEEE Trans. Knowl. Data Eng.*, 16(8), 2004.

[Reshef *et al.*, 2011] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062), 2011.

[Roth and Lange, 2003] Volker Roth and Tilman Lange. Feature selection in clustering problems. In *NIPS*, 2003.

[Sequeira and Zaki, 2004] Karlton Sequeira and Mohammed Javeed Zaki. SCHISM: A new approach for interesting subspace mining. In *ICDM*, 2004.

[Ye *et al.*, 2009] Mao Ye, Xue Li, and Maria E. Orlowska. Projected outlier detection in high-dimensional mixed-attributes data set. *Expert Syst. Appl.*, 36(3), 2009.

[Zhang *et al.*, 2008] Xiang Zhang, Feng Pan, Wei Wang, and Andrew B. Nobel. Mining non-redundant high order correlations in binary data. *PVLDB*, 1(1), 2008.

Difference-based Estimates for Generalization-aware Subgroup Discovery

Extended abstract of a paper originally published on ECML/PKDD 2013

Florian Lemmerich, Martin Becker, and Frank Puppe
University of Würzburg
{lemmerich, becker, puppe}@informatik.uni-wuerzburg.de

Abstract

In this work, we approach the topic of efficient subgroup mining with interestingness measures, which also take statistics on generalizations of the subgroup into account. For this setting we develop new optimistic estimate bounds, which allow to safely prune large parts of the search space. In contrast to previous approaches, the bounds are not only based on the anti-monotonicity of the number of covered instances of a pattern, but also on the number of instances, by which a pattern differs in comparison to its generalizations. Incorporating these bounds in an efficient algorithms leads to runtime improvements of up to an order of magnitude.

1 Problem Setting

Subgroup discovery [5] is a key technique for descriptive data mining. It aims at identifying descriptions of subsets of the data that show an interesting behavior with respect to a certain target concept. This is accomplished by using an interestingness measure to assign a score to all candidate patterns in the search space of all conjunctive descriptions. Traditional measures are based on the statistics of the evaluated subgroup and the entire dataset. In particular, the most popular family of interestingness measures weights between the number of instances covered by the subgroup, and the difference of the target share (or target mean value in a numeric target setting) in the subgroup a the target share in the total population. In recent research [1; 2; 3] these measure have been adapted to obtain more interesting and less redundant results: Generalization-aware measures replace the comparison with the target share (mean value) in the overall dataset with a comparison to the maximum target share of all generalizations of the subgroup. E.g., to compute the interestingness score of the subgroup $A \wedge B$, the target share for the three subgroup patterns $\emptyset$, A, B are compared to the target share of $A \wedge B$. In this paper, we focus on the most important families of interestingness measures for nominal and numeric target concepts in this direction:

$$r^a_{bin}(P) = i^a_P \cdot (\tau_P - \max_{H \subset P} \tau_H), a \in [0;1]$$

$$r^a_{num}(P) = i^a_P \cdot (\mu_P - \max_{H \subset P} \mu_H), a \in [0;1]$$

Here, i_P is the number of instances covered by the subgroup P, $\tau_P(\mu_P)$ is the target share (target mean value) in the subgroup P and $\max_{H \subset P} \tau_H (\max_{H \subset P} \mu_H)$ is the maximum target share (target mean value) in all generalizations of P. a is an user-specified parameter that allows to weight between the two factors.

This paper does not argue about the usefulness of these adaptations, but focuses on efficient subgroup mining in this setting. In particular, we propose novel, tighter *optimistic estimate bounds* [5] that allow to prune parts of the search space without losing the optimality of the results: The basic idea of optimistic estimates the following: if one can guarantee that no specialization of the currently evaluated pattern will have an interestingness score which is good enough to include the respective pattern into the result set then we can safely omit these patterns from the search. In this regard, we aim at the strictest bounds possible to reduce the remaining search space and thus to speed up the search process.

2 Difference-based estimates

Previous approaches to compute optimstic estimates are almost exclusively based on the anti-monotonicity of covered (positive) instances: For instance, if the subgroup A covers 10 positive examples, then each specialization of A, e.g., $A \wedge X$ covers also at most 10 positive examples. In addition to the statistics of the currently evaluated subgroup, our approach also takes into account statistics of generalizations in order to to determine the interestingness score. This additional information is used to determine tighter optimistic estimates.

For this end, the following lemma is proposed:

Lemma. *Let $P = A \wedge B$ be any pattern with A, B potentially being a conjunction of patterns themselves and $B \neq \emptyset$. Then for any specialization $S \supset P$ there exists a generalization $\gamma(S) \subset S$, such that $\Delta(\gamma(S), S) \subseteq \Delta(A, B)$.*

The lemma exploits, what can be described as an *anti-monotonicity of differences* in comparison to generalizations. For example, assume there are 5 instances, which are covered by U, but not by $U \wedge V$. Then the lemma guarantees, that for each specialization $S = U \wedge V \wedge X \wedge \ldots \wedge Y$ there exists a generalization, such that the difference between this generalization and S is also at most 5 instances (cf. also [4]).

Now, consider the interestingness score of such a specialization S: If S covers only few instances, then by the definition of the used interestingness measures, S is of low interestingness. On the other hand, if S covers more instances, the increase of the target share

d	3		4		5		6	
pruning	dpb	std	dpb	std	dpb	std	dpb	std
adults	1.0	1.1	0.9	1.8	1.6	8.1	1.7	30.2
audiology	0.1	0.1	0.1	2.8	0.6	51.7	-	-
census-kdd	17.9	20.6	37.2	99.8	107.9	2954.3	267.5	-
colic	0.1	0.2	0.3	1.1	0.4	5.1	0.4	16.4
credit-a	0.1	0.1	0.3	0.7	1.2	3.6	1.2	12.9
credit-g	0.2	0.2	1.5	4.0	4.0	35.2	7.0	-
diabetes	0.1	0.1	0.5	1.3	1.2	9.3	2.0	67.1
hepatitis	<0.1	0.1	0.2	0.6	0.8	3.3	0.3	11.9
hypothyroid	0.1	0.2	0.5	2.7	1.7	39.0	-	-
spammer	1.3	1.6	5.7	15.5	29.3	172.2	88.3	-
vehicle	1.0	1.3	4.8	57.8	15.6	-	-	-

Table 1: Runtime comparison (in s) of the base algorithm with traditional pruning based on the positives (std) and the novel algorithm with additional difference-based pruning (dbp) using different maximum numbers d of describing selectors in a pattern. As quality functions the generalization-aware mean test $r^{0.5}_{bin}$ was used. "-" indicates that the algorithm did not finish due to lack of memory.

in comparison to its generalization $\gamma(S)$ is limited by the lemma, since it states that $\gamma(S)$ only covers at most 5 more negative instances than S. As a consequence S is also not interesting in this case.

These considerations are exploited in formal theorems, which allow to determine optimistic estimates based on the difference of instances in generalizations:

Theorem. *Consider the pattern P with p_P positive instances. $P' \subseteq P$ is either P itself or one of its generalizations and $P'' \subset P'$ a generalization of P'. Let $n_\Delta = n_{P''} - n_{P'}$ be the difference in coverage of negative instances between these patterns. Then, an optimistic estimate of P for r^a_{bin} is given by:*

$$oe_{r^a_{bin}}(P) = \begin{cases} \frac{p_P \cdot n_\Delta}{p_P + n_\Delta}, & \text{if } a = 1 \\ \frac{n_\Delta}{1+n_\Delta}, & \text{if } a = 0 \\ \frac{\hat{p}^a \cdot n_\Delta}{\hat{p}+n_\Delta}, \text{ with } \hat{p} = \min(\frac{a \cdot n_\Delta}{1-a}, p_P), & \text{else} \end{cases}$$

This theorem provides optimistic estimate bounds, which are tight (low), if either (1) the number positives covered by a subgroup is low, or (2) if the difference of negatives between the subgroup and a generalization is low, or (3) if the difference of negatives between a generalization of the subgroup and another generalization of this generalization is low.

Another theorem (not shown in this abstract) introduces optimistic estimate bounds for the setting with a numeric target setting and mean-based generalization-aware interestingness measures r^a_{num}. These bounds also exploit the difference of the minimum target value removed in a specialization step to the maximum target value remaining in the subgroup.

3 Algorithm

Although the proposed optimistic estimate bounds can in principal be applied with any search strategy, we focus in this work on adapting Apriori, which is also employed by the current state-of-the-art algorithm of this problem setting [1]. For each candidate pattern additional information is stored, e.g., the minimum number of negatives in a generalization, the minimum difference in coverage between two generalizations and the maximum target share in a generalization of this pattern. The information is propagated efficiently during candidate generation and updated during the evaluation of the subgroup.

4 Evaluations

The effectiveness of the difference-based optimistic estimate bounds and its incorporation in an algorithm was evaluated in several series of experiments. Exemplary results are shown in Table 1. It can be observed that the novel approach improves the runtime often of more than an order of magnitude. Further investigation showed that the runtime improvement is particularly large, if the dataset contains many selectors that cover large parts of the dataset (see e.g., the audiology dataset). In can also be seen, that out-of-memory errors occur less often using the improved bounds, since less candidates are generated.

The full paper includes formal proofs, a more detailed algorithm description and more experimental results. It has been published as: Florian Lemmerich, Martin Becker, Frank Puppe: Difference-Based Estimates for Generalization-Aware Subgroup Discovery. In: Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, Filip Zelezný (Eds.): Proceedings of ECML/PKDD 2013, Part III, pages 288-303.

References

[1] Batal, I., Hauskrecht, M.: A concise representation of association rules using minimal predictive rules. Machine Learning and Knowledge Disc. pp. 87–102 (2010)

[2] Grosskreutz, H., Boley, M., Krause-Traudes, M.: Subgroup discovery for election analysis: a case study in descriptive data mining. Disc. Science pp. 57–71 (2010)

[3] Lemmerich, F., Puppe, F.: Local Models for Expectation-Driven Subgroup Discovery. 2011 IEEE 11th International Conference on Data Mining pp. 360–369 (2011)

[4] Webb, G.I., Zhang, S.: Removing trivial associations in association rule discovery. Proceedings of the First International NAISO Congress on Autonomous Intelligent Systems, p. NAISO Academic Press: Geelong, 2002 (2002)

[5] Wrobel, S.: An algorithm for multi-relational discovery of subgroups. Principles of Data Mining and Knowledge Discovery (1997)

Towards Distributed Emerging Pattern Mining on Itemset Streams

Björn Jacobs and Henrik Grosskreutz
Fraunhofer IAIS
{firstname.lastname}@iais.fraunhofer.de
Schloss Birlinghoven
53754 Sankt Augustin

Abstract

In this paper, we present a new approach for mining emerging patterns in itemset streams. Unlike in the classical setting, we do not assume that two datasets are given and that the task is to find itemsets occurring more often in one dataset than in the other. Instead, we look for itemsets that were seldom in the stream previously, but recently occur much more often. Our approach differs from earlier approaches in that it employs a distributed representation of the candidate space, which is scattered over a cluster of machines. This paper describes work in progress, meaning that it concentrates on the approach but lacks experimental evaluation.

1 Introduction

In this paper, we present a new approach to perform *Emerging Pattern Mining* on *data streams*. The task of emerging pattern mining [Dong and Li, 1999] considers the classical itemset setting, where the data to be analyzed takes the form of sets of items. A classical example is market basket analysis, where every data record, or "transaction", consists of the set of items bought together by a particular customer. Another example is Twitter, where a transaction corresponds to an individual tweet and the items correspond to the words within that tweet.

The goal of emerging pattern mining is to "capture significant changes between datasets", and to "capture emerging trends in business or demographic data" [Dong and Li, 1999]. In the market basket example, the emerging patterns would hence be sets of items that are often bought together in one dataset (for example, in recent sales data), but were only seldom bought in another dataset (older sales data). In the Twitter example, the emerging patterns would be words that co-occur often in one dataset (recent tweets), but seldom in another dataset (older tweets).

While classical algorithms operate on a static dataset, we consider the setting where the data takes the form of a stream. That is, we assume that the input is a potentially infinite sequence of transactions – as in the twitter example. Our goal is then to find, at any time, the current set of emerging patterns, that is, the itemsets that occur much more frequently in recent transactions than in older transactions. Ultimately, this allows identifying new emerging trends or topics.

Our approach is based on the idea to keep, in memory, a representation of the whole space of candidate patterns. With every candidate pattern, we store a set of statistics, which are updated whenever a new transaction comes in. As the space of candidate patterns can be extremely large, we aim for a *distributed computation approach*, where the candidate space (and the computation) can be distributed on a cluster of machines. The main intention of this approach is not to speed up the process (which nonetheless can be a positive side-effect), but to scale out horizontal in order to realize very large candidate spaces. To this end, we base our implementation a framework supporting the distribution of data and computation to multiple nodes - in our case the Akka toolkit.

This paper presents work-in-progress. That is, it describes the problem and the intended approach, but it lacks conclusive results. The finalization of the implementation and the evaluation of the system are the topic of an ongoing master thesis. Nevertheless, this paper already provides important contributions. In particular:

- We present our approach and show that its memory requirements are only logarithmic in the number of incoming transactions and linear in the size of the candidate pattern space (Section 4.1);
- We describe how our approach can be distributed on a cluster of computing nodes, which allows dealing with very large candidate spaces (Section 4.2).

The remainder of this paper is structured as follows: After a brief discussion of related work in Section 2 and a review of the task and the standard approaches in Section 3, we present our new approach in Section 4. Subsequently, we describe our prototypical implementation in Section 5, before we conclude in Section 6.

2 Related work

In recent years, a lot of research has been done to deal with change in data streams (also called *concept drift*). A large share of the proposed approaches, however, do not fit to our setting because they make the assumption that the incoming data streams contain a *label* and their goal is to find a classifier for future data (e.g. [Alhammady and Ramamohanarao, 2005; Wang *et al.*, 2005]).

An approach very similar to ours is the work of [Kifer *et al.*, 2004], who introduce a meta algorithm based on a 2-window-approach on unlabeled data streams. The algorithm measures the distance between the probability distributions of the items in each window and announces change if it lies above a certain threshold. The main difference between this approaches and ours is that we distribute the task over a cluster of machines, exploiting distributed memory to deal with very large candidate spaces, and using parallel computation for speed-up. This also distinguishes our approach from the numerous approaches to the related task of

itemset mining over data streams [Cheng *et al.*, 2008]. Another approach for parallel item-set mining is the work of [Li *et al.*, 2008]. It differs from our approach in two central ways. First, the algorithm is not designed to work on data streams and second, it only implicitly represents the search space by distributing the database, where our approach represents the search space explicitly.

3 Preliminaries

In this section, we provide a formal definition of the task of emerging pattern mining, and describe existing approaches.

3.1 Emerging Patterns and Supervised Descriptive Rule Discovery

Emerging pattern mining [Dong and Li, 1999] belongs to a family of tasks known as *supervised descriptive rule discovery* [Kralj Novak *et al.*, 2009], which also includes the tasks of subgroup discovery and contrast set mining. The input consists of a sequence of transactions $T_1, \ldots, T_m$. Every transaction consists of a set of items, i.e. $T_i = (i_{i,1}, \ldots, i_{i,N_i})$ where every item stems from a fixed universe of items.

A pattern P is also a subset of items, i.e. $P = i_1, \ldots, i_n$. We say that a transaction T_i *contains* a pattern P if and only if $P \subseteq T_i$. The (relative) *support* of a pattern P in a sequence of transactions $DB = T_1, \ldots, T_m$, denoted by $supp(P, DB)$ refers to the share of transactions in the sequence containing P.

In the classical setting, one is given two datasets DB_1 and DB_2, and the goal is to find patterns that have a noticeably higher support in DB_2 than in DB_1. This difference in support is measured using some *quality function*, which assigns a real-valued figure to any given pattern. The higher the figure, the more salient the difference in supports is considered. Different quality functions have been proposed in the supervised descriptive rule mining community. One example is the weighted relative accuracy (WRACC) [Lavrac *et al.*, 2004], defined as follows:

$$WRACC(P, DB_1, DB_2) = supp(P, DB_2) \times (1 - p_0) - supp(P, DB_1) \times p_0$$

Here p_0 is defined as $\frac{|DB_2|}{|DB_1 \cup DB_2|}$.

Based on these definitions, the task of *top-k emerging pattern mining* is to find the k patterns having highest quality (or, in case of ties, a set of k maximum-quality patterns).

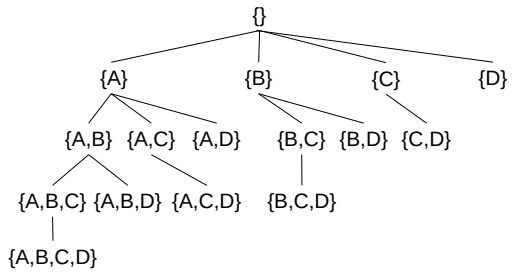
3.2 The Classical Approach To Supervised Descriptive Rule Discovery

The classical computational approach to supervised descriptive rule discovery is to load the whole dataset into memory and to traverse the search space of candidate patterns. Figure 1 will help illustrating this approach: it shows an example built from the 4 items A, B, C and D. Figure 1(a) shows the dataset and the corresponding search space is visualized in Figure 1(b). In this figure, the candidate patterns are arranged in levels, where every pattern in a level has the same cardinality, i.e. is built from the same number of items.

The different supervised descriptive rule discovery algorithms explore the search space in different ways. Some approaches apply some heuristic search and only explore a subset of the search space [Lavrac *et al.*, 2004], while other approaches exhaustively traverse the complete search space, e.g. relying on some tree traversal algorithm

Transactions
A
AC
B
BD
ABCD
C
AC
BC
...

(a) dataset



(b) candidate space

Figure 1: An example dataset and its corresponding candidate space

[Grosskreutz *et al.*, 2008]. Whenever a node is visited, the algorithm computes statistics about the pattern, in particular the support, which allows evaluating the quality of that candidate. The algorithms keeps track of the highest-quality patterns during traversal. Once the traversal ends, the top-quality pattern(s) are returned as result.

From the perspective of this paper, the details of these approaches are less important than the fact that for every candidate, these approaches have to compute statistics based on the *whole sequence of itemsets*. Typically, the algorithms will keep the dataset in memory to speedup this task, possibly relying on some efficient data structures [Han *et al.*, 2000; Atzmüller and Puppe, 2006]. This approach is, however, not applicable with *streams* of potentially infinite length.

3.3 The Sampling Approach of Scheffer and Wrobel

A completely different approach was proposed by Scheffer and Wrobel [Scheffer and Wrobel, 2002]. The main goal of this approach was to reduce the computation time by trading the exact-solution-guarantee for probabilistic guarantees with fixed bounds on confidence and error. To this end, the paper proposes a randomized algorithm which iteratively (1) draws a sample record, and (2) uses that sample to update the statistics of *all* candidate patterns. Once the algorithm can guarantee, with sufficiently high probability, that a candidate will have low or high quality, it is discarded respectively accepted at runtime. The iterations continue until with sufficiently high probability the k best-quality patterns are found.

4 Mining Emerging Patterns in Streams

We will first specify the task we consider, i.e. mining emerging patterns in the setting of data stream: The input consists of

$[t_0$ - $t_1)$	$[t_1$ - $t_2)$	$[t_2$ - $t_3)$	...	$[t_{m-1}$ - $t_m)$
5	8	13	...	27

Figure 2: Statistics stored for every candidate pattern

- an itemset stream, that is a (potentially infinite) sequence of pairs (T, t), where T is a transaction and t a timestamp. We assume the timestamps to be monotonically increasing.
- an integer k, and two timeframes W_{old} and W_{new}.

The output is a mapping, from every time t to the emerging patterns wrt. $DB_{new}(t, W_{new})$ and $DB_{old}(t, W_{new}, W_{old})$. Here, $DB_{new}(t, W_{new})$ consists of all "new" transactions having a timestamp in the interval $(t - W_{new}, t]$, and $DB_{new}(t, W_{new})$ consists of all "old" transaction occurring in the interval $(t - W_{new} - W_{old}, t - W_{new}]$.

4.1 Our Approach

Our approach is based on the idea of [Scheffer and Wrobel, 2002], namely to store a representation of the whole search space, together with statistics about the support of every candidate pattern. Unlike them, however, we are not concerned with probabilistic guarantees, as we do not aim for a randomized algorithm.

Another difference is that here, we are concerned with streaming data, and that our goal is to compute the emerging patterns with respect to two time windows. To this end, we do not just store, for every pattern, the number of occurrences in two datasets, but instead have to store all information required to continuously calculate the support in the two windows.

As storing all incoming transactions together with their timestamp would result in costs which are at least linear in the number of incoming transactions, instead we only store a discretized representation, which only stores the number of occurrences within m smaller time windows. Figure 2 illustrates the idea. In the first time window, the pattern occurred 5 times; in the second, it occurred 8 times, etc. As we are only interest in the frequency of patterns in the interval $(t - W_{new} - W_{old}, t]$, our representation can discard old time windows, which ensures that the size of the representation is limited.

The overall memory requirements for every entry in this data structure is bounded logarithmically in the number of incoming transactions. Overall, this approach has worst-case memory requirements bounded by $O(\log(|\mathcal{T}|) \cdot |\mathcal{P}| \cdot m)$, where $|\mathcal{T}|$ denotes the maximum number of incoming transactions within a timeframe of $(W_{old} + W_{new})$, m denotes the number of time windows used to cover the interval $(t - W_{new} - W_{old}, t]$, and $|\mathcal{P}|$ denotes the size of the candidate pattern space.

4.2 Distributed Computation

One issue with the approach to store a representation of the search space is that this can result in large memory requirements. It is obvious that the size of the space search is exponential in the number of items, and that hence its representation may exceed the main memory of a single machine. For this reason, it would be desirable to have a *distributed* algorithm that can run on a *cluster* of machines. In particular, we aim at an algorithm that can (almost) uniformly distribute the representation of the search space on any given cluster of N machines.

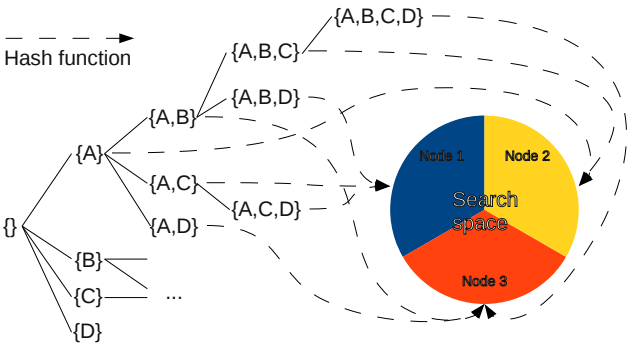


Figure 3: Distribution of the Candidate Space on a Cluster of Machines

Our solution for splitting the search space is illustrated in Figure 3 and works as follows: On the left, the figure shows the candidate pattern space, while on the right the different machines in the cluster are presented. Every pattern from the candidate pattern space is mapped to one of the nodes in the cluster by a hash function. The hash function deterministically determines the target node for each candidate pattern. In the given example the cluster consists of three separate nodes where each node is responsible for storing the statistics of a subset of all patterns. Together, the nodes represent the entire search space. Whenever a certain candidate pattern needs to be updated, the algorithm can quickly identify the responsible node which subsequently updates the pattern's statistics in the search space.

5 Implementation

We will now provide a overview of our implementation, which is based on the Akka toolkit.

5.1 Akka

The Akka toolkit (http://akka.io/) provides a high-level framework for developing distributed applications following the actor model. The user can create actors that implement different behavior. These actors can be distributed to remote nodes and communicate with each other by passing messages. With this, the user can create an topology of data-source and compute nodes, define their relations and the functions that are applied on the data.

Akka was chosen over other frameworks like Hadoop or Storm for different reasons. In contrast to Strom, Akka proved to have a more mature code base and documentation. Hadoop's focus on the other hand lies more on batch data processing than on streaming data processing.

5.2 A topology for Emerging Pattern Mining

Figure 4 describes the central topology of our system for mining Emerging Patterns. The control flow goes from left to right. The leftmost node represents a source of transactions. Regarding the Twitter example, this node would emit a flow of tweets. A transaction dispatcher node connects to the data stream and distributes the received transactions to a set of candidate pattern generator nodes by a broadcast. Every generator node receives the transaction *T*and subsequently creates a set of patterns that consists of all those patterns that are affected by the transaction and are element of its partition of the candidate space.

To control the amount of memory and computation required, we additionally restrict the cardinality of the patterns considered. This is a standard approach in supervised descriptive rule discovery (e.g. [Grosskreutz *et al.*,

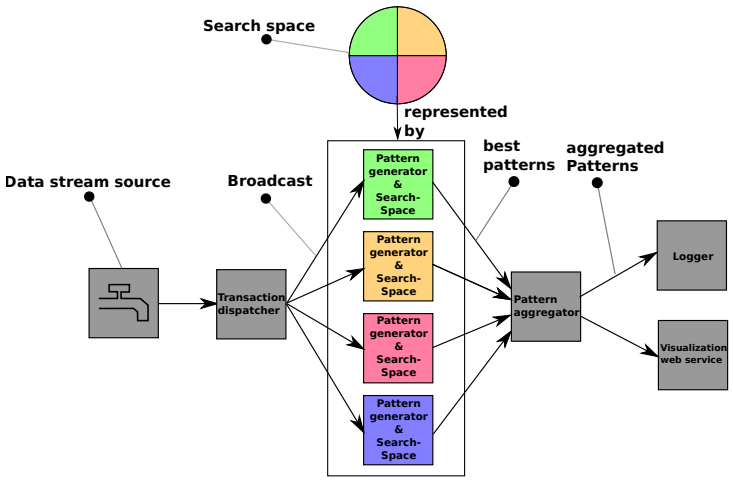


Figure 4: Topology of the system for performing Emerging Patterns Mining

2008]). Lets say we have a set of items in a transaction $T = \{i_1, \ldots, i_n\}$ and an integer value c denoting the maximal cardinality of the patterns to be considered. Then the power-set with limited cardinality c of that transaction is defined as follows:

$$P_c(T) = \{s \subseteq T : |s| < c\}$$

The generator has two central functions. It is responsible for generating patterns and for keeping track of the patterns' statistics. Following to the generation, it updates the statistics in an local data structure.

The node periodically calculates which of the patterns it takes care of have the highest quality value. It transfers this list of best patterns to a subsequent node that receives these lists from all nodes, creating an aggregated list of all global patterns. The two rightmost nodes represent example applications for the output, like logging to a file or presenting the result in a web page.

6 Discussion

In this paper, we have motivated a variation of the task of emerging pattern mining, which operates on data streams. The idea is to search for itemsets which occur more often in recent transactions than in older transactions. We have presented a new approach to this task, which allows distributing the computation and the representation of the candidate pattern space on a cluster of machines. This distributed representation is arguably the most distinguishing feature of our approach, compared to existing approaches to related tasks like frequent itemset mining on streams.

This paper presents work in progress and much remains to be done. In particular, we still lack an empirical evaluation of the approach. This will be the topic of an on-going master thesis, which will answer questions about the scalability of the approach and possible limitations.

Acknowledgments

This publication has been produced in the context of the EU Collaborative Project P-Medicine, which is funded by the European Commission under the contract ICT-2009-6-270089.

References

[Alhammady and Ramamohanarao, 2005] Hamad Alhammady and Kotagiri Ramamohanarao. Mining emerging patterns and classification in data streams. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, 2005.

[Atzmüller and Puppe, 2006] Martin Atzmüller and Frank Puppe. SD-map - a fast algorithm for exhaustive subgroup discovery. In *PKDD*, pages 6–17, 2006.

[Cheng *et al.*, 2008] James Cheng, Yiping Ke, and Wilfred Ng. A survey on algorithms for mining frequent itemsets over data streams. *Knowl. Inf. Syst.*, 16(1):1–27, 2008.

[Dong and Li, 1999] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. pages 43–52, 1999.

[Grosskreutz *et al.*, 2008] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. Tight optimistic estimates for fast subgroup discovery. In *ECML/PKDD (1)*, pages 440–456, 2008.

[Han *et al.*, 2000] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *SIGMOD Conference*, pages 1–12, 2000.

[Kifer *et al.*, 2004] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. VLDB '04, pages 180–191. VLDB Endowment, 2004.

[Kralj Novak *et al.*, 2009] Petra Kralj Novak, Nada Lavrač, and Geoffrey I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.

[Lavrac *et al.*, 2004] N. Lavrac, B. Kavsek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5(Feb), 2004.

[Li *et al.*, 2008] Haoyuan Li, Yi Wang, Dong Zhang, Ming Zhang, and Edward Y Chang. Pfp: parallel fp-growth for query recommendation. In *ACM conference on Recommender systems*. ACM, 2008.

[Scheffer and Wrobel, 2002] Tobias Scheffer and Stefan Wrobel. Finding the most interesting patterns in a database quickly by using sequential sampling. *Journal of Machine Learning Research*, 3:833–862, 2002.

[Wang *et al.*, 2005] Peng Wang, Haixun Wang, Xiaochen Wu, Wei Wang, and Baile Shi. On reducing classifier granularity in mining concept-drifting data streams. ICDM '05. IEEE Computer Society, 2005.

Spatio-Temporal Random Fields: Compressible Representation and Distributed Estimation

Nico Piatkowski, Sangkyun Lee and Katharina Morik
TU Dortmund University, Dortmund 44227, Germany
{nico.piatkowski,sangkyun.lee,katharina.morik}@tu-dortmund.de

Abstract

Modern sensing technology allows us enhanced monitoring of dynamic activities in business, traffic, and home, just to name a few. The increasing amount of sensor measurements, however, brings us the challenge for efficient data analysis. This is especially true when sensing targets can interoperate – in such cases we need learning models that can capture the relations of sensors, possibly without collecting or exchanging all data. Generative graphical models namely the Markov random fields (MRFs) fit this purpose, which can represent complex spatial and temporal relations among sensors, producing interpretable answers in terms of probability. The only drawback will be the cost for inference, storing and optimizing a very large number of parameters – not uncommon when we apply them for real-world applications.

In this paper, we investigate how we can make discrete probabilistic graphical models practical for predicting sensor states in a spatio-temporal setting. A set of new ideas allows keeping the advantages of such models while achieving scalability. We first introduce a novel alternative to represent model parameters, which enables us to compress the parameter storage by removing uninformative parameters in a systematic way. For finding the best parameters via maximal likelihood estimation, we provide a separable optimization algorithm that can be performed independently in parallel in each graph node. We illustrate that the prediction quality of our suggested methods is comparable to those of the standard MRFs and a spatio-temporal k-nearest neighbor method, while using much less computational resources.

1 Introduction

Sensor-based monitoring and prediction has become a hot topic in a large variety of applications. According to the slogan *Monitor, Mine, Manage* [1], series of data from heterogeneous sources are to be put to good use in diverse applications. A view of data mining towards *distributed sensor measurements* is presented in the book on ubiquitous knowledge discovery [11]. There are several approaches to distributed stream mining based on work like, e.g., Wolff *et al.* [21] or Sagy *et al.* [15]. The goal in these approaches is a general model (or function) which is built on the basis of local models while restricting communication costs. Most often, the global model allows to answer threshold queries, but also clustering of nodes is sometimes handled. Although the function is more complex, the model is global and not tailored for the prediction of measurements at a particular location. In contrast, we want to predict some sensor's state at some point in time given relevant previous and current measurements of itself and other sensors.

Since his influential book, David Luckham has promoted *complex event processing* successfully [9]. Detecting events in streams of data has accordingly been modeled, e.g. in the context of monitoring hygiene in a hospital [18]. However, in our case, the monitoring does not imply certain events. We do not aim at finding patterns that define an event, although they may show up as a side effect. We rather want to predict a certain state at a particular sensor or set of sensors taking into account the context of other locations and points in time. Although related, the tasks differ.

Let us illustrate the task of *spatio-temporal state prediction* by an example from traffic modeling. The structure of the model is given by a street network, which represents spatial relationships. Nodes within the network represent places, where the traffic is measured over time. The state of a node is the congestion at this street segment. At training time, we do not know which place at which time needs to be predicted as "jam". Given observations of the state variables at the nodes, a model is trained. The model must answer queries for all parts of the network and all points in time. For example:

- Given the traffic densities of all roads in a street network at discrete time points t_1, t_2, t_3 (e.g., Monday, Tuesday, Wednesday 8 o'clock): indicate the probabilities of traffic levels on a particular road A at another time point, not necessarily following the given ones (e.g., Thursday 7 o'clock).

One particular interest lies in learning probabilistic models for answering such queries in resource constrained environments. This addresses huge amounts of data on quite fast compute facilities as well as a rather moderate data volume on embedded or ubiquitous devices.

1.1 Previous Work

In this section, an overview of previous contributions to spatio-temporal modeling is given. The task of *traffic forecasting* is often solved by simulations [10]. This presupposes a model instead of learning it. In the course of urban traffic control, events are merely propagated that are already observed, e.g., a jam at a particular highway section results in a jam at another highway section, or the prediction is based on a physical rule that predicts a traffic jam based on a particular congestion pattern [3]. Many approaches apply statistical time series methods like auto-regression and moving average [20]. They do not take into account spatial relations but restrict themselves to the prediction of the state at one location given a series of observations at this particular location. An early approach is presented by Whittaker *et al.* [19], using a street network topology that represents spatial relations. The training is done using simply Kalman filters, which is not as expressive as is necessary for queries like the ones above. A statistical relational learning approach to traffic forecasting uses explicit rules for modeling spatio-temporal dependencies like $congestion(+s_1, h) \wedge next(s_1, s_2) \Rightarrow congestion(+s_2, h+1)$ [8]. Training is done by a Markov Logic Network delivering conditional probabilities of congestion classes. The discriminative model is restricted to binary classification tasks and the spatial dependencies need to be given by hand-tailored rules. Moreover, the model is not sparse and training is not scaleable. Even for a small number of sensors, training takes hours of computation. When the estimation of models for spatio-temporal data on ubiquitous devices is considered, e.g. learning to predict smartphone usage patterns based on time and visited places, minutes are the order of magnitude in demand. Hence, also this advanced approach does not yet meet the demands of the spatio-temporal prediction task in resource constrained environments.

Some geographically weighted regression or non-parametric k-Nearest Neighbour (kNN) methods model a task similar to spatio-temporal state prediction [23, 12]. The regression expresses the temporal dynamics and the weights express spatial distances. Another way to introduce the spatial relations into the regression is to encode the spatial network into a kernel function [7]. The kNN method by Lam *et al.* [6] models correlations in spatio-temporal data not only by their spatial but also by their temporal distance. As stated for spatio-temporal state prediction task, the particular place and time in question need not be known in advance, because the lazy learner kNN determines the prediction at question-time. However, also this approach does not deliver probabilities along with the predictions. For some applications, for instance, traffic prognoses for car drivers, a probabilistic assertion is not necessary. However, in applications of disaster management, the additional information of likelihood is wanted.

As is easily seen, generative models fit the task of spatio-temporal state prediction. For notational convenience, let us assume just one variable x. The *generative model* $p(x, y)$ allows to derive $p(y|x) = p(x, y)/p(x)$ as well as $p(x|y) = p(x, y)/p(y)$. In contrast, the *discriminative model* $p(y|x)$ must be trained specifically for each y. In our example, for each place, a distinct model would need to be trained. Hence, a huge set of discriminative models would be necessary to express one generative model. A discussion of discriminative versus generative models can be found in a study by Ng and Jordan [13]. Here, we refer to the capability of interpolation (e.g., between points in time) of generative models and their informativeness in delivering probability estimates instead of mere binary decisions.

Spatial relations are naturally expressed by *graphical models*. For instance, discriminative graphical models – as are Conditional Random Fields (CRFs) – have been used for object recognition over time [2], but also generative graphical models such as Markov Random Fields (MRFs) have been applied to video or image data [22, 4]. The number of training instances does not influence the model complexity of MRFs. However, the number of parameters can exceed millions easily. In particular when using MRFs for spatio-temporal state prediction, the many spatial and temporal relations soon lead to inefficiency.

1.2 Graphical Models

The formalism of probabilistic graphical models provides a unifying framework for capturing complex dependencies among random variables, and building large-scale multivariate statistical models [17]. Let $G = (V, E)$ be an undirected graph with the set of vertices V and the set of edges $E \subset V \times V$. For each node (or vertex) $v \in V$, let X_v be a random variable, taking values x_v in some space $\mathcal{X}_v$. The concatenation of all $n = |V|$ variables yields a multivariate random variable $\boldsymbol{X}$ with state space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$. Training delivers a full probability distribution over the random variable $\boldsymbol{X}$. Let $\boldsymbol{\phi}$ be an indicator function or *sufficient statistic* that indicates if a configuration $\boldsymbol{x}$ obeys a certain event $\{\boldsymbol{X}_\alpha = \boldsymbol{x}_\alpha\}$ with $\alpha \subseteq V$. We use the short hand notation $\{\boldsymbol{x}_\alpha\}$ to denote the event $\{\boldsymbol{X}_\alpha = \boldsymbol{x}_\alpha\}$. The functions of $\boldsymbol{x}$ defined in the following can be also considered as functions of $\boldsymbol{X}$ – we replace $\boldsymbol{x}$ by $\boldsymbol{X}$ when it makes their meaning clearer. Restricting α to vertices and edges, one gets

$$\boldsymbol{\phi}_{\{v=x\}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x}_v = x \\ 0 & \text{otherwise,} \end{cases}$$

$$\boldsymbol{\phi}_{\{(v,w)=(x,y)\}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } (\boldsymbol{x}_v, \boldsymbol{x}_w) = (x, y) \\ 0 & \text{otherwise} \end{cases}$$

with $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x}_v \in \mathcal{X}_v$ and $y \in \mathcal{X}_w$. Let us now define vectors for collections of those indicator functions:

$$\begin{aligned} \boldsymbol{\phi}_v(\boldsymbol{x}) &:= \left[\boldsymbol{\phi}_{\{v=x\}}(\boldsymbol{x})\right]_{x \in \mathcal{X}_v}, \\ \boldsymbol{\phi}_{(v,w)}(\boldsymbol{x}) &:= \left[\boldsymbol{\phi}_{\{(v,w)=(x,y)\}}(\boldsymbol{x})\right]_{(x,y) \in \mathcal{X}_v \times \mathcal{X}_w}, \\ \boldsymbol{\phi}(\boldsymbol{x}) &:= \left[\boldsymbol{\phi}_v(\boldsymbol{x}), \boldsymbol{\phi}_e(\boldsymbol{x}) : \forall v \in V, \forall e \in E\right]. \end{aligned} \tag{1}$$

The vectors are constructed for fixed but arbitrary orderings of V, E and $\mathcal{X}$. The dimension of $\boldsymbol{\phi}(\boldsymbol{x})$ is thus $d = \sum_{v \in V} |\mathcal{X}_v| + \sum_{(v,u) \in E} |\mathcal{X}_v| \times |\mathcal{X}_u|$. Now, consider a data set $\mathcal{D} = \left\{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\right\}$ with instances $\boldsymbol{x}^i$. Each $\boldsymbol{x}^i$ consists of an assignment to every node in the graph. It defines a full joint state of the random

variable $\boldsymbol{X}$. The quantities

$$\hat{\boldsymbol{\mu}}_{\{v=x\}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}_{\{v=x\}}(\boldsymbol{x}^i),$$

$$\hat{\boldsymbol{\mu}}_{\{(v,w)=(x,y)\}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}_{\{(v,w)=(x,y)\}}(\boldsymbol{x}^i)$$

are known as *empirical moments* and they reflect the empirical frequency estimates of the corresponding events. We say that a given probability mass function p with base measure ν and expectations $\mathbb{E}_p\left[\boldsymbol{\phi}_{\{\boldsymbol{x}_\alpha\}}(\boldsymbol{x})\right]$ is *locally consistent* with data $\mathcal{D}$ if and only if p satisfies the *moment matching condition*

$$\mathbb{E}_p\left[\boldsymbol{\phi}_{\{\boldsymbol{x}_\alpha\}}(\boldsymbol{x})\right] = \hat{\boldsymbol{\mu}}_{\{\boldsymbol{x}_\alpha\}}, \forall \alpha \in V \cup E,$$

i.e. the density p is consistent with the data w.r.t. the empirical moments. This problem is underdetermined, in that there are many densities p that are consistent with the data, so that we need a principle for choosing among them. The principle of maximum entropy is to choose, from among the densities consistent with the data, the densities p^* whose *Shannon entropy* $\mathcal{H}(p)$ is maximal. It can be shown that the optimal solution p^* takes the form of an exponential family of densities

$$p_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \exp[\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x})\rangle - A(\boldsymbol{\theta})],$$

parametrized by a vector $\boldsymbol{\theta} \in \mathbb{R}^d$. Note that the parameter vector $\boldsymbol{\theta}$ and the sufficient statistics vector $\boldsymbol{\phi}(\boldsymbol{x})$ have the same length d. The term $A(\boldsymbol{\theta})$ is called the *log partition function*,

$$A(\boldsymbol{\theta}) := \log \int_{\mathcal{X}} \exp[\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x})\rangle]\nu(d\boldsymbol{x}),$$

which is defined with respect to a reference measure $d\nu$ such that $P[X \in S] = \int_S p_{\boldsymbol{\theta}}(x)\nu(dx)$ for any measurable set S. Expanding $\boldsymbol{\phi}(\boldsymbol{x})$ by means of (1) reveals the usual density of pairwise undirected graphical models, also known as *pairwise Markov random field*

$$p_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{\Psi(\boldsymbol{\theta})}\prod_{v\in V}\psi_v(\boldsymbol{x})\prod_{(v,w)\in E}\psi_{(v,w)}(\boldsymbol{x}).$$

Here, $\Psi = \exp A$ is the cumulant-generating function of $p_{\boldsymbol{\theta}}$, and ψ_α are the so-called *potential functions*.

If the data set contains solely fully observed instances, the parameters may be estimated by the maximum likelihood principle. The estimation of parameters in the case of partially unobserved data is a challenging topic on its own. Here, we assume that the data set $\mathcal{D}$ contains only fully observed instances. The *likelihood* $\mathcal{L}$ and the *average log-likelihood* ℓ of parameters $\boldsymbol{\theta}$ given a set of i.i.d. data $\mathcal{D}$ are defined as

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) := \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}^i) \quad \text{and}$$

$$\ell(\boldsymbol{\theta}; \mathcal{D}) := \frac{1}{N}\sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^i) = \langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}\rangle - A(\boldsymbol{\theta}).$$

The latter is usually maximized due to numerical inconveniences of $\mathcal{L}$.

2 From Linear Chains to Spatio-Temporal Models

Sequential undirected graphical models, also known as linear chains, are a popular method in the natural language processing community [5, 16]. There, consecutive words or corresponding word features are connected to a sequence of labels that reflects an underlying domain of interest like entities or part of speech tags. If we consider a sensor network G that generates measurements over space as a word, then it would be appealing to think of the instances of G at different timepoints, like words in a sentence, to form a temporal chain $G_1 - G_2 - \cdots - G_T$. We will now present a formalization of this idea followed by some obvious drawbacks. Afterwards we will discuss how to tackle those drawbacks and derive a tractable class of generative graphical models for the spatio-temporal state prediction task.

We first define the part of the graph corresponding to the time t as the *snapshot graph* $G_t = (V_t, E_t)$, for $t = 1, 2, \ldots, T$. Each snapshot graph G_t replicates a given *spatial graph* $G_0 = (V_0, E_0)$, which represents the underlying physical placement of sensors, i.e., the spatial structure of random variables that does not change over time. We also define the set of spatio-temporal edges $E_{t-1;t} \subset V_{t-1} \times V_t$ for $t = 2, \ldots, T$ and $E_{0;1} = \emptyset$, that represent dependencies between adjacent snapshot graphs G_{t-1} and G_t, assuming a Markov property among snapshots, so that $E_{t;t+h} = \emptyset$ whenever $h > 1$ for any t. Note that the actual time gap between any two time frames t and $t+1$ can be chosen arbitrarily.

The entire graph, denoted by G, consists of the snapshot graphs G_t stacked in order for time frames $t = 1, 2, \ldots, T$ and the temporal edges connecting them: $G := (V, E)$ for $V := \cup_{t=1}^{T} V_t$ and $E := \cup_{t=1}^{T}\{E_t \cup E_{t-1;t}\}$.

The spatial graph G_0 and the sizes of the vertex state spaces $\mathcal{X}_v$ determine the number of model parameters d. In order to compute this quantity, we consider the exemplary construction of G. from G_0. First, all vertices v and all edges (u, v) from G_0 are copied exactly T times and added to $G = (V, E)$, whereas each copy is indexed by time t, i.e. $v \in V_0 \Rightarrow v_t \in V, 1 \le t \le T$ and likewise for the edges. Then, for each vertex $v_t \in V$ with $t \le T-1$, a temporal edge (v_t, v_{t+1}) is added to G. Finally, for each edge $(v_t, u_t) \in E$ with $t \le T-1$, the two spatio-temporal edges (v_t, u_{t+1}) and (v_{t+1}, u_t) are also added to G. The number of parameters per vertex v is $|\mathcal{X}_v|$ and accordingly $|\mathcal{X}_v||\mathcal{X}_u|$ per edge (v, u). If we assume that all vertices $v, u \in V$ share a common state space and that state spaces do not change over time, i.e. $\mathcal{X}_{v_t} = \mathcal{X}_{u_{t'}}, \forall v, u \in V, 1 \le t, t' \le T$, the total number of parameters is

$$d = T|V_0|\,|\mathcal{X}_{v_t}| + \left[(T-1)(|V_0| + 3|E_0|) + |E_0|\right]|\mathcal{X}_{v_t}|^2$$

with some arbitrary but fixed vertex v_t. Note that the last two assumptions are only needed to simplify the computation of d, the spatio temporal random field that is described in the following section is not restricted by any of these assumptions.

This model now truly expresses temporal and spatial relations between all locations and points in time for all features. However, the memory requirements of

such models are quite high due to the large problem dimension. Even loading or sending models may cause issues when mobile devices are considered as a platform. Furthermore, the training does not scale well because of stepsize adaption techniques that are based on sequential (i.e., non-parallel) algorithms.

The derivation and empirical evaluation of the compressible representation and distributed estimation can be found in [14].

Acknowledgments

Work on this paper has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", projects A1 and C1.

References

[1] David Campbell. Is it still Big Data if it fits in my pocket? In *Proceedings of the VLDB Endowment*, volume 4, page 694, 2011.

[2] Bertrand Douillard, Dieter Fox, and Fabio T. Ramos. A spatio-temporal probabilistic model for multi-sensor object recognition. In *IEEE/RSJ International Conference on IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2402–2408, 2007.

[3] S. F. Hafstein, R. Chrobok, A. Pottmeier, and M. Schreckenbergand F. Mazur. A high-resolution cellular automata traffic simulation model withapplication in a freeway traffic information system. *Computer-Aided Civil and Infrastructure Engineering*, 19(5):338–350, 2004.

[4] Rui Huang, Vladimir Pavlovic, and Dimitris Metaxas. A new spatio-temporal mrf framework for video-based object segmentation. In *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis*, 2008.

[5] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[6] W. H. K. Lam, Y. F. Tang, and M. Tam. Comparison of two non-parametric models for daily traffic forecasting in hong kong. *Journal of Forecasting*, 25(3):173–192, 2006.

[7] Thomas Liebig, Zhao Xu, Michael May, and Stefan Wrobel. Pedestrian quantity estimation with trajectory patterns. In *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 629–643. Springer, 2012.

[8] Marco Lippi, Matteo Bertini, and Paolo Frasconi. Collective traffic forecasting. In *Machine Learning and Knowledge Discovery in Databases*, volume 6322 of *Lecture Notes in Computer Science*, pages 259–273. Springer, 2010.

[9] David Luckham. *The Power of Events - An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison Wesley, 2002.

[10] Sigurdur F. Marinosson, Roland Chrobok, Andreas Pottmeier, Joachim Wahle, and Michael Schreckenberg. Simulation framework for the autobahn traffic in North Rhine-Westphalia. In *Cellular Automata – 5th Int. Conf. on Cellular Automata for Research and Industry*, pages 2977–2980. Springer, 2002.

[11] Michael May and Lorenza Saitta, editors. *Ubiquitous Knowledge Discovery*, volume 6202 of *Lecture Notes in Artificial Intelligence*. Springer, 2010.

[12] M. May, D. Hecker, C. Körner, S. Scheider, and D. Schulz. A vector-geometry based spatial knn-algorithm for traffic frequency predictions. *Data Mining Workshops, International Conference on Data Mining*, 0:442–447, 2008.

[13] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, 14:841–848, 2002.

[14] Nico Piatkowski, Sangkyun Lee, and Katharina Morik. Spatio-temporal random fields: Compressible representation and distributed estimation. *Machine Learning Journal*, 93(1):115–139, 2013 2013.

[15] Guy Sagy, Daniel Keren, Izchak Sharfman, and Assaf Schuster. Distributed threshold querying of general functions by a difference of monotonic representation. In *Proceedings of the VLDB Endowment*, volume 4, 2011.

[16] Charles Sutton and Andrew McCallum. An introduction to conditional random fields for relational learning. In *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[17] Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2007.

[18] Di Wang, Elke A. Rundensteiner, and Richard T. Ellison. Active complex event processing of event streams. In *Procs. of the VLDB Endowment*, volume 4, 2011.

[19] Joe Whittaker, Simon Garside, and Karel Lindveld. Tracking and predicting a network traffic process. *International Journal of Forecasting*, 13(1):51–61, March 1997.

[20] B.M. Williams and L.A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, 129(6):664–672, 2003.

[21] Ran Wolff, Kanishka Badhuri, and Hillol Kargupta. A generic local algorithm for mining data streams in large distributed systems. *IEEE Transactions on Knowledge and Data Engineering*, 21(4):465–478, 2009.

[22] Zhaozheng Yin and Robert Collins. Belief propagation in a 3D spatio-temporal MRF for moving object detection. *IEEE Computer Vision and Pattern Recognition*, 2007.

[23] F. Zhao and N. Park. Using geographically weighted regression models to estimate annual average daily traffic. *Journal of the Transportation Research Board*, 1879(12):99–107, 2004.

Tag Recommendations for SensorFolkSonomies

Resubmission from RSWeb Workshop at RecSys 2013

Juergen Mueller
L3S Research Center
University of Kassel
Wilhelmshöher Allee 73
Kassel, Germany
mueller@cs.uni-kassel.de

Stephan Doerfel
University of Kassel
Wilhelmshöher Allee 73
Kassel, Germany
doerfel@cs.uni-kassel.de

Martin Becker
L3S Research Center
University of Würzburg
Am Hubland
Würzburg, Germany
becker@informatik.uni-wuerzburg.de

Andreas Hotho
L3S Research Center
University of Würzburg
Am Hubland
Würzburg, Germany
hotho@informatik.uni-wuerzburg.de

Gerd Stumme
L3S Research Center
University of Kassel
Wilhelmshöher Allee 73
Kassel, Germany
stumme@cs.uni-kassel.de

Abstract

With the rising popularity of smart mobile devices, sensor data-based applications have become more and more popular. Their users record data during their daily routine or specifically for certain events. The application WideNoise Plus allows users to record sound samples and to annotate them with perceptions and tags. The app documents and maps the soundscape all over the world. The procedure of recording, including the assignment of tags, has to be as easy-to-use as possible. We therefore discuss the application of tag recommender algorithms in this particular scenario. We show, that this task is fundamentally different from the well-known tag recommendation problem in folksonomies as users do no longer tag fix resources but sensory data and impressions. The scenario requires efficient recommender algorithms that are able to run on the mobile device, since Internet connectivity is not always available. Therefore, we evaluate the performance of ten tag recommendation algorithms and discuss their applicability in the mobile sensing use-case.

Labelwise versus Pairwise Decomposition in Label Ranking

Weiwei Cheng, Sascha Henzgen and Eyke Hüllermeier
Computational Intelligence Group, University of Marburg
Marburg, Germany
{cheng,henzgen,eyke}@mathematik.uni-marburg.de

Abstract

Label ranking is a specific type of preference learning problem, namely the problem of learning a model that maps instances to rankings over a finite set of predefined alternatives (labels). State-of-the-art approaches to label ranking include decomposition techniques that reduce the original problem to binary classification; ranking by pairwise comparison (RPC), for example, constructs one binary problem for each pair of alternatives. In general, each classification example refers to the *pairwise* comparison of two alternatives in a ranking. In this paper, we introduce a new (meta) learning technique for label ranking, which is based on a *labelwise* instead of a pairwise decomposition. The basic idea is to train one model per class label, namely a model that maps instances to ranks. Instead of a quadratic number of binary problems, like in RPC, this obviously gives rise to a linear number of ordinal classification problems. We propose a generalization of this approach for the practically relevant case in which the training data only contains incomplete rankings, that is, rankings of some but not all alternatives; in this case, only imprecise information about the rank of individual labels can be derived. Moreover, we provide an experimental study, in which the pairwise and the labelwise decomposition techniques are compared in a systematic way.

1 Introduction

Preference learning is an emerging subfield of machine learning, which deals with the induction of preference models from observed or revealed preference information [10]. Such models are typically used for prediction purposes, for example, to predict context-dependent preferences of individuals on various choice alternatives. Depending on the representation of preferences, individuals, alternatives, and contexts, a large variety of preference models are conceivable, and many such models have already been studied in the literature.

A specific type of preference learning problem is the problem of *label ranking*, namely the problem of learning a model that maps instances to rankings (total orders) over a finite set of predefined alternatives (labels). Several methods for label ranking have already been proposed in the literature [18]. Most of these methods are *reduction techniques* transforming the original learning task into one or several binary classification tasks. Moreover, all existing methods are *relational* in so far as they seek to learn from *relative* or *comparative* preferences, such as pairwise comparisons between alternatives [15]. Since a ranking of alternatives, by its very nature, does indeed inform about *relative* and not about *absolute* preferences, the prevalence of the relational approach is of course completely understandable.

On the other hand, since the number of alternatives in a label ranking problem is fixed, a ranking is uniquely defined by the position (rank) of each of the alternatives, which can be seen as *absolute* preference information. Admittedly, as will be explained in more detail later on, this positional information is not always readily available for training. Yet, it is arguably a bit surprising that, to the best of our knowledge, an approach focused on the learning and prediction of absolute preferences has not even been tried so far.

In this paper, we introduce an approach of that kind, namely a new meta-learning technique for label ranking, which is based on a *labelwise* instead of a *pairwise* decomposition. The basic idea is to train one model per class label, namely a model that maps instances to ranks. In other words, given a new query instance, the idea is to predict the rank of each individual label right away. Unlike existing decomposition techniques, in which the reducts are binary classification problems, this approach leads to a linear number of ordered multi-class problems.

The paper is organized as follows. The next section provides some background of the label ranking problem, and Section 3 reviews existing methods for tackling this problem. Our new approach based on labelwise decomposition (LWD) is introduced in Section 4. Section 5 is devoted to a general discussion of similarities and differences between reduction techniques for label ranking. In Section 6, we provide an experimental study, in which LWD is compared with existing decomposition techniques in a systematic way. The paper ends with some concluding remarks in Section 7.

2 Label Ranking

Let $\mathcal{Y} = \{y_1, \ldots, y_K\}$ be a finite set of (choice) alternatives; adhering to the terminology commonly used in supervised machine learning, and accounting for the fact that label ranking can be seen as an extension of multi-class classification, the y_i are also called *class labels*. We consider total order relations $\succ$ on $\mathcal{Y}$, that is, complete, transitive, and antisymmetric relations, where $y_i \succ y_j$ indicates that y_i precedes y_j in the order. Since a ranking can be seen as a special type of preference relation, we shall also

say that $y_i \succ y_j$ indicates a preference for y_i over y_j.

Formally, a total order $\succ$ can be identified with a permutation $\bar{\pi}$ of the set $[K] = \{1, \ldots, K\}$, such that $\bar{\pi}(i)$ is the position of y_i in the order. We denote the class of permutations of $[K]$ (the symmetric group of order K) by $\mathbb{S}_K$. By abuse of terminology, though justified in light of the above one-to-one correspondence, we refer to elements $\bar{\pi} \in \mathbb{S}_K$ as both permutations and rankings.

In the setting of label ranking, preferences on $\mathcal{Y}$ are "contextualized" by instances $\boldsymbol{x} \in \mathbb{X}$, where $\mathbb{X}$ is an underlying instance space. Thus, each instance $\boldsymbol{x}$ is associated with a ranking $\succ_{\boldsymbol{x}}$ of the label set $\mathcal{Y}$ or, equivalently, a permutation $\bar{\pi}_{\boldsymbol{x}} \in \mathbb{S}_K$. More specifically, since label rankings do not necessarily depend on instances in a deterministic way, each instance $\boldsymbol{x}$ is associated with a probability distribution $\mathbf{P}(\cdot \mid \boldsymbol{x})$ on $\mathbb{S}_K$. Thus, for each $\bar{\pi} \in \mathbb{S}_K$, $\mathbf{P}(\bar{\pi} \mid \boldsymbol{x})$ denotes the probability to observe the ranking $\bar{\pi}$ in the context specified by $\boldsymbol{x}$.

As an illustration, suppose $\mathbb{X}$ is the set of people characterized by attributes such as sex, age, profession, and marital status, and labels are music genres: $\mathcal{Y} = \{\texttt{Rock}, \texttt{Pop}, \texttt{Classic}, \texttt{Jazz}\}$. Then, for $\boldsymbol{x} = (m, 30, \text{teacher}, \text{married})$ and $\bar{\pi} = (2, 1, 4, 3)$, $\mathbf{P}(\bar{\pi} \mid \boldsymbol{x})$ denotes the probability that a 30 years old married man, who is a teacher, prefers Pop music to Rock to Classic to Jazz.

2.1 The Label Ranking Problem

The goal in label ranking is to learn a "label ranker", that is, a model

$$\mathcal{M} : \mathbb{X} \longrightarrow \mathbb{S}_K$$

that predicts a ranking $\hat{\pi}$ for each instance $\boldsymbol{x}$ given as an input. More specifically, seeking a model with optimal prediction performance, the goal is to find a risk (expected loss) minimizer

$$\mathcal{M}^* \in \operatorname*{argmin}_{\mathcal{M} \in \mathbf{M}} \int_{\mathbb{X} \times \mathbb{S}_K} D(\mathcal{M}(\boldsymbol{x}), \bar{\pi}) \, d\mathbf{P} \ ,$$

where $\mathbf{M}$ is the underlying model class, $\mathbf{P}$ is the joint measure $\mathbf{P}(\boldsymbol{x}, \bar{\pi}) = \mathbf{P}(\boldsymbol{x})\mathbf{P}(\bar{\pi} \mid \boldsymbol{x})$ on $\mathbb{X} \times \mathbb{S}_K$ and D is a loss function on $\mathbb{S}_K$; common choices of D will be introduced below.

As training data $\mathbb{D}$, a label ranker uses a set of instances $\boldsymbol{x}_n$ ($n \in [N]$), together with information about the associated rankings π_n. Ideally, complete rankings are given as training information, i.e., a single observation is a tuple of the form $(\boldsymbol{x}_n, \pi_n) \in \mathbb{X} \times \mathbb{S}_K$; we call an observation of that kind a *complete* example. From a practical point of view, however, it is important to allow for incomplete information in the form of a ranking of some but not all of the labels in $\mathcal{Y}$:

$$y_{\tau(1)} \succ_{\boldsymbol{x}} y_{\tau(2)} \succ_{\boldsymbol{x}} \cdots \succ_{\boldsymbol{x}} y_{\tau(J)} \ , \tag{1}$$

where $J < K$ and $\{\tau(1), \ldots, \tau(J)\} \subset [K]$. For example, for an instance $\boldsymbol{x}$, it might be known that $y_2 \succ_{\boldsymbol{x}} y_1 \succ_{\boldsymbol{x}} y_5$, while no preference information is given about the labels y_3 or y_4.

In the following, we will write complete rankings $\bar{\pi}$ with an upper bar (as we already did above). If a ranking π is not complete, then $\pi(j)$ is the position of y_j in the incomplete ranking, provided this label is contained, and $\pi(j) = 0$ otherwise; thus, if $\bar{\pi}$ is a "completion" of π, then $\bar{\pi}(k) \geq \pi(k)$ for all $k \in [K]$. In the above example (1), $\pi = (2, 1, 0, 0, 3)$. We denote by $|\pi| = \{j \mid \pi(j) > 0\}$ the size of the ranking; thus, π is complete if $|\pi| = K$.

2.2 Prediction Accuracy

The prediction accuracy of a label ranker is assessed by comparing the true ranking $\bar{\pi}$ with the prediction $\hat{\pi}$, using a distance measure D on rankings. Among the most commonly used measures is the Kendall distance, which is defined by the number of inversions, that is, index pairs $\{i, j\} \subset [K]$ such that the order of y_i and y_j in $\bar{\pi}$ is inverted in $\hat{\pi}$:

$$D(\bar{\pi}, \hat{\pi}) = \sum_{1 \leq i < j \leq K} [\![(\bar{\pi}(i) - \bar{\pi}(j))(\hat{\pi}(i) - \hat{\pi}(j)) < 0]\!] \tag{2}$$

The well-known Kendall rank correlation measure is an affine transformation of (2) to the range $[-1, +1]$. Besides, the sum of L_1 or L_2 losses on the ranks of the individual labels are often used as an alternative distance measures:

$$D_1(\bar{\pi}, \hat{\pi}) = \sum_{i=1}^{M} |\bar{\pi}(i) - \hat{\pi}(i)| \tag{3}$$

$$D_2(\bar{\pi}, \hat{\pi}) = \sum_{i=1}^{M} (\bar{\pi}(i) - \hat{\pi}(i))^2 \tag{4}$$

These measures are closely connected with two other well-known rank correlation measures: Spearman's footrule is an affine transformation of (3) to the interval $[-1, +1]$, and Spearman's rank correlation (Spearman's rho) is such a transformation of (4).

3 Label Ranking Methods

The arguably most straightforward way to addressing the label ranking problem is to treat it as a classification problem with $K!$ classes, considering each ranking $\bar{\pi} \in \mathbb{S}_K$ as a separate (meta-)class; this is to some extent comparable to the *label powerset* approach to multilabel classification [17], which considers each subset Y of the original label set $\mathcal{Y}$ as a new meta-class. Obviously, however, this approach comes with a number of disadvantages, making it likely to fail in practice. First of all, the number of meta-classes is even larger than for multilabel classification. For example, with only $K = 6$ labels, the resulting classification problem would consist of 720 meta-classes—there is no classifier that can handle such a number of classes in a reasonable way. Second, it is not clear how to apply this approach in the case of incomplete observations (1). Third, by treating each meta-class as a separate category, this approach fails to exploit the structure on the output space $\mathbb{S}_K$, which is induced by the underlying distance measure D.

Indeed, label ranking can be seen as a specific type of *structured output* prediction [1], namely the problem to predict structures in the form of permutations. In the literature, several methods for label ranking have been proposed that try to exploit the structure on $\mathbb{S}_K$ in one way or the other, including generalizations of standard machine learning methods such as nearest neighbor estimation [3] and decision tree learning [6], as well as statistical inference based on parametrized models of rank data [5].

Here, we are specifically interested in *reduction techniques*, that is, meta-learning techniques that reduce the original label ranking problem into one or several classification problems that are easier to solve. Among the techniques proposed so far, there are two approaches that both reduce label ranking to binary classification, albeit in a different way. Whereas the first technique, *constraint classification* (CC), produces a single "large" classification

problem, the second one, *ranking by pairwise comparison* (RPC), yields a quadratic (in K) number of "small" binary problems. In the following, both approaches will be presented in more detail.

3.1 Constraint Classification

Constraint classification [12] is based on the idea of learning value functions $f_k : \mathbb{X} \longrightarrow \mathbb{R}$, one for each label y_k ($k \in [K]$), that estimate a (latent) degree of utility of y_k in the context specified by an instance. Given such functions, a prediction $\hat{\pi}$ for a new query instance $\boldsymbol{x}$ is then simply obtained by sorting the labels in decreasing order of their (estimated) utility:

$$\hat{\pi} = \underset{k \in [K]}{\operatorname{argsort}} f_k(\boldsymbol{x}) \tag{5}$$

More specifically, assuming $\mathbb{X} = \mathbb{R}^d$, the value functions are taken as linear functions of the form

$$f_k(\boldsymbol{x}) = f_k(x_1, \ldots, x_d) = \sum_{i=1}^{d} \alpha_{k,i} x_i \tag{6}$$

with label-specific coefficients $\alpha_{k,i}$ ($i \in [d]$).

Now, a pairwise preference $y_k \succ_{\boldsymbol{x}} y_j$ between two labels translates into the constraint $f_k(\boldsymbol{x}) - f_j(\boldsymbol{x}) > 0$ or, equivalently, $f_j(\boldsymbol{x}) - f_k(\boldsymbol{x}) < 0$. Both constraints, the positive and the negative one, can be expressed in terms of the sign of an inner product $\langle \boldsymbol{z}, \boldsymbol{\alpha} \rangle$, where

$$\boldsymbol{\alpha} = (\alpha_{1,1}, \ldots, \alpha_{1,d}, \alpha_{2,1}, \ldots, \alpha_{2,d}, \ldots, \alpha_{K,1}, \ldots, \alpha_{K,d})$$

is a concatenation of all label-specific coefficients. Correspondingly, the vector $\boldsymbol{z}$ is constructed by mapping the original d-dimensional training example $\boldsymbol{x} = (x_1, \ldots, x_d)$ into an $(K \times d)$-dimensional space: For the positive constraint, $\boldsymbol{x}$ is copied into the components $((k-1) \times d + 1), \ldots, (k \times d)$ and its negation $-\boldsymbol{x}$ into the components $((j-1) \times d+1), \ldots, (j \times d)$; the remaining entries are filled with 0. For the negative constraint, a vector is constructed with the same elements but reversed signs. Both constraints can be considered as training examples for a conventional binary classifier in a $(K \times d)$-dimensional space: The first vector is a positive and the second one a negative example.

CC constructs training examples of that kind by splitting observed rankings into pairwise preferences. More specifically, an incomplete ranking (1) is split into $J - 1$ preferences $y_{\tau(k)} \succ_{\boldsymbol{x}} y_{\tau(k+1)}$ ($k \in [J-1]$), and each of these preferences is turned into a positive and a negative example for the binary classifier as described above. The corresponding binary classification problem can then be tackled by standard methods for fitting a separating hyperplane in this space, that is, a suitable vector $\boldsymbol{\alpha}$ satisfying as many as possible constraints.

3.2 Ranking by Pairwise Comparison

Ranking by pairwise comparison [15] is an extension of pairwise classification [9], an established technique for reducing multi-class to binary classification. In the setting of label ranking, RPC trains one model $\mathcal{M}_{i,j} : \mathbb{X} \longrightarrow [0, 1]$ for each pair of labels $\{y_i, y_j\}$; thus, $K(K-1)/2$ such models are needed in total. Given instance $\boldsymbol{x}$ as input, the model $\mathcal{M}_{i,j}$ is supposed to predict the probability of $y_i \succ_{\boldsymbol{x}} y_j$, i.e., $\mathcal{M}_{i,j}(\boldsymbol{x})$ is an estimation of the probability $\mathbf{P}(\pi(i) < \pi(j) \,|\, \boldsymbol{x})$.

The data $\mathbb{D}_{i,j}$ used to train $\mathcal{M}_{i,j}$ is constructed from the original data $\mathbb{D}$ as follows: If $\boldsymbol{x}_n$ is an instance in $\mathbb{D}$ that has been observed together with a possibly incomplete ranking of labels in $\mathcal{Y}$, then

- $\boldsymbol{x}_n$ is added as a positive example to $\mathbb{D}_{i,j}$ if the ranking contains both y_i and y_j, and the former precedes the latter;
- $\boldsymbol{x}_n$ is added as a negative example to $\mathbb{D}_{i,j}$ if the ranking contains both y_i and y_j, and the latter precedes the former;
- $\boldsymbol{x}_n$ is ignored if either y_i or y_j (or both) are missing in the ranking.

Once $\mathbb{D}_{i,j}$ has been constructed, any method for (probabilistic) binary classification can be used to induce the model $\mathcal{M}_{i,j}$.

At prediction time, when a ranking $\hat{\pi}$ needs to be predicted for a new instance $\boldsymbol{x}$, this instance is first submitted to each of the models $\mathcal{M}_{i,j}$ ($1 \leq i < j \leq K$), and the predictions of these models are combined into a (weighted) preference relation

$$P = \begin{bmatrix} - & p_{1,2} & p_{1,3} & \cdots & p_{1,K} \\ p_{2,1} & - & p_{2,3} & \cdots & p_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{K,1} & p_{K,2} & p_{K,3} & \cdots & - \end{bmatrix}, \tag{7}$$

where

$$p_{i,j} = \begin{cases} \mathcal{M}_{i,j}(\boldsymbol{x}) & \text{if } i < j \\ 1 - \mathcal{M}_{j,i}(\boldsymbol{x}) & \text{if } j < i \end{cases}.$$

The preference relation (7) does normally not suggest a ranking $\hat{\pi}$ in an unequivocal way: Since the binary models $\mathcal{M}_{i,j}$ are trained independently of each other, and the predictions $p_{i,j}$ are not necessarily perfect, P may exhibit inconsistencies such as preferential cycles. What is needed, in general, is a *ranking procedure* that turns P into a ranking $\hat{\pi}$.

The standard approach in RPC is to apply a weighted voting procedure, in which the labels are sorted according to the sum of weighted votes in their own favor:

$$\hat{\pi} = \underset{k \in [K]}{\operatorname{argsort}} s_k(\boldsymbol{x}), \tag{8}$$

where

$$s_k(\boldsymbol{x}) = \sum_{1 \leq j \neq k \leq K} p_{k,j}.$$

Under certain technical assumptions (the $p_{i,j}$ are independent and unbiased estimations of $\mathbf{P}(\pi(i) < \pi(j) \,|\, \boldsymbol{x})$), it can be shown that the prediction (8) is minimizing the expected loss with respect to (4). For other loss functions, other ranking procedures might be optimal.

4 Labelwise Decomposition

In this section, we introduce a new meta-learning technique for label ranking, which is based on the idea of reducing the original problem to standard classification problems in a *labelwise* manner.

4.1 The Case of Complete Training Information

If the training data $\mathbb{D}$ consists of complete examples $(\boldsymbol{x}_n, \bar{\pi}_n)$, then each such example informs about the rank $\bar{\pi}(k)$ of the label y_k in the ranking associated with $\boldsymbol{x}_n$. Thus, a quite natural idea is to learn a model

$$\mathcal{M}_k : \mathbb{X} \longrightarrow [K]$$

that predicts the rank of y_k, given an instance $\boldsymbol{x} \in \mathbb{X}$ as an input. Indeed, such a model can be trained easily on the data

$$\mathbb{D}_k = \Big\{ (\boldsymbol{x}_n, r_n) \,|\, (\boldsymbol{x}_n, \bar{\pi}_n) \in \mathbb{D},\, r_n = \bar{\pi}_n(k) \Big\} \subset \mathbb{X} \times [K] \tag{9}$$

It is important to note, however, that the classification problem thus produced is not a binary one, like in CC and RPC. Instead, we need to solve a multi-class problem with K classes, where each class corresponds to a possible rank. More specifically, since these ranks have a natural order, we are facing an *ordinal classification* problem.

Like in RPC, we assume that a probabilistic approach is used to train the models $\mathcal{M}_k$ ($k \in [K]$). For example, if the (ordinal) classifiers are specified by a parameter $\theta \in \Theta$, $\mathcal{M}_k$ could be identified by the maximum likelihood estimate

$$\theta_k = \operatorname*{argmax}_{\theta \in \Theta} \prod_{n=1}^{N} \mathbf{P}(r_n \,|\, \boldsymbol{x}_n, \theta) \ . \tag{10}$$

Then, given a new query instance $\boldsymbol{x}$, each of these models is supposed to predict a probability distribution

$$\mathcal{M}_k(\boldsymbol{x}) = \big(p_{k,1}, p_{k,2}, \ldots, p_{k,K}\big) \in [0,1]^K \ , \tag{11}$$

where $p_{k,j} = \mathbf{P}(\bar{\pi}(k) = j \,|\, \boldsymbol{x})$ is the (predicted) probability that y_k is on rank j.

4.2 Aggregation

As we have seen in previous sections, each reduction techniques also involves an *aggregation procedure*, which is responsible for combining the predictions of the classification models into a ranking $\hat{\pi}$. In the case of CC and RPC, these aggregations are given by the sorting procedures (5) and (8), respectively.

Consider a loss function D on $\mathbb{S}_K$ that is labelwise decomposable, i.e., which can be written in the form

$$D(\bar{\pi}, \hat{\pi}) = \sum_{k=1}^{K} D_k(\bar{\pi}(k), \hat{\pi}(k)).$$

Obviously, the L_1 and L_2 loss (3) and (4) are both of this type. Then, given probabilities of the form (11), the expected loss caused by a prediction $\hat{\pi}$ can be written as

$$\begin{aligned} \mathbb{E}\big(D(\bar{\pi}, \hat{\pi})\big) &= \sum_{k=1}^{K} \mathbb{E}\big(D_k(\bar{\pi}(k), \hat{\pi}(k))\big) \qquad (12) \\ &= \sum_{k=1}^{K} \sum_{j=1}^{K} D_k(j, \hat{\pi}(k)) \cdot p_{k,j} \\ &= \sum_{k=1}^{K} L_k(\hat{\pi}(k)) \ , \end{aligned}$$

where $L_k(r)$ is the cost of putting y_k on position r, namely the loss expected on y_k when assigning this label to position r in the ranking $\hat{\pi}$. In the case of (3), for example, this cost is given by

$$L_k(r) = \sum_{j=1}^{K} |j - r| \cdot p_{k,j} \ .$$

Thus, an optimal solution would consists of assigning y_k the position $\hat{\pi}(k) = r$ for which $L_k(r)$ is minimal. However, noting that each position $r \in [K]$ must be assigned at most once, this approach is obviously not guaranteed to produce a feasible solution. Instead, the minimization of (12) requires the solution of an *optimal assignment problem* [4]:

- labels $y_k \in \mathcal{Y}$ must be uniquely assigned to ranks $r = \hat{\pi}(k) \in [K]$;
- assigning y_k to rank r causes a cost of $L_k(r)$;
- the goal is to minimize the sum of all assignment costs.

Assignment problems of that kind have been studied extensively in the literature, and efficient algorithms for their solution are available. The well-known Hungarian algorithm [16], for example, solves the above problem in time $O(K^3)$. Such algorithms can be used to produce a (risk minimizing) prediction $\hat{\pi}$ on the basis of probabilistic predictions (11).

4.3 The Case of Incomplete Training Information

As mentioned before, the original training data $\mathbb{D}$ is not necessarily supposed to contain complete rank information; instead, for a training instance $\boldsymbol{x}_n$, only an incomplete ranking π_n of a subset of the labels in $\mathcal{Y}$ might have been observed, while the complete ranking $\bar{\pi}_n$ is not given. In this case, the above method is not directly applicable: If at least one label is missing, i.e., $|\pi_n| < K$, then none of the true ranks $\bar{\pi}_n(k)$ is precisely known; consequently, the training data (9) cannot be constructed.

Nevertheless, even in the case of incomplete rankings, non-trivial information can be derived about the rank $\bar{\pi}(k)$ for at least some of the labels y_k. In fact, if $|\pi| = J$ and $\pi(k) = r > 0$, then

$$\bar{\pi}(k) \in \big\{r, r+1, \ldots, r+K-J\big\} \ .$$

Of course, if $\pi(k) = 0$ (i.e., y_k is not present in the ranking), only the trivial information $\bar{\pi}(k) \in [K]$ can be derived. Yet, more precise information can be obtained under additional assumptions. For example, if π is known to be the top of the ranking $\bar{\pi}$, then

$$\begin{cases} \bar{\pi}(k) = \pi(k) & \text{if } \pi(k) > 0 \\ \bar{\pi}(k) \in \{J+1, \ldots, K\} & \text{if } \pi(k) = 0 \end{cases} \ . \tag{13}$$

This scenario is highly relevant, since top-ranks are observed in many practical applications.

In general, the type of training data that can be derived for a label y_k in the case of incomplete rank information are examples of the form

$$\big(\boldsymbol{x}_n, R_n\big) \in \mathbb{X} \times 2^{[K]} \ , \tag{14}$$

that is, an instance $\boldsymbol{x}_n$ together with a set of possible ranks R_n. The problem of learning from data with *imprecise* class information has recently been studied in the literature, where it is called learning from *ambiguously labeled examples* [13] or learning from *partial labels* [11; 7]. As explained in [13], a reasonable approach to learning from imprecise data is to combine model identification and *data disambiguation*, that is, trying to fit an optimal model while simultaneously finding the "true data". Again adopting the principle of maximum likelihood inference, one way to realize this idea is to maximize a generalized likelihood function:

$$\theta_k = \operatorname*{argmax}_{\mathbf{r} \in \mathcal{R},\, \theta \in \Theta} \prod_{n=1}^{N} \mathbf{P}(r_n \,|\, \boldsymbol{x}_n, \theta) \ , \tag{15}$$

where $\mathcal{R}$ is the set of all selections of the rank information (14), that is, the set of all vectors $\mathbf{r} = (r_1, \ldots, r_N) \in [K]^N$ such that $r_n \in R_n$.

4.4 Probabilistic Modeling of Missing Label Information

Under additional assumptions about the process that eliminates labels from a complete ranking $\bar{\pi}$, this approach can be further refined. For example, under the "missing at random" assumption, according to which the $K - J$ labels that are missing have been selected uniformly at random from the set of all K labels, the probability to observe $\pi(k) = j > 0$ is given by

$$\sum_{r=j}^{j+K-J} \mathbf{P}(r \mid \boldsymbol{x}, \theta_k) \frac{\binom{r-1}{r-j}\binom{K-r}{K-J-r+j}}{\binom{K}{J}} \quad . \tag{16}$$

Each term in (16) expresses the probability that the true rank of y_k in $\bar{\pi}$ is r, $r - j$ labels are removed "above" y_k (thus bringing it to position j), and the other $K - J - r + j$ labels are removed "below" y_k. The probability to observe $\pi(k) = 0$ is given by

$$\sum_{r=1}^{K} \mathbf{P}(r \mid \boldsymbol{x}, \theta_k) \frac{J}{K} = \frac{J}{K} \quad ,$$

i.e., by a constant that can be ignored in likelihood maximization. Thus, estimation of θ_k can be accomplished as follows:

$$\hat{\theta}_k = \operatorname*{argmax}_{\theta \in \Theta} \prod_{n \in [N], \pi_n(k) > 0} \sum_{r=\pi_n(k)}^{\pi_n(k)+K-|\pi_n|} \mathbf{P}_{r,\pi_n(k)} \tag{17}$$

with

$$\mathbf{P}_{r,\pi_n(k)} = \mathbf{P}(r \mid \boldsymbol{x}, \theta) \frac{\binom{r-1}{r-\pi_n(k)}\binom{K-r}{K-|\pi_n|-r+\pi_n(k)}}{\binom{K}{|\pi_n|}}$$

Under the top-rank model (13), the probability to observe $\pi(k) = j$ is given by

$$\begin{cases} \mathbf{P}(j \mid \boldsymbol{x}, \theta_k) & \text{if } j > 0 \\ \sum_{r=|\pi_n|+1}^{K} \mathbf{P}(r \mid \boldsymbol{x}, \theta_k) & \text{if } j = 0 \end{cases} \quad , \tag{18}$$

and θ_k can be estimated as follows:

$$\hat{\theta}_k = \operatorname*{argmax}_{\theta \in \Theta} \prod_{n \in [N], \pi_n(k) > 0} \mathbf{P}(\pi_n(k) \mid \boldsymbol{x}, \theta) \tag{19}$$

$$\times \prod_{n \in [N], \pi_n(k) = 0} \sum_{r=|\pi_n|+1}^{K} \mathbf{P}(r \mid \boldsymbol{x}, \theta)$$

For our experiments in Section 6, we implemented (17) and (19) using a corresponding extension of ordinal logistic regression. Thus, the probabilities $\mathbf{P}(r \mid \boldsymbol{x}, \theta)$ are expressed in terms of log-linear functions. More specifically, ordinal logistic regression models ratios of the *cumulative distribution*:

$$\log\left(\frac{c_k(\boldsymbol{x})}{1 - c_k(\boldsymbol{x})}\right) = \beta_k + \boldsymbol{w}^\top \boldsymbol{x} \tag{20}$$

for $k \in [K-1]$, where $c_k(\boldsymbol{x}) = \mathbf{P}(r \leq k \mid \boldsymbol{x})$ is the (conditional) probability of a rank $\leq k$ (hence $\mathbf{P}(r \mid \boldsymbol{x}, \theta) = c_r(\boldsymbol{x}) - c_{r-1}(\boldsymbol{x})$). The parameter vector θ is here given by $\theta = (\boldsymbol{w}, \beta_1, \ldots, \beta_{K-1})$. Note that, since the left-hand side in (20) is non-decreasing in k, the β_k need to satisfy the condition $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{K-1}$.

5 Comparison of Reduction Techniques

Different reduction techniques are not easily comparable, especially because the performance of a meta-technique also depends on the base learner that is used to instantiate this technique. In this section, we nonetheless make an attempt at elaborating on commonalities and differences between existing reduction techniques (namely RPC and CC) and our new proposal (LWD), albeit not in much detail and not on a very technical level.

5.1 Complexity

If the original label ranking data consists of $|\mathbb{D}| = N$ complete examples, then the total number of examples generated by RPC is $NK(K-1)/2$. Since CC generates an example (actually even two) for each pairwise comparison, too, the same (or even twice this) number of examples can be produced for this method. However, whereas RPC distributes these examples over $K(K-1)/2$ instance spaces $\mathbb{X}_{i,j}$, which are all identical to the original space $\mathbb{X}$, CC combines them in a single expanded feature space $\overline{\mathbb{X}}$ whose dimensionality is K times as high, and solves a single problem in this space. In any case, even when leaving the dimensionality of the input space aside, RPC is theoretically more efficient than CC if the underlying base learner has a superlinear complexity, say, $O(N^\alpha)$ with $\alpha > 1$. In fact, in that case, solving $K(K-1)/2$ problems of size N is less expensive than solving a single problem of size $NK(K-1)/2$—the complexity of the former is $O(K(K-1)N^\alpha)$, while the latter is in $O((K(K-1)N)^\alpha)$.

It should also be mentioned that, in its original version, CC only constructs pairwise comparisons between *consecutive* labels in a ranking, not between all labels (hoping to capture the other relations implicitly via transitivity). In this case, the total number of examples reduces to $N(K-1)$. Of course, the same approach could be applied to any other pairwise method, including RPC. In terms of prediction performance, however, it turns out that the redundancy of the full encoding has significant advantages.

LWD constructs K classification problems of size N, thus KN examples in total; like in RPC, each of these problems uses the original input space $\mathbb{X}$. The complexity is not directly comparable, however, since LWD solves ordinal classification problems, whereas RPC and CC solve binary problems. Using decomposition techniques like those proposed by Frank and Hall [8], each ordinal problem could again be reduced to $K - 1$ binary problems of the same size. Then, the overall complexity would be $O(K(K-1)N^\alpha)$, the same as for RPC.

Needless to say, a comparison becomes even more difficult in the case of incomplete training information. In that case, LWD requires methods for learning from imprecise data, such as (15). Therefore, the underlying base learners are no longer comparable.

In terms of space efficiency and complexity at prediction time, LWD may have an advantage in comparison to RPC, as it only needs to store and query a linear instead of a quadratic number of models. Again, however, since the LWD models are ordinal and the RPC models are binary classifiers, a direct comparison is not completely straightforward.

5.2 Loss of Information

Every reduction technique involves a certain loss of information. This can be seen most clearly from the fact

that, from the information preserved on the level of the decomposition, the original probability distribution $\mathbf{P}(\cdot) = \mathbf{P}(\cdot \mid \boldsymbol{x})$ on $\mathbb{S}_K$ cannot always be recovered. For example, the uniform distribution $\mathbf{P}(\bar{\pi}) \equiv (K!)^{-1}$ and the bimodal distribution $\mathbf{P}'(\bar{\pi}) = 1/2$ for $\bar{\pi} = (1, 2, \ldots, K)$ and $\bar{\pi} = (K, K-1, \ldots, 1)$ (and $= 0$ otherwise) both induce the distribution $\mathbf{P}(y_i \succ y_j) \equiv 1/2$ on the level of pairwise comparisons. Thus, even if these pairwise probabilities were learned correctly, there is no chance to predict the true ranking from them. Obviously, the reason for this loss of information is the decomposition process itself: Decomposing a set of complex objects (in our case rankings) into a set of simple objects (e.g., pairwise preferences), the latter does not necessarily allow to recover the former.

As an important consequence, risk minimizing predictions cannot be produced for all loss functions. For example, as shown in [14], RPC is able to minimize (in expectation) the Kendall loss (2) and the Spearman loss (4) but not the L_1 loss (3). LWD, on the other hand, is able to minimize both L_1 and L_2, just like any other labelwise decomposable loss—this can be seen immediately from (12). It cannot minimize losses like Kendall, however, since probabilities of label inversions cannot be recovered from rank-probabilities on individual labels.

5.3 Modeling Incomplete Rank Information

As mentioned before, training information will normally not be provided in the form of complete rankings $\bar{\pi} \in \mathbb{S}_K$; instead, only incomplete examples (1) are available as training data. For a label ranking method, the ability to handle such information in a proper way is therefore of utmost importance.

Methods based on pairwise comparisons, such as CC and RPC, do have this ability and can handle missing label information in a quite straightforward way. In RPC, for example, if a label y_k is missing for a training instance $\boldsymbol{x}_n$, then none of the pairwise learners $\mathcal{M}_{i,j}$ with $k \in \{i, j\}$ will get $\boldsymbol{x}_n$ as an example. Similarly, missing labels reduce the number of training examples in CC. Yet, the examples that are produced from the observed labels are still *precise*. In other words, although missing labels reduce the number of examples, they do not affect the type and information content of those examples that are still produced. Correspondingly, the same learning algorithms can be used, and since they are applied to smaller data sets, the learning process will even become more efficient.

This is an important difference to LWD. Here, even a single missing label may affect all examples that are produced for a training instance $\boldsymbol{x}_n$—the class information (position of the label) will become imprecise and/or uncertain. Correspondingly, standard methods for ordinal classification are no longer applicable; instead, generalized methods for learning from imprecisely labeled examples must be used. Thus, missing label information may affect the quality of all examples that are derived from an instance $\boldsymbol{x}_n$ and, moreover, tend to increase the complexity of the learning problem instead of reducing it. Seen from this perspective, learning from *comparative* preferences does indeed appear to be advantageous to learning from *absolute* preferences.

6 Experiments

In this section, we experimentally compare LWD with RPC and CC in terms of prediction accuracy. All three meta-techniques are implemented using logistic regression as a base learner; RPC and CC get along with the basic binary version, whereas LWD requires an extended ordinal variant (cf. Section 4.4).

6.1 Data

We used several benchmark data sets for label ranking that have also been used in previous studies [15]; these are semi-synthetic data sets, namely label ranking versions of (real) UCI multi-class data. Moreover, we used two real label ranking data sets: The Sushi data[1] consists of 5000 instances (customers) described by 11 features, each one associated with a ranking of 10 types of sushis. The Students data [2] consists of 404 students (each characterized by 126 attributes) with associated rankings of five goals (want to get along with my parents, want to feel good about myself, want to have nice things, want to be different from others, want to be better than others). See Table 1 for a summary of the data.

Two missing label scenarios were simulated, namely the missing-at-random setting (16) and the top-rank setting (13). In the first case, a biased coin is flipped for every label in a ranking to decide whether to keep or delete that label; the probability for a deletion is specified by a parameter $p \in [0, 1]$. Thus, $p \times 100\%$ of the labels will be missing on average. Similarly, in the second case, only the J top-labels in a ranking are kept, where J has a binomial distribution with parameters K and $1 - p$.

Table 1: Properties of the data sets.

data set	# inst. (N)	# attr. (d)	# labels (K)
authorship	841	70	4
glass	214	9	6
iris	150	4	3
pendigits	10992	16	10
segment	2310	18	7
vehicle	846	18	4
vowel	528	10	11
wine	178	13	3
sushi	5000	11	10
students	404	126	5

6.2 Results

The results in Tables 2 and 3 are presented as averages of 5×10-fold cross validation in terms of the Kendall correlation measure; other measures such as (3) and (4) led to similar results. These tables support the following conclusions: (i) LWD and RPC perform much better than CC, which is not competitive. (ii) Overall, the drop in performance due to missing labels is more pronounced in the missing-at-random than in the top-rank setting. (iii) Compared with RPC, LWD is quite competitive if rankings are (almost) complete—in this case, it tends to be even a bit better; on the other hand, it drops in performance more quickly in the case of missing label information (the difference was found significant for 30% and 60% missing rate in the missing-at-random setting, using a two-tailed sign test at the 5% level).

7 Summary and Conclusion

In this paper, we introduced and analyzed labelwise decomposition (LWD) as a new meta-learning technique for label ranking. In contrast to existing techniques, which are

[1] http://kamishima.new/sushi/

Table 2: Performance in terms of Kendall's tau on synthetic data: missing-at-random (above) and top-rank setting (below).

	complete ranking			30% missing labels			60% missing labels		
	LWD	RPC	CC	LWD	RPC	CC	LWD	RPC	CC
authorship	.913±.01	.910±.02	.594±.04	.860±.02	.888±.03	.559±.05	.682±.02	.874±.03	.357±.06
glass	.883±.04	.882±.04	.834±.06	.837±.04	.854±.04	.825±.06	.760±.04	.790±.06	.748±.07
iris	.928±.06	.885±.07	.828±.06	.809±.06	.875±.07	.802±.07	.712±.08	.772±.10	.729±.11
pendigits	.928±.00	.932±.00	.584±.01	.914±.00	.932±.00	.534±.01	.895±.00	.929±.00	.506±.01
segment	.943±.01	.934±.01	.628±.05	.923±.01	.932±.01	.560±.06	.895±.01	.919±.01	.556±.10
vehicle	.867±.02	.854±.02	.839±.02	.828±.02	.834±.03	.823±.03	.759±.03	.778±.03	.759±.05
vowel	.674±.02	.647±.02	.577±.03	.656±.02	.643±.02	.548±.03	.609±.02	.612±.02	.525±.02
wine	.908±.06	.921±.05	.847±.10	.882±.06	.894±.07	.790±.07	.743±.07	.855±.10	.775±.12
Avg. Rank	1.25	1.75	3	1.875	1.125	3	2.375	1	2.625
authorship	.913±.01	.910±.02	.594±.04	.913±.02	.903±.02	.582±.04	.909±.02	.893±.03	.544±.04
glass	.883±.04	.882±.04	.834±.06	.872±.05	.880±.04	.824±.06	.812±.11	.845±.04	.819±.05
iris	.928±.06	.885±.07	.828±.06	.924±.05	.884±.07	.811±.07	.902±.09	.850±.09	.797±.06
pendigits	.928±.00	.932±.00	.584±.01	.919±.00	.931±.00	.535±.01	.863±.01	.920±.00	.507±.00
segment	.943±.01	.934±.01	.628±.05	.932±.01	.932±.01	.555±.07	.891±.03	.916±.01	.529±.12
vehicle	.867±.02	.854±.02	.839±.02	.859±.02	.850±.02	.828±.03	.841±.03	.832±.03	.812±.03
vowel	.674±.02	.647±.02	.577±.03	.665±.03	.645±.02	.567±.02	.619±.03	.645±.02	.527±.02
wine	.908±.06	.921±.05	.847±.10	.904±.05	.917±.06	.822±.10	.896±.07	.916±.05	.783±.10
Avg. Rank	1.25	1.75	3	1.5	1.5	3	1.75	1.375	2.875

Table 3: Performance in terms of Kendall's tau on real-world data: missing-at-random (above) and top-rank setting (below).

sushi	0%	10%	20%	30%	40%	50%	60%	70%
LWD	.329±.010	.328±.009	.329±.010	.328±.009	.328±.010	.327±.009	.325±.010	.321±.010
RPC	.329±.010	.329±.010	.328±.009	.328±.009	.327±.009	.327±.010	.325±.009	.322±.010
CC	.075±.011	.072±.012	.072±.011	.072±.013	.070±.013	.069±.012	.065±.012	.060±.013
LWD	.329±.010	.329±.010	.329±.010	.329±.010	.328±.010	.325±.010	.323±.010	.319±.010
RPC	.329±.010	.329±.010	.329±.010	.329±.010	.328±.010	.326±.010	.324±.010	.321±.010
CC	.075±.011	.069±.013	.071±.012	.071±.013	.072±.012	.069±.012	.068±.011	.065±.009
students	0%	10%	20%	30%	40%	50%	60%	70%
LWD	.500±.046	.474±.053	.459±.055	.431±.050	.411±.054	.391±.054	.376±.059	.389±.066
RPC	.477±.037	.471±.052	.458±.052	.458±.056	.443±.063	.445±.044	.446±.052	.445±.045
CC	.455±.064	.424±.068	.339±.073	.304±.056	.316±.062	.284±.058	.274±.064	.268±.058
LWD	.500±.046	.497±.048	.499±.044	.496±.044	.481±.048	.451±.042	.420±.057	.397±.056
RPC	.477±.056	.460±.053	.456±.056	.452±.059	.445±.058	.441±.058	.449±.052	.445±.048
CC	.455±.064	.457±.067	.448±.069	.438±.064	.378±.065	.378±.058	.268±.072	.162±.073

based on decomposing training information into *comparative* preferences, this approach is based on *absolute* preference information in the form of ranks. The idea is quite simple: For each individual label, a model is learned that, given a query instance as an input, predicts the rank of the label in the associated ranking.

Technically, LWD reduces label ranking to ordinal classification problems with imprecise class information. Moreover, the aggregation step, which is responsible for combining the predictions of these classifiers into a complete label ranking, can be realized by means of an optimal assignment problem—this way, each labelwise decomposable loss function can be minimized in expectation.

Comparing LWD with state-of-the-art reduction techniques for label ranking, we did not find any systematic improvements in terms of prediction accuracy. On the contrary, although improvements could be achieved on several data sets in the case of (almost) complete training data, LWD seems to be more sensitive to missing label information. Actually, these results fully confirm our expectations, and can be explained by the fact that absolute preference information is more strongly affected by missing labels than relative preference information.

Overall, however, and especially in light of the unambitious expectations we started with, we found LWD to be surprisingly competitive. Moreover, one should keep in mind that LWD is a meta-learning technique whose performance is strongly influenced by the base learner. Since the implementation of this base learner is non-trivial, and the version used in this paper not necessarily optimal, there is certainly scope to improve this part of the method. Besides, LWD has other interesting properties. For example, while its performance is competitive to RPC, it only needs a linear instead of a quadratic number of models, which might not only be advantageous from a complexity point of view but also interesting with regard to the comprehensibility of a label ranker.

All things considered, we therefore believe that our results, despite not (yet) advancing the state-of-the-art in terms of performance, are promising enough to justify a further investigation of LWD as an alternative learning technique for label ranking. For future work, we therefore plan to explore this approach in more depth and to develop it further, with the goal to fully exploit its potential.

References

[1] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, editors. *Predicting structured data*. MIT Press, 2007.

[2] M. Boekaerts, K. Smit, and F.M.T.A. Busing. Salient goals direct and energise students' actions in the classroom. *Applied Psychology: An International Review*, 4(S1):520–539, 2012.

[3] K. Brinker and E. Hüllermeier. Case-based label ranking. In *Proceedings ECML–06, 17th European Conference on Machine Learning*, pages 566–573, Berlin, September 2006. Springer-Verlag.

[4] R.E. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. SIAM, 2009.

[5] W. Cheng, K. Dembczynski, and E. Hüllermeier. Label ranking based on the Plackett-Luce model. In

J. Fürnkranz and T. Joachims, editors, *Proceedings ICML–2010, International Conference on Machine Learning*, Haifa, Israel, 2010.

[6] W. Cheng, J. Hühn, and E. Hüllermeier. Decision tree and instance-based learning for label ranking. In *Proceedings ICML–2009, 26th International Conference on Machine Learning*, Montreal, Canada, 2009. [27% acceptance rate].

[7] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

[8] E. Frank and M. Hall. A simple approach to ordinal classification. In *Proc. ECML–2001, 12th European Conference on Machine Learning*, pages 145–156, Freiburg, Germany, 2001.

[9] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

[10] J. Fürnkranz and E. Hüllermeier, editors. *Preference Learning*. Springer-Verlag, 2011.

[11] Y. Grandvalet. Logistic regression for partial labels. In *IPMU–02, Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1935–1941, Annecy, France, 2002.

[12] Sariel Har-Peled, Dan Roth, and Dav Zimak. Constraint classification for multiclass classification and ranking. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS-02)*, pages 785–792, 2003.

[13] E. Hüllermeier and J. Beringer. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–440, 2006.

[14] E. Hüllermeier and J. Fürnkranz. On predictive accuracy and risk minimization in pairwise label ranking. *Journal of Computer and System Sciences*, 76(1):49–62, 2010.

[15] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1917, 2008.

[16] H.W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

[17] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Int. J. of Data Warehousing and Mining*, 3(3):1–13, 2007.

[18] S. Vembu and T. Gärtner. Label ranking: a survey. In J. Fürnkranz and E. Hüllermeier, editors, *Preference Learning*. Springer-Verlag, 2010.

Analysing the Potential Impact of Labeling Disagreements for Engineering Sensor Data

Andre Busche and **Ruth Janning** and **Lars Schmidt–Thieme**
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim
{busche, janning, schmidt-thieme}@ismll.uni-hildesheim.de

Abstract

We present the hyperbola recognition problem in Ground Penetrating Radar – GPR – data as an example for pattern recognition in complex engineering sensor data. Traditionally, GPR data are analyzed manually by human experts in a tedious and time-consuming process, e.g., to deduce the positioning of linear object underneath roads just before reconstruction works take place. For supporting this process using Machine Learning methods, one needs to have accurate ground truth data to derive models out of it. As an accurate acquisition of such annotated data is impossible even for a quasi-ideal case, we annotated 700 radargram images manually. This paper presents and discusses the outcomes of this study and concludes, that using just a single evaluation criteria to compare performances of GPR-focused Machine Learning methods might not be enough.

1 Introduction

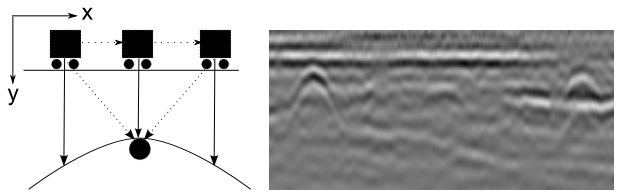
Ground Penetrating Radar (GPR) is used to investigate the shallow surface, e.g., to find buried landmines [Wilson *et al.*, 2007] or pipes and cables underneath (road) surfaces. Our current data is measured using an on-site vehicle equipped with a multi-channel array (multiple 'channels' are recorded at different frequencies and relative positions while the vehicle is moving) and illuminates structures in subsoil down to about 3-4 meters in depth. We are aiming at assisting the analysis process by means of probabilistic methods, while a special focus is put on the identification of pipes and cables of various types (e.g., PE, metal, stoneware) which are represented as hyperbola-like structures on measured radargram images (see Figure 1).

One ultimate goal of GPR data analysis is the derivation of supply maps, that is, maps of buried objects of a certain kind. The creation of those maps is required, e.g., as municipalities, according to one of our project partners and at least in Germany, seldomly have a single map of their buried structures. Instead, maps and plans of buried pipes and cables are cluttered and only available in a distributed manner, and can only be partially gathered and combined when requested by all parties owning buried objects, such as water supply companies, power supply companies, and telecommunication companies. In any case, those existing maps may be inaccurate and not recent, causing additional problems when highly accurate maps are required. This paper represents a first step towards to (semi-)automated creation of such supply maps, by means of developing supervised Machine Learning methods for an automated detection of such buried objects. The overall process can be split up into two distinct aspects: (a) the detection of individual objects in radargram images, and (b) the creating of supply maps out of individually detected object locations. This work focuses on the first aspect, whereas the latter aspect (b) can be tackled, e.g., by solutions as presented in [Chen and Cohn, 2011]. The collection of individual radargrams, resp. cross-sections, e.g., of a road, is done as follows: A specialized measurement vehicle drives at a constant speed along the x-axis (cf. Figure 1a) and measures a radargram image. This image visualizes reflected energies / intensities at discrete time points (y-axis). While in theory, one is able to *induce* the appearance of a radargram image out of known subsoil structures (e.g., by means of a numerical simulation software, see [Giannopoulos, 2003]), the reverse action - the *deduction of subsoil structures* out of radargram images - is a highly non-trivial task in real-world situations for a variety of reasons: (a) hyperbola reflections get distorted by supplementary reflections from horizontal layer breaks (see Figure 1, on the top right), (b) the signals' energy decays with increasing depth, resulting in lower (visual) contrasts (see our preprocessing in Figure 2 at the bottom of area 7), (c) the pipes' type and fillings (e.g., PE pipes filled with water, or being empty) causes multiple vertical reflections, (d) the depth–dependent energy decay of emitted radar wave requires preprocessing techniques and cause reflections to transition into background clutter at increasing depths.

Supporting the hyperbola recognition task by means of supervised models requires us of having a labeled set of radargram images at hand.

We will take the scenario of applying patch-based image-classification techniques as a running example. This requires us of having an a priori labeled dataset which contains patches (fractions of a radargram image) being labeled positive (patches containing hyperbola shapes) and negative (clutter; background noise).

For our data being measured on a test-site, GPS measurements exist for all pipes buried therein. Though one intuitively assumes that this GPS information helps for the creation of ground truth data (inducing the positions of hyperbola apexes out of the known subsoil structure), the opposite situation is the case, as (a) the heterogeneity of soil makes a consistent estimation of the actual pixel-*depth* impossible, (b) an unsteady movement of the vehicle needs to be interpolated and aligned to the radar traces, resulting in inaccurate *horizontal* pixel positions, (c) given the pipe identification task, *multiple reflections* occur, e.g., for water-filled pipes, while only the top-most reflection hyperbola can be derived from the GPS ground truth data. What



(a) GPR Data Collection Process (b) Real-world radargram (excerpt)

Figure 1: Hyperbolas are caused by steady movement of a radar vehicle across a buried pipe (a). At each horizontal position, an A-Scan (column) is measured, whose stacked visualization results in so-called radargrams, or B-Scans, (b).

immediately follows is an inaccuracy in pixel position in the ground truth data obtained from the given GPS data. This is the aspect we aim to improve by manually altering the apex positions derived from GPS annotations to match visual phenomena in the underlying radargram image. This paper is only concerned with obtaining and analyzing accurate hyperbola annotations at pixel level; the real-world deduction of pipe positions after identifying 'enough' candidate apex positions is not considered and can be achieved, e.g., utilizing methods as given in [Chen and Cohn, 2011].

Instead, this paper analyzes the outcome of these manual annotation sessions according to the following aspects:

1. **Is there a measurable difference between human annotation behavior for GPR data, and if so, can it be related to a certain aspect of the data?**

   While some reflections are clearly visible in the measured data, other cases (e.g., multiple reflections caused by the material of the pipe) exist, for which it is less intuitive to decide, whether or not those should be annotated. The question is now, to what extend humans agree on the visibility of these visual phenomena.

2. **Is it possible to gain an annotation set for complex engineering data which is inter-humanly agreed, or is the quality of annotations subject to personal taste?**

   If the annotation task can be identified as being related to personal taste, the question arises how the suitability and correctness of automatized methods trained on human tastes should be addressed.

2 Related Work

Relevant work is summarized from both the Machine Learning and Psychological perspective, with special emphasis on applications in real-world scenarios and possibilities for automatization, for which the human factor was identified to influence final results on the applicability of a technique.

2.1 Machine Learning Techniques for GPR Data Interpretation

The ultimate goal of GPR analysis is the derivation of complete and accurate tomographies based on usually just a small set of radargrams [Simi *et al.*, 2008; Chen and Cohn, 2011]. Before radargram images are fed into an automated algorithm, they are usually preprocessed. This process is usually visually [Pasolli *et al.*, 2009; Busche *et al.*, 2012] or methodologically [Chen and Cohn, 2010; 2011; Janning *et al.*, 2012a] driven.

One out of three different approaches for hyperbola detection in radargram images can be distinguished: (a) Estimation from sparse data [Chen and Cohn, 2010; Janning *et al.*, 2012a], (b) Brute-force methods, e.g., the Hough Transform, [Simi *et al.*, 2008], and (c) Supervised machine-learning, e.g., Neural Networks for patch-based classification, for which training data needs to be carefully collected beforehand (as we do here) [Al–Nuaimy*et al.*, 2000; Birkenfeld, 2010]. Our analysis discussed here influences each of these approaches, as those approaches need to be evaluated against some ground truth knowledge which, as we will show, is not obvious to obtain.

2.2 Psychological Aspects covered in this paper

For conducting manual annotation tasks, generally two groups of people can be distinguished: non-specialized humans having weak prior knowledge [Nowak and Rüger, 2010], and domain experts [Mello-Thoms, 2006; McCarley *et al.*, 2004; Volkmer *et al.*, 2005]. Many works were identified for other domains, e.g., videos / keyframes [Volkmer *et al.*, 2005] or texts [Nowak and Rüger, 2010], the one most similar being the one presented in [Klebanov *et al.*, 2008] for the text domain. No such work were identified for partial image annotations in the domain of complex engineering sensor data.

In the image domain, validating and adjusting previously defined and given annotations requires human annotators to establish a best-matching hypothesis, explaining which annotations correspond best to which structure in the raw data [Gregory, 1980]. Making final judgments on the suitability of an interpretation (being ones own or someone else's) is a non-trivial problem [Cavanagh, 2011; Nowak and Rüger, 2010], as mistakes may easily affects human life (e.g., wrong interpretations in medical screening [Mello-Thoms, 2006] or x-ray luggage screening at airports [McCarley *et al.*, 2004]). Multilayer interpretation of image (patches) by means of Neural networks [Sermanet *et al.*, 2009; Birkenfeld, 2010] is well known in the Machine Learning Community, while a thorough survey on their motivation based on the functioning of the brain is presented in [Rolls, 2012].

3 GPR Data Analyzer & Annotator

We first need to discuss GPR data preprocessing steps, as those might have an effect on the later visibility of subsoil structures (cf. Figure 2 on the left). The discussion continues with the presentation of our specialized GPR annotation software as shown in Figure 2 on the right which human annotators used to perform the manual annotation task.

3.1 Data Preprocessing

A multi-stage filter chain (presented in Figure 2 on the left) was used in the following way: the intensity values were modified to (1) have zero mean at each position (A-Scan) and (2) zero mean at each depth. Intensities are increased in a depth-dependent manner (3) to compensate the wave propagation loss. The characteristic reflection pattern was augmented using a pattern correlation filter (4): A sliding window of fixed length was moved along an A-Scan, calculating a correlation score against the sequence $(0_4, 1_5, 0_4)$ (the subscript denotes the number of repetitions), resulting in a new intensity value used for the following analy-

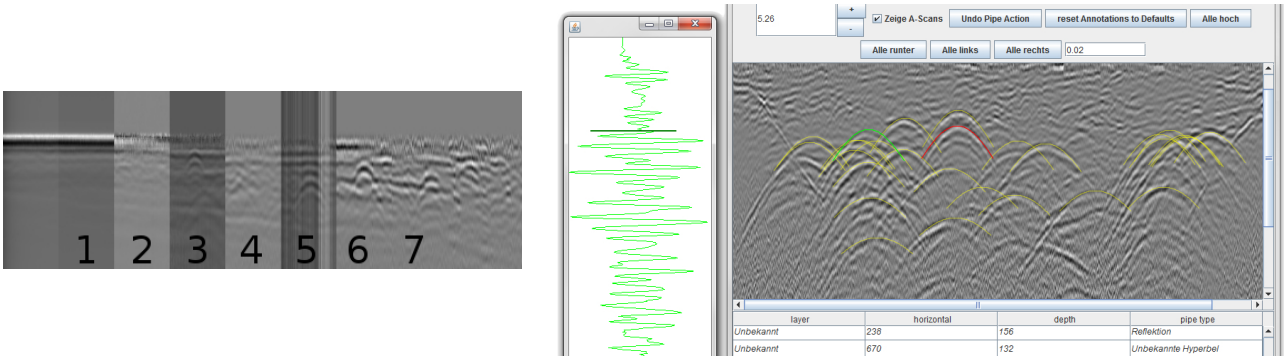


Figure 2: Our GPR Analyzer and Annotator preprocesses data as shown on the left. After preprocessing, a GUI shows basic information on individual pipes / hyperbolas and is used to interactively perform CRUD - **C**reate, **R**etrieve, **U**pdate and **D**elete - operations for hyperbolas.

sis. Intensity values for an A-Scan are adjusted (5) to have variance 1. Step (6) repeats step (1). Finally, (7) a depth-dependent windowed moving average subtraction filter is applied.

All but the pattern correlation filter are non-parametric; the pattern correlation filter corresponds to a rectangular function which approximates the vertical reflection pattern of a pipe [Busche *et al.*, 2012].

3.2 Interactive Controls

Given an initial seed annotation set (a list of pixel positions) and constant soil permittivity, our GPR Data Annotator stacks an interactive layer to create, modify and delete hyperbola annotations (yellow / bright on the Figures) on top of a raw radargram visualization panel. On the top pane, it is both possible to adjust the soil permittivity (using a homogeneity assumption; used to adjust the curvature of hyperbola annotations) and to jointly adjust the initial positions of all annotations (since close-by pipes cause distortions which are hard to distinguish). Hyperbola annotations are highlighted based on the mouse position, can be drag-dropped and contain further information (e.g., its type) shown in tabular form at the bottom. Optionally, A-Scans may be visualized (left).

4 Manually Annotating GPR Data

The annotation process of 350 radargram images in total was designed while having potential inter-human disagreements in mind. We aimed at having at least two annotation sets per radargram at the end, thus finally resulting in 700 images to be annotated. For the annotation process being repetitive, tedious, and long-lasting (an average of 4 minutes per image results in estimated 46 hours) and thus error-prone, we splitted the process into two phases, also allowing to incorporate feedback after round 1.

4.1 Annotation Protocol

The annotation protocol for the human annotators was as follows: Initially, a radargram image along with a seed set of annotations ('annotation set' in the following) obtained by the GPS measurements, were visualized on the user interface. Though an initial inspection of GPS annotation sets showed inaccuracies of up to $15cm$ in both horizontal and depth / vertical image direction ($1cm$ does not scale equally on both axes), those GPS seed annotations were shown, because the test site contains many near-by located pipes being represented by interfering and intersecting reflection patterns whose distinction is challenging even for human experts (see the right area in the radargram in Figure 2). The exact location of those near-by pipe apexes was therefore determined by a 'radargram-wise global best visual match' of all hyperbola annotations being present, as some hyperbola reflections were clearly visible (see the center of the radargram therein). After globally adjusting the annotation locations, individual positions of hyperbola annotations were altered.

4.2 Round 1

In the first phase, 6 human annotators were annotating 484 radargrams in total, spending 2 days on this task. The group of human annotators was composed of both domain experts (researchers working on the data on a daily basis) as well as semi-professionals (students) working with the data. 142 radargrams were annotated once, while two annotation sets were gained for 171 images. On average, 35 (± 3) radargrams per channel (1 out of 14 antenna configurations measured by the vehicle per measurement run) were annotated, while each human annotator saw each channel at least once (on average, 5.9 ± 3.8 radargrams / channel). The absolute amount of radargrams per human annotator ranged from 31 to 162.

4.3 Qualitative Results for Round 1

Feedback on the quality of annotation sets was gathered using a structured online feedback questionnaire, showing radargrams row-wise either in a comparison-style view for two existing annotation sets, or in a single radargram view, if just one annotation set existed. Three experts acting as human 'judges', two of which were also annotating the radargrams beforehand, the other one being a highly skilled GPR data analyst working at our project partners company, were visually investigating the quality of the annotation sets as follows. When comparing two annotation sets, (a) a vote could be casted, ranking one set over the other. Marking (b) a pair of annotation sets as being 'problematic' examples triggered a discussion with the human annotators before round 2. For all annotation sets, assigning the *incomplete* label to a radargram marked them as being a candidate for re-annotation in phase 2, if at least two votes ranked that set higher than its competitor. A *wrong* label removed the set from further analysis.

The fact, that an annotation set is 'accurate' given its radargram, was inferred if neither an incomplete nor wrong

Annotators' quality	1	2	3
accurate	26	16	15
incomplete	57	45	37
total annotations	227	250	229

Table 1: Final counts of annotation sets along with their judgement ('accurate', 'incomplete') obtained for radargram images by three human annotators 1, 2 and 3 after both annotation rounds.

label was assigned by any judge (this held true for 39 annotation sets, representing 8% of all annotation sets). We fed all 113 annotation sets (23%) which were marked as being incomplete but superior over the other into the set of radargram images to be annotated in phase 2.

From this first feedback cycle we got first evidence that obtaining a single, consistent and inter-humanly agreed set of annotated radargrams is much more challenging than previously expected. Therefore, we decided to assign annotation quotas, a fixed amount of radargrams to be annotated, to each human.

4.4 Round 2

After further analysis on quality and comparability of the judged results from round 1, we invited only 3 (of all 6) humans to annotate the remaining part, each having a different quota. The set of radargrams contained all remaining and incompletely labeled radargrams from round 1. On average 27 ($\pm$3.2) radargrams per channel were annotated, while each human annotator saw at least 2 images per channel (9 on average, $\pm$4).

4.5 Feedback and results from Round 2

We used the same feedback mechanism as used in round 1 for gathering feedback from the same judges, while this time only one of them also annotated the radargrams. 31 annotation sets (8%) were inferred to be 'accurate', while still 69 annotation sets (18%) are marked as being incomplete, but preferable over the other.

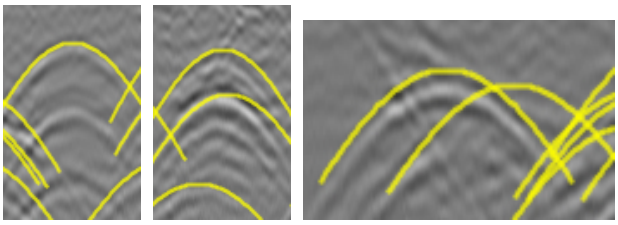
4.6 Result for both Rounds

Taking jointly both annotation rounds for just the three annotators from round 2 into account, we finally ended up with the counts of annotation sets per human annotator as shown in Table 1. As an example, annotator 3 labeled 229 radargram images in total, for which 15 (6.5%) images were judged accurate, 37 (16.2%) were judged incomplete.

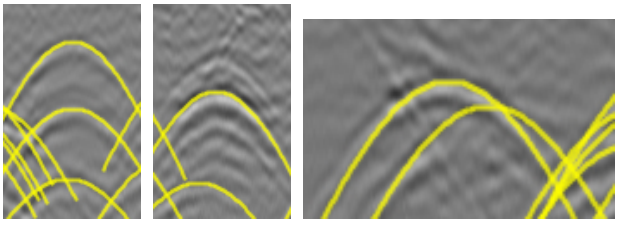
The peak of accurate annotations of annotator 1 is in line with Gregory's [Gregory, 1980] hypothesis who argues that prior experience strongly influences the perception and, in direct consequence, the quality of annotations (annotator 1 is more familiar with diverse sets of radargram images than annotators 2 and 3).

5 Analyzing and Discussing Human Annotation Behavior

To answer our questions stated in the beginning, we will now have a closer look at the results of the visual judgments with respect to its potential later automatization. The following error types are qualitatively introduced after having investigated both, all obtained annotation sets, and prior GPS annotations, and shall guide both our current analysis and future ideas and developments while developing algorithms for automatic GPR image data interpretation.



(a) Type A (b) Type B (c) Type B



(d) Type C (e) Type A (f) Type C

Figure 3: Analysis of humans' annotation behavior and qualitative categorization to the proposed error types.

1. **Type A Errors** are defined as being *non-annotated hyperbolas* in radargram images. Their counter phenomena, annotated hyperbolas without visual evidence, also falls in this category.
2. **Type B Errors** denote annotated hyperbolas having a *locational apex error* of a few pixels compared to the underlying radargram image.
3. **Type C Errors** correspond to a *wrongly assigned curvature* (estimated soil permittivity) information.

Most related work refers to Type A errors by measuring accuracy [Wilson *et al.*, 2007; Chen and Cohn, 2010], while Type B errors relate to RMSE – Root Mean Squared Error – scores on the apex positions (for which a prior matching of identified apex positions to the existing ground truth data is required) [Janning *et al.*, 2012a; Pasolli *et al.*, 2009; Janning *et al.*, 2012b]. Type C errors correspond to soil permittivity estimation [Simi *et al.*, 2008]. We presented a viable solution for its estimation once the apex is found [Busche *et al.*, 2012] and will not further focus on this error type / task here.

Our current, GPS-based annotation sets contain all three kinds of errors: (a) Type A errors are present through reflections, (b) Type B and Type C errors correspond to soil heterogeneity.

Figure 3 demonstrates a qualitative categorization of errors which were still contained in our human annotation sets after phase 1. In the Figure, each row corresponds to the same human annotator, while the same radargram patches are each used to visualize different annotation phenomena.

5.1 Assessing Differences in Annotation Behavior

To answer our first question, *whether or not a measurable difference between human annotation behavior exist*, we will first have a look at the quantity of pipes annotated (Type A errors), while thereafter having a closer look at Type B errors.

As our human annotations were derived from seed GPS-based annotation sets (annotations denoting previously known objects), we are able to measure two characteristic

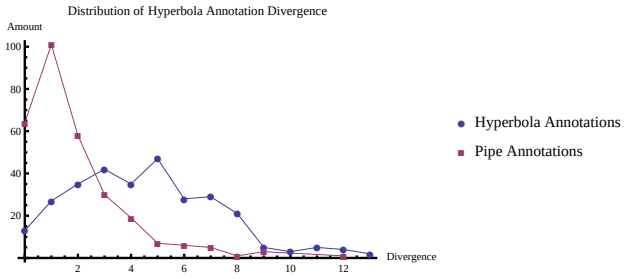


Figure 4: Inter human analysis of annotation behavior: The absolute differences exceed 12 hyperbolas (not shown here), resulting in diverging model performances at a later stage.

Visibility Loss	PE	Stoneware	Steel
Round 1	83%	41%	57%
Round 2	72%	36%	50%

Table 2: Percentage of pipe (type) annotations which are still present after the individual annotation rounds in the manual annotation sets (humans were able to identify visual evidence for this pipe type).

quantities for a given radargram, if two annotation sets are present: (a) For previously *existing* annotations, calculating the difference between the amount of removed pipes in the both annotation sets (by two different human annotators) indicates how well the final annotation sets match the data / the visual presence of reflection phenomenas. As an example, if both humans agreed on removing *the same* 5 annotations (denoting the same pipe annotations) from their annotation sets, the absolute difference of both annotation set sizes is 0. One can deduce that they both agreed on the fact that for 5 pipes, no visual evidence exists, whereas for all other pipe annotations, a corresponding visual evidence existed in the radargram. (b) For previously *unknown* hyperbolic shapes that have been *added* to the annotation sets, e.g., to mark reflections, measuring their absolute difference gives insights in whether or not these are identifiable by humans (distinguishable for background clutter). Take, as another example, a pair of annotation sets whose additional hyperbola annotations differ by 3: Then, one annotator was able to identify 3 more hyperbolic shapes on the radargram image.

Figure 4 visualizes the counts of differences between the size of two annotations sets for the same radargram image. Therein, *Pipe Annotation* denotes case (a) from above, and *Reflection Annotation* denotes case (b). We deduce from the high counts for low absolute differences in the *pipe reflections* case that *direct* pipe reflections are quite noticeable for humans. Contrary, the rather high counts for larger differences for *Reflection Annotations* (cf. Figure 3 (a) and (d)) indicate that the annotation of reflections is either more likely subject to personal taste, or is subject to ambiguity (e.g., the contrast is too low for humans to reliably distinguish them from the background).

As we have now identified a general difference depending on the type of annotation, a closer look at the fraction of remaining annotations per pipe type (the amount / kind of annotations *not* being removed) reveals certain notable characteristics, as shown in Table 2.

As can be seen, e.g., only 41% of the stoneware annotations being present in the GPS-based annotation sets were still existing after the first annotation round. Since humans were asked to retain only visible annotations, one may either deduce that (a) identifying these types is more challenging, (b) the current preprocessing techniques are not optimal for those pipe types, or (c) their absolute depth and filling or surrounding material causes masking effects, which are, compared with Figure 4, rather consistent for the individual pipe types.

	GPS vs. human	human vs. human	round
PE	18.3 ± 11.1	8.2 ± 11.0	1
	18.7 ± 13.7	7.3 ± 8.8	2
Stone-ware	21.6 ± 16.6	6.6 ± 4.9	1
	20.5 ± 19.0	7.3 ± 6.4	2
Steel	23.2 ± 15.7	9.7 ± 14.6	1
	22.7 ± 13.8	12.2 ± 16.8	2

Table 3: Averaged locational distances between pipe annotations in different annotation sets. Clearly visible is an inter-human agreement that pipes are not located at the GPS positions (distances to GPS are large), but at other locations (inter-human distances are small)

For sure, some of these differences are also influenced by experience and familiarity of the human annotators while working on GPR data. Before answering the second question, we are having a closer look at the individual locational differences when manually adjusting the apex positions for two cases (shown in Table 3): (a) The *GPS vs. human* comparison computes for all humanly created annotation sets for all annotations therein the RMSE / euclidean distance against the apex positions from the ground truth derived from the GPS measurements. (b) The *human vs. human* comparison computes RMSE differences on retained pipe annotations to compare how the individual manipulation of hyperbola apexes matches between humans. In other words, for both cases, we compare the distances of hyperbola apexes (type B errors) by assuming that lower distances correspond to more accurate annotation qualities, as annotations more closely match visual phenomena. This already gives first insights in inter-human agreements, that is, how close individual apex positions in two annotation sets created by humans for the same radargram are.

Even though these numbers are biased against the smaller quantity of pairs of annotation sets contained in the *human vs. human* comparison, a clear trend towards more consistent and accurate apex estimations for the inter-human comparison case is visible.

5.2 Gaining Inter-Human Agreed Annotation Sets

To answer our second question, *whether it is possible to get an inter-humanly agreed annotation set for complex engineering data*, we may both refer to table 3, indicating that there is a rather low average pixel-distance between human annotation sets, and present an indicative result when comparing the rankings of judges on their own annotation sets, that is, having a look at whether a judge favourizes his own annotation set over an annotation set of another human annotator.

Even tough we have only limited data (there are just three cases (2 annotators have also been judges in phase 1, whereas only 1 annotator was a judge in phase 2), we may take the following results as an indication: Ones own

annotations in round 1 were preferred by the first annotator in $77/96$ (80%), the other in $46/66$ (70%) of all cases. Less indicative is round 2, for which $16/28$ (57%) annotations were preferred.

Even though we are able to derive that humans tend to favourize their own annotation sets, relating these to the appropriateness or suitability of the annotation sets to the *hyperbola identification task* is not easily possible: As we have seen above, a rather high deletion rate of annotations, esp. for certain pipe types, could be observed, even though it was well known to all participants that those pipes actually exist. The only valid conclusion which may be drawn here is that ones own *interpretation* on how to solve the pipe annotation task differs in a constant way between humans.

6 Impact of the Labeling Accuracy for Machine Learning Algorithms

We evaluated the quality of our annotation sets by performing a simple classification experiment using a state of the art classifier for patch-based classification, namely a Convolutional Neural Network as implemented in the eblearn library [Sermanet *et al.*, 2009]. The network structure is the well-known Lenet-5 network. We note that we are not primarily seeking for an optimal classification result, but aim at validating the suitability of the annotation sets, that is, whether the annotated apex locations obey an underlying structure in the radargram image which is easy to generalize.

Using the set of radargram images for which we obtained 'accurate' annotations, we created a dataset and splitted it in a leave-one-out fashion per individual radargram image. One radargram image was used for validating the classifier, whereas performance scores are reported on a test radargram image.

We used a grid search to determine an optimal hyperparameter combination for patches of size 32×32 as follows: The set of learning rates was set to $\{5, 1, 0.1, 0.5, 0.05\} \cdot 10^{-3}$, while different $l1$ and $l2$ regularizations, each being set to $\{0, 10^{-2}, 10^{-3}, 10^{-4}\}$, were tested as well. Positive training instances were created by using all patches being centered at annotated hyperbola apex positions, as well as using their neighbourhood, given that the amount of neighboring pixels in those patches were overlapping by 95%. Patches with an overlap between 95% and 30% were discarded to not introduce class boundary ambiguities. Negative training instances were randomly sampled from the remainder of the radargram image at a 2% rate, resulting in a class imbalance of approx. $1:7$, that is, seven times as much negative training instances than positive training instances, resulting in a baseline accuracy for a constant classifier being about 87.5%.

Table 4 shows the performance assessments for two dataset variants for two techniques to derive an optimal model given the performance scores on the validation set. The patches used to create the dataset may either be normalized, resulting in a pixel contrast range per individual patch over the whole greyscale from $[0, 255]$, or not normalized, for which the raw patches as present in a preprocessed radargram images are used. The algorithm as implemented in eblearn is trained by minimizing an 'energy' value. We determine a model to be used for evaluating its performance on the test set on both, the minimal energy value on the validation set (Accuracy (energy)), as well as on the maximal accuracy score on the validation set (Accuracy (correct)). Performance scores in Table 4 are both showing accuracy scores for comparability.

Our main aim is to compare the lift of the accurate annotations over those ones obtained from the GPS measurements. What can be seen is that for all four combinations, when combining both dataset variants with both performance assessment scores, our manual annotations increase the classifier performance. We conclude that our manually obtained annotation sets more closely match characteristic patterns within the radargram images, compared to those ones as obtained by the GPS measurement.

7 Conclusion and Future Work

This paper presented our methodology in annotating 700 GPR images, representing one example for the annotation task of complex engineering sensor data. Based upon the initial finding that the derivation of an accurate ground truth from a priori measured (GPS-) data is impossible for our current task at hand, we successfully showed that humans are able to improve the overall annotation quality. Anyhow, due to the still large fraction of 'inaccurately' annotated radargram images, we need to note that this overall process is costly, so that the question on the generalizability of this approach to other domains arises.

For our specific use case, we are now able to define certain subsets of the data, representing different 'agreement levels' between humans, that is, proportions of annotation sets for which a majority of the jugdes agree on their quality, as follows:

1. **Validation Dataset**: 57 'accurate' radargram images being well balanced between 3 human annotators compose a small dataset for which inter-human agreement exists on a high-quality annotation
2. **Scale-Up Dataset**: 139 incomplete, but not wrongly, annotated radargram images allow us to simulate automated analysis in semi-observed scenarios for evaluating performances under presence of noisy and partially ambiguous annotations.
3. **Inter- / Intra-Human Consistency checks**: For either set, models can be tested on their ability to generalize over either human annotators, or radargram images, in a controlled environment.

Besides of having an accurately labeled GPR data corpus, we showed that the annotated proportions within the underlying radargram images are more easily to generalize compared to the annotations derived from the GPS data. For reaching our goal of deriving supply maps, we are now able to proceed with improving state of the art Machine Learning Models for the detection of patches containing hyperbolic structures.

For similar use cases (in other application domains), the two main outcomes of this study are as follows: (a) Annotation quality improves at the cost of introducing inter-human disagreements to the annotations, and (b) Multiple evaluation metrices are desired to assess model performance. Though the quality and thus the suitability increases in general, the 'human factor' introduces ambiguities in the ground truth data. These need to be taken into account by designing and using evaluation measures which consider these aspects, e.g., by using an accuracy score that considers a locational displacement of a few pixels still as being correct.

normalized?	annotation	Accuracy (correct)	Accuracy (energy)
true	GPS	88.58 (3.72)	91.20 (1.80)
	accurate	89.55 (2.90)	92.57 (1.27)
false	GPS	87.00 (3.93)	90.70 (2.03)
	accurate	90.21 (2.35)	92.41 (2.19)

Table 4: The high quality of our manually obtained annotation sets is validated by performing a simple classification experiment comparing the annotations obtained by the GPS measurements with the ones obtained during the manual annotation rounds. Figures show accuracy scores and their variances in brackets.

Acknowledgments

Special thanks go to Mr. Marek Naser acting as our external expert to judge the annotations. This work is co-funded by the European Regional Development Fund project AcoGPR (Adaptive Contactless Ground Penetrating Radar) under the grant agreement no. WA3 80122470 (`http://acogpr.ismll.de`)

References

[Al–Nuaimy*et al.*, 2000] W. Al–Nuaimy, Y. Huang, M. Nakhkash, M.T.C Fang, V.T. Nguyen, and A. Eriksen. *Automatic detection of buried utilities and solid objects with GPR using neural networks and pattern recognition*. Journal of Applied Geophysics, Vol 43 Nr. 24, pp. 157–165, 2000.

[Birkenfeld, 2010] S. Birkenfeld *Automatic Detection of Reflexion Hyperbolas in GPR Data with Neural Networks*. World Automation Congress, pp. 1–6, 2010.

[Busche *et al.*, 2012] A. Busche, R. Janning, T. Horvath, L. Schmidt-Thieme *A Unifying Framework for GPR Image Reconstruction*. Proc. of the 36nd Annual Conference of the Gesellschaft fr Klassifikation (GfKl 2012), 2012.

[Cavanagh, 2011] P. Cavanagh *Visual cognition*. Vision Res. 51(13), pp. 1538-1551, 2011.

[Gregory, 1980] R.L. Gregory *Perceptions as Hypotheses*. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, Vol. 290, No. 1038, The Psychology of Vision, pp. 181–197, 1980

[Janning *et al.*, 2012a] R. Janning, T. Horvath, A. Busche, and L. Schmidt-Thieme *GamRec: a Clustering Method Using Geometrical Background Knowledge for GPR Data Preprocessing*. 8th AIAI, 2012

[Janning *et al.*, 2012b] R. Janning, T. Horvath, A. Busche, and L. Schmidt-Thieme *Pipe Localization by Apex Detection*. Proc. of the IET Intl. Conf. on radar systems (Radar), Glasgow, Scotland, 2012.

[Klebanov *et al.*, 2008] B.B. Klebanov, E. Beigman, and D. Diermeier *Analyzing disagreements*. Proc. of the WS on Human Judgements in Computational Linguistics (HumanJudge), pp. 2–7, 2008.

[Chen and Cohn, 2010] H. Chen, and A.G. Cohn *Probabilistic robust hyperbola mixture model for interpreting ground penetrating radar data*. IJCNN IEEE, pp. 1–8, 2010.

[Chen and Cohn, 2011] H. Chen, and A.G. Cohn *Buried Utility Pipeline Mapping Based on Multiple Spatial Data Sources: A Bayesian Data Fusion Approach*. IJCNN IEEE, 2411-2417, 2011.

[Giannopoulos, 2003] A. Giannopoulos *Modeling ground penetrating radar by GprMax*. Non Destructive Testing: Selected papers from Structural Faults and Repair, 755-762, 2003.

[McCarley *et al.*, 2004] J.S. McCarley, A.F. Kramer, C.D. Wickens, E.D. Vidoni, and W.R. Boot *Visual Skills in Airport-Security Screening*. Psychological Science, vol. 15 no. 5 pp. 302–306, 2004

[Mello-Thoms, 2006] C. Mello-Thoms *The problem of image interpretation in mammography: effects of lesion conspicuity on the visual search strategy of radiologists*. The British Journal of Radiology, 79, pp. 111-116, 2006

[Nowak and Rüger, 2010] S. Nowak, and S. Rüger *How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation*. In: Proc. of the Intl. Conf. on Multimedia Information Retrieval (MIR '10), pp. 557–566, 2010.

[Pasolli *et al.*, 2009] E. Pasolli, F. Melgani, and W. Donelli *Automatic Analysis of GPR Images: A Pattern-Recognition Approach*. IEEE Transactions Geoscience and Remote Sensing. Col 46, No. 7, pp. 2206–2217, 2009.

[Rolls, 2012] E. Rolls *Invariant visual object and face recognition: neural and computational bases, and a model, VisNet*. Front. Comput. Neurosci., 19, 2012

[Sermanet *et al.*, 2009] P. Sermanet, K. Kavukcuoglu, and Y. LeCun *EBLearn: Open-Source Energy-Based Learning in C++*. Proc. Intl. Conf. on Tools with Artificial Intelligence, IEEE, 2009

[Simi *et al.*, 2008] A. Simi, A. Bracciali, and G. Manacorda *Hough transform based automatic pipe detection for array GPR: Algorithm development and on-site tests*. Proc. of Radar, pp. 1–6, 2008.

[Volkmer *et al.*, 2005] T. Volkmer, J.R. Smith, and A. Natsev *A web-based system for collaborative annotation of large image and video collections: an evaluation and user study*. Proc. of the 13th annual ACM intl. conf. on Multimedia (MULTIMEDIA '05), pp. 892–901, 2005.

[Wilson *et al.*, 2007] J.N. Wilson, P. Gader, L. Wen-Hsiung H. Frigui, and K.C. Ho *A Large-Scale Systematic Evaluation of Algorithms Using Ground-Penetrating Radar for Landmine Detection and Discrimination*. IEEE Transactions on Geoscience and Remote Sensing, vol.45, no.8, pp. 2560–2572, 2007.

Towards Multilabel Rule Learning

Eneldo Loza Mencía and Frederik Janssen
Knowledge Engineering Group, TU Darmstadt
Darmstadt, Germany
eneldo@ke.tu-darmstadt.de, janssen@ke.tu-darmstadt.de

Abstract

In this position paper, we provide first insights into possible schemes to utilize rule learning algorithms to solve the task of multilabel classification. The main idea is to exploit specific properties of symbolic rule representations to build models that consist of high-quality multilabel rules. To this end, novel ideas which rely on the adaptation of conventional inductive rule learners to multilabel data are presented. Their expected advantages and disadvantages, opportunities and limitations are reviewed and discussed.

1 Introduction

Rule learning has a very long history and is a well-known problem in the machine learning community. Over the years many different algorithms to learn a set of rules were introduced. The main advantage of rule-based classifiers are interpretable models as rules can be easily comprehended by humans. Also, the structure of a rule offers the calculation of overlapping of rules, more specific, and more general relations. Thus, the rule set can be easily modified as opposed to most statistical models such as SVMs or neural networks. However, most rule learning algorithms are prone to multi-class classification.

On the other hand, many problems involve assigning more than one single class to an object. These so-called multilabel problems can be often found when text is classified into topics or tagged with keywords, but there are also many examples from other media such as the recognition of music instruments or emotions in audio recordings or the classification of scenes in images and from the domain of biology and gene function classification.

It is widely accepted that one major issue in learning from multilabel data is the exploitation of label dependencies. Learning approaches may greatly benefit from considering label correlations, and we believe that rule induction algorithms provide a good base for this. Firstly, label dependencies can directly be modeled and expressed in form of rules. Secondly, such rules are directly interpretable and comprehensible for humans. Even if complex and long rules are generated, the implication between classes can be estimated more easily than with other approaches by focusing on the part of the rules considering the classes.

In this paper, we present current work in progress and perspectives towards multilabel rule learning. Relatively little work exist regarding rule learners taking into account the popularity of multilabel classification. An overview of related work shows the current possibilities and limitations of such approaches. The challenges in rule induction and multilabel learning are reviewed and two general directions are proposed and discussed.

2 Related Work

Many rule-based approaches to multilabel learning rely on association rules. This is an obvious choice as this kind of rules is capable of having more than one condition in the head of the rule. However, as the goal of all classification algorithms is to assign classes to examples, usually Classification Association Rules (CARs) are used, instead of regular association rules that are induced in an unsupervised fashion. Often, these single-label association rules are introduced as a first step and then are combined to yield multilabel association rules or are used to directly predict the labels of a given test instance. The latter works by using all single-label association rules that cover the example and predict all labels that are in the head of these rules. However, in this case, the model does not consist of multilabel association rules.

The literature shows only a few approaches to multilabel rule learning. Most of them utilize association rule learning to induce the set of rules. As mentioned above, often the capability of the algorithms to handle multilabel data does not stem from the representation of the model (i.e., by using multilabel rules) but is reached by employing certain classification schemes. The approach of Arunadevi and Rajamani (2011) operates on spatial data. Single-label association rules are learned by using a multi-objective genetic algorithm. Then, the rules are sorted by a weighted combination of support, confidence and J-measure, and the final classifiers is produced according to this ranking.

In the same manner as Arunadevi and Rajamani (2011), Ávila *et al.* (2010) use a genetic algorithm to induce the single-label association rules. However, they use a decision list for classification of single labels. The multilabel prediction is also built by using a combination of all covering rules of the different rule sets. They also account for a good distribution of the labels by using a token-based recalculation of the fitness values of each rule.

Li *et al.* (2008) also learn single-label association rules and the test data is labeled by setting exactly those labels that have a probability greater than 0.5 in the covering rules.

Another method that can be applied to tackle multilabel data are the so-called multilabel alternating decision trees (De Comité *et al.*, 2003). The idea is to adapt boosting techniques to multilabel classification. As a result, the algorithm yields rules that have only one decision (similar to decision stumps) and that predict confidence values for

each label.

A different idea is to change the model representation to make it suitable for multilabel data. Consequently, the rule representation has to be generalized to multilabel, i.e., a label vector instead of a single value in the head of the rules. In the work of Allamanis *et al.* (2013), such a generalized rule format is introduced. Interestingly, the proposed rules also allow for postponing the classification by offering a "*don't care*"-value. As there may be cases where the rule is not confident enough or simply when no rule covers the example such a value may be beneficial. In this work, a Michigan-style Learning Classifier System (LCS) is used in combination with a genetic algorithm. The classification is done by using a weighted voting scheme (the fitness of the rules is used as weight) as many multilabel rules may cover the example.

Another algorithm that also finds multilabel rules is *MMAC* (Thabtah *et al.*, 2004). The idea here is to use a multi-class, multilabel associative classification approach by not only generating from all frequent itemsets the rules that pass the confidence threshold but also include the second best rules and so on. These single-label association rules then are merged to create multilabel rules. The algorithm proceeds by deriving the frequent itemsets, generating the association rules, removing the covered instances, and repeat these steps on the remaining instances. Hence, rules that have the same conditions in the body then can be merged by using their single-label classes in the multilabel vector in the head of the rule. In this manner it is possible to create a total ranking of all labels for each test instance.

Another associate multilabel rule learner with several possible labels in the head of the rules was developed by Thabtah *et al.* (2006). These labels are found in the whole training set, while the multilabel lazy associative approach of Veloso *et al.* (2007) generates the rules from the neighborhood of a test instance during prediction. The advantage then is that fewer training instances are used to compute the coverage statistics which is beneficial when small disjuncts are a problem as they are often predicted wrong due to whole training set statistics. Another important aspect mentioned in this work is that essentially one assumes dependencies between the labels. Otherwise, multilabel data can be simply solved by decomposing it into single-label datasets by using schemes such as binary relevance. Surprisingly, Veloso *et al.* (2007) was the only work that mentioned this problem. Their solution is simple as they use the prediction of a first iteration as additional attribute in the dataset for a second iteration. This lasts as long as labels remain unused in the attribute section of the dataset.

In summary, most of the relevant work is based on classification association rules (CARs). Often, evolutionary algorithms are used to derive a high-quality rule set. Label dependencies are not tackled explicitly though they might be taken into account by algorithm-specific properties.

3 Multilabel Classification

Multilabel classification refers to the task of learning a function that maps instances $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}$ to label subsets or label vectors $\mathbf{y} = (y_1, \ldots, y_n) \subset \{0,1\}^n$, where $\mathcal{L} = \{\lambda_1, \ldots, \lambda_n\}$, $n = |\mathcal{L}|$ is a finite set of predefined labels and where each label attribute y_i corresponds to the absence (0) or presence (1) of label λ_i . Thus, in contrast to multiclass learning, alternatives are not assumed to be mutually exclusive, such that multiple labels may be associated with a single instance.

Potentially, there are 2^n different allowed allocations of $\mathbf{y}$, which is a dramatic growth compared to the n possible states in the multiclass setting. This, and especially the resulting correlations and dependencies between the labels in $\mathcal{L}$, make the multilabel setting particularly challenging and interesting compared to the classical field of binary and multiclass classification.

From a probabilistic point of view, one of the main differences between multilabel and binary or multiclass classification are the possible dependencies in the label output space. In binary and multiclass problems the only observable probabilistic dependence is between the input variables, i.e. the attributes x_j, and the label variables y_i. A learning algorithm tries to learn exactly this dependence in form of a classifier function h. In fact, if a classifier provides a score or confidence for its prediction $\hat{\mathbf{y}}$, this is often regarded as an approximation of $P(\mathbf{y} = \hat{\mathbf{y}} \mid \mathbf{x})$, i.e. the probability that $\hat{\mathbf{y}}$ is true given a document $\mathbf{x}$.

As mentioned above, we may additionally observe dependencies between labels in multilabel classification. I.e. we may observe that the occurrence or absence of single labels under certain circumstances *correlate* with each other. From the early beginning of multilabel classification, there have been attempts to exploit these types of *label correlations* (cf. e.g. McCallum, 1999; Ghamrawi and McCallum, 2005; Zhu *et al.*, 2005). A middle way is followed by Read *et al.* (2009) and Dembczyński *et al.* (2010a) and their (probabilistic) classifier chains by stacking the underlying binary relevance classifiers with the predictions of the previous ones. However, only recently Dembczyński *et al.* (2010b) provided a clarification and formalization of label dependence in multilabel classifications. Following their argumentation, one must distinguish between unconditional and conditional label dependence. Roughly speaking, *unconditional dependence* or independence of labels does not depend on a specific given input instance (the condition) while *conditional dependence* does. An example may illustrate this.

Suppose a label space indicating topics from news articles, and suppose further that λ_u is the topic *politics* and λ_v corresponds to *foreign affairs*. Especially if the topics are organized in a hierarchy and λ_v is a sub-topic of λ_u, there will obviously be a dependency between both labels. We will hence observe y_u with a different probability $P(y_u = 1) < 1$ as if y_v was also observed, since then it holds $P(y_u = 1 | y_v = 1) = 1$. The probability $P(y_v = 1 | y_u = 1)$ of seeing an article about *foreign affairs* on a page in the politics section will in turn be also much higher than by just randomly opening the newspaper, which corresponds to $P(y_v = 1)$. These probabilities are *unconditional* since they do not depend on a particular document. Suppose now that a news article is about the *Euro crisis*. The *conditional* probabilities $P(\lambda_u = 1|\mathbf{x})$, $P(\lambda_v = 1|\mathbf{x})$ and $P(y_v = 1 | y_u = 1, \mathbf{x})$ would likely increase and hence be different from the unconditional ones. However, if an article was about the *cardiovascular problems of Ötzi*, we would observe that both labels are *conditionally independent*, since (very likely) $P(y_u = a | y_v = b, \mathbf{x}) = P(y_u | \mathbf{x}) = 1 - a$ for all $a, b \in \{0,1\}$ and interchanged u and v.

4 Inductive Rule Learning

Inductive rule learning is researched very well. Over the years the community has introduced a bunch of algorithms that are still in use (cf., *Ripper* (Cohen, 1995) as one of the

popular examples). However, most multilabel rule learning algorithms rely on association rule mining (cf., Section 2), the combination of inductive rule learners and multilabel data is yet to be evaluated.

A rule learning algorithm has a set of rules as result. These rules are of the form

$$body \rightarrow head$$

where the body consists of a number of conditions (attribute-value tests) and, in the regular case, the head has only one single condition of the form $y_i = 0$ or 1. However, multilabel rules may have several of such conditions.

Most inductive rule learning algorithms for classification employ a separate-and-conquer (SeCo) strategy (Fürnkranz, 1999). Its basic idea is to find a rule that covers a part of the example space that is not explained by any learned rule yet (the conquer step). The possible candidates are evaluated according to a quality function (heuristic) defined on statistics of covered positive and negative examples. After such a rule is found, it is added to the current set of rules, and all examples that are covered by this rule are removed from the data set (the separate step). Then, the next rule is searched on the remaining examples. This procedure is repeated until no more (positive) examples are left. In order to prevent overfitting, the two constraints that all examples have to be covered (*completeness*) and that no negative example has to be covered in the binary case (*consistency*) can be relaxed so that some positive examples may remain uncovered and/or some negative examples may be covered by the set of rules.

Obviously, there are some shortcomings when the SeCo strategy should be employed on multilabel data. First of all, there is no direct and intuitive notion of positive and negative examples (cf. also Section 5) This affects the computation of the heuristics for selecting the candidate conditions.

Secondly, a SeCo algorithm is usually learned in order to subsequently cover the examples of each possible class (ordered one-against-all). This is obviously not longer possible in the multilabel setting, since an example may belong to different classes. Hence different decomposition approaches and stopping criterions have to applied in the multilabel settings.

5 Multilabel Rule Learning

The predominant approach in multilabel classification is *binary relevance* learning Tsoumakas and Katakis (cf. e.g. 2007). It tackles a multilabel problem by learning one classifier for each class, using all objects of this class as positive examples and all other objects as negative examples. There exists hence a strong connection to concept learning, which is dedicated to infer a model or description of a target concept from specific examples of it (see e.g. Domingos, 1997, Sec. 2.2). When several target concepts are possible or given for the same set of instances, we formally have a multilabel problem.

The problem of this approach is that each label is considered independently of each other, and as we have seen by the example given before, this can lead to loss of useful information for classification. This problem is commonly shared by all approaches mentioned in Section 2 which can contain only one condition, i.e. one label in the head of a rule.

5.1 Labelsets Approach

A rule induction approach which may consider several conditions in the head seem hence more appropriate for the multilabel setting. A possible simple solution is to use the label powerset transformation (Tsoumakas and Katakis, 2007), which decomposes the initial problem into a multiclass problem with $\{P_{\mathbf{x}} \mid \mathbf{x} \in \text{training set}\} \subseteq 2^{\mathcal{L}}\}$ as possible classes. This problem can then be processed with common rule induction algorithms, which will thus produce rules with several labels in the head.

In general, we can state that this approach is able to consider conditional dependency between labels of high order when using a separate and conquer approach, since rules are learned locally on subsets of the instances. However, an obvious disadvantage is that we only can only predict label relations and combinations which were seen in the training data. Hence, no new relationships can be discovered, and we may miss the correct labelsets in unknown test data.

We propose to modify the SeCo iteration as explained in the following: Firstly, we learn so-called multiclass decision lists, which allows to use different heads in the rules of the the decision list. If we limit ourselves to labelsets seen in the training data, this corresponds to using label powerset transformation with a multiclass decision list learner, with the mentioned shortcomings. In addition, the evaluation for each possible labelset can be very expensive ($\mathcal{O}(\min(2^n, m))$ in contrast to $\mathcal{O}(n)$). Hence, we propose a greedy approach. It starts by evaluating the current added condition with respect to all labels independently in order to determine the best covered label. If we add an additional label to our head, we can only stay the same or get worse, since the number of covered examples remain the same and the number of covered *positives*, for which the head applies, can not increase. Hence, we can safely prune great part of the label combinations as soon as the heuristic becomes worse.

Several aspects of this approach have to be analyzed. Firstly, it is not clear whether the greedy refinement step leads to mostly single label heads. Secondly, an interesting issue is the effect of allowing negative predictions, i.e. heads of the type $y_i = 0$. This is somehow contrary to the notion of concept learning, where we are interested in finding convenient representations of *the concept*, but it is in line with the label symmetry assumption of binary relevance and many other multilabel approaches. And thirdly, it has to be analyzed if this approach is indeed effective in predicting labelsets which could not be observed in the training set.

5.2 Chaining and Bootstrapping

An effective approach for exploiting conditional label dependencies showed to be classifier chains (Read *et al.*, 2009). Classifier chains (CC) make use of stacking the previous binary relevance predictions in order to implement the chain rule in probability theory $P(y_1, \ldots, y_n) = P(y_n \mid y_1, \ldots, y_{n-1})$, since they learn the binary classifiers h_i with training examples of the form $(x_1, \ldots, y_1, \ldots, y_{i-1})$ (Dembczyński *et al.*, 2010a). One drawback of CC is the predetermined, fixed order of the classifiers (and hence the labels) in the chain, which makes it impossible to learn dependencies in the contrary direction.

Thus, we propose to use a bootstrapping approach in order to benefit from the chaining rule effect but also in order to overcome the main disadvantage of CC, the fixed order. As we will see, our version of bootstrapping is particularly adequate for rule induction.

Like in binary relevance, we learn one theory for each label, but we expand our training instances by the label information of the other labels, i.e. the training examples vectors for learning label y_i is $(x_1, \ldots, y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. Hence, we obtain theories with label attributes in the body, like in CC. The prediction for a test instance begins with empty label attributes, which means that they are set to *unknown*. Here we benefit from the natural support for such attribute states (*missing*, *don't care*, etc.) of symbolic approaches. Hence, in the first iteration, only rules apply which do not include any label attribute in the body. These rules were generated during the training process because there was enough local evidence and support for such a decision, which is only based on the instance attributes. This would be hardly reasonably and justifiable if we were using approaches like SVMs, which are in general not excluded from being used in similar bootstrapping settings. The prediction is then used in the next iteration to set the corresponding label attribute for the other classifiers. However, if no appropriate rule was found we prefer to absent from classifying instead of applying the default rule (predicting the majority class) so that the attribute may be filled up in consequent iterations. Again, rule induction algorithm naturally provide this option.

A deadlock may of course occur if no rules apply at all. We are currently investigating this issue also with respect to using different heuristics, but the overall preliminary results are very promising.

Nevertheless, the next natural step is to skip the binary relevance decomposition and to (virtually) apply bootstrapping directly in the SeCo training phase, hence to learn one single theory with rules with label conditions in the body.

6 Conclusions

This work deals with the challenges and chances of using rule induction in multilabel learning. We have presented two main perspectives. The first one addresses the fact that multilabel learning has to deal with sets of classes rather than single classes. The second one addresses the problem of label dependencies by using bootstrapping. In essence, both issues are solved by extending the formulation of the head and the body of a rule with additional conditions on the labels. First experiments with the bootstrapping approach make us confident about the potential of multilabel rule induction. However, we are still at the beginning of implementing all the presented ideas.

Moreover, many other aspects have still to be addressed: The right selection of the heuristic was already a complex issue in traditional rule induction and has to be reviewed for multilabel learning. Also, unordered and multiclass decision lists gain new relevance. And of course, a combination of both approaches, leading to global rules describing multilabel data, is also worth to be investigated.

References

M. Allamanis, F. Tzima, and P. Mitkas. Effective Rule-Based Multi-label Classification with Learning Classifier Systems. In *Adaptive and Natural Computing Algorithms, 11th International Conference, ICANNGA 2013*, pages 466–476, 2013.

J. Arunadevi and V. Rajamani. An evolutionary multi label classification using associative rule mining for spatial preferences. *IJCA Special Issue on Artificial Intelligence Techniques - Novel Approaches and Practical Applications*, (3):28–37, 2011.

J. Ávila, E. Galindo, and S. Ventura. Evolving Multi-label Classification Rules with Gene Expression Programming: A Preliminary Study. In *Hybrid Artificial Intelligence Systems*, volume 6077, pages 9–16. Springer, 2010.

W. W. Cohen. Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning (ICML-95)*, pages 115–123, 1995.

F. De Comité, R. Gilleron, and M. Tommasi. Learning multi-label alternating decision trees from texts and data. In *Proceedings of the 3rd international conference on Machine learning and data mining in pattern recognition*, MLDM'03, pages 35–49, Berlin, Heidelberg, 2003. Springer-Verlag.

K. Dembczyński, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 279–286. Omnipress, June 2010.

K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence in multi-label classification. In *Proceedings of the ICML-10 Workshop on Learning from Multi-Label Data*, pages 5–12, June 2010.

P. Domingos. A Unified Approach to Concept Learning. Dissertation, University of California, Irvine, 1997.

J. Fürnkranz. Separate-and-Conquer Rule Learning. *Artificial Intelligence Review*, 13(1):3–54, February 1999.

N. Ghamrawi and A. McCallum. Collective multi-label classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200. ACM, 2005.

B. Li, H. Li, M. Wu, and P. Li. Multi-label Classification based on Association Rules with Application to Scene Classification. In *Proceedings of the 2008 The 9th International Conference for Young Computer Scientists*, pages 36–41. IEEE Computer Society, 2008.

A. K. McCallum. Multi-label text classification with a mixture model trained by EM. In *AAAI 99 Workshop on Text Learning*, 1999.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier Chains for Multi-label Classification. In *Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer, 2009.

F. Thabtah, P. Cowling, and Y. Peng. MMAC: A New Multi-Class, Multi-Label Associative Classification Approach. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 217–224. IEEE Computer Society, 2004.

F. Thabtah, P. Cowling, and Y. Peng. Multiple labels associative classification. *Knowledge and Information Systems*, 9(1):109–129, 2006.

G. Tsoumakas and I. Katakis. Multi Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

A. Veloso, W. Meira, M. A. Gonçalves, and M. Zaki. Multi-label lazy associative classification. In *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 605–612. Springer, 2007.

S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 274–281. ACM, 2005.

Generalizing Generalized Cores – An Analysis of Tag-Recommender Evaluation Procedures

Stephan Doerfel
University of Kassel
Wilhelmshöher Allee 73
34121 Kassel, Germany
doerfel@cs.uni-kassel.de

Robert Jäschke
L3S Research Center
Appelstraße 4
30167 Hannover, Germany
jaeschke@l3s.de

Abstract

Since the rise of collaborative tagging systems on the web, the *tag recommendation task* – suggesting suitable tags to users of such systems while they add resources to their collection – has been tackled. However, the (offline) evaluation of tag recommendation algorithms usually suffers from difficulties like the sparseness of the data or the cold start problem for new resources or users. Previous studies therefore often used so-called *post-cores* (specific subsets of the original datasets) for their experiments. In this paper, we generalize the notion of a core by introducing the new notion of a *set-core* – that is independent of any graph structure – to overcome a structural drawback in the construction of post-cores. We complement the theoretical results with a large-scale experiment in which we analyze different tag recommendation algorithms on different classes of cores on three real-world datasets.

1 Acknowledgements

Part of this research was funded by the DFG in the project "Info 2.0 – Informationelle Selbstbestimmung im Web 2.0".

Towards Optimal Active Learning for Matrix Factorization in Recommender Systems

Rasoul Karimi, Christoph Freudenthaler, Alexandros Nanopoulos, Lars Schmidt-Thieme
University of Hildeshiem
Samelsonplatz 1, University of Hildesheim, D-31141 Hildesheim, Germany
karimi,freudenthaler,nanopoulos,schmidt-thieme@uni-hildesheim.ismll.de

Abstract

This is the extended abstract of the paper which has already been published in the proceeding of IEEE ICTAI 2011 Conference (http://www.cse.fau.edu/ictai2011/). It applies active learning technique to new user problem in recommender systems.

1 Introduction

Recommender systems help web users to address information overload in a large space of possible options [1]. In many applications, such as in e-commerce, users have too many choices and too little time to explore them all. Moreover, the exploding availability of information makes this problem even tougher.

Collaborative filtering is the popular technique for recommender systems. Nevertheless, recent research (especially as has been demonstrated during the Netflix challenge[1]) indicates that Matrix Factorization (MF) is a superior prediction model compared to other approaches [2].

Evidently, the performance of collaborative filtering depends on the amount of information that users provide regarding items, most often in the form of ratings. However, a well identified problem is that users are reluctant to provide information for a large amount of items [3; 4]. This fact impacts negatively the quality of generated recommendations. A simple and effective way to overcome this problem, is by posing queries to new users in order that they express their preferences about selected items, e.g., by rating them. Nevertheless, the selection of items must take into consideration that users are not willing to answer a lot of such queries. To address this problem, *active learning* methods have been proposed to acquire those ratings from users, that will help most in determining their interests [4; 3].

2 Proposed Method

In this paper, we propose a novel method for applying active learning in recommender systems. Due to the rapidly increasing interest in MF as a powerful prediction model in recommender systems, the proposed method introduces an active learning approach designed to take into account the characteristics of MF in order to improve its accuracy. The proposed method is inspired from optimal active learning for regression problem. Assuming the distribution of the test data is known, it is possible to find the optimal active learning algorithm for specific regression models [5]. As MF is actually a regression problem, it makes sens to use the same approach for active learning in MF. Given the test items are known, we develop a method which approximates the optimal active learning for MF. It capitalizes on the updating mechanism of MF and allows us to formulate a new criterion for the selection of the queried items, in terms of reducing the expected prediction error. A detailed experimental evaluation is performed, whose results demonstrate the superiority of the proposed method. Our results provide insight into the effectiveness of the proposed criterion for selecting the queried items, as it compares favorably to methods that use MF but are based on simplistic criteria.

3 Experimental Result

In this section, we examine experimentally the performance of the proposed method.

3.1 Experimental set up

The main challenge in applying active learning for recommender systems is that users are not willing to answer many queries in order to rate the queried items. For this reason, we report the performance of all examined methods in terms of prediction error (MAE) versus the number of queried items, which is simply denoted as the *number of queries*. Non-myopic active learning [6], and random selection are used as the baseline.

We use the MovieLens(100K)[2] dataset in our experiments. MovieLens contains 943 users and 1682 items. The dataset was randomly split into training and test sets. The training dataset consists of 343 users (the same number used in [4]) and the rest of users are in the test dataset. Each test user is considered as a new user. The latent features of the new user are initially trained with three random ratings. 20 rated items of each test user are separated to compute the error. The test items are not new item and already appeared in the training data. The remaining items are in the pool dataset, i.e the dataset that is used to select a query. For simplicity, we assume that the new user will always be able to rate the queried item. In our experiment, 10 queries are asked from each new user. Therefore, the pool dataset should contain at east 10 items which exist in the training data. Considering 10 queries and 20 test items, each test user has given ratings to at least 30 items.

3.2 Results

Figure 1 illustrates the comparison between the proposed method, non-myopic active learning [6], and random selection in terms of MAE as a function of the number of

[1] www.netflixprize.com

[2] www.grouplens.org/system/files/ml-data0.zip

queried items. The non-myopic method works well in the first queries but it finally converges to the random selection. This convergence also happens for active learning in AM [3]. Generally, this evidence holds for active learning methods aiming to improve the new user parameters using some heuristics. In the optimization theory, usually the heuristics provide a good performance only if the difference between current solution and optimal solution is high. At first, as the new user has provided a few ratings, the new user parameters are inaccurate and are far away from the optimal parameters. But as more ratings are provided by the new user, the accuracy of the estimated parameters also increases and the heuristic-based methods do not gain much improvement. However, the proposed method in this paper has a different approach. It aims to directly optimize the test error. That is why its performance continues and does not converge to the random selection. Therefore, if the new user is ready to provide more ratings, the proposed method can efficiently use them to improve the accuracy.

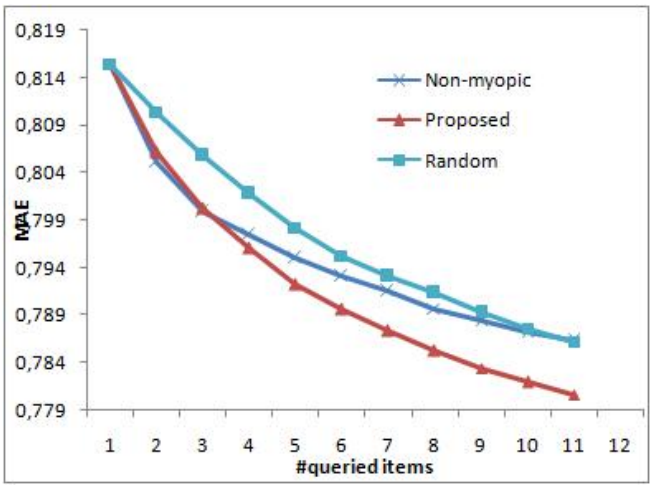


Figure 1: MAE results of the proposed active learning, non-myopic and random. Smaller MAE means better accuracy

References

[1] R. Burke, "Hybrid recommender systems," *User Modeling and User Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, pp. 30–37, 2009.

[3] R. Jin and L. Si, "A bayesian approach toward active learning for collaborative filtering," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.

[4] A. S. Harpale and Y. Yang, "Personalized active learning for collaborative filtering," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 91–98.

[5] D. A. Cohn, G. Z., and M. Jordan, "Active learning with statistical models," in *Advances in Neural Information Processing Systems(NIPS)*, 1995.

[6] R. Karimi, C. Freudenthaler, A. Nanopoulos, and L. Schmidt-Thiemee, "Non-myopic active learning for recommender systems based on matrix factorization," in *IEEE Information Reuse and Integration (IRI)*. IEEE, 2011.

Rectifying Classifier Chains for Multi-Label Classification

Robin Senge[1], **Juan José del Coz**[2] **and Eyke Hüllermeier**[1]
[1]:Department of Mathematics and Computer Science, University of Marburg, Germany
[2]:Artificial Intelligence Center, University of Oviedo at Gijón, Spain

Abstract

Classifier chains have recently been proposed as an appealing method for tackling the multi-label classification task. In addition to several empirical studies showing its state-of-the-art performance, especially when being used in its ensemble variant, there are also some first results on theoretical properties of classifier chains. Continuing along this line, we analyze the influence of a potential pitfall of the learning process, namely the discrepancy between the feature spaces used in training and testing: While true class labels are used as supplementary attributes for training the binary models along the chain, the same models need to rely on estimations of these labels at prediction time. We elucidate under which circumstances the attribute noise thus created can affect the overall prediction performance. As a result of our findings, we propose two modifications of classifier chains that are meant to overcome this problem. Experimentally, we show that our variants are indeed able to produce better results in cases where the original chaining process is likely to fail.

1 Introduction

Multi-label classification (MLC) has attracted increasing attention in the machine learning community during the past few years. Apart from being interesting theoretically, this is largely due to its practical relevance in many domains, including text classification, media content tagging and bioinformatics, just to mention a few. The goal in MLC is to induce a model that assigns a *subset* of labels to each example, rather than a single one as in multi-class classification. For instance, in a news website, a multi-label classifier can automatically attach several labels—usually called tags in this context—to every article; the tags can be helpful for searching related news or for briefly informing users about their content.

Current research on MLC is largely driven by the idea that optimal predictive performance can only be achieved by modeling and exploiting *statistical dependencies* between labels. Roughly speaking, if the relevance of one label may depend on the relevance of others, then labels should be predicted *simultaneously* and not *separately*. This is the main argument against simple *decomposition techniques* such as binary relevance (BR) learning, which splits the original multi-label task into several independent binary classification problems, one for each label.

Until now, several methods for capturing label dependence have been proposed in the literature. They can be categorized according to two major properties: (i) the size of the subsets of labels for which dependencies are modeled and (ii) the type of label dependence they seek to capture. Looking at the first property, there are methods that only consider pairwise relations between labels [5; 6; 14; 19] and approaches that take into account correlations among larger label subsets [12; 13; 17]; the latter include those that consider the influence of all labels simultaneously [2; 8; 11]. Regarding the second criterion, it has been proposed to distinguish between the modeling of *conditional* and *unconditional label dependence* [3; 4], depending on whether the dependence is conditioned on an instance [3; 11; 13; 16] or describing a kind of global correlation in the label space [2; 8; 19].

In this paper, we focus on a method called *classifier chains* (CC) [13]. This method enjoys great popularity, even though it has been introduced only lately. As its name suggests, CC selects an order on the label set—a *chain* of labels—and trains a binary classifier for each label in this order. The difference with respect to BR is that the feature space used to induce each classifier is extended by the previous labels in the chain. These labels are treated as additional attributes, with the goal to model conditional dependence between a label and its predecessors. CC performs particularly well when being used in an ensemble framework, usually denoted as *ensemble of classifier chains* (ECC), which reduces the influence of the label order.

Our study aims at gaining a deeper understanding of CC's learning process. More specifically, we address an issue that, despite having been noticed [4], has not been picked out as an important theme so far: Since information about preceding labels is only available for training, this information has to be replaced by estimations (coming from the corresponding classifiers) at prediction time. As a result, CC has to deal with a specific type of attribute noise: While a classifier is learnt on "clean" training data, including the true values of preceding labels, it is applied on "noisy" test data, in which true labels are replaced by possibly incorrect predictions. Obviously, this type of noise may affect the performance of each classifier in the chain. More importantly, since each classifier relies on its predecessors, a single false prediction might be propagated and possibly even reinforced along the whole chain.

The contribution of this paper is twofold. First, we analyze the above problem of classifier chains in more detail. Using both synthetic and real data sets, we design experiments in order to reveal those factors that influence the effect of error propagation in CC. Second, we propose and

evaluate modifications of the original CC method that are intended to overcome this problem.

The rest of the paper is organized as follows. The next section introduces the setting of MLC more formally, and Section 3 explains the classifier chains method. Section 4 is devoted to a deeper discussion of the aforementioned pitfalls of CC, along with some first experiments for illustration purposes.[1] In Section 5, we introduce modifications of CC and propose a method called *nested stacking*. An empirical study, in which we experimentally compare this method with the original CC approach, is presented in Section 6. The paper ends with a couple of concluding remarks in Section 7.

2 Multi-Label Classification

Let $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ be a finite and non-empty set of class labels, and let $\mathcal{X}$ be an instance space. We consider a MLC task with a training set $S = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, generated independently according to a probability distribution $\mathbf{P}(\mathbf{X}, \mathbf{Y})$ on $\mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{Y}$ is the set of possible label combinations, i.e., the power set of $\mathcal{L}$. To ease notation, we define $\boldsymbol{y}_i$ as a binary vector $\boldsymbol{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,m})$, in which $y_{i,j} = 1$ indicates the presence (relevance) and $y_{i,j} = 0$ the absence (irrelevance) of λ_j in the labeling of $\boldsymbol{x}_i$. Under this convention, the output space is given by $\mathcal{Y} = \{0, 1\}^m$. The goal in MLC is to induce from S a hypothesis $\boldsymbol{h} : \mathcal{X} \longrightarrow \mathcal{Y}$ that correctly predicts the subset of relevant labels for unlabeled query instances $\boldsymbol{x}$.

The most straightforward and arguably simplest approach to tackle the MLC problem is *binary relevance* (BR) learning. The BR method reduces a given multi-label problem with m labels to m *binary classification* problems. More precisely, m hypotheses $h_1, h_2, \ldots, h_m$ are induced, each of them being responsible for predicting the relevance of one label, using $\mathcal{X}$ as an input space:

$$h_j : \mathcal{X} \longrightarrow \{0, 1\} \tag{1}$$

In this way, the labels are predicted independently of each other and no label dependencies are taken into account.

In spite of its simplicity and the strong assumption of label independence, it has been shown theoretically and empirically that BR performs quite strong in terms of decomposable loss functions [3], including the well-known *Hamming loss*:

$$L_H(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h_i(\boldsymbol{x})]\!] \tag{2}$$

The Hamming loss averages the standard 0/1 classification error over the m labels and hence corresponds to the proportion of labels whose relevance is incorrectly predicted. Thus, if one of the labels is predicted incorrectly, this accounts for an error of $\frac{1}{m}$. Another extension of the standard 0/1 classification loss is the *subset 0/1 loss*:

$$L_{ZO}(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = [\![\boldsymbol{y} \neq \boldsymbol{h}(\boldsymbol{x})]\!] \tag{3}$$

Obviously, this measure is more drastic and already treats a mistake on a single label as a complete failure. The necessity to exploit label dependencies in order to minimize the generalization error in terms of the subset 0/1 loss has been shown in [3].

[1] This section is partly based on [15]

3 Classifier Chains

While following a similar setup as BR, classifier chains (CC) seek to capture label dependencies. CC learns m binary classifiers linked along a chain, where each classifier deals with the binary relevance problem associated with one label. In the training phase, the feature space of each classifier in the chain is extended with the actual label information of all previous labels in the chain. For instance, if the chain follows the order $\lambda_1 \rightarrow \lambda_2 \rightarrow \ldots \rightarrow \lambda_m$, then the classifier h_j responsible for predicting the relevance of λ_j is of the form

$$h_j : \mathcal{X} \times \{0, 1\}^{j-1} \longrightarrow \{0, 1\} \ . \tag{4}$$

The training data for this classifier consists of instances $(\boldsymbol{x}_i, y_{i,1}, \ldots, y_{i,j-1})$ labeled with $y_{i,j}$, that is, original training instances $\boldsymbol{x}_i$ supplemented by the relevance of the labels $\lambda_1, \ldots, \lambda_{j-1}$ preceding λ_j in the chain.

At prediction time, when a new instance $\boldsymbol{x}$ needs to be labeled, a label subset $\boldsymbol{y} = (y_1, \ldots, y_m)$ is produced by successively querying each classifier h_j. Note, however, that the inputs of these classifiers are not well-defined, since the supplementary attributes $y_{i,1}, \ldots, y_{i,j-1}$ are not available. These missing values are therefore replaced by their respective predictions: y_1 used by h_2 as an additional input is replaced by $\hat{y}_1 = h_1(\boldsymbol{x})$, y_2 used by h_3 as an additional input is replaced by $\hat{y}_2 = h_2(\boldsymbol{x}, \hat{y}_1)$, and so forth. Thus, the prediction $\boldsymbol{y}$ is of the form

$$\boldsymbol{y} = \big(h_1(\boldsymbol{x}),\, h_2(\boldsymbol{x}, h_1(\boldsymbol{x})), \ldots \big)$$

Realizing that the order of labels in the chain may influence the performance of the classifier, and that an optimal order is hard to anticipate, the authors in [13] propose the use of an ensemble of CC classifiers. This approach combines the predictions of different random orders and, moreover, uses a different sample of the training data to train each member of the ensemble. *Ensembles of classifier chains* (ECC) have been shown to increase predictive performance over CC by effectively using a simple voting scheme to aggregate predicted relevance sets of the individual CCs: For each label λ_j, the proportion $\hat{w}_j$ of classifiers predicting $y_j = 1$ is calculated. Relevance of λ_j is then predicted by using a threshold t, that is, $\hat{y}_j = [\![\hat{w}_j \geq t]\!]$.

4 The Problem of Attribute Noise in Classifier Chains

The learning process of CC violates a key assumption of supervised learning, namely the assumption that the training data is representative of the test data in the sense of being identically distributed. This assumption does not hold for the chained classifiers in CC: While using the *true* label data y_j as input attributes during the training phase, this information is replaced by *estimations* $\hat{y}_j$ at prediction time. Needless to say, y_j and $\hat{y}_j$ are not guaranteed to follow the same distribution; on the contrary, unless the classifiers produce perfect predictions, these distributions are likely to differ in practice (in particular, note that the $\hat{y}_j$ are deterministic predictions whereas the y_j normally follow a non-degenerate probability distribution).

From the point of view of the classifier h_j, which uses the labels $y_1, \ldots, y_{j-1}$ as additional attributes, this problem can be seen as a problem of *attribute noise*. More specifically, we are facing the "clean training data vs. noisy test data" case, which is one of four possible noise scenarios that have been studied quite extensively in [20]. For CC,

this problem appears to be vital: Could it be that the additional label information, which is exactly what CC seeks to exploit in order to gain in performance (compared to BR), eventually turns out to be a source of impairment? Or, stated differently, could the additional label information perhaps be harmful rather than useful?

This question is difficult to answer in general. In particular, there are several factors involved, notably the following:

- *The length of the chain*: The larger the number $j-1$ of preceding classifiers in the chain, the higher is the potential level of attribute noise for a classifier h_j. For example, if prediction errors occur independently of each other with probability ϵ, then the probability of a noise-free input is only $(1-\epsilon)^{j-1}$. More realistically, one may assume that the probability of a mistake is not constant but will increase with the level of attribute noise in the input. Then, due to the recursive structure of CC, the probability of a mistake will be reinforced and increase even more rapidly along the chain.
- *The order of the chain*: Since some labels might be inherently more difficult to predict than others, the order of the chain will play a role, too. In particular, it would be advantageous to put simpler labels in the beginning and harder ones more toward the end of the chain.
- *The accuracy of the binary classifiers*: The level of attribute noise is in direct correspondence with the accuracy of the binary classifiers along the chain. More specifically, these classifiers determine the input distributions in the test phase. If they are perfect, then the training distribution equals the test distribution, and there is no problem. Otherwise, however, the distributions will differ.
- *The dependency among labels*: Perhaps most interestingly, a (strong enough) dependence between labels is a prerequisite for both, an improvement and a deterioration through chaining. In fact, CC cannot gain (compared to BR) in case of no label dependency. In that case, however, it is also unlikely to loose, because a classifier h_j will most likely[2] ignore the attributes $y_1, \ldots, y_{j-1}$. Otherwise, in case of pronounced label dependence, it will rely on these attributes, and whether or not this is advantageous will depend on the other factors above.

In the following, we present two experimental studies that are meant to illustrate the above issues. Based on our discussion so far and these experiments, two modifications of CC will then be introduced in the next sections, both of them with the aim to alleviate the problems outlined above.

4.1 First Experiment

Our intuition is that attribute noise in the test phase can produce a propagation of errors through the chain, thereby affecting the performance of the classifiers depending on their position in the chain. More specifically, we expect classifiers in the beginning of the chain to systematically perform better than classifiers toward the end. In order to verify this conjecture, we perform the following simple experiment: We train a CC classifier on 500 randomly generated label orders. Then, for each label order and each position, we compute the performance of the classifier on that position in terms of the relative increase of classification error compared to BR. Finally, these errors are averaged *position-wise* (not label-wise). For this experiment, we used three standard MLC benchmark data sets whose properties are summarized in Table 1 (shown in Section 5).

[2]The possibility to ignore parts of the input information does of course also depend on the type of base classifier used.

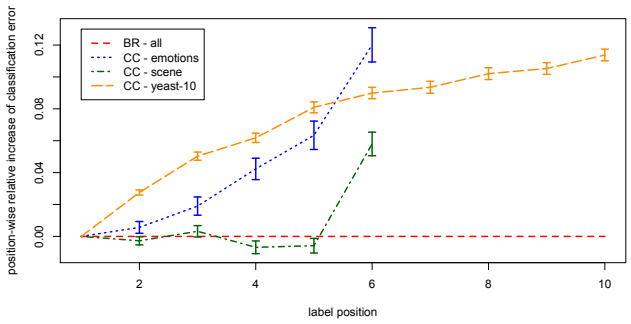


Figure 1: Results of the first experiment: position-wise relative increase of classification error (mean plus standard error bars). The *yeast-10* data set used here is a reduced yeast data set containing only the ten most frequent labels and their instances.

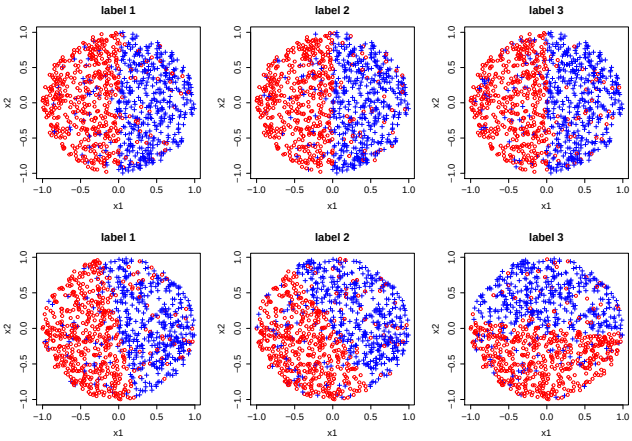


Figure 2: Example of synthetic data: the top three labels are generated using $\tau = 0$, the three at the bottom with $\tau = 1$.

The results in Figure 1 clearly confirm our expectations. In two cases, CC starts to loose immediately, and the loss increases with the position. In the third case, CC is able to gain on the first positions but starts to loose again later on.

4.2 Second Experiment

In a second experiment, we use a synthetic setup that was proposed in [4] to analyze the influence of label dependence. The input space $\mathcal{X}$ is two-dimensional and the underlying decision boundary for each label is linear in these inputs. More precisely, the model for each label is defined as follows:

$$h_j(\boldsymbol{x}) = \begin{cases} 1 & a_{j,1}x_1 + a_{j,2}x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The input values are drawn randomly from the unit circle. The parameters $a_{j,1}$ and $a_{j,2}$ for the j-th label are set to

$$a_{j,1} = 1 - \tau r_1, \quad a_{j,2} = \tau r_2 \quad , \tag{6}$$

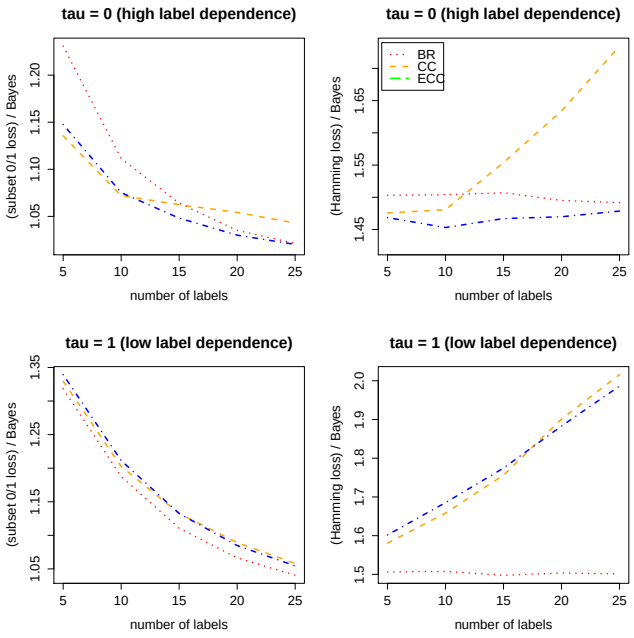


Figure 3: Results of the second experiment for $\tau = 0$ (top—high label dependence) and $\tau = 1$ (bottom—low label dependence).

with r_1 and r_2 randomly chosen from the unit interval. Additionally, random noise is introduced for each label by independently reversing a label with probability $\pi = 0.1$. Obviously, the level of label dependence can be controlled by the parameter τ. Figure 2 shows two example data sets with three labels. The first one (pictures on the top) is generated with $\tau = 0$, the second one (bottom) with $\tau = 1$. As can be seen, the label dependence is quite strong in the first case, where the model parameters (6) are the same for each label. For the second case, the model parameters are different for each label. There is still label dependence, but certainly less pronounced.

For different label cardinalities $m \in \{5, 10, 15, 20, 25\}$, we run 10 repetitions of the following experiment: We created 10 different random model parameter sets (two for each label) and generated 10 different training sets, each consisting of 50 instances. For each training set, a model is learnt and evaluated (in terms of Hamming and subset 0/1 loss) on an additional data set comprising 1000 instances.

Figure 3 summarizes the results in terms of the average loss divided by the corresponding Bayes loss (which can be computed since the data generating process is known); thus, the optimum value is always 1. Apart from BR and CC, we already include the performance curve for the method to be introduced in the next section (NS); this should be ignored for now. Comparing BR and CC, the big picture is quite similar to the previous experiment: The performance of CC tends to decrease relative to BR with an increasing number of labels. In the case of low label dependence, this can already be seen for only five labels. The case of high label dependence is more interesting: While CC seems to gain from exploiting the dependency for a small to moderate number of labels, it cannot extend this gain to more than 15 labels.

5 Nested Stacking

A first very simple idea to mitigate the problem of attribute noise in CC is to let a classifier h_j use predicted labels $\hat{y}_1, \ldots, \hat{y}_{j-1}$ as supplementary attributes for training instead of the true labels $y_1, \ldots, y_{j-1}$. This way, one could make sure that the data distribution is the same for training and testing. Or, stated differently, the situation faced by a classifier during training does indeed equal the one it will encounter later on at prediction time. Since then a classifier is trained on the predictions of other classifiers, this approach fits the stacked generalization learning paradigm [18], also simply known as *stacking*.

5.1 Stacking versus Nested Stacking

The idea of stacking has already been used in the context of MLC by Godbole and Sharawagi [8]. In the learning phase, their method builds a stack of two groups of classifiers. The first one is formed by the standard BR classifiers: $\boldsymbol{h}^1(\boldsymbol{x}) = (h_1^1(\boldsymbol{x}), \ldots, h_m^1(\boldsymbol{x}))$. On a second level, also called meta-level, another group of binary models (again one for each label) is learnt, but these classifiers consider an augmented feature space that includes the binary outputs of all models of the first level: $\boldsymbol{h}^2(\boldsymbol{x}, \boldsymbol{y}') = (h_1^2(\boldsymbol{x}, \boldsymbol{y}'), \ldots, h_m^2(\boldsymbol{x}, \boldsymbol{y}'))$, where $\boldsymbol{y}' = \boldsymbol{h}^1(\boldsymbol{x})$. The idea is to capture label dependencies by learning their relationships in the meta-level step. In the test phase, the final predictions are the outputs of the meta-level classifiers, $\boldsymbol{h}^2(\boldsymbol{x})$, using the outputs of $\boldsymbol{h}^1(\boldsymbol{x})$ exclusively to obtain the values of the augmented feature space.

Mimicking the chain structure of CC, our variant of stacking is a *nested* one: Instead of a two-level architecture as in standard stacking, we obtain a nested hierarchy of stacked (meta-)classifiers. Hence, we call it *nested stacking* (NS). Moreover, each of these classifiers is only trained on a subset of the predictions of other classifiers. Like in CC, m models need to be trained in total, while $2m$ models are trained in standard stacking.

5.2 Out-of-Sample versus Within-Sample Training

To make sure that the distribution of the labels $\hat{y}_1, \ldots, \hat{y}_{j-1}$, which are used as supplementary attributes by the classifier h_j, is indeed the same at training and prediction time, these labels should be produced by means of an out-of-sample prediction procedure. For example, an internal leave-one-out cross validation procedure could be implemented for this purpose.

Needless to say, a procedure of that kind is computationally complex, even for classifiers that can be trained and "detrained" incrementally (such as incremental and decremental support vector machines [1]). In our current version of NS, we therefore implement a simple within-sample strategy. In several experimental studies, we found this strategy to perform almost as good as out-of-sample training, while being significantly faster. In fact, methods such as logistic regression, which are not overly flexible, are hardly influenced by excluding or including a single example.

5.3 A First Experiment

To get a first impression of the performance of NS, we return to the experiment in Section 4.2. As can be seen in Figure 3, NS does indeed gain in comparison to CC with an increasing number of labels; only if the labels are few, CC is still a bit better. This tendency is more pronounced

in the case of strong label dependency, whereas the differences are rather small if label dependence is low.

To explain the competitive performance of CC if the number of labels is small, note that replacing "clean" training data $y_1, \ldots, y_{j-1}$ by possibly more noisy data $\hat{y}_1, \ldots, \hat{y}_{j-1}$, as done by NS, may not only have the positive effect of making the training data more authentic. In fact, it may also make the problem of learning h_j more difficult (because the dependency $y_1, \ldots, y_{j-1} \rightarrow y_j$ might be "easier" than the dependency $\hat{y}_1, \ldots, \hat{y}_{j-1} \rightarrow y_j$). Apparently, this effect plays an important role if the number of labels is small, whereas the positive effect dominates for longer label chains.

5.4 Subset Correction

Our second modification is motivated by the observation that the number of label combinations that are commonly observed in MLC data sets is only a tiny fraction of the total number $|\mathcal{Y}| = 2^m$ of possible subsets; see Table 1, which reports the value $|\mathcal{Y}_D|2^{-m}$, where $\mathcal{Y}_D$ is the set of unique label combinations contained in the data D, as the "observation rate" in the last column. Moreover, if a label combination $\boldsymbol{y}$ has an occurrence probability of $\epsilon > 0$, then the probability that it has never be seen in a data set of size n reduces to $(1-\epsilon)^n$. Thus, by contraposition, one may argue that such a label combination is indeed unlikely to exist at all (at least for large enough n).

Our idea of "subset correction", therefore, is to restrict a learner to the prediction of label combinations whose existence is testified by the (training) data. More precisely, let $\mathcal{Y}_S$ denote the set of label subsets $\boldsymbol{y}$ that have be seen in the training data S. Then, given a prediction $\hat{\boldsymbol{y}}$ produced by a classifier $\boldsymbol{h}$, this prediction is replaced by the "most similar" subset $\boldsymbol{y}^* \in \mathcal{Y}_S$:

$$\boldsymbol{y}^* \in \operatorname*{argmin}_{\boldsymbol{y}' \in \mathcal{Y}_S} L_H(\hat{\boldsymbol{y}}, \boldsymbol{y}') \tag{7}$$

Thus, $\boldsymbol{y}^*$ is eventually returned as a prediction instead of $\hat{\boldsymbol{y}}$. If the minimum in (7) is not unique, those label combinations with higher frequency in the training data are preferred.

In principle, the Hamming loss could of course be replaced by other MLC loss functions in (7). Its use here is mainly motivated by the fact, that it is used for a similar purpose, namely decoding, in the framework of *error correcting output codes* (ECOC). As such, it has been applied in multi-class classification [**?**] and lately also in MLC [9; 7].

6 Nested Stacking vs. Classifier Chains

In this section, we compare NS and CC, both with and without subset correction, on real MLC benchmark data. As can be seen in Table 1, the data sets differ quite significantly in terms of the number of attributes, examples, labels, cardinality (number of labels per example) and the observation rate.

Logistic regression was used as a base learner for binary prediction in all MLC methods [10]. Unlike [13], we do not apply any threshold selection procedure; instead, we simply used $t = 0.5$ for deciding the relevance of a label. In fact, our goal is to study the behavior of CC and NS without the influence of other factors that may bias the results.

Since CC's main goal is to detect conditional label dependence, we used example-based metrics for evaluation. In addition to Hamming and subset 0/1 loss introduced earlier, we also applied the F_1 and Jaccard index defined, respectively, as follows (note that these are accuracy measures instead of loss functions):

$$F_1(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \frac{2\sum_{i=1}^m [\![y_i = 1 \text{ and } h_i(\boldsymbol{x}) = 1]\!]}{\sum_{i=1}^m ([\![y_i = 1]\!] + [\![h_i(\boldsymbol{x}) = 1]\!])} \tag{8}$$

$$Jaccard(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \frac{\sum_{i=1}^m [\![y_i = 1 \text{ and } h_i(\boldsymbol{x}) = 1]\!]}{\sum_{i=1}^m [\![y_i = 1 \text{ or } h_i(\boldsymbol{x}) = 1]\!]} \tag{9}$$

The value for a test set is defined as the average over all instances. The scores reported in Tables 2 and 3 were estimated by means of 10-fold cross-validation, repeated three times. We used a paired t-test for establishing statistical significance on each data set.

Table 4: The effect of subset correction in terms of statistical significance. The corresponsing loss/accuracy values can be found in Tables 2-3. $\uparrow\uparrow$ ($\downarrow\downarrow$) means that NS_{SC} (CC_{SC}) is significantly better (worse) than NS (CC) at level $p < 0.01$ ($\uparrow$ and $\downarrow$ at level $p < 0.05$) in a paired t-test.

NS vs. NS_{SC}

no.	m	Hamming	Subset 0/1	Jaccard	F_1
1	159	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
2	6		$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
3	53	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
4	27		$\uparrow\uparrow$		
5	5	$\downarrow\downarrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
6	101	$\uparrow\uparrow$	$\uparrow\uparrow$		$\downarrow\downarrow$
7	45	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow$	$\downarrow\downarrow$
8	7	$\downarrow\downarrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
9	6	$\downarrow\downarrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
10	22	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	
11	14	$\downarrow\downarrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$

CC vs. CC_{SC}

no.	m	Hamming	Subset 0/1	Jaccard	F_1
1	159	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
2	6		$\uparrow$		
3	53	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
4	27		$\uparrow\uparrow$		
5	5				
6	101	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
7	45	$\uparrow\uparrow$	$\uparrow\uparrow$		$\downarrow\downarrow$
8	7	$\downarrow\downarrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$
9	6	$\downarrow$	$\uparrow\uparrow$		$\downarrow\downarrow$
10	22	$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$	
11	14		$\uparrow\uparrow$	$\uparrow\uparrow$	$\uparrow\uparrow$

Looking at the comparison between CC and NS (without subset correction) as shown in Table 2), the first thing to mention is the strong performance of NS in terms of Hamming loss (8 significant wins and 3 losses). In terms of their properties, the three data sets on which NS looses do indeed seem to be favorable for CC: Since slashdot, medical and genbase all have a rather low Hamming loss, the danger of error propagation is limited. Thus, the results are completely in agreement with our expectations.

For Jaccard and F1, the picture is not as clear. In both cases, NS wins 6 times. Again, like for Hamming loss, NS

Table 1: Properties of the data sets used in the experiments.

no.	Data set	Attributes	Examples	Labels	Cardinality	Observation Rate
1	bibtex	1836	7395	159	2.40	3.9E-45
2	emotions	72	593	6	1.87	4.0E-1
3	enron	1001	1702	53	3.38	8.3E-14
4	genbase	1185	662	27	1.25	2.3E-7
5	image	135	2000	5	1.24	6.0E-1
6	mediamill	120	5000	101	4.27	2.5E-27
7	medical	1449	978	45	1.25	2.6E-12
8	reuters	243	7119	7	1.24	1.9E-1
9	scene	294	2407	6	1.07	2.3E-1
10	slashdot	1079	3782	22	1.18	3.7E-5
11	yeast	103	2417	14	4.24	1.2E-2

Table 2: Experimental results of NS and CC on benchmark data sets. $\uparrow\uparrow$ ($\downarrow\downarrow$) means that NS is significantly better (worse) than CC at level $p < 0.01$ ($\uparrow$ and $\downarrow$ at level $p < 0.05$) in a paired t-test.

		F_1			JACCARD INDEX		
no.	m	CC	NS		CC	NS	
1	159	0.1697±.0071	0.1747±.0077	$\uparrow\uparrow$	0.1098±.0060	0.1133±.0064	$\uparrow\uparrow$
2	6	0.5883±.0534	0.6028±.0500	$\uparrow$	0.5003±.0521	0.5144±.0514	$\uparrow$
3	53	0.3483±.0191	0.3729±.0214	$\uparrow\uparrow$	0.2474±.0163	0.2693±.0178	$\uparrow\uparrow$
4	27	0.9863±.0090	0.9854±.0085	$\downarrow$	0.9804±.0115	0.9789±.0109	$\downarrow$
5	5	0.5556±.0284	0.4780±.0299	$\downarrow\downarrow$	0.5196±.0271	0.4460±.0278	$\downarrow\downarrow$
6	101	0.5326±.0054	0.5619±.0053	$\uparrow\uparrow$	0.4280±.0052	0.4459±.0052	$\uparrow\uparrow$
7	45	0.6462±.0331	0.6444±.0340		0.5828±.0343	0.5804±.0356	
8	7	0.8599±.0128	0.8570±.0116	$\downarrow\downarrow$	0.8336±.0138	0.8302±.0129	$\downarrow\downarrow$
9	6	0.5969±.0403	0.6031±.0348		0.5745±.0405	0.5766±.0344	
10	22	0.3278±.0185	0.3259±.0186		0.2747±.0176	0.2726±.0180	
11	14	0.5836±.0182	0.6068±.0172	$\uparrow\uparrow$	0.4848±.0198	0.4990±.0183	$\uparrow\uparrow$

		HAMMING LOSS			SUBSET 0/1 LOSS		
no.	m	CC	NS		CC	NS	
1	159	0.0724±.0020	0.0672±.0016	$\uparrow\uparrow$	0.9837±.0052	0.9833±.0052	
2	6	0.2367±.0268	0.2169±.0253	$\uparrow\uparrow$	0.7578±.0575	0.7477±.0633	
3	53	0.1233±.0051	0.1050±.0051	$\uparrow\uparrow$	0.9565±.0135	0.9510±.0133	$\uparrow$
4	27	0.0019±.0011	0.0020±.0010	$\downarrow$	0.0408±.0211	0.0443±.0213	$\downarrow$
5	5	0.2104±.0127	0.1962±.0119	$\uparrow\uparrow$	0.5857±.0269	0.6468±.0249	$\downarrow\downarrow$
6	101	0.0303±.0004	0.0291±.0004	$\uparrow\uparrow$	0.8752±.0049	0.8969±.0048	$\downarrow\downarrow$
7	45	0.0248±.0031	0.0249±.0031		0.5890±.0425	0.5934±.0463	
8	7	0.0506±.0046	0.0483±.0043	$\uparrow\uparrow$	0.2454±.0173	0.2499±.0175	$\downarrow$
9	6	0.1470±.0143	0.1397±.0124	$\uparrow\uparrow$	0.4918±.0434	0.5019±.0355	$\downarrow$
10	22	0.0908±.0027	0.0913±.0028	$\downarrow$	0.8652±.0185	0.8678±.0198	
11	14	0.2242±.0093	0.2069±.0087	$\uparrow\uparrow$	0.8104±.0229	0.8469±.0231	$\downarrow\downarrow$

outperforms CC on data sets with many labels (bibtex, enron, mediamill) or a relatively high Hamming loss (yeast), whereas CC is better for data sets with only a few labels (image, reuters) or with high accuracy (genbase).

The picture for CC and NS with subset correction (denoted CC_{SC} and NS_{SC}, respectively) is quite similar (Table 3), although the performance differences tend to decrease in absolute size. On subset 0/1 loss, for which the original CC performs quite strong and typically outperforms NS, the corrected version NS_{SC} even achieves 3 significant wins over CC_{SC}.

To analyze the effect of subset correction in more detail, Table 4 provides a summary of a comparison of Table 2 and Table 3. Interestingly enough, subset correction yields improvements on almost every experiment, regardless of the performance measure, and most of these improvements are even significant. More specifically, counting the number of significant wins, subset correction appears to be most beneficial for subset 0/1 loss and least beneficial for Hamming loss. In fact, for Hamming loss, subset correction looses for data sets with only a few labels (reuters, scene, yeast and image) and a relatively high observation rate. Comparing NS and CC, the former seems to benefit even more from subset correction than the latter, except for Hamming loss, on which NS is already strong in its basic version. In terms of subset 0/1 loss, however, significant improvements can be seen on every single data set. In light of the simplicity of the idea, these effects of subset correction are certainly striking.

7 Conclusions

This paper has thrown a critical look at the classifier chains method for multi-label classification, which has been adopted quite quickly by the MLC community and is now commonly used as a baseline when it comes to comparing methods for exploiting label dependency. Notwith-

Table 3: Experimental results of NS_{SC} and CC_{SC} on benchmark data sets. $\uparrow\uparrow$ ($\downarrow\downarrow$) means that NS_{SC} is significantly better (worse) than CC_{SC} at level $p < 0.01$ ($\uparrow$ and $\downarrow$ at level $p < 0.05$) in a paired t-test.

		F_1			JACCARD INDEX		
no.	m	CC_{SC}	NS_{SC}		CC_{SC}	NS_{SC}	
1	159	0.2026±.0119	0.2090±.0113	$\uparrow\uparrow$	0.1528±.0099	0.1582±.0100	$\uparrow\uparrow$
2	6	0.5905±.5905	0.6132±.6132	$\uparrow\uparrow$	0.5027±.0521	0.5239±.0525	$\uparrow\uparrow$
3	53	0.3843±.3843	0.4016±.4016	$\uparrow\uparrow$	0.2821±.0190	0.3005±.0238	$\uparrow\uparrow$
4	27	0.9843±.9843	0.9838±.9838		0.9807±.0129	0.9802±.0125	
5	5	0.5557±.5557	0.5315±.5315	$\downarrow\downarrow$	0.5197±.0272	0.4972±.0304	$\downarrow\downarrow$
6	101	0.5328±.0054	0.5610±.0052	$\uparrow\uparrow$	0.4282±.0052	0.4457±.0050	$\uparrow\uparrow$
7	45	0.6220±.6220	0.6231±.6231		0.5898±.0435	0.5900±.0460	
8	7	0.8624±.8624	0.8639±.8639		0.8367±.0142	0.8382±.0126	
9	6	0.5921±.5921	0.6105±.6105	$\uparrow\uparrow$	0.5739±.0423	0.5873±.0370	$\uparrow\uparrow$
10	22	0.3271±.3271	0.3248±.3248		0.2843±.0186	0.2818±.0202	
11	14	0.5889±.5889	0.6141±.6141	$\uparrow\uparrow$	0.4890±.0200	0.5104±.0200	$\uparrow\uparrow$

		HAMMING LOSS			SUBSET 0/1 LOSS		
no.	m	CC_{SC}	NS_{SC}		CC_{SC}	NS_{SC}	
1	159	0.0282±.0008	0.0270±.0006	$\uparrow\uparrow$	0.9592±.0080	0.9568±.0082	$\uparrow$
2	6	0.2363±.0268	0.2190±.0266	$\uparrow\uparrow$	0.7555±.0581	0.7404±.0652	$\uparrow$
3	53	0.0819±.0023	0.0766±.0030	$\uparrow\uparrow$	0.9491±.0130	0.9346±.0156	$\uparrow\uparrow$
4	27	0.0019±.0012	0.0019±.0012		0.0332±.0176	0.0337±.0172	
5	5	0.2104±.0127	0.2199±.0140	$\downarrow\downarrow$	0.5855±.0270	0.6027±.0277	$\downarrow\downarrow$
6	101	0.0302±.0004	0.0291±.0003	$\uparrow\uparrow$	0.8750±.0049	0.8925±.0051	$\downarrow\downarrow$
7	45	0.0210±.0025	0.0210±.0027		0.5017±.0465	0.5037±.0514	
8	7	0.0513±.0049	0.0506±.0042		0.2403±.0177	0.2391±.0167	
9	6	0.1479±.0147	0.1441±.0130	$\uparrow$	0.4802±.0449	0.4815±.0386	
10	22	0.0840±.0026	0.0842±.0028		0.8348±.0186	0.8380±.0201	
11	14	0.2243±.0093	0.2089±.0097	$\uparrow\uparrow$	0.8073±.0230	0.8097±.0237	

standing the appeal of the method and the plausibility of its basic idea, we have argued that, at second sight, the chaining of classifiers begs an important flaw: A binary classifier that has learnt to rely on the values of previous labels in the chain might be misled when these values are replaced by possibly erroneous estimations at prediction time. The classification errors produced because of this attribute noise may subsequently be propagated or even reinforced along the entire chain. Roughly speaking, what looks as a gift at training time may turn out to become a handicap in prediction.

Our results have shown that the problem of error propagation is highly relevant, and that it may strongly impair the performance of CC. In order to avoid this problem, the method of nested stacking proposed in this paper uses predicted instead of observed label relevances as additional attribute values in the training phase. Our experimental studies clearly confirm that, although NS does not consistently outperform CC, it seems to have advantages for those data sets on which error propagation becomes an issue, namely data sets with many labels or low (label-wise) prediction accuracy.

There are several lines of future work. First, it is of course desirable to complement this study by meaningful theoretical results supporting our claims. Second, it would be interesting to investigate to what extent the problem of attribute noise also applies to the probabilistic variant of classifier chains introduced in [3]. Last but not least, given the interesting effects that are produced by the simple idea of subset correction, this approach seems to be worth further investigation, all the more as it is completely general and not limited to specific MLC methods such as those considered in this paper.

References

[1] Gert Cauwenberghs and Tomaso Poggio. Incremental and decremental support vector machine learning. *Proc. NIPS*, pages 409–415, 2001.

[2] W. Cheng and E. Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3):211–225, 2009.

[3] K. Dembczyński, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *ICML*, pages 279–286, 2010.

[4] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, To appear, 2012.

[5] A. Elisseeff and J. Weston. A Kernel Method for Multi-Labelled Classification. In *ACM Conf. on Research and Develop. in Infor. Retrieval*, pages 274–281, 2005.

[6] J. Fürnkranz, E. Hüllermeier, E.L. Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73:133–153, 2008.

[7] Johannes Fürnkranz and Sang-Hyeun Park. Error-correcting output codes as a transformation from multi-class to multi-label prediction. In *Proc. Discovery Science*, pages 254–267. 2012.

[8] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Pacific-Asia Conf. on Know. Disc. and Data Mining*, pages 22–30, 2004.

[9] Tomasz Kajdanowicz and Przemysław Kazienko. Multi-label classification using error correcting out-

put codes. *International Journal of Applied Mathematics and Computer Science*, 22(4):829–840, 2012.

[10] C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region Newton method for logistic regression. *Journal of Machine Learning Research*, 9(Apr):627–650, 2008.

[11] E. Montañés, J. R. Quevedo, and J. J. del Coz. Aggregating independent and dependent models to learn multi-label classifiers. In *Proc. ECML/PKDD*, 2011.

[12] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *IEEE Int. Conf. on Data Mining*, pages 995–1000. IEEE, 2008.

[13] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

[14] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. In *Machine Learning*, pages 135–168, 2000.

[15] Robin Senge, Juan Jos del Coz, and Eyke Hüllermeier. On the problem of error propagation in classifier chains for multi-label classification. In *Conference of the German Classification Society*, 2012.

[16] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. 2010.

[17] G. Tsoumakas and I. Vlahavas. Random k-Labelsets: An Ensemble Method for Multilabel Classification. In *Proc. ECML/PKDD*, LNCS, pages 406–417. Springer, 2007.

[18] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:214–259, 1992.

[19] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. on Knowl. and Data Eng.*, 18:1338–1351, 2006.

[20] X. Zhu and X. Wu. Class noise vs. attribute noise: a quantitative study of their impacts. *Artificial Intelligence Review*, 22(3):177–210, 2004.

Integrating User Preferences Into Distance Metrics

Mark Kröll
Know-Center Graz, Austria
mkroell@know-center.at

Vedran Sabol
Know-Center Graz, Austria
vsabol@know-center.at

Roman Kern
Graz University of Technology, Austria
rkern@tugraz.at

Michael Granitzer
University of Passau, Germany
michael.granitzer@uni-passau.de

Abstract

Many practical scenarios such as search, classification or clustering benefit from better understanding their users, for instance, to deliver more relevant search results. Instead of committing ourselves to a specific field of research, e.g. by generating user profiles to enhance information retrieval, we seek to incorporate user preferences into the distance metric itself which lies at the heart of many algorithms including Information Retrieval and Machine Learning. The two approaches we explore in this paper allow users to directly convey their preferences in an intuitive way. The first approach adheres to the idea that just stating whether two documents are similar or not is more intuitive for a user than, for instance, assigning them to a broad spectrum of topics. The second approach seeks to take into account a user's mental construct of the world being provided with a user-specific concept hierarchy. To evaluate our two approaches, we perform a text classification task. In the classification setting we use the Reuters RCV1 corpus to simulate user preferences. Our results indicate the principal feasibility of these two approaches and encourage further investigations.

1 Introduction

Calculating the similarity between textual resources lies at the heart of many algorithms including Information Retrieval, Text Mining or Machine Learning algorithms. Traditional approaches such as TF-IDF [Salton and McGill, 1986] often apply weighting schemes to adapt the impact of certain terms. Yet, a drawback these parametric approaches suffer from is that they are not capable of taking into account user interests.

Practical scenarios such as *search* benefit from better understanding their users. To provide more relevant documents, information retrieval applications aim to personalize search results, e.g. by integrating user interests (cf. [Qiu and Cho, 2006]) or by actively learning search result rankings (cf. [Radlinski and Joachims, 2007]). Other approaches choose a more direct approach by allowing user interaction to convey their preferences. In that sense, users are often required to tune parameters, e.g. decide on cluster size or on the number of neighbors, which affect an algorithm's internal functionality. Yet, adapting these parameters might be counter-intuitive or might require expert knowledge in the sense of a deeper understanding of the algorithm.

We therefore seek to incorporate user preferences into the similarity calculation in a more intuitive manner. Our first approach adheres to the idea that just stating whether two documents are similar or not is more intuitive for a user than, for instance, assigning them to topics (cf. [Saaty, 2008]). In psychology, the idea of using paired comparisons to gain ranking information is a long-established one (cf. [Thurstone, 1927]). In a second approach we seek our distance metric to reflect a user's mental construct of the world by exploiting information from a user-specific concept hierarchy. In this paper, we raise awareness of intuitively incorporating user preferences into the computation of document similarity. In addition, we provide implementations of these two approaches and discuss their characteristics as well as lessons learnt. Finally, we evaluate them in a practical application scenario, i.e. text classification.

2 Related Work

In the following, we review work from two fields of research, (i) semantic representation of textual resources and (ii) learning semantic similarity metrics for textual resources.

2.1 Semantic Representation

Introducing semantic similarity between features often refers to introducing dependencies amongst formerly unrelated feature dimensions. Attempts to incorporate semantic knowledge into the classical vector space representations include semantic networks, latent semantic indexing or co-occurrence analysis where a semantic relation is assumed between terms whose occurrence patterns in the documents of a corpus are correlated [Cristianini *et al.*, 2002]. Especially kernel-based methods represent an attractive choice

for inferring relations from textual documents since they enable a document-by-document setting rather than a term-by-term setting. [Basili *et al.*, 2005] accessed WordNet as external lexical knowledge base to include semantics into the description of textual resources. In their setting they analysed the performance of small-sized training sets for the task of text classification. External knowledge was also used by [Gabrilovich and Markovitch, 2007] which represented the meaning of texts in a high-dimensional space of concepts derived from Wikipedia .

2.2 Learning Semantic Similarity

Parametric approaches suffer from the drawback that they do not adapt to particular domains or do not take into account users' personal requirements. [Metzler and Zaragoza, 2009] overcame the rigidity of parametric weighting schemes by introducing semi-parametric and non-parametric weighting schemes. In supervised learning settings, for instance, nearest neighbor classification (cf. [Weinberger and Saul, 2009]), numerous attempts have been made to define or learn either local or global metrics for classification. A number of researchers have demonstrated that nearest neighbor classification can be greatly improved by learning an appropriate distance metric from labeled examples. [Shalev-Shwartz *et al.*, 2004], for instance, optimized the Mahalanobis distance via linear transformations in order to boost the accuracy of a k-NN classification algorithm, which can be seen as implicit application of a weighting scheme.

3 User Preference Integration

The integration of user preferences into the similarity computation can be regarded as some form of semantic enrichment. In that sense, semantically enriching the documents' content allows influencing their similarity by introducing dependencies amongst formerly unrelated feature dimensions, as for instance a semantic kernel does (cf. [Cristianini *et al.*, 2002]). We explore two approaches to incorporate user preferences in a more intuitive way and describe implementation details, i.e. how we accordingly adapt underlying distance metrics. For evaluation purposes, we perform a text classification task, i.e. classifying documents from the Reuters RCV1 corpus, a well-known benchmark dataset. In both approaches we use Reuters RCV1 document-to-topic mapping to simulate user preferences.

3.1 Similar Document Pairs (SDP)

To adhere to the idea of stating whether two documents are similar or not, we process and merge document pairs to generate new samples. A positive sample is formed by two documents belonging to same category; a negative one by taking two documents belonging to different categories. In our experiments we use a component-wise multiplication (Hadamard product) which results in strengthening common dimensions.

In a first step, the documents' input space is transformed into a higher dimensional space by including bigrams and named entity information, i.e. a concatenation of several feature types. To generate a new sample, we merge two documents by performing a component-wise multiplication. This multiplication results in a new sample vector exhibiting the same dimensionality. In the training phase, we perform an offline processing of the Reuters RCV1 corpus and store relevant information in Lucene[1] indices for fast feature engineering. We then generate new training/test data splits by merging pairs of documents (Hadamard multiplication). From preliminary experiments we learnt that some sort of "intelligent sampling" is required, i.e. "sampling" to keep the number of training/test data manageable in the optimization step and "intelligent" to choose appropriate negative examples. From a class distribution point of view, these negative examples lie close to the boundary of the positive class. To perform this intelligent sampling, we utilize Lucene's search functionality. For every selected document, we search the index for the top n most similar samples once bearing the same class label and once bearing a different class label. These samples are considered for the merging procedure. We then apply Vowpal Wabbit[2], an optimization toolkit, to learn the importance of feature dimensions, i.e. to learn regression weights which are optimized with respect to the new binary classification problem. We remark that in this setting the prior multi-class classification problem is transformed into a binary one. The testing phase handles previously unseen data items, i.e. generating feature types on the fly to calculate similarity values. The same processing steps have to be applied, i.e. multiplying two input documents to determine whether they are similar or dissimilar. Lastly, this new vector is then "informed" by the learnt weights.

3.2 Personal Concept Hierarchy (PCH)

In this approach the user provides us with her personal concept hierarchy whose semantic concepts are representative for a certain domain. We then map documents onto these semantic concepts and then apply standard similarity metrics such as cosine distance. Semantic concepts may correspond to categories in a taxonomy as for instance in our case to Reuters RCV1 topics or to Wikipedia concepts as it is done in [Gabrilovich and Markovitch, 2007].

In the training phase, we perform an offline processing of the Reuters RCV1 corpus and store relevant information in Lucene indices for fast feature engineering. The mapping of Reuters documents onto semantic concepts is achieved by generating and applying a classification model exhibiting a multi-class, multi-label functionality.

We deliberately do not apply any threshold, so potentially a document could be assigned to all classes with varying degrees of confidences. The returned class/confidence vector is the new, low-dimensional representation of the document. We decided on Mallet's[3] Naive Bayes implementation to construct classification models for various feature representations including token n-grams. The testing phase handles new data items, i.e. generate feature types, on the fly for calculating similarity values. Before applying the similarity calculation, the new documents need to be mapped onto the semantic concepts. So each document undergoes a process of feature engineering first and is then classified by the trained Naive Bayes model which corresponds to the mapping onto the semantic concepts. Similarity values are then calculated by applying a standard similarity metric, e.g. cosine similarity, to document pairs represented by their affinity to semantic concepts.

[1] http://lucene.apache.org/

[2] http://hunch.net/ vw/

[3] http://mallet.cs.umass.edu/

4 Experiments

To evaluate our two approaches to integrate user preferences into the similarity computation, we perform a text classification task using the Reuters RCV1 corpus, a well-known benchmark dataset. The RCV1 dataset ([Lewis *et al.*, 2004]) was drawn from one of the news agency Reuters online databases. The dataset consists of English language stories produced by Reuters journalists between August 20, 1996, and August 19, 1997. To simulate user preferences, we use the stories' topic codes assigned to capture their major subjects. They were organized in four hierarchical groups: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). Each group is further divided into subgroups providing a more detailed categorization. For the classification task, only documents assigned to exactly one group are considered thereby avoiding a multi-label setting. The annotation process was conducted in a thorough manner - Reuters employed 90 people to handle the annotation of 5.5 million stories per year. We therefore considered the Reuters RCV1 dataset to be an adequate candidate to simulate user preferences.

In both approaches, we experimented with different feature combinations to represent the Reuters documents including unigrams, bigrams, part-of-speech information and named entity information. Sanitization steps included (i) a removal of invalid English words, e.g. a combination of literals and digits, (ii) a removal of stop words and (iii) token stemming using the Porter stemmer [Porter, 1997]. For sentence delimiting and named entity recognition we used Ling Pipe[4] and Apache's OpenNLP[5] natural language processing toolkit. We applied the Stanford part-of-speech tagger to obtain part-of-speech information.

4.1 Results

Similar Document Pairs

To learn a weight vector optimized to separate two classes, we used about 10000 Reuters documents for each of the four main categories, i.e. CCAT, GCAT, ECAT, MCAT. Representative documents were stratified for the positive and negative class. 80% of the documents were used for training, 20% for testing. We point out that by merging two documents with each other, we generate a new example and thus transform the instance space as well. Two documents from the same class are merged into a positive example reflecting a user's decision that these two documents are similar. We handed them over to Vowpal Wabbit's internal linear regression framework. We experimented with different feature representations to learn the weight vector including unigrams, bigrams, nouns, verbs, named entities and combinations thereof. Using the regression framework's performance criteria we compared different feature representations and eventually decided to use only unigrams. To evaluate the discrimination quality of the learnt weight vector on our overall multi-class problem, we used Weka's[6] machine learning framework to compare two settings: once with the learnt weights and once without them. Due to Weka's memory consumption, we used 850 Reuters documents in our classification setting.

Table 1 contrasts the accuracy results (10-fold cross validation) for two classification models, i.e. a Nearest-Neighbor classifier and a linear Support Vector Machine.

[4]http://alias-i.com/lingpipe/

[5]http://opennlp.apache.org/

[6]http://www.cs.waikato.ac.nz/ml/weka/

	k = 1	k = 5	k = 10	SVM (lin)
Accuracy (unweighted)	0.64	0.58	0.60	0.89
Accuracy (weighted)	0.65	0.59	0.58	0.86

Table 1: Accuracy results for the Nearest Neighbor classifier and the linear Support Vector Machine(SVM) - once with and once without applying the learnt weights. (10-fold cross validation)

We chose the Nearest Neighbor classification model because it does not apply any additional optimization steps as the Support Vector Machine does. The resulting values state that the learnt weights do not add any additional information regarding the classification problem. We hypothesize that the merging procedure itself strenghtens or weakens the respective dimensions that further weighting is not necessary.

As a second observation we learn that additional processing, e.g. optimization in case of the Support Vector Machine, does allow an increase in classification accuracy. From a theoretical perspective it would be interesting to compare the optimization strategies of (i) using Vowpal Wabbit to learn a weight vector and (ii) using a linear Support Vector Machine to learn Lagrange coefficients - to a certain extent both strategies aim to identify discriminant dimensions in the input space and yet the latter is by far more successful.

Personal Concept Hierarchy

The second approach's idea is to transform the documents' input space into a space of semantic concepts, i.e. creating a semantic concept representation. To map the documents onto concepts, we first generated a classification model exhibiting a multi-class, multi-label functionality. We decided on Mallet's Naive Bayes implementation to train models for various feature representations including unigrams, bigrams, nouns, verbs and named entities. We used 20000 Reuters documents for each of the four primary-level categories, i.e. CCAT, GCAT, ECAT, MCAT.
Using three of the learnt models, i.e. unigrams, bigrams and named entities, we mapped the Reuters documents onto semantic concepts and performed the multi-class problem with the new semantic concept representation using WEKA's Nearest Neighbour implementation. We evaluated 9000 documents by a 10 fold cross-evaluation - evaluation results for different numbers of neighbours are shown in Table 2.

	Unigrams	Bigrams	NEs
k = 1	0.93	0.95	0.82
k = 5	0.95	0.96	0.85
k = 10	0.95	0.97	0.86

Table 2: Accuracy results for the 4-class classification task based on different number of neighbors and different feature types.

These results show that the semantic concept representation preserves the information and performs well in the simple 4-class classification setting. To create a more realistic setting, we extended the number of concepts by focusing on Reuters secondary level categories. As with the four primary level categories, we used Mallet's classification framework to generate a model for 54 Reuters categories.

Since some categories contained only few documents, we decided to use only 100 documents per category as training samples. We used WEKA's Nearest Neighbour classifier implementation to perform the classification task. We evaluated 7500 documents by a 10 fold cross-evaluation - evaluation results are shown in Table 3.

	Unigrams	Bigrams	NEs
k = 1	0.83	0.74	0.51
k = 5	0.83	0.74	0.47
k = 10	0.82	0.73	0.48

Table 3: Accuracy results for the 54-class classification task based on different number of neighbors and different feature types.

In the following, we compared the Nearest Neighbour classifier with two other standard classification schemes - a Naive Bayes classifier and a linear Support Vector Machine.

	k = 1	NB	SVM(lin)
Unigrams	0.83	0.74	0.83

Table 4: Accuracy values for a Nearest Neighbor classifier (k = 1), a Naive Bayes(NB) classifier and a linear Support Vector Machine(SVM).

Table 4 shows similar performance values for the Nearest Neighbor classifier and the Support Vector Machine which indicates that additional optimization does not yield further gains for the classification task.

5 Conclusion

In this work we explore two approaches to intuitively integrate user preferences into the similarity computation of textual documents and provide implementation details. Both approaches directly affect the distance metric which has the advantage of being to a certain extent algorithm-independent. Instead of being bound to a certain research field, our approaches can be adopted by algorithms across such fields including Machine Learning or Information Retrieval. The results encourage further engagement and analysis of the underlying ideas. A first direction is to investigate why the learnt optimization weights in the "Similar Document Pairs" approach have so little effect on the resulting accuracy values. From a theoretical perspective a comparison to the optimization strategies of a Support Vector Machine would be interesting. An advantage of the SDP approach certainly is that adding additional document/personal classes is simple. In contrast, the "Personal Concept Hierarchy" approach cannot handle the adding of classes so easily. It has to re-compute the classification models for the mapping operation. As to the requirement of a concept hierarchy for the approach to work, we remark that this information can to a certain degree be automatically generated by taking into account a person's tagging, searching or reading behavior. A natural next step represents the application of both approaches in a real-world setting having persons (i) providing personal information, e.g. in form of decisions, and (ii) evaluating the results and giving feedback.

Acknowledgments

The work presented in this paper was developed within the DIVINE project (www.weblyzard.com/divine), funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT (www.ffg.at/fit-it). The Know-Center is funded within the Austrian COMET Program (Competence Centers for Excellent Technologies) under the auspices of BMVIT, the Austrian Ministry of Economics and Labor, and by the State of Styria.

References

[Basili *et al.*, 2005] R. Basili, M. Cammisa, and R. Moschitti. A semantic kernel to classify texts with very few training examples. In *Proceedings of the Workshop on Learning in Web Search, at the International Conference on Machine Learning*, 2005.

[Cristianini *et al.*, 2002] N. Cristianini, J. Shawe-Taylor, and H. Lodhi. Latent semantic kernels. *Journal of Intelligent Information Systems*, 18(2-3):127–152, 2002.

[Gabrilovich and Markovitch, 2007] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artifical intelligence*, pages 1606–1611, 2007.

[Lewis *et al.*, 2004] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[Metzler and Zaragoza, 2009] D. Metzler and H. Zaragoza. Semi-parametric and non-parametric term weighting for information retrieval. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval*, ICTIR'09, pages 42–53, 2009.

[Porter, 1997] M. Porter. An algorithm for suffix stripping. In *Readings in Information Retrieval*, 1997.

[Qiu and Cho, 2006] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *Proceedings of the International Conference on World Wide Web*, WWW'06, pages 727–736, 2006.

[Radlinski and Joachims, 2007] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 570–579, 2007.

[Saaty, 2008] T. Saaty. Relative measurement and its generalization in decision making. *RACSAM - Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales.*, 102(2):251–318, 2008.

[Salton and McGill, 1986] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[Shalev-Shwartz *et al.*, 2004] S. Shalev-Shwartz, Y. Singer, and A. Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the International Conference on Machine learning*, ICML'04, pages 94–, New York, NY, USA, 2004.

[Thurstone, 1927] L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927.

[Weinberger and Saul, 2009] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

A Machine Learning Approach to Drought Stress Level Classification of Tobacco Plants

C. Stocker[1], F. Uhrmann[1], O. Scholz[1], C. Hügel[1], T. Rademacher[2], M. Siebers[3], U. Schmid[3]
[1]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
{stockech, Franz.Uhrmann, Oliver.Scholz}@iis.fraunhofer.de
[2]Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Aachen, Germany
[3]University of Bamberg, Cognitive Systems Group, Bamberg, Germany
{michael.siebers, ute.schmid }@uni-bamberg.de

Abstract

We show that a model for drought stress level classification of tobacco leaves can be learned from measurement data. The data was acquired using a sheet-of-light measurement system developed at the *Fraunhofer Institute for Integrated Circuits IIS*. Spatial attributes like length, width or bending were extracted by fitting a parameterized leaf model to the measurement data. The attributes were transformed to simple attribute vectors describing relevant aspects of plant growth and stress evidence. The resulting attribute vectors were used to train decision trees, neural networks and linear regression classifiers. To provide a broad range of data, plants were assessed in a planned measurement campaign. Stress was induced by cutting off the water supply to simulate drought. Evidence for drought stress could be recognized from the data. Classification of whole plants yielded better results than classification of single leaves.

1 Introduction

A highly controlled production of plants in greenhouses or phytotrons (e.g. automated production of plants for pharmaceutical applications or high-throughput plant phenotyping for breeding) requires fully automated systems for the continuous monitoring of the growth conditions and the plant status. While simple factors like climate, nutrition and water supply can be controlled with simple reactive systems, more complex aspects like the detection of stress, diseases or pest infestation require intelligent systems which are able to detect anomalies in plant growth. To accomplish this, a measuring system must be designed which is able to capture the necessary features of plant growth in a non-destructive manner. Furthermore, a classification model is required which provides information about how to assess the measurement data.

The aim of our work was to evaluate whether it is possible to construct such a classification model directly from measurement data without additional expert knowledge. As part of its internal funding program the Fraunhofer Future Foundation is currently promoting the Malaria-Vaccines project of the *Fraunhofer IME*, Aachen. Through the participation of two additional Fraunhofer Institutes (IPT and IIS) the project synergistically combines expertise from the life sciences, engineering and medical technology fields. One of the major project goals is to develop an automated production facility for the GMP-compliant manufacturing of IMEs novel malaria vaccine candidates in tobacco plants. The task for our work was to model the behavior of tobacco plants when exposed to drought stress. From a machine learning point of view, this is a classification task of distinguishing stressed plants from regularly watered plants. Drought was chosen as the stress type of interest since it is easy to simulate by cutting off irrigation. In a measurement campaign designed for this work, a set of tobacco plants was measured over the course of one week. To trigger drought stress, the plants were cut off from water supply according to a fixed time schedule. Plant data was acquired using a sheet-of-light measurement system developed at the IIS in the *Department for Contactless Test and Measuring Systems*.

Furthermore, a biologist was asked to assess the plants' stress level according to the measurement data. These ratings served as classification labels for supervised machine learning techniques. The measurement data was reduced to attribute vectors describing essential features of the physical shape of a plant. This was done using a parameterized leaf model developed at the *Fraunhofer IIS*.

Combined with the labels provided by the expert rating, these attribute vectors form an input data which is compatible with standard machine learning techniques. Decision trees, neural networks and linear regression were used for classification to evaluate which technique is suited best for the data provided.

2 Acquisition of plant data

2.1 Measurement campaign setup

The data used for this work was acquired in a measurement campaign carried out at the IME. Over the course of one week, tobacco plants of different stress states were measured on a regular basis. The test group consisted of 50 Nicotiana tabacum plants grown hydroponically in stonewool blocks. They were sowed in five groups of ten plants in a weekly sequence. The plants were cultivated in a phytotron under LED light in a nutrient film technique (NFT) system at 25°C during the light phase (16h) and 22°C during the dark phase (8h) with a constant relative humidity of 70%. During the light phase, the plants in the gullies were periodically supplied with nutrient solution (15 min flow / 45 min off). The measurements were started when the last group reached an age of three weeks. Thus the plants ranged in three to seven weeks of age at the beginning of the measurements. Each day, all plants were measured two times. The first measurement was carried out in the morning, the second one in the afternoon.

To accomplish objective states of drought stress, single plants were cut off from water supply at fixed points in

time during the measurement campaign. At each cut-off point, two more plants of each age group were exposed to drought stress. Once a plant was separated from irrigation, it was kept unwatered until the measurement campaign was finished. The exact water cut-off points were distributed over the week in such a way that every three measurements, two more plants were cut off from irrigation. Since there were 14 measurements, this scheme was set off by one measurement to define objective start and end points. This means that at the first measurement no plant was exposed to drought stress to acquire an unstressed measurement of each plant individual as a reference point for further changes over the course of the following measurements. At the last measurement, there was no additional cut-off point included to keep two watered plants as a reference with respect to the stressed individuals.

2.2 Measurement system

The measurements were carried out using a sheet-of-light measuring system developed at the *Fraunhofer IIS*. This system projects laser light onto the plant, which is then captured by several cameras. The cameras are positioned below, above and in front of the plant. In the measuring process, the plant is turned about 360 degrees to expose all plant parts to the cameras. During the rotation, the distance from the plant to the camera is measured by tracking the positions of the points were the laser light was reflected on the plant's surface. The result is a 3D point cloud of the plant surface. An example for the result is given in figure 1.

Figure 1: Picture of a 52 days old tobacco plant (left side) and the corresponding 3D reconstruction (right side). Each leaf is shown in a different color.

2.3 A parametric leaf model

Since most conventional machine learning algorithms require data in the form of attribute vectors, the 3D point clouds were further processed to acquire relevant aspects of plant growth in the form of attribute-value pairs. Each attribute stands for a certain spatial feature of a leaf, e.g. the length, the width, the bending angle towards the ground or its widthwise bending. Figure 2 illustrates one of the bending attributes.

The attribute extraction was done in two steps. In the first step, the plant was segmented into leaves using a spatial clustering algorithm. In the second step, attribute extraction was done using a leaf model fitting algorithm. In the course of this algorithm, a model leaf is transformed until it fits the segmented leaf. From the resulting transformation the values of the attribute vector can be calculated. Further details of the attribute extraction methods can be found in [Uhrmann *et al.*, 2013]. An example of the result can be seen in figure 3.

Figure 2: Example showing the default model with no bending (left side) and a model with a slight bend towards the ground (right side).

Figure 3: The result of the model fitting algorithm. Each leaf (shown in different colors) has a model leaf fitted to it.

2.4 Expert ratings

A biologist rated the stress level of the measured plants to create class labels for supervised classification algorithms from human expert knowledge. To avoid external influences the ratings were performed in a controlled experiment situation. The expert was asked to assign stress classes to each plant measurement. The classes were *no stress*, *moderate stress* and *strong stress*. This simple 3-choice distinction was chosen to keep the ratings comparable and as objective as possible. A more complex scale, for instance an estimation of stress measured in days of exposure, would suffer from personal rating preferences of the expert.

For each measurement the expert was presented a side view photo of the plant. The expert was asked to rate each of the 700 measurements. To conceal the pattern to which the plants were stressed during the measurements, the plant images were shown in a random order. Additionally, the age of the plant in days was provided to the expert. The resulting class labels were assigned to the corresponding leaf attribute vectors by extending each vector by a class attribute whose value was the classification of the respective plant.

3 Classification of tobacco leaves

3.1 Preprocessing

Processing the raw measurement data to attribute vectors describing the leaf shape consists of three steps: Plant measuring and reconstruction, leaf segmentation, and model fitting. Each of these processing steps may induce noise into the data, which is described in the following.

1. In the measuring and reconstruction step it can happen that parts of a leaf cannot be captured. The main reason

for this are occlusions, e.g. upper leaves that cover up parts of lower leaves and prevent the laser beams from reaching all parts of the plant surface. This might result in gaps or clipped leaves which is challenging for the following processing steps.

2. In the segmentation step there is a possibility that a leaf is not recognized, e.g. because it is to close to another one. In that case, the resulting leaf mesh would contain two leaves. This is problematic because the model fitting algorithm is designed for an input mesh which contains only one single leaf. The opposite might also be the case: More than one leaf is detected where there should be only a single one. This might happen if there are big gaps in the point cloud, which virtually split the leaf into several parts. The resulting meshes would contain parts of a single leaf, which are all assumed to be whole leaves. Both effects add noise into the data since the model fitting algorithm is not able to detect inconsistent input meshes.

3. In the model fitting step, the major source for errors is invalid input data from the previous processing steps. An example for this is given in figure 4. The point cloud data of the small leaf in the front contains gaps, which causes the model fit to fail. The resulting attribute vector for this leaf will therefore contain errors and reduce the quality of the classification.

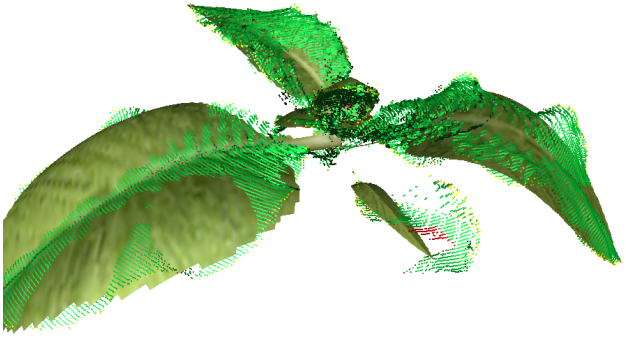


Figure 4: An example for an invalid model fit. The small leaf in the front was not fitted correctly due to gaps in the point cloud data.

Avoiding these errors in advance is difficult. The pivot is the measuring and reconstruction step, because errors in this step propagate through all subsequent steps. However, measuring and reconstruction of plants is very challenging. Due to the complex shape of plants there is no way to avoid occlusions in all cases. Therefore it is probably impossible to design reconstruction algorithms which are able to avoid gaps and clippings completely.

Consequently, data cleaning is required to filter out erroneous attribute vectors before classification. In the course of our work, two types of cleaning were applied to the data.

Firstly, the distance between the model leaf and the original point cloud was taken as an error indicator. This was measured as the accumulated distance between each point of the model leaf mesh and the nearest point of the reconstructed leaf mesh. Each attribute vector which showed a very large distance value was deleted from the data.

Secondly, each attribute vector was checked for inconsistent values with respect to correlated attributes. As an example, figure 5 shows a plot of all leaf attribute vectors, showing the leaf length on the x-axis and leaf area on the y-axis. There is a correlation between the scale and the area of a leaf. All leaf attribute vectors which exceeded a fixed threshold with respect to the ratio between leaf length and area were considered as outliers. The failed model fit in figure 4 is a typical example for how this outliers emerge. In that case, the model leaf was deformed to a needle-like shape. Therefore the ratio between scale and area is too small and the attribute vector can easily be identified as an outlier.

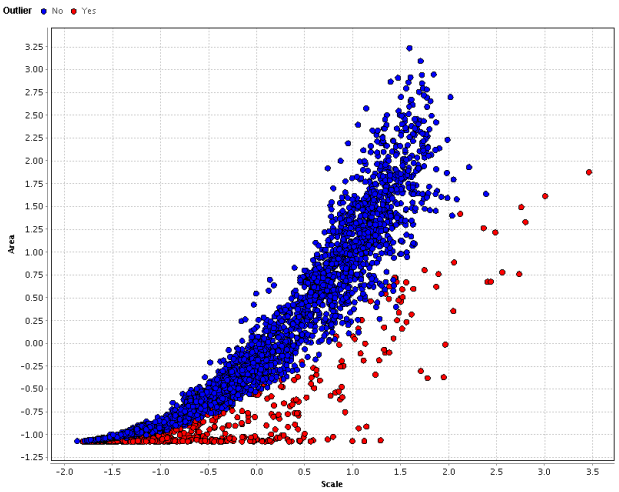


Figure 5: The correlation between the area and the length (scale) of a leaf. Outliers are marked in red.

3.2 Classification and evaluation

Decision trees, neural networks and linear regression were used for classification to evaluate which of these classifiers is suited best for the data. We used RapidMiner 5.3 for the data mining. Each classifier was tested using cross-validation with three subsets of validation. The input data was weighted by stratification, since healthy plants outnumbered moderately and strongly stressed plants.

Different approaches to learning have been tested, which differed in two aspects.

Firstly, the approaches differed in the number of classification classes. Tertiary classification approaches included the classes *healthy*, *moderately stressed* and *strongly stressed*. Moreover, binary classification approaches were tested, in which the class *moderately stressed* was omitted.

Secondly, the approaches used different input data. Approaches on leaf level were based on the attribute vectors corresponding to the respective leaves. On plant level, some adaptations were required, since no global plant attributes were provided by the preprocessing steps. Therefore the attribute vectors of single leaves were transformed to attribute vectors describing whole plants. This was done by creating plant attributes containing the mean values of all corresponding leaf attributes. Furthermore, global plant attributes were calculated, e.g. height, radius and the total leaf surface area.

Regarding all aspects, there are four different approaches which were carried out. For each approach, the already mentioned classifiers were applied. Therefore 12 classification results were achieved. Table 6 shows the respective accuracy rates.

The first approach was performed using leaf attribute vectors as input data and performing a tertiary classification. The best accuracy rate was 52.04% using linear re-

	tertiary, leaves	binary, leaves	tertiary, plants	binary, plants
decision tree	43.77%	52.28%	75.07%	93.69%
neural network	45.40%	72.02%	82.46%	97.08%
linear regression	52.04%	81.81%	82.35%	95.53%

Figure 6: The results of the classification, showing the accuracy rates for each learning approach.

gression. However, this poor result is reasonable. As it was explained in section 2.4, leaf attribute vectors were labeled with the rating of the whole plant. However, tobacco leaves show different behavior when exposed to drought stress, depending on the type of the leaf. A tobacco plant tries to retain young, strong leaves as long as possible. Consequently, older leaves show stress symptoms earlier than younger leaves. Since the leaf data was labeled with the ratings of the whole plant, the poor result of the first approach can therefore be explained because of discrepancies between the labels and the data.

The second approach was carried out by performing binary classification on the leaf attribute vectors. The reason was to assure that stress is recognizable in the leaf data at all. If the intermediate class is omitted, there is more tolerance for the definition of class boundaries, which allows for an easier classification. This second approach yielded much better results. Accuracy rates up to 81.81% could be reached using linear regression. Consequently stress symptoms can be recognized from the data, but the labels provided are not suitable for tertiary classification on the leaf level.

In the third approach tertiary classification was performed on the level of whole plants, which yielded much better results. An accuracy rate of 82.35% could be reached, using linear regression. Since this was a tertiary classification, this is a considerable improvement compared to the first and second approach. Furthermore this confirms that the stress state of a plant cannot be classified by considering single leaves only.

The fourth approach was a binary classification on the plant level. Is was carried out to check whether the binary classification yields considerably better results compared to the tertiary classification. The task was almost solved by neural networks, with an accuracy rate of 97.08%. Consequently, tertiary classification is more challenging than binary classification. One of the possible reasons for this is explained in section 4.

It is worth to note that decision trees performed worse than any other classifier in all approaches. This indicates that decision trees seem to have a learning bias which is disadvantageous for the classification of this kind of data. This might be due to the fact that the input data contains only continuous values.

4 Discussion and Outlook

Since our work in this field is still in progress, there are plenty of open points and ways to proceed. In order to further increase accuracy rates, some effort must be put into the reduction of errors in the preprocessing steps, since they are propagated through all subsequent processing steps and are difficult to be recognized in the cleaning step.

Furthermore, other approaches to extract plant attributes from the measurement data might yield different results than our model fit approach. For example, Lin et al. [2013] use a simple function model describing the shape of the leaf margins. It is possible that such an approach to leaf modeling might also provide suitable data for classification tasks.

Moreover, there are also different ways in which the transition from leaf to plant level classification could be realized. In our approach, plants were mainly classified based on the mean values of the corresponding leaf attributes. Another possible approach would be to perform pre-classification on the leaf level, and a second classification on the plant level. This might yield better results since not every leaf of a plant would be considered individually, whereas the smoothing effects of calculating mean values is avoided.

Another critical point is the reliability of the expert ratings. As it was stated in section 2.4, the classification labels stem from interviewing a human expert. These ratings might vary in precision due to the subjectivity of human judgments. As our fourth classification approach showed, adding an intermediate class between healthy and stressed plants adds plenty of complexity to the classification task. If this is due to vague intermediate classifications by the expert, a machine learner which is trained with this data might therefore never be able to yield optimal results.

Consequently the reliability of the expert ratings must be validated. A possible approach would be to repeat the ratings with the same data but different experts. If the ratings match, this is an indication that human experts are reliably able to distinguish stressed from healthy plants.

5 Related work

As far as we are aware, there is only little related work in the field of stress classification of plants by machine learning methods. In [Wu *et al.*, 2007], a leaf recognition algorithm is described using probabilistic neural networks based on leaf images acquired by scanner or digital cameras. Such an approach might be adapted to distinguish stressed from healthy plants.

In [Chaerle and Van Der Straeten, 2001], a survey of several techniques for monitoring plant health is provided, including fluorescence imaging, thermal imaging and others. These methods have the benefit that stress can be detected earlier than with visual measurement systems since visible changes in plant shape are already effects of biochemical processes, which can be detected earlier with the described methods. Therefore it would be worth to apply machine learning techniques on the data provided by these methods and compare the results with our work.

However, some of this methods are not applicable with our framework since high throughput of single plants must be assured. For instance, systems based on hyperspectral imaging, like they are used in [Römer *et al.*, 2012], are not feasible in our context although they have successfully been applied in the detection of drought stress.

6 Summary

We have shown that constructing a model for the impact of drought stress on plant growth can be inferred from measured geometric leaf features using machine learning techniques. These features are acquired using a sheet-of-light measurement system. Such a system could be used to monitor plant growth in greenhouses, as they are used in the production of pharmaceutical products.

To build our model, we set up a measurement campaign to acquire a broad range of data, extracted attributes of interest from the spatial data and applied several machine learning techniques to achieve a number of comparable results. In this measurement campaign, tobacco plants of the species *Nicotiana tabacum* were measured on a regular basis and stressed according to a fixed schedule. Drought was chosen as the stress type of interest since it is easy to simulate by cutting off irrigation. The measurement data was reduced to vectors of attribute-value-pairs describing essential features of the physical shape of a plant. This was done using a parameterized leaf model developed at the *Fraunhofer IIS*. Combined with labels provided by expert ratings, these attribute vectors form input data which is compatible with standard machine learning techniques.

Classification on the level of single leaves yields poor results (lowest accuracy: 43.77%) because the labels were not appropriate for single leaf data. However, classification on the level of whole plants yields good results with accuracy rates up to 97.08%.

There are several ways to further increase classification performance. For instance, further effort could be put into the reduction of errors in the preprocessing step. Using different data sources or using other methods for the transition from leaf to plant level might also yield better results.

Acknowledgments

This study was mainly funded by the Fraunhofer Future Foundation within its "Malaria-Vaccines" project.

References

[Chaerle and Van Der Straeten, 2001] L. Chaerle and D. Van Der Straeten. Seeing is believing: imaging techniques to monitor plant health. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1519(3):153–166, 2001.

[Lin *et al.*, 2013] F. Lin, X. Wang, H. Wang, X. Shi, and D. Li. Modeling the blade shape of landscape trees. In *Proceedings of the 7th International Conference on Functional-Structural Plant Models, Saariselkä, Finland*, June 2013.

[Römer *et al.*, 2012] C. Römer, M. Wahabzada, A. Ballvora, F. Pinto, M. Rossini, C. Panigada, J. Behmann, J. Lon, C. Thurau, C. Bauckhage, K. Kersting, U. Rascher, and L. Plümer. Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis. *Functional Plant Biology*, 39:878–890, 2012.

[Uhrmann *et al.*, 2013] F. Uhrmann, C. Hügel, S. Paris, O. Scholz, M. Zollhöfer, and G. Greiner. A model-based approach to extract leaf features from 3d scans. In *Proceedings of the 7th International Conference on Functional-Structural Plant Models, Saariselkä, Finland*, June 2013.

[Wu *et al.*, 2007] S.G. Wu, F.S. Bao, E.Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang. A leaf recognition algorithm for plant classification using probabilistic neural network. In *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, pages 11–16, 2007.

Evaluation der Qualität lexikalischer Ressourcen zur Stimmungserkennung in literarischen Texten

Dirk Reinel
Hochschule für Angewandte Wissenschaften Hof
Alfons-Goppel-Platz 1
95028 Hof, Deutschland
dreinel@iisys.de

Zusammenfassung

In dieser Veröffentlichung wird die Qualität bzw. Effektivität von lexikalischen Ressourcen zur automatischen Stimmungserkennung in literarischen Texten evaluiert. Dazu werden die drei unterschiedlichen Wortlisten Sentiment Phrase List (SePL), NRC Emotion Lexicon und SentimentWortschatz (SentiWS) sowie ein manuell klassifiziertes Referenzset verwendet. Der Testkorpus besteht aus 20 ausgewählten Märchen aus der Sammlung der schönsten Kinder- und Hausmärchen der Brüder Grimm, online bezogen von der Website des Projekt Gutenberg-DE. Durch die Berechnung bestimmter Maßzahlen wird gezeigt, dass die Werte der automatisch klassifizierten Texte mit den Werten des Referenzkorpus korrelieren. Es wird jedoch deutlich, dass die berechneten Korrelationen, bedingt durch die begrenzte Anzahl der Texte des Testkorpus, sehr instabil sind. Zudem wird gezeigt, dass die manuelle Klassifikation (positiv / negativ) auf Satzebene durch zwei voneinander unabhängige Personen bei dieser Art von literarischen Texten zufriedenstellend funktioniert.

1 Einleitung

Literarische Texte im Allgemeinen und Märchen im Speziellen sind und waren schon immer Träger von verschiedenen Emotionen und Stimmungen. Durch gezielte Variationen und Schwankungen im Verlauf der Handlung, wie beispielsweise durch den Tod handelnder Figuren, erzeugen Autoren unterschiedliche Stimmungen, die die Leserschaft „fesseln“ und mitfühlen lassen sollen.
Bislang beschränken sich Informationen über diese Texte meist auf verschiedene Metadaten (z.B. Autorennamen), quantitative Größen (z.B. Anzahl der Seiten), (semi-) professionelle Kritiken und, in den Zeiten des Web 2.0, Nutzerbewertungen.

Unsere Forschungsarbeit, die sich mit dem Thema „Analyse von literarischen Texten mit Methoden aus den Bereichen Sentiment Analysis und Opinion Mining“ beschäftigt, knüpft an diesem Punkt an. Durch verschiedene Untersuchungen und Experimente, z.B. mit Wortlisten oder Methoden des maschinellen Lernen, sollen Emotionen und Stimmungen in literarischen Werken automatisch identifiziert und entsprechend verarbeitet bzw. visualisiert werden. Somit werden die momentan verfügbaren Informationen zu Texten um eine neue Dimension erweitert, wodurch eine neue Möglichkeit zur Suche entsteht, eine Suche auf Basis von Stimmungen und Emotionen.

Ziel dieser laufenden Arbeit, die einen Teil der erwähnten Forschungsarbeit darstellt, ist einerseits die generelle Evaluierung bereits vorhandener Typen von Wortlisten für die Verwendung zur Stimmungserkennung in literarischen Texten. Andererseits soll eine Methode zur Erzeugung eines geeigneten Referenzkorpus getestet und dieser Korpus anschließend auf Plausibilität geprüft werden. Dazu wird ein Testkorpus, bestehend aus verschiedenen Märchen der Brüder Grimm, zusammengestellt, entsprechend vorverarbeitet und anschließend von zwei voneinander unabhängigen Personen manuell klassifiziert. Nach Prüfung der Plausibilität (Übereinstimmung), durch Berechnung des statistischen Maßes *Cohens Kappa*, wird einer der beiden manuell klassifizierten Korpora als Referenz ausgewählt. Anschließend wird der ursprüngliche Testkorpus von drei ausgewählten Wortlisten automatisch klassifiziert und die Ergebnisse werden mit dem Referenzkorpus verglichen.

2 Verwandte Arbeiten

In den vergangenen Jahren wurde in den Bereichen *Opinion Mining* und *Sentiment Analysis* intensive Forschungsarbeit geleistet. Eine gute Übersicht über die verschiedenen Themen und Forschungsbereiche gibt die Veröffentlichung von Liu und Zhang [Liu and Zhang, 2012]. Ergänzend sind die beiden, erst kürzlich erschienenen, Artikel von Feldman [Feldman, 2013] und Cambria und Kollegen [Cambria *et al.*, 2013] zu nennen, die ebenfalls einen aktuellen Überblick über dieses Forschungsfeld liefern.

In der Veröffentlichung von Mohammad und Turney wird die Erstellung eines englischsprachigen Lexikons mit emotionstragenden Unigrammen (*NRC Emotion Lexicon*) beschrieben [Mohammad and Turney, 2010] und dessen Anwendung unter anderem anhand von Versuchen mit den ins Englisch übersetzten Märchen der Brüder Grimm gezeigt [Mohammad, 2011]. In einem aktuelleren Artikel [Mohammad, 2012] führt Mohammad dieses Thema weiter aus und beschreibt die Suche nach Emotionen in E-Mails und Büchern.
Klenner präsentiert in seiner Arbeit [Klenner, 2009] einen regelbasierten Ansatz zur Sentimentanalyse für die deutsche Sprache und evaluiert diesen anhand eines literarischen Textes.

Neben den in der Arbeit behandelten Lexika *Sentiment Phrase List (SePL)* [Rill *et al.*, 2012a] und *SentimentWortschatz (SentiWS)* [Remus *et al.*, 2010], existieren zudem noch weitere lexikalische Ressourcen für die deutsche Sprache. Dabei handelt es sich zum einen um das *Polarity Lexicon* [Clematide and Klenner, 2010] mit ca. 8.000 meinungstragenden Worten, zum anderen um *GermanPolari-*

tyClues [Waltinger, 2010] mit mehr als 10.000 Einträgen. Beide Ressourcen beinhalten Unigramme sowie einen dazugehörigen Meinungswert.

3 Testkorpus und lexikalische Ressourcen

3.1 Erstellung eines geeigneten Testkorpus

Bei der Auswahl geeigneter Texte für die Durchführung der Versuche entschieden wir uns für die Märchen der Brüder Grimm[1]. Diese literarischen Texte kommen aus den folgenden Gründen zum Einsatz. Zum einen sind es kurze Texte, die schnell gelesen und klassifiziert werden können. Außerdem sind sie leicht verständlich und weisen eine klare Struktur auf, wodurch sie für erste Experimente bestens geeignet scheinen. Des Weiteren beinhalten sie eine Vielzahl verschiedener Emotionen und Figuren, die größtenteils stereotypisch handeln und somit in Klassen eingeteilt werden können. Ein letzter und wichtiger Punkt der für den Einsatz dieser Märchen spricht, ist die Verwendung in anderen Forschungsarbeiten (siehe z.B. [Mohammad, 2011]). Dies erlaubt es weitere, beispielsweise sprachübergreifende, Vergleiche durchzuführen.
Um die Texte in digitaler Form zu erhalten, wurde ein Werkzeug entwickelt, mit dessen Hilfe es möglich war, den gesamten Korpus automatisch von der Website des Projekt Gutenberg-DE[2] zu beziehen. Da der Märchenkorpus insgesamt aus 194 Texten besteht, eine Menge die ohne großen Zeitaufwand nicht gelesen und manuell klassifiziert werden kann, wurden daraus 20 Texte ausgewählt. Diese Texte wurden anschließend manuell so aufbereitet, dass jeder Satz bzw. jede zusammengehörende Konstellation auf einer eigenen Textzeile stand. Diese Vorarbeit erleichterte einerseits den Taggern die Identifikation und manuelle Klassifikation der Sätze und ermöglichte andererseits die reibungslose Durchführung der automatischen Klassifikation auf Basis der Wortlisten. Die Vorverarbeitung der 20 Texte wurde vollständig manuell durchgeführt, da erste Versuche mit automatischen Methoden zeigten, dass die strikte Satztrennung durch Pattern, wie z.B. Satzzeichen, nicht zufriedenstellend funktionierte. Ein Problem waren dabei fehlende Satzzeichen innerhalb der Texte. Des Weiteren wurden durch die automatische Satztrennung Textfragmente, die logisch zusammengehören, getrennt (z.B. „ ‚Soll die dumme Gans bei uns in der Stube sitzen!‘ sprachen sie.“).

3.2 Verwendete Ressourcen

Für die Versuche verwenden wir drei lexikalische Ressourcen, die unterschiedliche Strukturen und Einträge aufweisen und somit auf verschiedene Weise auf die Daten des Testkorpus angewendet werden müssen (siehe Kapitel 4.2). Eine dieser Ressourcen enthält zudem ausschließlich Einträge in englischer Sprache, wodurch eine Vorverarbeitung in Form einer Übersetzung notwendig wird.

Die *Sentiment Phrase List* (SePL) enthält in der aktuellen Version 1.1 15.142 meinungstragende Adjektiv- und Substantiv-Phrasen mit einer Phrasenlänge bis zu $N = 5$. Jeder Eintrag der Liste enthält zudem einen Wert (Opinion Value), der die Meinung der Phrase auf einer stetigen Skala im Bereich $[-1, +1]$ widerspiegelt. Die Liste beinhaltet ausschließlich Phrasen in deutscher Sprache, die Einträge sind zudem lemmatisiert. Der generische Ansatz zur sprachunabhängigen Erstellung einer solchen Liste wurde in der Veröffentlichung von Rill und Kollegen [Rill *et al.*, 2012b] vorgestellt.

Das *NRC Emotion Lexicon* enthält 141.820 Einträge mit englischsprachigen Unigrammen (Version 0.92). Dabei existieren jedoch pro eindeutigem Unigramm zehn Einträge, die binär die Zugehörigkeit zu [keiner, einer, mehreren] der 8 Basisemotionen [Plutchik, 1980] sowie die Zuordnung zu „positiv“ oder „negativ“ anzeigen. Dadurch verringert sich die tatsächliche Anzahl zunächst auf 14.182 Wörter, wobei nicht jedem Wort ein Wert zugeordnet ist.
Um diese Ressource für die Versuche verwenden zu können, musste sie entsprechend vorbereitet werden. Dazu wurden im ersten Schritt alle Wörter entfernt, die keine Zuordnung zu „positiv“ oder „negativ“ hatten. Im zweiten Schritt wurde die, nun deutlich kleinere, Liste automatisch mit Hilfe der Google Translate API[3] übersetzt. Dabei kam ein einfacher „first-best“ - Ansatz zum Einsatz, d.h. es wurde immer der jeweils erste Übersetzungsvorschlag übernommen. Einige Unigramme konnten nicht automatisch übersetzt werden, wodurch sich die Größe der Liste auf letztendlich 5.255 Einträge verringerte.
Tabelle 1 zeigt die schrittweise Selektion durch die verschiedenen Verarbeitungsschritte und die entsprechende Anzahl der Einträge, die den jeweiligen Schritt passieren konnten.

Selektionsschritte	**Anzahl der Einträge**	
	Absolut	Relativ
Vollständige Liste	141.820	-
Eindeutige Wörter	14.182	100%
Wörter mit Zuordnung	5.555	39,17%
Übersetzte Wörter	5.255	37,05%

Tabelle 1: Übersicht der Selektionsschritte beim *NRC Emotion Lexicon*.

Die Ressource *SentiWS* enthält in der verwendeten Version 1.8c 3.468 meinungstragende deutsche Unigramme in ihrer Grundform mit einem dazugehörigen Meinungswert auf einer stetigen Skala im Bereich $[-1, +1]$. Eine Besonderheit dieser Liste ist die generelle Separation positiver und negativer Wörter durch die Auflistung in zwei verschiedenen Sublisten. Zu allen Einträgen werden zudem Part-of-Speech Tags sowie, falls existent, verschiedene Beugungsformen gelistet. Diese Ressource enthält neben Adjektiven und Adverbien auch meinungstragende Verben und Substantive.

4 Experimente

Im Folgenden wird die manuelle Klassifikation von Sätzen zur Erzeugung des Referenzkorpus sowie die automatische Klassifikation auf Basis der Wortlisten beschrieben. Bei letzterem werden ebenfalls kurz die notwendigen Schritte der Vorverarbeitungen der Märchen (Testkorpus), die Annahmen für Wortlisten und die Algorithmen zur Erkennung bzw. Extraktion beschrieben. Am Ende des Kapitels wird schließlich der für die automatische Klassifikation der Sätze zuständige Algorithmus erläutert.

4.1 Manuelle Klassifikation von Sätzen

Um eine Bewertungsgrundlage für die Ergebnisse der späteren automatischen Klassifikation zu erhalten, mussten alle Texte des Testkorpus manuell klassifiziert werden.

[1] Brüder Grimm: Die schönsten Kinder- und Hausmärchen (http://gutenberg.spiegel.de/buch/6248/1)

[2] http://www.projekt.gutenberg.de/

[3] https://developers.google.com/translate/

Dafür wurden zwei voneinander unabhängige Personen beauftragt. Diese Personen, im Folgenden Tagger 1 (T1) und Tagger 2 (T2) genannt, wurden nicht explizit über die Arbeitsweise der Algorithmen oder den Einsatz von Wortlisten informiert. Beide Tagger erhielten alle Texte des Testkorpus sowie den Auftrag emotions- bzw. stimmungstragende Sätze mit entsprechenden Tags zu markieren. Sätze mit positiven Emotionen sollten mit dem Tag „<p>“, Sätze mit negativen Emotionen mit Tag „<n>“ am Zeilenende versehen werden. Sätze, die keine Emotionen tragen oder in denen sich positive und negative Stimmungen gegenseitig „aufheben“, sollten nicht markiert werden. Insgesamt sollte jeder der beiden Tagger 1.604 Sätze lesen und klassifizieren.
Zur Überprüfung der Übereinstimmung der klassifizierten Korpora von Tagger 1 und Tagger 2 wurde für jeden Text das statistische Maß *Cohens Kappa* berechnet. Dabei stellten wir fest, dass die Übereinstimmung bei 70% der Texte zufriedenstellend war ($\kappa = 0,57$). Im gesamten Durchschnitt lag die Übereinstimmung bei $\kappa = 0,51$. Dieser Wert erscheint für Texte dieser Art durchaus akzeptabel. Werden nun ausschließlich Sätze betrachtet die von beiden Personen als positiv oder negativ klassifiziert wurden, ergibt sich eine Übereinstimmung von 87%.
Für die Evaluation der verschiedenen Wortlisten wurden die Ergebnisse von Tagger 1 als Referenz ausgewählt.

4.2 Automatische Klassifikation von Sätzen

Um den erstellten Testkorpus (siehe Kapitel 3.1) für die automatische Klassifikation verwenden zu können, sind einige Vorverarbeitungsschritte notwendig. Diese Schritte sowie die verwendeten Algorithmen zur Extraktion stimmungstragender Wörter und Phrasen werden in Abbildung 1 schematisch dargestellt.

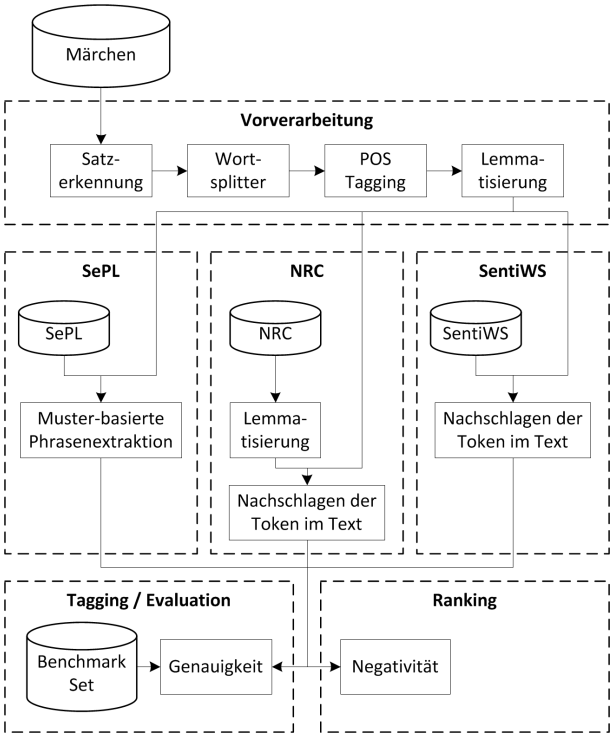


Abbildung 1: Schematische Darstellung des Ablauf der Experimente.

Vorverarbeitung
Im ersten Schritt müssen alle Sätze identifiziert und separiert werden. Dieser Schritt wurde bereits manuell durchgeführt, wodurch ein simpler Algorithmus, der das jeweilige Zeilenende in Texten erkennt an dieser Stelle ausreichend ist. Anschließend werden Apache OpenNLP[4] Tokenizer sowie Apache OpenNLP Part-of-Speech Tagger angewendet, um einzelne Wörter zu extrahieren und mit einem korrekten POS-Tag zu versehen. Zum Schluss werden alle Wörter lemmatisiert. Dazu wird das auf Morphy[5] [Lezius *et al.*, 1998] basierende deutsche Morphologie-Lexikon[6] verwendet.

Sentiment Phrase List
Für die Anwendung dieser Liste, die größtenteils Phrasen enthält, wird ein musterbasierter Algorithmus für die Extraktion meinungstragender Wörter und Phrasen verwendet. Dazu werden die Texte nach bestimmten Adjektivphrasen ([adverbiales Adjektiv | Adverb | Partikel | Verb im Partizip Perfekt | unbestimmtes Pronomen] & Adjektiv) und Substantivphrasen ([Adjektiv | Adverb | unbestimmtes Pronomen] & Substantiv) durchsucht und anschließend mit den 15.142 Einträgen der Wortliste verglichen. Für die anschließende Klassifizierung der Sätze in „positiv“ und „negativ“ wurden bestimmte Schwellenwerte für Phrasen festgelegt. Phrasen mit einem Opinion Value $(OV) \leq -0,3$ werden negativ gewertet. Phrasen mit einem $OV \geq 0,3$ werden positiv gewertet. Alle Phrasen mit einem Opinion Value zwischen diesen beiden Schwellenwerten werden als neutral angesehen.

NRC Emotion Lexicon
Bei dieser übersetzten Wortliste, die größtenteils Unigramme enthält, kommt ein einfacher „Lookup - Algorithmus“ zum Einsatz. Das bedeutet, dass im Text nach einzelnen NRC Token gesucht wird und diese entsprechend ihrer binären Zuordnung bewertet werden. Um die Trefferquote zu erhöhen werden die Wörter dieser Liste zudem noch lemmatisiert.

SentimentWortschatz
Da diese Liste ebenfalls ausschließlich Unigramme enthält, die zudem bereits in ihrer Grundform vorhanden sind, wird wiederum ein „Lookup - Algorithmus“ verwendet. Für die Einteilung der Unigramme in die Klassen „positiv“ und „negativ“ wird die bereits vorhandene Vorgabe durch die Liste verwendet (siehe Kapitel 3.2).

Algorithmus zur Klassifizierung der Sätze
Durch die entsprechenden Vorarbeiten und Festlegungen liefern alle Listen für stimmungstragende Wörter oder Phrasen eine Klassifizierung in „positiv“ und „negativ“. Um damit Sätze des Testkorpus klassifizieren zu können, wird ein einfacher Algorithmus verwendet, der die Anzahl gefundener positiver und negativer Wörter und Phrasen pro Satz zählt und auf Basis der Mehrheit eine Entscheidung trifft. Ist beispielsweise die Anzahl der negativ klassifizierten Wörter in einem Satz größer als die Anzahl positiv klassifizierter Wörter, wird dieser Satz „negativ“ klassifiziert. Sollten in einem Satz gleich viele positive und negative Phrasen vorkommen wird dieser neutral eingestuft.
Durch diese simple Methode wird die Qualität der Listen, im Bezug auf die Erkennung stimmungstragender Wörter in Texten dieser Art, direkt miteinander vergleichbar.

[4] http://opennlp.apache.org/
[5] http://www.wolfganglezius.de/doku.php?id=cl:morphy
[6] http://www.danielnaber.de/morphologie/

5 Erste Ergebnisse

Tabelle 2 zeigt einen direkten Vergleich der relativen Häufigkeit von positiv oder negativ klassifizierten Sätzen im Verhältnis zu allen Sätzen pro Text. Verglichen werden die Ergebnisse von Tagger 1 mit den Ergebnissen der drei Ressourcen, wobei die Tabelle nach den Ergebnissen von Tagger 1 absteigend sortiert ist. Es zeigt sich, dass die beiden Wortlisten *Sentiment Phrase List* (mit 30%) und *SentimentWortschatz* (mit 37%) deutlich weniger Sätze eindeutig zuordnen können als Tagger 1. Die Ergebnisse des *NRC Emotion Lexicon* liegen mit durchschnittlich 60% eindeutig klassifizierter Sätze nur knapp hinter denen von Tagger 1 (68%). Auch wird deutlich, dass die Ergebnisse dieser Ressource erstaunlich stabil sind.

Märchen	**Relative Häufigkeiten**			
	T1	**SePL**	**NRC**	**SentiWS**
Drei Spinnerinnen	0,89	0,43	0,68	0,43
Wasser des Lebens	0,85	0,22	0,66	0,40
Hund und Sperling	0,84	0,24	0,65	0,20
Geschenke d. kl. V.	0,83	0,60	0,87	0,50
Hans mein Igel	0,79	0,31	0,78	0,31
Aschenputtel	0,79	0,22	0,73	0,38
Zwei Brüder	0,74	0,20	0,76	0,31
Zaunkönig und Bär	0,74	0,21	0,74	0,39
Zwölf Brüder	0,70	0,29	0,78	0,46
Nixe im Teich	0,69	0,39	0,78	0,43
Schneew. und RR	0,68	0,31	0,75	0,46
Das Rätsel	0,68	0,23	0,77	0,38
Eisenofen	0,68	0,37	0,67	0,42
Geist im Glas	0,57	0,23	0,80	0,25
Gevatter Tod	0,55	0,27	0,58	0,32
Jorinde / Joringel	0,55	0,30	0,64	0,43
Hans im Glück	0,53	0,38	0,75	0,29
Rotkäppchen	0,51	0,23	0,59	0,38
Spindel, Webers. ...	0,48	0,33	0,74	0,33
Die klugen Leute	0,47	0,17	0,68	0,29

Tabelle 2: Vergleich der relativen Häufigkeit positiv / negativ klassifizierter Sätze im Vergleich zu allen Sätzen.

Um die Ergebnisse der Listen vergleichen und bewerten zu können, wurden pro Text drei Negativitätswerte berechnet. Der erste Wert beschreibt das Verhältnis von negativ klassifizierten Sätzen im Vergleich zu allen Sätzen (neg/all), der zweite Wert das Verhältnis von negativ klassifizierten Sätzen im Vergleich zu positiv klassifizierten Sätzen (neg/pos). Der letzte Wert gibt schließlich das Verhältnis von negativ klassifizierten Sätzen im Vergleich zu allen positiv oder negativ klassifizierten Sätzen an ($neg/(neg+pos)$).
Tabelle 3 zeigt die berechneten Korrelationskoeffizienten der Ergebnisse der verschiedenen Negativitätswerte. Dabei wurde jeweils die Korrelation zwischen den Ergebnissen des Referenzkorpus (Tagger 1) mit denen der einzelnen Listen untersucht. Es wird deutlich, dass die Ergebnisse aller drei Listen mit denen des Referenzkorpus für alle Negativitätswerte korrelieren.

Durch den Einsatz von lediglich 20 Referenztexten ist diese Korrelation allerdings instabil. Werden die Ergebnisse eines einzigen Textes aus der Korrelationsberechnung entfernt, verändern sich die Korrelationskoeffizienten für alle Negativitätswerte bereits sehr stark. Besonders das Fehlen von extrem unilateral klassifizierten Texten führt dabei zu einer deutlichen Verschlechterung der Korrelation. Eine Korrelation der Ergebnisse der *Sentiment Phrase List* mit dem Referenzkorpus ist in diesem Fall nicht mehr erkennbar und auch die Korrelation zwischen den Ergebnissen des *SentimentWortschatz* und der Referenz nimmt stark ab (mit Ausnahme der Ergebnisse von neg/all). Lediglich die Ergebnisse des *NRC Emotion Lexicon* bleiben stabil und zeigen weiterhin eine Korrelation.

Kandidaten	**Korrelationskoeffizient**		
	neg/all	neg/pos	$neg/(neg+pos)$
T1-SePL	0,49	0,88	0,52
T1-NRC	0,63	0,77	0,76
T1-SentiWS	0,45	0,83	0,56

Tabelle 3: Korrelationskoeffizienten der Ergebnisse.

6 Zusammenfassung und Ausblick

In dieser laufenden Arbeit wurden Experimente mit drei verschiedenen lexikalischen Ressourcen mit dem Ziel durchgeführt, Aussagen über deren Qualität beim Einsatz zur Stimmungserkennung in literarischen Texten zu treffen. Dazu wurde ein Testkorpus erstellt, der 20 ausgewählte Märchen der Brüder Grimm beinhaltete. Alle Sätze in diesem Korpus wurden manuell von zwei voneinander unabhängigen Personen klassifiziert. Anschließend wurde ein Referenzkorpus erstellt. Zudem wurden alle Sätze des Testkorpus, basierend auf den Daten der drei Wortlisten, automatisch klassifiziert. Die Resultate der automatischen Klassifikation wurden mit denen des Referenzkorpus verglichen und erste Ergebnisse präsentiert.
Es zeigte sich, dass lexikalische Ressourcen durchaus dazu geeignet sind, Emotionen und Stimmungen in literarischen Texten zu erkennen. Die Erkennungsraten variieren jedoch von Liste zu Liste, sodass genauere Analysen erforderlich sind. Es wurde ebenfalls deutlich, dass der verwendete Testkorpus mit 20 Märchen deutlich zu klein dimensioniert war, um verlässliche Ergebnisse zu erhalten.

Die nächsten Schritte werden somit zum einen die Generierung eines größeren Referenzsets und zum anderen eine detaillierte Analyse der automatisch klassifizierten Daten sein. Dabei stehen vor allem Analysen zur Güte der automatischen Erkennung im Vordergrund (Precision und Recall) und basierend darauf die Suche und anschließende Auswertung von möglichen Problemen und Fehlerquellen.

Literatur

[Cambria *et al.*, 2013] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

[Clematide and Klenner, 2010] Simon Clematide and Manfred Klenner. Evaluation and Extension of a Polarity Lexicon for German. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 7–13, 2010.

[Feldman, 2013] Ronen Feldman. Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[Klenner, 2009] M Klenner. Süsse Beklommenheit und schmerzvolle Ekstase: Automatische Sentimentanalyse in den Werken von Eduard von Keyserling. In *Von*

der Form zur Bedeutung: Texte automatisch verarbeiten (Proceedings of the Biennial GSCL Conference 2009), pages 91–97. Narr Francke Attempto Verlag GmbH, 2009.

[Lezius *et al.*, 1998] Wolfgang Lezius, Reinhard Rapp, and Manfred Wettler. A Freely Available Morphological Analyzer, Disambiguator and Context Sensitive Lemmatizer for German. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, pages 743–748, 1998.

[Liu and Zhang, 2012] Bing Liu and Lei Zhang. A Survey of Opinion Mining and Sentiment Analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer US, 2012.

[Mohammad and Turney, 2010] Saif M. Mohammad and Peter D. Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, 2010.

[Mohammad, 2011] Saif Mohammad. From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, 2011.

[Mohammad, 2012] Saif M. Mohammad. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730 – 741, 2012.

[Plutchik, 1980] R. Plutchik. *A general psychoevolutionary theory of emotion*, pages 3–33. Academic press, 1980.

[Remus *et al.*, 2010] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Conference on Language Ressources and Evaluation*, pages 1168–1171, 2010.

[Rill *et al.*, 2012a] Sven Rill, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V. Zicari, and Nikolaos Korfiatis. A Phrase-Based Opinion List for the German Language. In *Proceedings of KONVENS 2012*, pages 305–313, 2012.

[Rill *et al.*, 2012b] Sven Rill, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, and Daniel Simon. A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In *Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, 2012.

[Waltinger, 2010] Ulli Waltinger. GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1638–1642, 2010.

Fully Self-Supervised Learning of an Arm Model

Martin V. Butz and **Armin Gufler** and **Konstantin Schmid** and **Fabian Schrodt**
Cognitive Modeling, Department of Computer Science,
Faculty of Science, Eberhard Karls University of Tübingen,
Sand 14, 72076 Tübingen, Germany
martin.butz@uni-tuebingen.de; armin.gufler@student.uni-tuebingen.de;
konstantin.schmid@tum.de; FSchrodt@gmx.de

Abstract

Performing a mere infinite number of different movements in our everyday life happens mostly in an automatic and unconscious way, requiring hardly any attention. One necessary requirement for generating different movements dependent on the current circumstances is knowledge about redundant behavioral alternatives and the capability to flexibly choose the current best one. In this paper, we evaluate an architecture that learns to represent such behavioral alternatives in the form of a modular body model from scratch. Moreover, the architecture is able to selectively choose between the behavioral alternatives, yielding kinematic control commands. The proposed architecture combines (temporal) Hebbian learning mechanisms for learning the body model with model-based reinforcement learning techniques for controlling the body. We evaluate the current system capabilities, comparing several configurations and parameter dependencies. Our results show that the architecture can robustly learn a highly flexible arm control system.

1 Introduction

Nicolai Bernstein has called behavior that is extremely flexible and adaptive *dexterous behavior* [Bernstein, 1967]. A key ingredient to succeed in the generation of dexterous behavior is (i) knowledge about the redundant behavioral interaction alternatives and (ii) the capability to flexibly choose between these alternatives task-dependently on the fly.

Few learning models so far have focused on utilizing redundant behavioral capabilities for generating flexible behavior. On the other hand, a huge number of learning architectures exist that learn to coordinate an arm without exploiting redundancy. Traditional Reinforcement learning (RL) algorithms [Berthier *et al.*, 2005] and various forms of policy gradients have been used [Peters and Schaal, 2008; Sigaud and Peters, 2010]. Also other direct learning methods have succeeded in controlling robot arms, such as direct inverse modeling approaches [Kuperstein, 1988] or resolved motion rate control and distal supervised learning [Whitney, 1969; Jordan and Rumelhart, 1992], each of which resolves redundancy during learning.

Here we focus on the SURE_REACH model [Butz *et al.*, 2007; Herbort and Butz, 2007], which is a sensorimotor unsupervised learning model that learns about the redundancy of a body and resolves this redundancy on the fly once the current goal and constraints are given. The model represents the end-effector location (i.e. location of the hand) in a *task space* and, additionally, the arm constellation (i.e. joint angles) in a *posture space*. Previously, hard-encoded population codes were used to cover both spaces by means of uniform neural grids. Hebbian and temporal Hebbian learning mechanisms were used to learn the (redundant) inverse kinematic mappings from task space to posture space. Model-based RL within posture space was used to execute dexterous, goal-directed behavior. Both, location goals or posture goals can be pursued, while potentially avoiding obstacles and considering additional task constraints. It was also shown that the model is able to anticipate subsequent task goals and consequently to incorporate those into the current behavior optimization [Herbort and Butz, 2007]. Moreover, SURE_REACH is able to incorporate uncertainties in its goal choice [Herbort *et al.*, 2007]. In sum, it has been shown that SURE_REACH is able to generate highly dexterous behavior.

Although SURE_REACH proved to be a useful model of dexterous human arm control, two challenges remained. First, the neural representation in joint space scales hyper-exponentially with the degrees of freedom controlled, so that a seven degree of freedom arm cannot be modeled with sufficient accuracy. Second, the neural population codes were pre-wired and not learned. While the first issue has been addressed by modularizing SURE_REACH, separating the posture space into individual joint spaces [Ehrenfeld and Butz, 2013], the second issue remained open. Thus, we developed a model that also learns the neural population codes. In the following we explain and evaluate this system, which is able to learn a model of an arm with three degrees of freedom (3-DOF) in two-dimensional (2D) space from scratch.

2 Arm Model Overview

Using kinematic motor commands a 3-DOF 2D arm is simulated. The simulation provides angles and end-effector location signals to the learner and is able to execute kinematic motor commands (small angular changes), respecting joint and torque constraints and adding some noise. Fig. 1 gives an overview over the implemented architecture.

2.1 Learning of a Kinematic Arm Representation

In contrast to SURE_REACH, the arm space representations are learned using growing, self-organizing neural network techniques. In particular, we use the Time Growing Neural Gas (TGNG) algorithm [Butz *et al.*, 2010]. TGNG

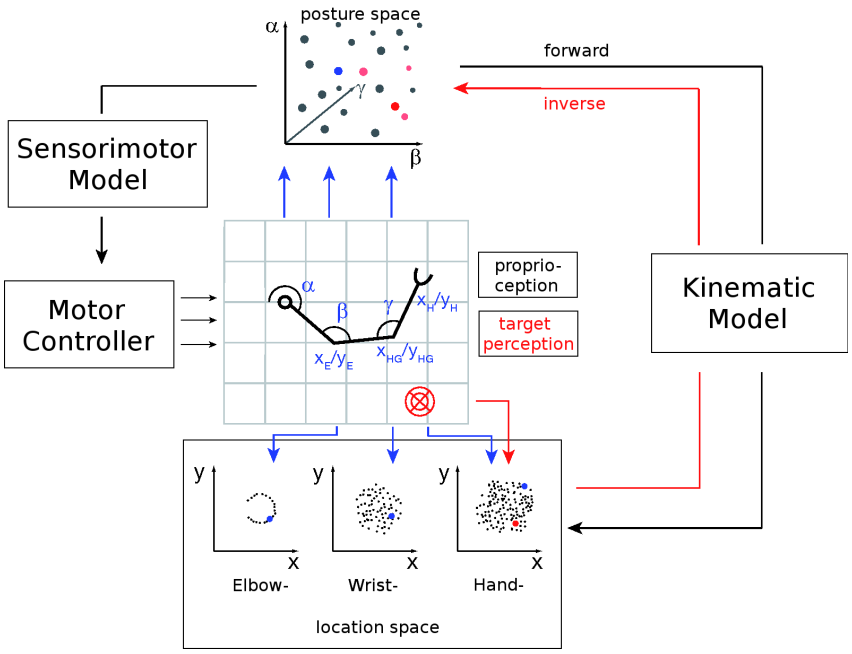


Figure 1: Overview of the arm model

grows neurons on demand, given the current sensory input differs more from the closest neuron than a threshold, specified by parameter θ. Moreover, TGNG grows neural connections between neurons that were the closest neurons in temporal succession. The connections are associated with a motor code that approximates the average motor command executed when traversing this connection. TGNG is used to learn neural representations of the task space given (x,y) locations of the end-effector, elbow, and wrist, and of the posture space given 3D angular vectors. Concretely, learning is achieved by a random walk, executing random arm movements and learning from the consequences. In this way, all accessible locations in the posture and task spaces will be observed, so that neural population codes can be distributed across the respective spaces by means of TGNG.

The second key component of the architecture is the mapping between task and posture spaces. Similar to SURE_REACH, we use a Hebbian learning mechanism [Carpenter and Grossberg, 1991] for learning the inverse kinematic mappings. More details about this learning process are provided below.

2.2 Goal Directed Behavior

The final central feature of the system is its ability to behave in a goal directed manner. Either a target in location space can be specified or a target in posture space. Pursuing a particular posture is rather easy: in this case, first, a posture goal activates the closest neurons in posture space. Next, this goal activity is propagated backwards throughout the posture space by means of model-based RL [Sutton and Barto, 1998]. Finally, control is invoked by deducing the neuron in posture space closest to the current arm posture, extracting the connection to the most strongly activated neighboring neuron, and executing the motor code that is associated with the neural connection. Given a well-connected network of neurons, behavior is guaranteed to reach the goal-activated neuron. When pursuing a goal location, however, this goal location first activates the closest neurons in the corresponding task space. Next, this activation is projected into the posture space by the inverse kinematic connection matrix (established by the Hebbian learning mechanism) between task space and posture space. The resulting goal manifold in posture space is then propagated throughout posture space and control is invoked as before.

To get more specific, we use the following notation: nodes of a neural network are denoted by small letters, $\mathcal{P}$ denotes the set of all posture space neurons, and $\mathcal{L}$ the set of all hand location space neurons. Furthermore, the terms "node" and "neuron" will be used in an interchangeable fashion, as nodes are part of neuronal networks. The activity of node n is denoted by a_n.

Location Goal Reaching in Detail

For a given target $X \in \mathbb{R}^2$, the neuron $l^* \in \mathcal{L}$ being nearest to X is determined. Then, the correlations (weights from the learned inverse kinematic) are used to activate corresponding posture nodes. As the mapping encodes full redundancy, there may be many posture neurons having significant correlations. Our reference approach considers all posture neurons having a correlation weight greater or equal to the correlation threshold τ ($\tau = 0.15$ has shown to be a good choice) and induces external activity a_p^{ext} to posture neuron $p \in \mathcal{P}$ by

$$a_p^{ext} = w_{p,l^*} \cdot \xi + 1, \quad \text{if } w_{p,l^*} \geq \tau, \tag{1}$$

with $w_{p,l} \in [0,1]$ being the learned correlation weight (mapping location- with posture-space neurons) between p and l. Thus, a good portion of neurons is ignored and a manifold of interesting postures is activated. The constant factor ξ applied to the weight is set to 0.3.

The induced activity is then propagated through the network using the following model-based state value learning rule (according to [Butz *et al.*, 2010])

$$a_p \leftarrow \max(a_p^{ext}; \gamma \cdot \max_{q \in N(p)} (a_q)), \tag{2}$$

where $N(p)$ denotes the set of neurons being neighbors of neuron p. The learning rule thus encodes either the own ex-

Table 1: Performance of location goal reaching (702 start-target combinations) using the setups STD (standard setting), *1ACT* (activating only the best node in location space), *ALLACT* (activating all nodes in location space), *NNEM* (check for new node after every movement). See text for details about the configurations.

Measurement	*STD*	*1ACT*	*ALLACT*	*NNEM*
Successfully reached	**676.7**	597.5	679.0	628.1
Quality of path (QoP)	**2.46**	4.09	10.50	3.10
Average #steps needed	**44.7**	65.8	246.2	56.6
Median #steps	**36.5**	61.8	126.0	38.3
Standard deviation #steps	**52.1**	43.8	336.8	128.3

ternal activity a_p^{ext} or the activity of the most active neighboring neuron. The parameter γ is a constant discount factor, set to 0.7 to accord with ξ, which is set to 0.3 in (1). In this way, propagated activity will always be smaller than externally induced activity from task space.

The reaching process uses the proprioception of the arm to determine which posture neuron is currently the closest. This is denoted as the arms' current node. Given the arm's current node, the system considers all neighboring nodes and attempts to move in the direction of the most active neighbor node. This is done until the arm has reached the actual target.

Inhibition Once a node has been visited by the arm, it gets inhibited, in order to make it less interesting for the arm to aim at the same posture again. Hence, loops where the arm moves back and forth are avoided effectively. Also in general it seems reasonable to avoid the reaching of the same postures with the arm within one goal-directed movement. A similar mechanism was also used in TGNG [Butz *et al.*, 2010].

Transitioning to neighbor nodes At every time step the arm is at a specific node and tries to move to the best neighbor node. An important concept is how to actually transition the state of the arm to the next node. The most straight forward approach is to check after each movement which node in the posture space network is nearest to the arms' new proprioception. Analyzing the behavior showed that in many cases the arm does not reach the neighbor node it originally aimed at. Often a node in between is encountered and becomes the new current node. In consequence, the arm has to do new planning again and the trajectories in turn become more turbulent. Such nodes in between are not necessarily connected to the node the arm came from and the neighbor it wants to move to, as connections are established depending on the movements made during the learning process. In effect, undesired erroneous behavior may occur.

To avoid this behavior, we estimate the approximate distance d the arm has to travel when starting at node v and aiming at neighbor node w. Searching for a new current node is now only done if the distance moved since starting at v is $\geq d$. The benefit of this approach is that movements are becoming much smoother overall and the arm jumps less between near nodes. Additionally, lots of searches for the current nearest node are omitted, which is particularly useful when the posture network is very dense.

3 Evaluation of the Current Performance

To evaluate the current performance we focus on reaching hand location goals. 702 different start-goal combinations were chosen randomly to test the behavior of the arm. Before trying tor reach targets, the system executes the self-supervised learning process for $T = 100,000$ time steps. Within each step one smooth movement of the arm is executed, the space representations (neural networks) are updated, and forward and inverse kinematics are learned. No obstacles were placed within the environment for the main parameter evaluations. In order to decrease the influence of random learning movements, if not reported differently, all results presented are average values from 100 runs, each of which was tested on the identical 20 sets of 702 start-goal combinations.

We evaluated the system with the aim of revealing parameter dependencies and robustness. Thus, we first vary several crucial parameter settings revealing the respective parameter influences. After that, we illustrate the capability of the arm to avoid obstacles. Finally, we show one slightly more involved run, in which case more than 99% of all location goals are reached successfully.[1]

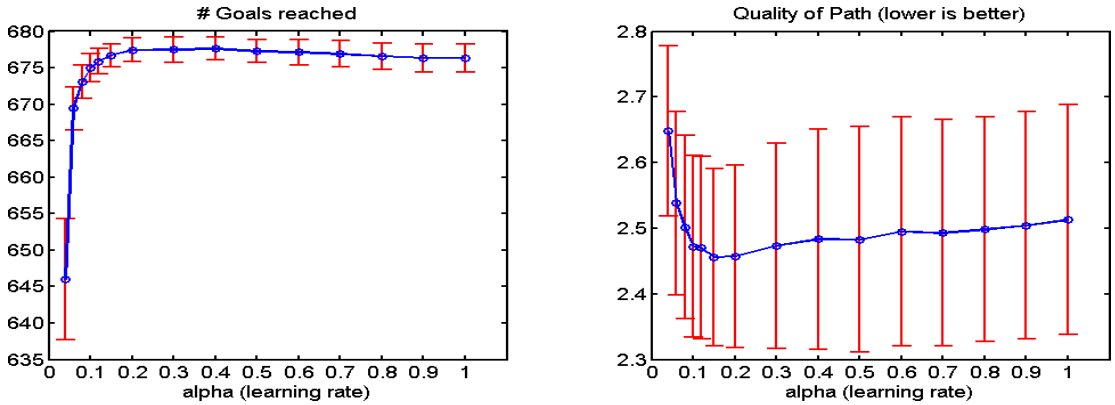
3.1 Goal-based Neural Activation

Table 1 shows the performance of goal directed behavior of the arm. We present the average number of successfully reached targets (of 702 in total) along with measurements regarding the quality and length of the path taken. The latter values are only considered for successful movements. The quality of path (QoP) is defined as the ratio of the actual distance moved to the optimal distance, with the optimal distance being the Euclidean distance from start to target location.

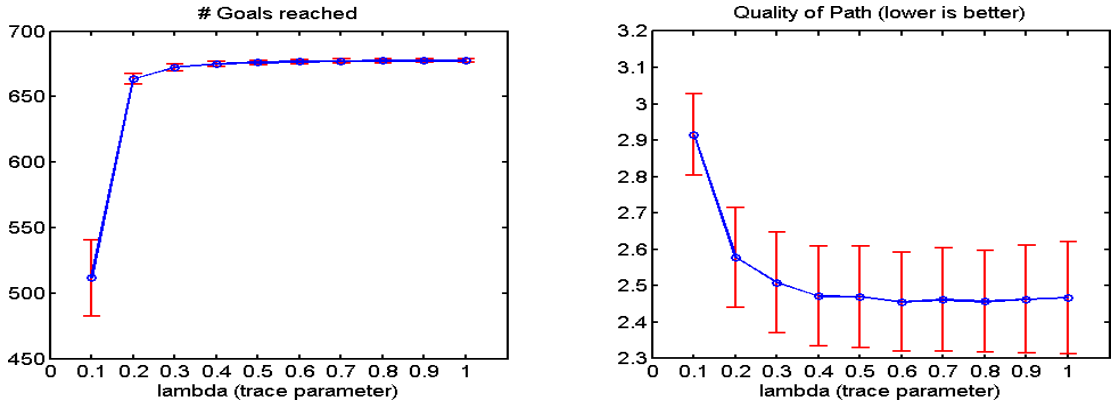
The reference approach (STD) induces activity to a manifold of suitable posture neurons, activating each of them with an external activity close to one (slightly graded dependent on the strength of the learned connection weights). Note that the system does not reach all targeted locations because we stop the goal reaching and count the trial as a failure when the highest activated node in posture space is reached but this node is not close-enough to the targeted goal. If the trial was not stopped in this case, the system typically reaches all 702 nodes eventually (due to the neural inhibition mechanism), but the quality of the path degrades strongly.

The baseline approach (1ACT) considers the results when only the strongest connected node in posture space is activated by the task space goal node. In this case, the redundant mapping is fully disregarded; the arm simply attempts to reach the neuron with the highest connection

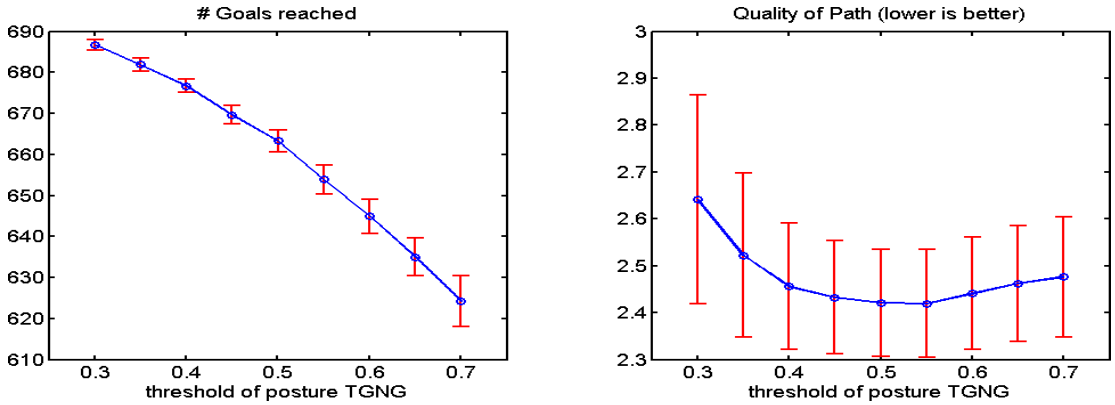
[1] If not stated differently, the parameters of the system were set as follows: TGNG node creation thresholds: $\theta_P = .4$; $\theta_L = .2$; TGNG node parameter adaptation value: $\epsilon_P = .075$; $\epsilon_L = .2$. Hebbian mapping learning parameters: learning rate $\alpha = .15$, trace value $\lambda = .6$, decelerating learning rate .99.



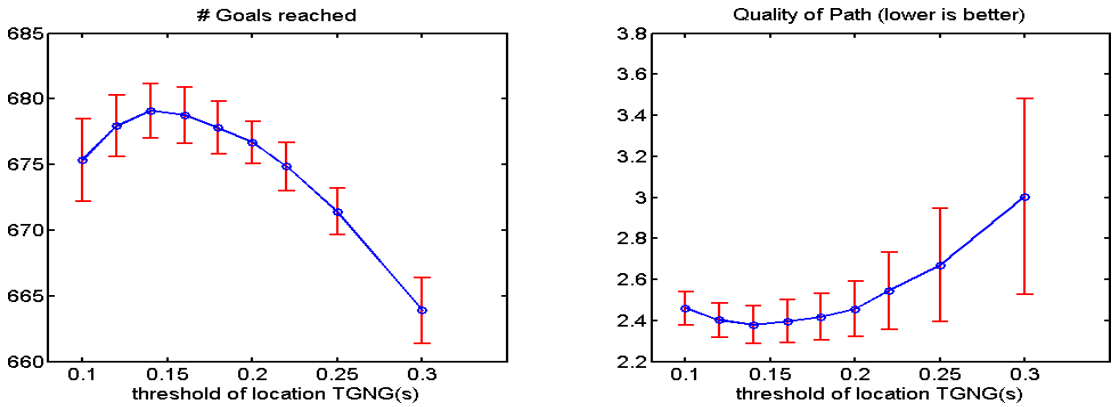
(a) Influence of learning rate parameter α.



(b) Influence of activity trace parameter λ



(c) Influence of posture space TGNG threshold parameter θ_P.



(d) Influence of location space TGNG threshold parameter θ_L.

Figure 2: Parameter influences on goal-reaching and quality of path performance.

weight from task space. The evaluation results show, that less goals are reached and the average distance moved to reach goals is larger using this configuration. This clearly shows, that the information encoded in the mapping is indeed useful. It activates a whole manifold of posture nodes and nicely handles redundancy.

In the all activated (ALLACT) case, all posture nodes that are connected to the goal location neuron are activated. The evaluation demonstrates (see Table 1) that the goal reaching performance of the arm gets worse using this configuration. Even though most goals are finally reached, the quality of the reaching movement is really bad compared to the other configurations. This is because neurons in joint space are activated that are hardly anywhere close to the goal location – thus the arm moves through this broad goal manifold in posture space rather randomly, yielding a much worse QoP.

As explained in Sec. 2.2 our reference model does not search for a new current posture node after each executed movement step. *NNEM* denotes the alternative approach estimating a new nearest posture node after each step (as was done previously in TGNG [Butz *et al.*, 2010]). The evaluations of Table 1 clearly show that this approach is leading to worse results. All quality measures get worse, particularly the much higher variance in the steps needed suggests that there are cases where the arm is getting lost in some loops or unsuitable detours. Note however, that the approximation of the distance to travel before making a transition may in other representation spaces not be that easy to estimate. For example many obstacles or frequently changing environments can be a problem.

3.2 Parameter Influence

Besides these goal-directed neural activations, we also pursued a more involved study of parameter learning influences. In the following, we explore influences of the learning rate and the trace parameter of the Hebbian learning mechanism, as well as of the threshold parameter of TGNG for learning the neural location and posture populations.

Hebbian Learning Parameters

The learning rate parameter α determines how fast and aggressive the mapping weights between task and posture space are adjusted. Learning the inverse mapping is done via the Hebbian feedback learning rule [Carpenter and Grossberg, 1991]:

$$\Delta_w = \alpha \cdot \overline{a_p} \cdot (a_l - w) \tag{3}$$

where w is the current weight between node $p \in \mathcal{P}$ and location node $l \in \mathcal{L}$ and Δ_w is the change of the weight in the current learning step. $\overline{a_p}$ denotes the activity trace of p and a_l the activity of l.

Varying the parameter between 0.04 (very slow learning process) and 1.0 (immediate weight changes) shows that a rather low learning rate between 0.1 and 0.3 is well-suited (cf. Fig. 2(a)). Higher values do not necessarily lead to much less goals reached, but the quality of path gets worse. Overall, however, the learning rate has a rather small influence on the whole system, if within a reasonable, moderate range.

The parameter λ defines the length of the activity trace, when learning the weights of a mapping between posture and location space. The rule to compute the activity trace $\overline{a_n}$ of node n is defined as

$$\overline{a_n} = (1 - \lambda) \cdot a_n^{old} + \lambda \cdot a_n^{new}, \tag{4}$$

where a_n^{old} is the activation of n in the previous step and a_n^{new} the newly induced activation in the current time step. The activity trace aims at encoding movement trajectories. Thus, nodes visited previously can also be correlated to current nodes (from other representation spaces). Setting λ to 1 means that there is no activation trace at all.

The results shown in Fig. 2(b) suggest, that a medium lambda parameter ≥ 0.4 is suitable, in order to have a good QoP. Having a really long trace (low λ), the resulting mappings are lacking quality. Having no trace on the other hand seems to be a reasonable option. At least the variables analyzed here are not suffering from the absence of the activity traces.

Thresholds of TGNG Networks

The threshold parameter(s) of the Time Growing Neural Gas networks (TGNGs) [Butz *et al.*, 2010] have a huge influence on the general performance. A lower (error)-threshold means that the network will have higher density because new nodes are inserted more frequently.

Fig. 2(c) shows the influence of the posture threshold θ_P and the location threshold θ_L. Generally, with a lower threshold parameter θ_P more goals can be reached successfully. But having $\theta_P < 0.4$ leads to a significantly worse average quality of path. Further evaluation showed, that for $\theta_P < 0.4$ the standard deviation of the average number of steps needed to reach targets also increases significantly. With more nodes in posture space the arm has more movement possibilities and the probability that it looses the optimal direction for some steps increases. Moreover, the number of updates per connection decreases when a denser network is grown. Another important consideration at this point is the proper choice of θ_L depending on θ_P and vice-versa. The densities of the networks should be somewhat comparable to ensure proper mappings between them. While this is a rather experimental task at this point, it appears that the densities can be increased significantly for both networks without getting bad behavioral results. However, the computation-wise efficiency worsens with decreasing θ_P as the number of nodes increases quite rapidly (approximately 4000 nodes for a threshold of 0.3 compared to 2000 nodes with the threshold set to 0.4). Moreover, the number of learning iterations necessary to develop a proper kinematic mapping and to associate sufficiently accurate motor codes increases with increasing density of the networks.

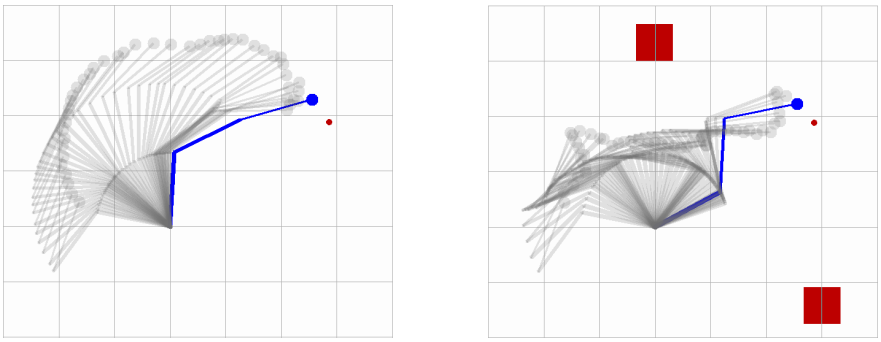
The location TGNG threshold parameter θ_L has a similar influence. With increasing θ_L, less goals can be reached. The quality of path gets worse with increasing θ_L too. But for lower values like 0.15 and 0.20 the difference is not really significant.

Obstacles

The system is also able to handle obstacles within the environment. To avoid crashing into them, neurons near and within obstacles are simply inhibited. Moreover, the inverse kinematics model provides information about postures that are possibly causing crashes, which are also inhibited so that no reward can be propagated through obstacle-based inhibitions. Fig. 3 shows typical trajectories of the arm when reaching a target, pointing out different trajectories chosen due to an obstacle.

Reliable Goal Location Reaching

Seeing that the presented results so far have not reached all goal locations robustly, we ran slightly more involved runs with the TGNG threshold parameters set to $\theta_P = .2$;



(a) Without obstacles (b) With two obstacles

Figure 3: Two reaching movements, (a) without obstacles and (b) with two obstacles (red squares). The arm has to move its end effector to the location goal target (small red circle). The trajectory taken by the arm is visualized in light gray color. The resulting posture when the target was (nearly) reached is shown in blue. Due to the obstacles in (b) the resulting trajectory differs, effectively avoiding crashing into the upper obstacle.

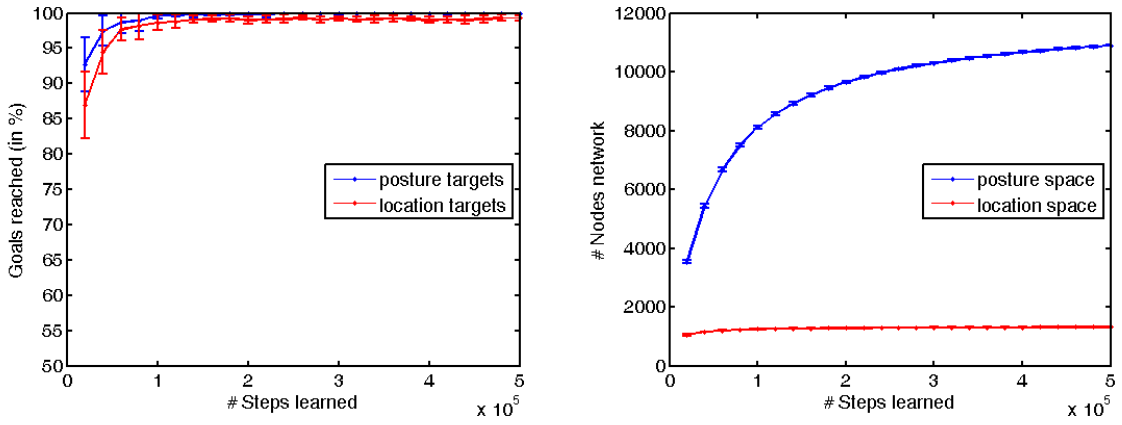


Figure 4: With even lower TGNG thresholds, the networks grow bigger and the target reaching performance also reaches close-to 100% even when reaching for location targets.

$\theta_L = .05$ for $500,000$ learning iterations. All other parameters were set to the standard settings. Fig. 4 shows that with these settings nearly all location goals are reached. The results also confirm that posture goals are easier to reach and are in all cases reached reliably. Moreover, the results also show that the network sizes grow significantly. Clearly the posture network is much larger due to the 3D angular space being covered in contrast to the 2D location space. When considering the path qualities achieved, Fig. 5 shows that the path does not become fully straight. The performance with respect to posture space goals indicates that the motor encodings in the neural connections are not perfect. Moreover, due to the focus on the next best node, the path cannot be fully straight. Low-pass filters in the motion generation may alleviate this problem. The location space performance is even worse. In this case, though, the system attempts to generate a straight path in posture space, not in location space. Thus, the measure confirms progressive learning but it is not an absolute performance measure.

In sum, the results show that the overall system is able to reach all goals in posture space as well as in location space, even though all representations and associations between these representations were learned from scratch learning from uncontrolled, random motions. While the path optimality may be improved further, it has to be kept in mind that the system currently blindly executes each movement

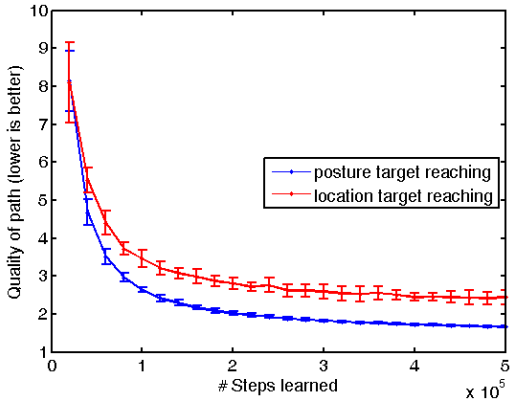


Figure 5: Also the path quality improves when a larger network is learned. Note that the system always attempts to move straight in posture space - thus the path quality for location goals has to be taken with a grain of salt. Nonetheless, there is definite room for improvement when considering the path quality results.

without considerations of the previous one. Thus, integrating successive motion vectors may be able to solve the challenge of generating more smooth and straight paths to goals. Nonetheless, the results concerning obstacle avoidance have confirmed that the system is indeed able to generate dexterous behavior, having learned its body model fully from scratch.

4 Conclusion

The results presented in this paper confirm that the combination of TGNG with Hebbian learning and model-based RL works rather effectively. However, clearly learning takes a rather long time and the final path quality is not optimal. Note, however, that learning was based on completely random movements. Others have shown that goal-babbling from early on can improve the speed of learning significantly [Rolf *et al.*, 2011]. Moreover, active information seeking, that is curious behavior [Oudeyer *et al.*, 2007], which was included in the TGNG algorithm [Butz and Reif, 2010], may be included in the current system to speed-up learning even further by essentially acting information-oriented. In this study, however, we refrained from utilizing such techniques to reveal a baseline system performance.

Besides curiosity, other motivations may be included to explore the external environment further once the arm model is sufficiently accurate. Forward kinematic mappings can also be learned along similar lines, allowing the anticipation of action consequences in posture space as well as in task space. Such anticipatory capabilities may be used for filtering incoming sensory information, acting based on internal expectations, as well as for forward planning. Finally, we intend to use this learning approach to learn the population encodings and mappings in the modular modality frame (MMF) model [Ehrenfeld and Butz, 2013]. MMF modularizes the SURE_REACH approach yielding a body model with maximally three dimensional spaces. However, currently no structural learning takes place in MMF. Due to this dimensional restriction in MMF, the combination of the utilized learning techniques with MMF promises to yield a system that is able to learn a full seven degree of freedom human arm model or even a full human body model in 3D space. Future research will need to investigate the capabilities of generating dexterous behavior within such a learned, distributed body model.

Acknowledgments

Funding from the Emmy Noether program (German Research Foundation, DFG, BU1335/3-1) is acknowledged. Moreover, the authors like to thank the Cognitive Modeling team for their support.

References

[Bernstein, 1967] N A Bernstein. *The co-ordination and regulation of movements*. Pergamon Press, Oxford, 1967.

[Berthier *et al.*, 2005] N. E. Berthier, M. T. Rosenstein, and A. G. Barto. Approximate optimal control as a model for motor learning. *Psychological Review*, 112:329–346, 2005.

[Butz and Reif, 2010] M. V. Butz and K. L. Reif. Motivated TGNG: Algorithm and performance evaluations. Technical Report CoboslabY2010N001, COBOSLAB, Department of Psychology, University of Würzburg, Würzburg, Germany, 2010. http://www.coboslab.psychologie.uni-wuerzburg.de.

[Butz *et al.*, 2007] M. V. Butz, O. Herbort, and J. Hoffmann. Exploiting redundancy for flexible behavior: Unsupervised learning in a modular sensorimotor control architecture. *Psychological Review*, 114:1015–1046, 2007.

[Butz *et al.*, 2010] M. V. Butz, E. Shirinov, and K. L. Reif. Self-organizing sensorimotor maps plus internal motivations yield animal-like behavior. *Adaptive Behavior*, 18(3-4):315–337, 2010.

[Carpenter and Grossberg, 1991] G. A. Carpenter and S. Grossberg. *Pattern Recognition by Self-Organizing Neural Networks*. MIT Press, Cambridge, MA, 1991.

[Ehrenfeld and Butz, 2013] S. Ehrenfeld and M. V. Butz. The modular modality frame model: Continuous body state estimation and plausibility-weighted information fusion. *Biological Cybernetics*, 107:61–82, 2013.

[Herbort and Butz, 2007] O. Herbort and M. V. Butz. Encoding complete body models enables task dependent optimal behavior. *Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007*, pages 1424–1429, 2007.

[Herbort *et al.*, 2007] O. Herbort, D. Ognibene, M. V. Butz, and G. Baldassarre. Learning to select targets within targets in reaching tasks. *6th IEEE International Conference on Development and Learning*, ICDL 2007:7 – 12, 2007.

[Jordan and Rumelhart, 1992] M. I. Jordan and D. E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

[Kuperstein, 1988] M. Kuperstein. Neural model of adaptive hand-eye coordination for single postures. *Science*, 239:1308–1311, 1988.

[Oudeyer *et al.*, 2007] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *Evolutionary Computation, IEEE Transactions on*, 11:265–286, 2007.

[Peters and Schaal, 2008] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21:682–697, 2008.

[Rolf *et al.*, 2011] M. Rolf, J. J. Steil, and M. Gienger. Online goal babbling for rapid bootstrapping of inverse models in high dimensions. *IEEE Int. Conf. on Development and Learning and on Epigenetic Robotics*, pages 1–8, 2011.

[Sigaud and Peters, 2010] O. Sigaud and J. Peters, editors. *From Motor Learning to Interaction Learning in Robots*. Springer, 2010.

[Sutton and Barto, 1998] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 1998.

[Whitney, 1969] D. E. Whitney. Resolved motion rate control of manipulators and human prostheses. *IEEE Transactions on Man-Machine Systems*, 10:47–53, 1969.

Three Handwriting Adaptation Approaches for Digit Recognition

Dieter Lutz and **Martin Toepfer** and **Frank Puppe**
Department of Computer Science VI, University of Würzburg
Am Hubland, Würzburg, Germany
dieter.lutz@googlemail.com, {toepfer, puppe}@informatik.uni-wuerzburg.de

Abstract

Handwritten digit recognition in applications like automatic exam grading is challenging because handwritings inherently differ between development data sets and real application scenarios. To overcome this issue, we propose three handwriting adaptation methods and compare them on a data set of 2860 samples of 26 different users. We explain preprocessing and feature extraction steps, and suggest different adaptation approaches: two methods are similar to bootstrapping, and one method uses dimensionality reduction. Experimentally, we show that adaptive approaches yield significantly better results than the standard classifier. Adaptation improved the precision of an already good baseline by about one to four percent depending on the size of the training set.

1 Introduction

High accuracy handwritten digit recognition systems typically build upon thousands to tens of thousands of annotated training images [DeCoste and Schölkopf, 2002]. However, even then classifiers appear to be *brittle* when applied on slightly different data [Seewald, 2012]. One of the reasons for this is that the same digit can look differently when it was written by different persons as depicted in Figure 1. The samples on each side were written by one student respectively, and we can see clearly that each handwriting has distinct characteristics. As a result, samples of one class are consistent for a certain student, but there may be inconsistencies summing over samples of different handwritings. Therefore, it is difficult to categorize a sample without annotated samples from the same person for a normal classifier. The proposed adaptation methods deal with this problem by classifying a whole set of samples from one user at once instead of one by one. By this means, they can utilize the similarities of these samples to adapt the recognition process to new handwriting styles. The evaluation of the techniques shows that this adaptation of a classifier on the handwriting of one user is possible and can improve the precision of the classifier.

The foundation of the customization approaches in this paper is a Support Vector Machine classifier [Cortes and Vapnik, 1995] which is also used as a baseline in the evaluation. However the introduced techniques are not restrained to this type of classifier, but they can also be used with every other classifier which is able to specify the likelihood of its classification for every class. A possible use case of the proposed methods is automatic exam grading, like in [Mandel *et al.*, 2011]. Hence the considered classes are digits (0-9) as well as commas. On the exam sheets, every digit and symbol is written into a separate box. The exams get scanned and the boxes are cut apart. The resulting gray-scale images are grouped for every exam because they originate from the same user. These sets of images are the input for the classifier.

Our paper has the following structure. Firstly, we discuss related work and give an introduction to preprocessing and feature extraction of the images. In Section 5, we describe three different types of handwriting customization techniques. In the evaluation section, the used data set is described, and the results are presented and discussed. Finally, we suggest topics for future work and give a summary of our findings.

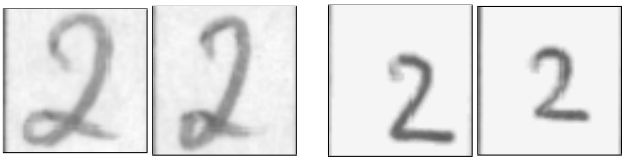


Figure 1: Two samples of digit “2”, written by 2 students with distinct handwritings.

2 Related Work

Handwritten digit recognition is a well examined field with many different approaches. One of the most successful classifiers in this domain are Support Vector Machines (SVM) [Cortes and Vapnik, 1995]. This is confirmed in [DeCoste and Schölkopf, 2002] where a Support Vector Machine approach with the lowest reported test error on the MNIST digit recognition task at that time was proposed.

Processing human and computer generated data has led researchers to adaptation techniques in different domains of artificial intelligence. For instance, context consistencies arise in text processing or speech recognition. Klügl et al. [2012], for example, collectively segment references of scientific papers using statistical graphical models. They exploit the homogeneity of formatting inside of sections originating from style guide usage. Intelligent speech recognition systems must also learn to map signals with speaker specific characteristics to general symbols. For instance, Leggetter and Woodland [1995] showed that linear transformations can be used to maximize the likelihood of the new data and thereby associate patterns across in-

dividual speakers. To adapt online[1] handwritten character recognition, Szummer and Bishop [2006] proposed to use a mixture of experts. They assume a supervised setting where some labeled examples of the new handwriting are always available. Classifiers are trained on clusters of similar handwriting styles, and combined to produce an adaptive model weighted by each classifiers' posterior probability on the labeled samples of a new handwriting. In this paper, however, we propose approaches for adaptation when no labeled data is present.

The subspace embedding technique in this paper is sustained by the assumption that identical digits are close to each other in a common subspace. This hypothesis is supported by [Chapelle *et al.*, 2002] where clustering of unlabeled instances in subspaces was utilized for kernel adaptation in Support Vector Machines, which improved their error rate.

Adaptation of digit recognition models to user characteristics is strongly related to semi-supervision. A very popular technique in this domain is bootstrapping [Yarowsky, 1995] which successively populates the training set with the most confident predictions on the unlabeled data. Two of the methods (Best-First-One/Two) that we propose in Section 5 can be regarded as bootstrapping methods. Best-First-One directly populates the training set with certain unlabeled instances, whereas Best-First-Two creates a new classifier with them and balances this classifier against the initial one.

3 Preprocessing

The first stage of the classification process aims to reduce the impact of different scales, positions and intensities of the symbols. Furthermore it tries to remove noise from the images. In the following, we illustrate all preprocessing steps for the example shown in Figure 2.

Horizontal Cropping As the first step the digit is cropped horizontal. This is done to remove potential vertical lines originating from the frame of the boxes. The detection and removal of lines at the borders is achieved by calculating the average gray-scale value of the two most left and right pixel columns. If the average value for the first column is below 230 the column is removed and the second column is taken into consideration in the same way. Otherwise the image is not modified. Additionally a variance-based heuristic is applied to try and crop the digit horizontally in order to get rid of lines in the interior.

Binarization The next preprocessing step is the previously mentioned binarization. The main goal is to make the input invariant to different intensities of the handwritten symbol. To put this into effect Otsu's method is used to compute a threshold based on the histogram of gray-scale values. The threshold is then applied to categorize the image into black and white pixels.

Centralization The symbol is moved to the center of the image by its center of mass in order to compensate translations.

Symbol-Cropping After centralization, we remove irrelevant and noisy parts of the image to reduce the effect of differently sized symbols. We apply two methods. Firstly a variance-based heuristic, similar to the one used earlier, but this time in horizontal and vertical directions. Secondly, the left and right boundaries of the symbol are estimated as the 3rd and the 97th percentile of the x-coordinates of the black pixels respectively. The bottom and top boundaries are identified analogously.

Resizing Finally, we resize the image to the uniform size of 20x30 pixels by antialiasing. By this means, identical symbols should have nearly the same size, and, essentially, we achieve scale invariance to some degree.

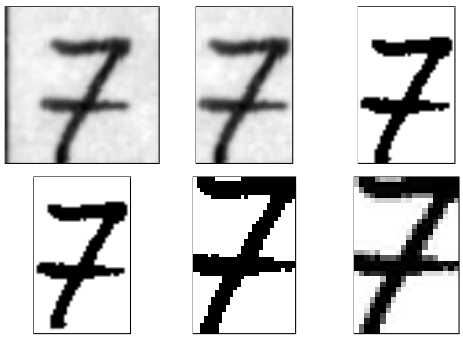


Figure 2: Preprocessing steps: Original, Horizontal Cropping, Binarization, Centralization, Symbol-Cropping, Resizing.

4 Feature Extraction

Feature extraction is the next stage of the classification process. The features used can be easily extracted due to the extensive preprocessing. There are also only two different types of features. These are explained in the following.

Pixel Gray-Scale Value The first type of features are the gray-scale values of the image resulting from the preprocessing. The images have 20x30 pixels, so there are 600 pixel features in total. These are intended to represent the general shape of the digits.

Zone Gray-Scale Value The second type of features are also gray-scale values. These are extracted during preprocessing after the symbol gets centralized. The basic idea is to lay a coarse grid over the image and use the average gray-scale value of the resulting zones as features. In the actual implementation, the image is resized to 14x14 pixels and their value is extracted. This amounts to 196 zones features. The purpose of this type of feature is to distinguish between similar shapes by properties removed through symbol cropping.

5 Approaches for Handwriting Adaptation

Handwriting adaptation is the final and central stage of our system. In contrast to static handwritten digit recognition systems, we utilize the similarities between samples of the handwriting of one user to improve the classifier's performance particularly on this user's set of samples.

One of the inherent challenges of adaptation is how to model and obtain handwriting specific information for a new person. The naive approach is clustering the samples and assigning a class to each cluster. However it is difficult to get an accurate clustering in a real-world application. So a similar method is used which places similar samples near each other in a low-dimensional space. This well-known method is called subspace embedding. Another way to utilize the similarities of a user's samples is bootstrapping. The idea of bootstrapping in this case is to classify only the certain samples at first and then use these samples as if they

[1] online handwriting recognition processes path trajectories rather than static scanned images

were annotated samples to improve the further classification. This method is utilized in both Best-First techniques.

Another challenge is how to integrate the similarity information into the classification process. One possibility is to model it as additional features. But it is not really clear how to make these features invariant to the user's specific differences and the weighting of the additional to the original features is difficult to handle. A simpler way is to directly modify the classifier's probabilities of class affiliation based on the similarity information. This is also the method utilized in all three adaptation techniques, which are proposed in the following.

A further common denominator is the use of an Support Vector Machine as an initial classifier derived from a training set of annotated samples. This classifier is mostly used to calculate the probabilities of class assignment for samples. So an additional challenge is to modify mainly the samples which the classifier is unsure about. We handle this by weighting the modifications of the following techniques by the certainty of the classifier. To determine this certainty a function based on the class probabilities is needed, which is the entropy function in our case.

5.1 Subspace Embedding

The first approach introduced is subspace embedding. The idea of subspace embedding is to reduce the dimensions of the feature vectors of the user's samples to a few concepts. This can be achieved through principal component analysis [Jolliffe, 1986]. Identical digits are likely to have the same concepts so they are also likely to be positioned near each other in the subspace. This is used to modify the sample's probabilities of class affiliation in order to shift the classification of unsure samples in the right direction. To be more concrete, for every sample a number of nearest neighbors are used for the modification. The influence of every neighbor decays based on their quadratic euclidean distance to the considered sample. The overall magnitude of the adjustments made to the probabilities also depend on the certainty of the sample's classification. To estimate how sure the classifier is about an assignment the normalized entropy over the probabilities of class affiliation is used. After these adjustments, the class with the highest probability is assigned to the sample.

5.2 Best-First-One

The Best-First-One method initially classifies all samples of the user set to gather the best samples, i.e. the ones which the classifier is most certain about. The certainty of a sample is identified by the entropy of it's probabilities of class assignment. A threshold is used to split the safe from the unsafe samples. However a specific fraction of the samples is always assumed as safe. These samples are treated as annotated examples and added to the training set. A classifier is created from the enhanced training set. The remaining samples are then classified with this new classifier.

5.3 Best-First-Two

The Best-First-Two technique is very similar to the Best-First-One method. The difference is that the safe samples are not added to the initial training set but used to create a new classifier only based on these samples. The classes for the remaining samples are then determined by the sum of weighted probabilities of both classifiers. The weights are appointed according to the entropies of the probabilities. This way the classifier, which is more sure about the class of a sample, is taken into account to a further extend. A direct advantage of this variation is that only the certain samples have to be learned and not the entire training set. This results in shorter runtimes, especially if the initial training set is large.

6 Evaluation

We evaluated the three different customization methods and the standard classifier, which does not use user specific information. Therefore, a user set was chosen as the test set, a training set was sampled from the remaining user sets, and the precision scores of the four techniques were measured on the test set. This process was repeated for each of the 26 user sets similar to a leave-one-out evaluation setting. The results were averaged over all repetitions for every method and training size. The implementation was done in Python. The libraries NumPy[2], SciPy[3], scikit-learn[4] and Pillow were utilized. In the following, the parameters used for each method are listed.

Support Vector Machine The Support Vector Machine classifier used a polynomial kernel with a degree of three. Probability calculations for class affiliation was enabled.

Subspace Embedding The feature vectors were reduced to three dimensions. The three nearest neighbors were used to adjust the sample's probabilities. The entropies were normalized with the factor $1/7$.

Best-First-One/Two The threshold chosen to separate the safe from the unsafe samples was 0.9. Ten percent of the samples of the user set were always assumed to be safe.

6.1 Data Set

A custom data set was used since most standard data sets do not provide user information associated with their samples. However this is necessary for the adaptation techniques to work. The data set was gathered using test forms with boxes for every symbol. It consists of 26 sets of different users. Every set contains 10 annotated examples for the digits 0-9 and the comma symbol. So there are 110 examples for each user in total and the data set amounts to 2860 samples overall.

6.2 Results

The results are plotted in Figure 3 with the size of the training set as the x-axis and the average precision as the y-axis. All adaptation methods exceed the standard classifier's curve regardless of the training set size. The distance between the standard classifier and the customization methods at a training set size of 100 amounts to about four percent. For the training set size 1000 the difference is still around one to two percent for all adaptation techniques.

Table 1 presents concrete numbers for these experiments. Values printed in bold type show the respective best approach for each training set size. Underlined values mark statistically significantly better precisions compared to the standard classifier (paired t-test, p-value <0.05).

Subspace embedding and Best-First-One provide very similar performance on all training set sizes. The precisions provided by Best-First-Two are also similar for the sizes 100, 200, 600 and 700. On the sizes 300 to 500 Best-First-Two was outperformed by the other two approaches by around 0.5%, but it surpasses them on the sizes 800 to 1000 by about the same amount.

[2]http://www.numpy.org/

[3]http://www.scipy.org/

[4]http://scikit-learn.org/

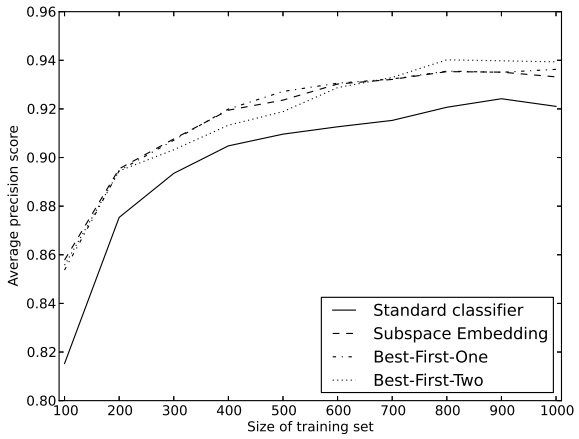


Figure 3: Average precision of the different methods for several training set sizes

Table 1: Table of average precisions. Bold values show the best approach for each training set size. Underlined values are significantly better compared to the standard classifier.

Size of training set	Standard classifier	Subspace Embedding	Best-First-One	Best-First-Two
100	0.815	**0.858**	0.854	0.856
200	0.875	**0.895**	0.895	0.894
300	0.893	**0.908**	0.907	0.903
400	0.905	0.919	**0.920**	0.913
500	0.910	0.924	**0.927**	0.919
600	0.913	0.930	**0.931**	0.929
700	0.915	0.932	0.932	**0.933**
800	0.921	0.935	0.936	**0.940**
900	0.924	0.935	0.935	**0.940**
1000	0.921	0.933	0.936	**0.939**

6.3 Discussion

Figure 3 shows that the adaptation methods provided consistently superior precision in comparison to the standard approach. The improvements were statistically significant in 20 out of 30 cases (see Table 1 for details). Hence it can be assumed that the proposed methods are able to adapt to handwritten digits and commas of a new handwriting in order to improve their performance. In summary, all techniques provide considerable improvements to the standard classifier by about the same extend and are recommended for further investigation.

7 Future Work

In this work, we applied a special subspace embedding technique, different methods of dimensionality reduction like Isomap projections could be considered for a dispersion of different symbols in the subspace. Future work can further investigate other approaches to weight the certainties of the initial recognition, or the distance functions used. For the Best-First methods one important parameter is the optimal threshold to separate certain samples from uncertain ones. The Best-First-Two approach could possibly be further enhanced by more precise weighting between the two classifiers, and incorporating more complex features, confer, for example, [Leibfried, 2012].

8 Summary

We proposed three approaches for handwriting adaptation of digit recognition: one approach that applies subspace embedding and two approaches that are similar to bootstrapping. Our experiments showed that adaptive techniques can enhance the precision of the standard classifier significantly which emphasizes the importance of adaptation in this domain. The overall precision scores of the three approaches were comparable. Their notable improvements over an already good baseline ranged from about one to four percent.

Acknowledgments

We thank Alexander Hörnlein and all participants for their help with the data set.

References

[Chapelle *et al.*, 2002] Olivier Chapelle, Jason Weston, and Bernhard Schölkopf. Cluster kernels for semi-supervised learning. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 585–592. MIT Press, 2002.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[DeCoste and Schölkopf, 2002] Dennis DeCoste and Bernhard Schölkopf. Training invariant support vector machines. *Machine Learning*, 46(1-3):161–190, 2002.

[Jolliffe, 1986] Ian T. Jolliffe. *Principal component analysis*, volume 487. Springer-Verlag New York, 1986.

[Klügl *et al.*, 2012] Peter Klügl, Martin Toepfer, Florian Lemmerich, Andreas Hotho, and Frank Puppe. Collective information extraction with context-specific consistencies. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *ECML/PKDD (1)*, volume 7523 of *Lecture Notes in Computer Science*, pages 728–743. Springer, 2012.

[Leggetter and Woodland, 1995] Chris J. Leggetter and Philip C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech and language*, 9(2):171, 1995.

[Leibfried, 2012] Felix Leibfried. Recognition of handwritten digits. Diplomarbeit, University of Würzburg, Würzburg, September 2012.

[Mandel *et al.*, 2011] Alexander Mandel, Alexander Hörnlein, Marianus Ifland, Edeltraud Lüneburg, Jürgen Deckert, and Frank Puppe. Cost analysis for computer supported multiple-choice paper examinations. *GMS Z Med Ausbild*, 28(4):Doc55, 2011.

[Seewald, 2012] Alexander K. Seewald. On the brittleness of handwritten digit recognition models. *ISRN Machine Vision*, 2012:10, 2012.

[Szummer and Bishop, 2006] Martin Szummer and Christopher M. Bishop. Discriminative writer adaptation. In *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2006.

[Yarowsky, 1995] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Hans Uszkoreit, editor, *ACL*, pages 189–196. Morgan Kaufmann Publishers / ACL, 1995.

Workshop "Knowledge and Experience Management" (FGWM-2013)

The annual workshop "Knowledge and Experience Management" is organized by the Special Interest Group on Knowledge Management of the German Informatics society (GI), that aims at enabling and promoting the exchange of innovative ideas and practical applications in the field of knowledge and experience management.
All submissions of current research from this and adjacent areas are welcome, in particular, work in progress contributions. The latter can serve as a basis for interesting discussions among the participants and provide young researchers with feedback. We also invite researchers to contribute to the workshop by resubmitting conference papers and share their ideas with the research community. Topics of Interest
Submissions from all areas contributing to the development and application of intelligent knowledge and experience management systems are welcome. We explicitly encourage paper submissions which are not mainstream but from communities within mathematics, social sciences or economics in order to obtain a more interdisciplinary view on the subject.

Topics of Interest

Topics of interest include but are not limited to:

- Experience & knowledge search and knowledge integration approaches: case-based reasoning, logic-based approaches, text-based approaches, semantic portals/wikis/blogs, Web 2.0, etc.
- Applications of knowledge and experience management (corporate memories, e-commerce, design, tutoring/e-learning, e-government, software engineering, robotics, medicine, etc.)
- Big Data and Knowledge Management (KM)
- (Semantic) Web Services for KM
- Agile approaches within the KM domain
- Agent-based & Peer-to-Peer KM
- Just-in-time retrieval and just-in-time knowledge capturing
- Knowledge representation (ontologies, similarity, retrieval, adaptive knowledge, etc.)
- Support of authoring and maintenance processes (change management, requirements tracing, (distributed) version control, etc.)
- Evaluation of KM systems
- Practical experiences ("lessons learned") with IT aided KM approaches
- Integration of KM and business processes
- Introspection and explanation capabilities of KM systems
- Application of Linked Data
- Combination of KM with other systems and concepts (e.g. Decision Support, Business Intelligence, etc.)

Program Chairs

- Dr. Andrea Kohlhase
- Prof. Dr.-Ing. Bodo Rieger

Program Committee

- Klaus-Dieter Althoff, DFKI Kaiserslautern / University of Hildesheim
- Kerstin Bach, Verdande Technology As (Trondheim)
- Joachim Baumeister, denkbares GmbH (Würzburg)
- Axel Benjamins, University of Osnabrück
- Mareike Dornhöfer, University of Siegen
- Norbert Gronau, University of Potsdam
- Constantin Jucovschi, Jacobs University Bremen
- Dimitris Karagiannis, University of Vienna
- Christoph Lange, University of Birmingham
- Ronald Maier, University of Innsbruck
- Mirjam Minor, University of Trier
- Ulrich Reimer, University of Applied Sciences St. Gallen
- Jochen Reutelshöfer, University of Würzburg
- Thomas Roth-Berghofer, University of West London
- Sonja Schulze, University of Osnabrück

Synthese aus Prozessmodellierungswerkzeug und Semantic Wiki: Nutzensystematisierung und Forschungsagenda

Michael Fellmann

Universität Osnabrück,
Institut für Informationsmanagement und Unternehmensführung,
Katharinenstraße 3, 49069 Osnabrück
Michael.Fellmann@uni-osnabrueck.de

Abstract

Die semiformale Modellierung von Geschäftsprozessen ist in Wissenschaft und Praxis verbreitet, ebenso wie die kollaborative Repräsentation und Kommunikation von Wissen mit (semantischen) Wikis. Einer Verbindung beider Bereiche auf Werkzeugebene wurde bisher wenig Beachtung geschenkt. Im vorliegenden Beitrag wird ein erster Schritt in Richtung dieser Synthese vollzogen, indem die Nutzenpotenziale systematisiert werden und eine Forschungsagenda vorgestellt wie auch ein Überblick über den relevanten Stand der Forschung gegeben wird.

1 Einleitung

Als „Synthese“ wird – der Bedeutung aus dem altgriechischen „σύνθεσις“ bzw. „sýnthesis“ folgend – die Zusammenfassung oder Verknüpfung von zwei oder mehr Elementen zu einer neuen Einheit verstanden. In diesem Beitrag wird die Kombination eines Prozessmodellierungswerkzeugs mit einem Semantic Wiki untersucht, die zu einem integrierten System zusammengefasst werden, das als *Plattform zur semantischen Prozessmodellierung* bezeichnet wird. Diese Plattform erlaubt die Modellierung von Geschäftsprozessmodellen. Im Unterschied zu herkömmlichen Prozessmodellierungswerkzeugen wird jedoch die Bedeutung der Modellelemente durch die Korrespondenz eines jeden Modellelements mit einer Seite in einem Semantic Wiki präzisiert. Das Wiki stellt hierzu vielfältige Möglichkeiten zur Diskussion und Versionierung bereit. Eine Aktivität „Prüfe Rechnung“ kann im Wiki etwa hinsichtlich der beteiligten Rollen und der verwendeten Ressourcen und Systeme beschrieben werden. Die mit Semantic Wikis umsetzbaren, erweiterten Strukturierungsmöglichkeiten und die darauf aufbauenden Suchfunktionen [Schaffert et al., 2009] gehen dabei wesentlich über die bisher vereinzelt in Prozessmodellierungswerkzeugen anzutreffenden Funktionalitäten zur Glossarverwaltung hinaus. Hierin liegt eine besondere Stärke der hier vorgestellten Synthese: Sie erlaubt eine Verknüpfung der ablauforientierten Sicht, die durch Geschäftsprozessmodelle repräsentiert wird, mit Domänenwissen, das in Form von Wikiseiten repräsentiert wird. Die Verknüpfung wird dadurch erreicht, dass Elementen der Prozessmodelle Wikiseiten hinterlegt werden. Die Verwendung eines Semantic Wikis erlaubt dabei eine formale Wissensrepräsentation. Diese ermöglicht es, Beziehungen zwischen Konzepten oder Konzept-Hierarchien (im Wiki-Bereich auch „Kategorien“ genannt) in einer maschinell verarbeitbaren Form zu speichern. Werden nun diese im Wiki definierten Konzepte zur Beschreibung der Prozesse verwendet, so kann hierdurch eine über syntaktische Prüfungen hinausgehende, inhaltliche Korrektheitsprüfung maschinell durchgeführt werden. Mit dieser Prüfung kann etwa festgestellt werden, ob die Geschäftsprozesse eines Unternehmens vorgegebenen Regeln und Richtlinien entsprechen.

Mit einer Reporting-Komponente kann darüber hinaus das im Wiki enthaltene Wissen genutzt werden, um fachliche Fragen zu beantworten. Eine von der Organisation der Daten unabhängige Beantwortung von Fragen kann hierbei durch die Verwendung einer im Vergleich zu relationalen Datenmodellen einfachen Tripelstruktur erreicht werden. Diese Struktur ist die Grundlage des sog. *Semantic Web* und wird von der Komponente *Semantic Kernel* der Plattform zur semantischen Prozessmodellierung als Grundlage der Speicherung von Prozessmodellen verwendet. Die besondere Stärke dieser Struktur liegt in einer Abfrage beliebiger Zusammenhänge und in der Möglichkeit, durch maschinelle Schlussfolgerungen neue Aussagen zu gewinnen (implizites Wissen), ohne dass hierfür Programmieraufwand erforderlich ist. So kann bspw. vorhergesagt werden, welche Prozesse beim Ausscheiden eines Mitarbeiters oder einer Gesetzesänderung angepasst werden müssen.

Abb. 1 zeigt die Struktur der Plattform zur semantischen Prozessmodellierung.

Die zu lösende Problemstellung besteht aus wissenschaftlicher und technischer Sicht darin, die verschiedenen Technologien zu einem Prototyp zusammenzufassen, mit dem der Nutzen der Synthese aus Modellierungswerkzeug und Wiki empirisch überprüft werden kann. Insbesondere folgende Aufgaben sind hierbei zu lösen:

- Erweiterung eines Prozessmodellierungswerkzeugs um Komponenten, die eine Verknüpfung von Prozessmodellelementen mit Ontologiebegriffen sowie eine semantische Korrektheitsprüfung erlauben.
- Schaffung einer Serverkomponente, die sowohl Funktionalitäten für das Prozessmodellierungswerkzeug bereitstellt als auch eine Schnittstelle zu einem Semantic Wiki aufweist.
- Ermittlung geeigneter Hilfestellungen und Assistenten zur Erstellung von Reports, die das durch die

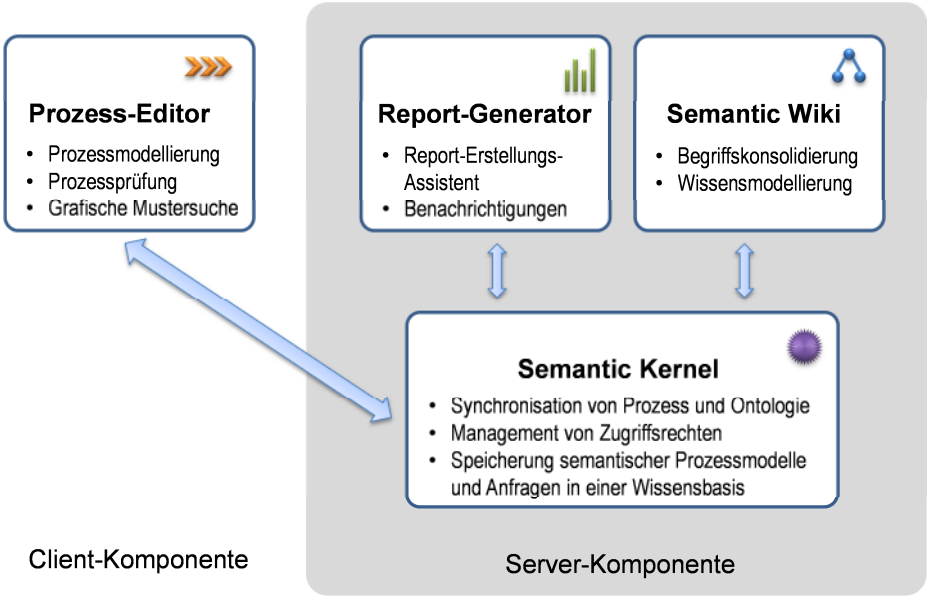


Abb. 1: Plattform zur semantischen Prozessmodellierung

Verknüpfung mit dem Semantic Wiki zur Verfügung stehende Wissen nutzen. Es sind unterschiedliche Ansätze wie Formulare, stichwortbasierte Anfragesprachen, und strukturierte Anfragen hinsichtlich ihrer Benutzbarkeit und Nutzerakzeptanz vergleichend zu evaluieren.

Im Rahmen dieses Beitrags wird ein erster Schritt zur Erreichung der angestrebten Synthese aus Prozessmodellierungswerkzeug und Semantic Wiki unternommen, indem zunächst der Nutzen dieser Verbindung systematisiert und charakterisiert wird. Anschließend wird eine Forschungsagenda vorgestellt, anhand derer das Werkzeug entwickelt werden soll.

2 Nutzentypen der Synthese

Abb. 2 zeigt die fünf wesentlichen Nutzenpotenziale der Synthese aus Prozessmodellierungswerkzeug und Semantic Wiki gegenüber den aktuell angebotenen Werkzeugen. Die Nutzenpotenziale (1-5) werden im Folgenden beschrieben und durch eine detailliertere Darstellung der Nutzenaspekte (N) konkretisiert.

2.1 Semantisch eindeutig definierte Prozesse

Ein Vorteil des hier beschriebenen Konzepts gegenüber existierenden Lösungen ist es, die Stärken eines Prozessmodellierungswerkzeugs, die in den Funktionalitäten zur Konstruktion und zum Layout von Modellen liegen, mit den Stärken eines Semantic Wikis zur Unterstützung von Begriffsdiskussionen und der Wissens-Externalisierung zu kombinieren. Durch eine Verknüpfung der Prozessmodellelemente mit Konzepten aus dem Semantic Wiki in Form einer sog. *semantischen Annotation* kann so eine semantisch eindeutige und für alle Akteure verständliche Beschreibung der Prozesse erreicht werden, deren Widerspruchsfreiheit zudem durch die integrierte formale Wissensrepräsentation sichergestellt wird. Diese Beschreibung begegnet dem in Praxisprojekten auftretenden Problem, dass die Festlegung gemeinsamer Begriffe eines der größten Hindernisse beim Einsatz von Informationsmodellen ist [Sarshar et al., 2006, S. 125], die zudem zunehmend arbeitsteilig erstellt werden [vom Brocke, 2003, S. 163 ff.; Thomas, 2006, S. 366 ff.].

Semantisch eindeutig definierte Prozesse sind nicht nur während der Modellierung, sondern auch in den nachgelagerten Phasen etwa bei der Konstruktion von Informationssystemen vorteilhaft. So wird die Entwicklung von Informationssystemen nach dem in der Literatur und Praxis etablierten Paradigma der Serviceorientierten Architektur in mehr als 50 % der Unternehmen durch 10 Teams gleichzeitig betrieben [ebizQ, 2008, S. 1]. Die durch die Synthese erzielbaren Verbesserungen in der Kommunikation zwischen Modellkonstrukteuren und Modellanwendern können einen wesentlichen Beitrag dazu leisten, die Qualität und Passgenauigkeit der entwickelten Systeme zu steigern. Die in den Kommunikationsprozessen erzielten Ergebnisse werden im Semantic Wiki gespeichert und in den Prozessmodellen referenziert, sodass eine Wiederverwendung von konsensualem Wissen erreicht wird. Dessen Bedeutung wird auch dadurch ersichtlich, dass sich besonders erfolgreiche Unternehmen durch eine gemeinsame Sprache auszeichnen (engl.: shared language community) [Rosenkranz, 2009 S. 206]. Über Verlinkungsmechanismen ermöglicht die Plattform zur semantischen Prozessmodellierung jederzeit, zwischen Prozessmodell und Wiki zu wechseln, sodass auch neue Mitarbeiter oder diejenigen, die nicht aktiv an der Modellkonstruktion beteiligt waren, sich in kurzer Zeit in die Prozesse einarbeiten können. Durch einen einfachen Zugriff per Webbrowser und eine nutzerfreundliche Oberfläche der Wiki-Komponente können die in die Prozesse involvierten Mitarbeiter die Beschreibung der Prozessschritte zudem selbstständig und unabhängig von dem Prozessmodell aktuell halten. Ebenso können hierbei benötigte Fachkenntnisse spezifiziert werden, die gegenwärtig nur in 3 % der Unternehmen für alle Aufgaben der Kerngeschäftsprozesse klar definiert sind [Wolf und Harmon, 2010, S. 21]. Die mit der in diesem Beitrag beschriebenen Synthese mögliche, verbesserte Dokumentation wirkt diesem Mangel entgegen. Darüber hinaus ist es mit der Wiki-Komponente möglich, die in den Anfangsphasen von Strategie- oder Reorganisationsprojekten erforderli-

chen Einigungsprozesse auf zentrale Konzepte und Begriffe durch die Bestückung des Wikis mit vordefinierten Begrifflichkeiten aus Referenz-Ontologien zu beschleunigen. Eine mit der Kombination aus Prozessmodellierungswerkzeug und Wiki realisierbare standardisierte, an internationalen Standards wie dem Process Classification Framework (PCF) (www.apqc.org/pcf) ausgerichtete Prozessbeschreibung erlaubt zudem eine größere semantische Kompatibilität der Prozesse.

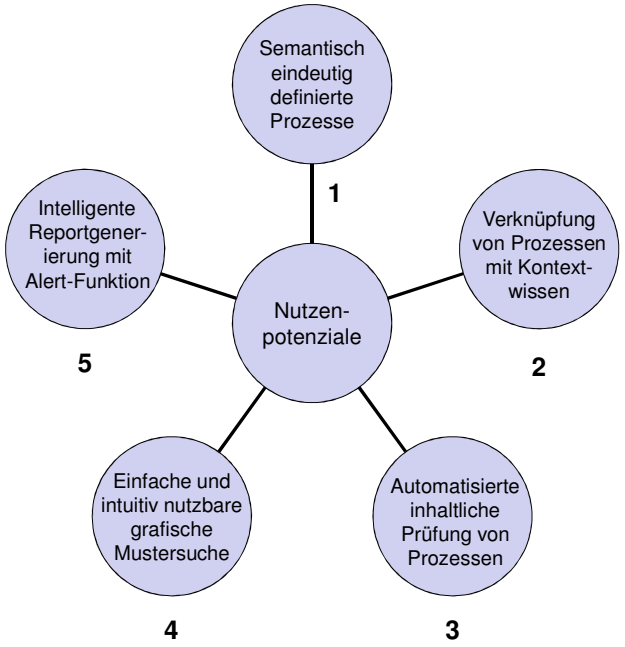


Abb. 2: Nutzenpotenziale der Synthese aus Prozessmodellierungswerkzeug und Wiki

Diese senkt insbesondere bei unternehmensübergreifenden oder internationalen Kooperationen den oft enormen Aufwand, der bei der Integration semantisch heterogener Prozesse und Datenstrukturen entsteht. Da gut 50 % des Aufwandes für IT-Projekte auch heute noch auf die Integration entfallen [Brodie et al., 2005, S. 98], ist eine Verbesserung in diesem Bereich besonders wirkungsvoll.

- N1.1 Wiederverwendung konsensualen Wissens zur Prozessmodellierung
- N1.2 Schnellere Einarbeitung neuer Mitarbeiter in die Prozesse
- N1.3 Beschleunigung von Einigungsprozessen durch Referenzontologien
- N1.4 Reduktion des Aufwandes zur Prozessintegration durch standardisierte Semantiken

2.2 Verknüpfung von Prozessen mit Kontextwissen

Bestehende Ansätze zur multiperspektivischen Modellierung erlauben es teilweise, Elemente eines Prozesses mit Elementen weiterer Modelle zu verbinden. Allerdings wird hierbei die Ebene der Modellierung nicht verlassen. Eine Verknüpfung von Prozessen mit dem weiteren organisationalen Umfeld, in dem diese ablaufen, wird von bestehenden Werkzeugen kaum realisiert, jedoch für die Zukunft gefordert [Recker et al., 2009, S. 344]. Genau dies ermöglicht die Plattform zur semantischen Prozessmodellierung, indem über eine Wiki-Komponente die Verknüpfung von Prozessen mit externen Inhalten erreicht wird. Beispiele hierfür sind etwa die Dokumentation der IT-Systeme, deren Verknüpfung mit Prozessen das Business/IT-Alignment verbessern kann oder die Verknüpfung mit den Unternehmenszielen, die zu einem verbesserten Strategic Alignment führen kann. Die mit dem Business/IT-Alignment verfolgte einheitliche Sicht auf IT-Systeme und Prozesse wird in der Praxis bisher nur unzureichend umgesetzt [Softlab Group, 2007, S. 1]. Darauf deutet auch die geringe Zahl von nur 21 % der in einer Studie befragten Unternehmen hin, die angibt, dass die eingesetzte Software ihre Geschäftsprozesse meistens oder immer unterstützt [Wolf und Harmon, 2010, S. 20]. Durch die Verknüpfung der verwendeten IT-Systeme mit Definitionen von Prozessaktionen im Semantic Wiki und die Verwendung dieser Definitionen zur Prozessmodellierung kann genau verfolgt werden, welche IT-Systeme welche Prozesse unterstützen. Analog kann das Strategic Alignment verbessert werden, indem Prozessaktionen mit den ausführenden Organisationseinheiten und betrieblichen Zielen verknüpft werden. Durch die Verbindung von Prozess-Editor und Semantic Wiki wird darüber hinaus eine Parallelisierung der Erarbeitung von Prozess- und Kontextwissen ermöglicht. Während die Fachabteilungen sich auf die Erarbeitung von Definitionen der Kontextobjekte konzentrieren können, sind die noch in der Diskussion befindlichen betrieblichen Objekte über eindeutige Kennungen bereits von Prozessanalysten oder der IT-Abteilung zur Prozessbeschreibung verwendbar.

- N2.1 Verbesserung des Business/IT-Alignment
- N2.2 Verbesserung des Strategic Alignment
- N2.3 Parallelisierung der Erarbeitung von Prozess- und Kontextwissen

2.3 Automatisierte inhaltliche Prüfung von Prozessen

Durch die eindeutige Spezifikation von Prozessen in Verbindung mit den formalen, im Semantic Wiki hinterlegten Wissensstrukturen wird eine automatisierte inhaltliche Prüfung der Prozesse ermöglicht. Hiermit geht die Plattform erheblich über die bei aktuellen Werkzeugen anzutreffenden syntaktisch orientierten Prüffunktionen hinaus. So könnte etwa im Bereich des E-Government eine Bedingung lauten, dass in einem Prozess, in dessen Verlauf eine gebührenfreie Bearbeitung entschieden wurde, keine Gebührenberechnung erfolgen darf. Durch die Verwendung der im Semantic Wiki sowohl natürlichsprachlich als auch formal definierten Begriffe zur Prozessmodellierung ist nun automatisiert überprüfbar, ob diese Restriktion eingehalten wird. Der Nutzen liegt somit in Modellen, die weniger Fehler enthalten und die darüber hinaus nachweisbar bestimmte Richtlinien und Gesetze einhalten. Dies ist zum einen bei hohen Fehler-Folgekosten relevant. Zum andern kommt zum Tragen, dass die Kosten der Fehlerverhütung bzw. -erkennung exponentiell auf jeder Stufe im Wertschöpfungsprozess steigen [Pfeifer, 1996, S. 11]. Weiter können Kosteneinsparungen erzielt werden, wenn bei häufigen Änderungen z.B. an Detailmodellen eine wiederholte manuelle Kontrolle des Gesamtmodells durch einen automatisierten Prüfvorgang ersetzt wird.

Die zu inhaltlichen Prüfungen erforderlichen, formal spezifizierten semantischen Korrektheitsbedingungen können hierbei angepasst an den Verwendungszweck und die Kompetenz der Mitarbeiter in mehreren Stufen spezi-

fiziert werden. Neben fest in die Plattform eingebauten, domänenspezifisch vordefinierten Korrektheitsbedingungen können über die im Prozess-Editor realisierte Mustersuche durch die Konstruktion einfacher Modelle ebenfalls Korrektheitsbedingungen erfasst werden. Komplexe Korrektheitsbedingungen können mittels strukturierter Anfragen mit der leicht erlernbaren [Fellmann und Thomas, 2011] und vom W3C normierten Anfragesprache SPARQL innerhalb der Reporting-Komponente spezifiziert werden.

- N3.1 Automatisierte inhaltliche Prüfung von Prozessmodellen
- N3.2 Variable und Nutzer-angepasste Spezifikation von Korrektheitsbedingungen

2.4 Einfache und intuitiv nutzbare grafische Mustersuche

Grundprinzip der Mustersuche ist es, Modelle als strukturierte Anfragen zu interpretieren. Somit kann das vorhandene Wissen zur Modellierung auch zur Suche im Modellbestand verwendet werden und damit ein weiteres Mal zur Anwendung kommen. Wird also beispielsweise ein Prozessmodell mit zwei durch eine Flussbeziehung verbundenen Funktionen konstruiert, so kann mit diesem Modell im gesamten Modellbestand nach Modellen gesucht werden, die dieses Muster enthalten. Durch Verfeinerungen wie Platzhalterzeichen und variable Pfadlängen sowie durch die Nutzung maschineller Schlussfolgerungen kann mit kompakten Modellen eine hohe Ausdrucksstärke der resultierenden Anfragen erzielt werden.

Ein dreifacher Nutzen der so erzeugten strukturierten Anfragen besteht in (a) der Suche nach Mustern oder Modellfragmenten, (b) der Verwendung der strukturierten Anfragen als Korrektheitsprüfungen sowie (c) der Verwendung der Anfragen zur Erstellung von Reports. Während die bisher in diesem Bereich entwickelten Anfragesprachen wie BPMN-Q [Awad et al., 2008] jeweils für genau eine Prozessmodellierungssprache entworfen wurden, wird mit der Plattform eine allgemeine Lösung angestrebt. Diese wird dadurch erreicht, dass die Prozessmodelle in einer verallgemeinerten Form (d.h. unabhängig von einer konkreten Sprache) im Semantic Kernel gespeichert werden.

- N4.1 Verwendung grafischer Modelle zur Spezifikation von Anfragen
- N4.2 Mehrfachnutzung von Anfragen zur Mustersuche, als Korrektheitsbedingung und in Reports

2.5 Intelligente Reporting-Komponente mit Alert-Funktion

Die Reporting-Komponente erlaubt eine variable und Nutzer-angepasste Spezifikation von Reports über verschiedene Assistenzsysteme. Ein wesentlicher Vorteil gegenüber bestehenden Werkzeugen ist, dass Reports ohne Programmierung erstellt werden können. Zur Report-Erstellung wird u.a. auch auf die Anfrage-Bausteine der grafischen Mustersuche zurückgegriffen sowie die Verwendung der vom W3C normierten Anfragesprache SPARQL.

Ein weiterer innovativer Aspekt ist die Nutzung von maschinellen Schlussfolgerungen, die durch eine Verknüpfung der Prozessmodelle mit den formalen Wissensstrukturen des Semantic Wikis ermöglicht wird. Die Reports können somit Fakten enthalten, die nicht explizit in den Modellen enthalten, aber aus ihnen logisch ableitbar sind. Ein Beispiel hierfür wäre etwa die Ausgabe von in den Modellen enthaltenen Funktionen, die den Lagerbestand reduzieren. Verbraucht eine Funktion Ressourcen, die in einem Lager bevorratet werden, so kann gefolgert werden, dass diese Funktion den Lagerbestand reduziert. Ebenso können in sehr einfacher Weise Abhängigkeiten und Zusammenhänge zwischen Funktionsbereichen eines Unternehmens aufgedeckt werden (bspw. „50 % aller Prozesse nutzen die Rechtsabteilung“), die ggf. zur internen Leistungsverrechnung herangezogen werden können. Weiter sind auch Prozess-Metriken ermittelbar, etwa wie häufig Abteilungssprünge auftreten oder wie hoch die Anzahl beteiligter Dokumente an einem Prozess ist.

Darüber hinaus wird der Nutzen der Report-Komponente durch eine Benachrichtigungsfunktion bei kritischen Ereignissen weiter gesteigert (Alert-Funktion). So kann die Report-Komponente proaktiv Nachrichten verschicken, wenn für einen bestimmten, als kritisch definierten Report neue Ergebnisse vorliegen. Ein solcher

Tabelle 1: Zusammenfassende Darstellung und Charakterisierung der Nutzentypen

		Nutzentyp-Charakterisierung						
ID	Nutzentyp	Zeit	Kosten	Güte	Qualitativ	Quantitativ	Kurzfristig	Langfristig
N 1.1	Wiederverwendung konsensualen Wissens zur Prozessmodellierung	x	x		x			x
N 1.2	Schnellere Einarbeitung neuer Mitarbeiter in die Prozesse	x				x	x	
N 1.3	Beschleunigung von Einigungsprozessen durch Referenzontologien	x	x			x	x	
N 1.4	Reduktion des Aufwandes zur Prozessintegration durch standardisierte Semantiken	x	x			x	x	
N 2.1	Verbesserung des Business/IT-Alignment			x	x		x	
N 2.2	Verbesserung des Strategic Alignment			x	x			x
N 2.3	Parallelisierung der Erarbeitung von Prozess- und Kontextwissen	x				x	x	
N 3.1	Automatisierte inhaltliche Prüfung von Prozessmodellen	x	x			x	x	
N 3.2	Variable und Nutzer-angepasste Spezifikation von Korrektheitsbedingungen			x		x	x	
N 4.1	Verwendung grafischer Modelle zur Spezifikation von Anfragen	x	x			x	x	
N 4.2	Mehrfachnutzung von Anfragen zur Mustersuche, als Korrektheitsbedingung und in Reports	x	x			x	x	
N 5.1	Unterstützung der Report-Generierung über verschiedene Assistenzsysteme	x	x			x	x	
N 5.2	Vollständigere Ergebnisse durch die Nutzung maschineller Schlussfolgerungen			x		x	x	
N 5.3	Proaktive Benachrichtigungsfunktion bei Änderungen			x	x		x	

Report könnte bspw. in der Ausgabe aller Server bestehen, für die kein zugreifender Prozess oder keine verantwortliche Organisationseinheit (mehr) bekannt ist.

- N5.1 Unterstützung der Report-Generierung über verschiedene Assistenzsysteme
- N5.2 Vollständigere Ergebnisse durch die Nutzung maschineller Schlussfolgerungen
- N5.3 Proaktive Benachrichtigungsfunktion bei Änderungen

3 Integration der Nutzentypen

Tabelle 1 zeigt die verschiedenen, bisher identifizierten Nutzentypen in einer Gesamtschau. Jeder Nutzentyp wird zusätzlich dadurch charakterisiert, ob die Zeit, Kosten oder Güte durch den Nutzen beeinflusst werden sowie ob diese Beeinflussung qualitativ oder quantitativ messbar ist und eher langfristig nach dem Einsatz der Plattform zu erwarten ist oder unmittelbar beim Einsatz. Besonders die Nutzenbereiche N1 und N4 erscheinen sehr relevant, da diese sowohl die Zeit als auch die Kosten betreffen, kurzfristig realisierbar sind und zu einer quantitativen Beurteilung geeignet.

4 Forschungsagenda und Prototypenbau

Das vorgestellte Konzept einer Synthese aus Prozessmodellierungswerkzeug und Semantic Wiki ist nicht ausschließlich theoriegeleitet durch deduktive Argumentationen evaluierbar. Vielmehr bedarf es zur empirischen Fundierung und Überprüfung der Nutzenbetrachtung einer praktischen Anwendung der Plattform in einem konkreten Szenario, die nur durch eine prototypische Realisierung erreicht werden kann. Die zur Erstellung eines Prototyps relevanten Schritte nebst einigen Forschungsfragen werden im Folgenden im Sinne eines Ausblicks auf die zukünftigen Arbeiten des Autors skizziert.

4.1 Analyse der Anforderungen an die Synthese

Anhand von Umfragen und Experteninterviews muss die Relevanz der einzelnen funktionalen Merkmale der Plattform festgestellt werden, sodass die Erforschung und prototypische Entwicklung der Teilkomponenten entsprechend priorisiert werden kann.

4.2 Gestaltung des Prozess-Editors

Es muss erforscht werden, wie eine Erweiterung eines bestehenden Modellierungswerkzeugs realisiert werden kann, die die beschriebene Synthese umsetzt. Insbesondere muss die Annotation mit Begriffen aus einer formalen Ontologie derart gestaltet und erforscht werden, dass der Prozess-Editor dem Modellkonstrukteur unterstützend passende Begriffe vorschlägt und somit der Aufwand zur Annotation nachweisbar gering gehalten werden kann.

4.3 Entwicklung des Semantic Kernel

Die prototypische Realisierung muss eine Software-Komponente umfassen, die eine interne Wissensbasis beinhaltet. Zur Strukturierung dieser Wissensbasis können sog. *Upper*-Ontologien herangezogen werden, die bereichsübergreifendes Wissen beinhalten, sodass die konkreten Wissensbasen zueinander kompatibel bleiben. In die Wissensbasis werden die vom Prozess-Editor erzeugten Modelle importiert. Der Semantic Kernel stellt Schnittstellen bereit, über die Modelle eingespeist, aktualisiert oder gelöscht werden können. Über weitere Schnittstellen ist es für den Prozess-Editor möglich, Informationen über die in der Wissensbasis befindlichen Modelle abzurufen bzw. noch in der Konstruktion befindliche Modelle einer Vorab-Prüfung vor der Einspeisung in die Wissensbasis zu unterziehen. Ein Forschungsaspekt ist hierbei, wie der Semantic Kernel eine Synchronisation der mit dem Prozess-Editor erstellten Modelle mit der im Semantic Wiki vorhandenen Ontologie leisten kann. Insbesondere ist zu erforschen, wie Änderungen an Prozessmodellen, die außerhalb der Plattform gespeichert werden, und den mit ihnen korrespondierenden semantischen Repräsentation in der Plattform synchronisiert werden können.

4.4 Integration des Semantic Wiki und Nutzertest

Es muss erforscht werden, wie ein Wiki zur Pflege und Weiterentwicklung der formal definierten Begriffe in das Gesamtsystem integriert werden kann. Eine prototypische Implementierung eines an den Semantic Kernel angebundenen Wikis, das zudem mit einer Ontologie ausgestattet ist, kann die Machbarkeit der Plattform zeigen und erlaubt eine Evaluation der Integration von Domänenwissen (Konzepte und Relationen im Semantic Wiki) und Prozesswissen (Prozessmodelle im Prozess-Editor) anhand praktischer Anwendungsbeispielen.

4.5 Entwicklung des Report-Generators

Es muss erforscht werden, wie ein Report-Generator im Hinblick auf eine flexible und schnelle (ohne Programmierung umsetzbare) Erstellung von Berichten auf der Basis der im Semantic Kernel gespeicherten Prozessmodelle zu gestalten ist. Grundlage für die Reports sind strukturierte Anfragen an die im Semantic Kernel enthaltene Wissensbasis, die unter Nutzung der im Semantic Kernel enthaltenen Inferenzmaschine beantwortet werden. Für die Erfassung von Anfragen durch Nutzer sind verschiedene Szenarien und Ansätze zu erforschen wie (a) eine Selektion und (Re-)Kombination einer Anfrage aus vorgefertigte Anfragen, (b) die Erstellung von Anfragen durch Modellfragmente im Prozess-Editor oder (c) die Erstellung von Anfragen über einen formularbasierten Assistenten, (d) mittels einer grafischen Anfragesprache oder (e) mittels einer Anfragesprache wie SPARQL.

4.6 Abschließende Evaluation

Das in den vorherigen Forschungsarbeiten entwickelte Gesamtsystem zur semantischen Prozessmodellierung wird einem Nutzertest unterzogen. Hierbei eventuell aufgedeckte Mängel werden dokumentiert und durch eine Anpassung des Konzepts mit erneutem Durchlauf der Evaluation behoben.

5 Stand des Wissens und verwandte Arbeiten

Semantik im Kontext von Modellierungssprachen. Untersuchungen zur Semantik von Modellierungssprachen haben sich bislang hauptsächlich auf die *formale* Semantik der zur Verfügung stehenden Sprachelemente – im Folgenden auch Sprachkonstrukte genannt – konzentriert. Die formale Semantik ist u. a. in der Theoretischen Informatik und der Logik verankert und beschäftigt sich mit der exakten Bedeutung künstlicher (d. h. konstruierter)

oder natürlicher Sprachen. Einen zentralen Stellenwert bei der Untersuchung der formalen Semantik von Modellierungssprachen nehmen mathematische Methoden ein (für die Sprache der Ereignisgesteuerten Prozesskette EPK vgl. stellvertretend Kindler 2006 und die dort zitierte Literatur). Arbeiten zur formalen Semantik im Bereich der Prozessmodellierung betreffen zumeist dynamische Aspekte (die sog. Ausführungssemantik von Modellen) und zielen auf die Untersuchung und Vermeidung bestimmter Anomalien wie etwa Verklemmungen (Deadlocks) ab [van der Aalst, 1999; Dijkman et al., 2007; Mendling, 2009, S. 7]. Die Semantik, die Modellelementen in Form von Modellelementbezeichnern hinzugefügt wird und gerade bei semiformalen Sprachen an die natürliche Sprache gebunden ist, wird in den Arbeiten nicht berücksichtigt.

Ein weiterer Bezugspunkt semantischer Analysen ist die Untersuchung der Bedeutung der Sprachkonstrukte von Modellierungssprachen [Green, 1996]. In diesem Kontext ist vor allem das Bunge-Wand-Weber-Modell hervorzuheben [Wand und Weber, 1995], das – vereinfacht gesprochen – als ein Ansatz zur Beschreibung von Informationssystemen verstanden werden kann. Die hiermit verbundene Beurteilung der Eignung und der Ansprüche von Informationsmodellen und den zu ihrer Konstruktion verwendeten Modellierungssprachen sowie die Herleitung von Kriterien zur Bestimmung der Güte der Artefakte ist nicht Gegenstand des in diesem Beitrag beschriebenen Konzepts.

Fachbegriffsmodelle zur Vereinheitlichung von Bezeichnungen. Zur Vereinheitlichung der in Modellen und Modellelementen verwendeten Terminologie können Fachbegriffsmodelle eingesetzt werden [Ortner, 1997; Rosemann und Schwegmann, 2002]. Die Verwendung von Ontologien zur formalen Repräsentation einer Domäne besitzt jedoch gegenüber den Fachbegriffsmodellen den Vorteil einer maschinellen Interpretation. Hierdurch können insb. nicht explizit repräsentierte Fakten durch Verfahren des maschinellen Schließens automatisiert ergänzt werden, um eine vollständige Interpretation der Semantik bspw. bei der Suche in oder Korrektheitsprüfung von Modellen zu ermöglichen.

Ontologien zur Formalisierung von Semantik. Die formale Repräsentation von Wissen wird im Forschungsgebiet der Künstlichen Intelligenz vorangetrieben. Zur Darstellung der komplexen Wissensbeziehungen werden hierbei häufig Ontologien verwendet, die aktuell durch die Bestrebungen, das World Wide Web zu einem Semantic Web zu erweitern [Fensel et al., 2003], an Bedeutung gewinnen. Unter einer Ontologie wird in der hier verwendeten informatiknahen Interpretation nach Gruber [1993, S. 199] eine explizite formale Spezifikation einer Konzeptualisierung (engl.: conceptualization) verstanden. Eine Konzeptualisierung ist dabei eine abstrakte, vereinfachte Sicht der Welt, die für bestimmte Zwecke repräsentiert werden soll. Neuere Definitionsversuche betonen darüber hinaus die intersubjektive Gültigkeit der Konzeptualisierung, sodass Ontologien auch als ein von mehreren Individuen entwickeltes Vokabular aufgefasst werden können, das gemeinsam in einer Gruppe akzeptiert und genutzt wird (engl.: shared conceptualization) [Studer et al., 1998, S. 186; Gómez-Pérez al., 2004, S. 8]. Die Wirtschaftsinformatik macht sich die zum Teil umfangreichen Vorarbeiten der Ontologieforschung zunutze. In diesen Untersuchungen werden struktural organisierte Artefakte (z.B. Dokumente oder Produktmodelle) durch Ontologien repräsentiert. Das in diesem Beitrag beschriebene Konzept baut auf diesen Überlegungen teilweise auf, indem Prozesse mit in Ontologien formalisiertem Wissen verknüpft werden.

Verknüpfung von semiformalen Modellen mit Ontologien. Die Potenziale einer Verknüpfung von Ontologien und Prozessmodellen werden seit geraumer Zeit in der Literatur erkannt [Hepp et al., 2005; Lin und Strasunskas, 2005; Ahlemann et al., 2006; Hepp und Roman, 2007]. Die Autoren versuchen i. d. R. mit einer ontologiebasierten Attributierung von Prozessmodellen die Grundlage für eine automatisierte Verarbeitung der Ablaufmodelle zu schaffen. Solche semantischen Erweiterungen für Prozessbeschreibungssprachen existieren u. a. für das Petri-Netz [**K**oschmider und Ried, 2005; Brockmans et al., 2006], die EPK [Bögl et al., 2008; Thomas und Fellmann, 2009], die BPMN [Abramowicz et al., 2007), die Demo Engineering Methodology for Organizations (DEMO) [Dietz, 2006] und die Extended Enterprise Modeling Language (EEML) [Lin und Ding, 2005]. Für das UML-Aktivitätsdiagramm liegt ein Konzept zur automatischen Synthese und Modifikation von Modellen nach Änderungen an Subprozessen vor [Lautenbacher und Bauer, 2006].

Ontologiebasierte Korrektheitsprüfung von Modellen. Eine Prüfung der inhaltlichen Korrektheit von Prozessbeschreibungen – im Rahmen der Arbeit auch als semantische Verifikation bezeichnet und verstanden im Sinne eines Nachweises, dass ein Modell spezifizierte Korrektheitskriterien einhält – wurde bisher vor allem im Bereich der Semantic Web Services erforscht. Diese setzen zur Verifikation semantisch annotierter Prozesse die Beschreibung von Vorbedingungen und Effekten ein, die durch Ontologien spezifiziert werden [Weber et al., 2010]. Im Gegensatz zu diesen Ansätzen wird mit der vorliegenden Arbeit eine über Vorbedingungen und Effekte hinausgehende Formalisierung der Semantik individueller Modellelemente angestrebt, deren Bezugspunkt ein semiformales Modell ist und nicht – wie im Bereich der Semantic Web Services – die Beschreibung einer Softwarekomponente bzw. eines Services. Zwar existieren mittlerweile vereinzelt auch im Bereich der semiformalen Modellierung Arbeiten, die Ontologien zur Korrektheitsprüfung heranziehen [El Kharbili und Stein, 2008; El Kharbili et al., 2008; Ly et al., 2009], allerdings wird der spezifische Beitrag von Ontologien nicht systematisch diskutiert und abgegrenzt. Mit dem in diesem Beitrag beschriebenen Konzept sollen die bereits in [Fellmann et al., 2010; Fellmann et al., 2011] beschriebenen Ansätze angewendet werden, die auf der Formalisierung von Semantik fußen.

Semantische Prozesswikis. Ansätze zur Prozessmodellierung in Wikis zielen darauf ab, Wikis um eine Visualisierung von Prozessen zu ergänzen. Die existierenden Systeme wie Wikiing Pro [Dengler et al., 2011], MoKi Wiki [Rospocher et al., 2009] und Makna [Dello et al., 2008] besitzen Funktionalitäten zur Visualisierung von Inhalten als Prozess, können allerdings nicht mit dem Funktionsumfang ausgereifter Modellierungswerkzeuge mithalten. Insbesondere unterstützen die Wikis keine Korrektheitsprüfungen der Modelle, wie sie in Desktop-Werkzeugen üblich sind.

6 Fazit

Im Rahmen dieses Beitrags wurde der Frage nachgegangen, welche Nutzenpotenziale eine Synthese aus Prozessmodellierungswerkzeug und Semantic Wiki besitzt. Die Verbindung aus Prozessmodellierung und kollaborativem Wissensmanagement, das mit (semantischen) Wikis ermöglicht wird, verspricht vielfältige Nutzen mit sich zu bringen und insbesondere auch durch die Integration des Wikis das Wissen von Akteuren mit einzubeziehen, die nicht direkt mit der Prozessmodellerstellung befasst sind. Durch die geplanten prototypischen Implementierungen im Rahmen der Forschungsagenda sollen ausgewählte Bestandteile einer empirischen Überprüfung unterzogen werden, um so schließlich die Grundlage für theoretisch abgesichertes Gestaltungswissen zu schaffen.

Literatur

[Abramowicz et al., 2007] W. Abramowicz, A. Filipowska, M. Kaczmarek, T. and Kaczmarek. Semantically enhanced Business Process Modelling Notation. In Hepp, M. et al. (Eds.): *Proceedings of SBPM 2007, Innsbruck, Austria, June 7, 2007 (CEUR Workshop Proceedings)*, S. 88–91.

[Ahlemann et al., 2006] F. Ahlemann, F. Teuteberg, und G. Brune. Ontologie-basierte Attributierung von Informationsmodellen: Grundlagen und Anwendungsgebiete. In F. Teuteberg und F. Ahlemann (Hrsg.): *ISPRI-Arbeitsbericht*, 01/2006, Universität Osnabrück.

[Awad et al., 2008] A. Awad, A. Polyvyanyy and M. Weske. Semantic querying of business process models. In *IEEE (Eds.): Proceedings of the 12th Internat. IEEE Enterprise Distributed Object Computing Conference (ECOC 2008)*, September 15–19, Munich, Germany. IEEE, S. 85–94.

[Bögl et al., 2008] A. Bögl, M. Schrefl, G. Pomberger, and N. Weber. Semantic Annotation of EPC Models in Engineering Domains by Employing Semantic Patterns. In J. Cordeiro and J. Filipe (Eds.): *Proc. of the 10th Int.l Conf. on Enterprise Information Systems (ICEIS 2008)*, Barcelona, Spain, June 12–16, S. 106–115.

[Brockmans et al., 2006] S. Brockmans, M. Ehrig, A. Koschmider, A. Oberweis, and R. Studer. Semantic Alignment of Business Processes. In Y. Manolopoulos, et al. (Eds.): *Proc. of the 8th International Conference on Enterprise Information Systems (ICEIS 2006),* Paphos, Cyprus : INSTICC Press, S. 191–196.

[Brodie et al., 2005] M. Brodie, C. Bussler, J. de Brujin, T. Fahringer, D. Fensel, M. Hepp, H. Lausen, D. Roman, T. Strang, H. Werthner, and M. Zaremba. Semantically Enabled Service Oriented Architectures: A Manifesto and a Paradigm Shift in Computer Science, Technical Report TR20051226, DERI – Digital Enterprise Research Institute.

[Dello et al., 2008] K. Dello, L. Nixon, and R. Tolksdorf. Extending the Makna Semantic Wiki to support Workflows. In C. Lange et al. (Eds.) *Proc. of the 3rd Semantic Wiki Workshop (SemWiki 2008)*, June 2nd, Tenerife, Spain.

[Dengler et al., 2011] F. Dengler, D. Vrandecic, and E. Simperl. Wikiing pro: semantic wiki-based process editor. In *Proceedings of the sixth international conference on Knowledge capture (K-CAP '11)*. New York : ACM.

[Dietz, 2006] J. L. G. Dietz. The deep structure of business processes. In *Communications of the ACM* 49, Nr. 5, S. 58–64.

[Dijkman et al., 2007] R. M. Dijkman, M. Dumas, and C. Ouyang. Formal semantics and automated analysis of BPMN process models. Preprint Technical Report 5969.

[ebizQ, 2008] ebizQ. SOA Market Pulse 2008. ebizQ. http://www.ebizq.net/white_papers/9172.html, Abruf am 16.05.2009

[El Kharbili und Stein, 2008] M. El Kharbili und S. Stein. Policy-Based Semantic Compliance Checking for Business Process Management. In P. Loos et al. (Eds.): *Modellierung betrieblicher Informationssysteme (MobIS 2008)*. RWTH Aachen (CEUR Workshop Proceedings), S. 165–177.

[El Kharbili et al., 2008] M. El Kharbili, S. Stein, S., I. Markovic, and E. Pulvermüller. Towards a Framework for Semantic Business Process Compliance Management. In *Proc. of the 1st International Workshop on Governance, Risk and Compliance – Applications in Informaton Systems (GRCIS2008)*, June 17, Montpellier, France.

[Fellmann et al., 2010] M. Fellmann, F. Hogrebe, M. Nüttgens, and O. Thomas. An ontology-driven approach to support semantic verification in business process modeling. In *Tagungsband der Tagung Modellierung betrieblicher Informationssysteme (MoBIS)*, 15.–17. September, Dresden, 2010, S. 99–110.

[Fellmann und Thomas, 2011] M. Fellmann und O. Thomas. Semantic Verification of Business Process Models: An Ontology-based Query-driven Approach. In *Proceedings of the 14th International Conference on Business Information Systems (BIS 2011)*, Poznan, Poland June 15–17, 2011.

[Fellmann und Thomas, 2011] M. Fellmann und O. Thomas. Process Model Verification with SemQuu. In M. Nüttgens et al. (Eds.): *Enterprise Modelling and Information Systems Architectures (EMISA 2011)*, Hamburg, Germany, September 22–23, 2011. Bonn : Köllen (GI LNI, P-190), S. 231–236.

[Fensel et al., 2003] D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster (Eds.). *Spinning the Semantic Web : Brining the World Wide Web to Its Full Potential.* 1 Aufl. Cambridge, London : MIT Press.

[Gómez-Pérez et al., 2004] A. Gómez-Pérez, M. Fernández-López, and O. Corcho. *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web.* London : Springer.

[Green, 1996] P. F. Green. *An Ontological Analysis of Information Systems Analysis and Design (ISAD) Grammars in Upper Case Tools*. Brisbane, Australia, University of Queensland, PhD Thesis.

[Gruber, 1993] T. R. Gruber. A Translation Approach to Portable Ontology Specifications. In *Knowledge Acquisition* 5, Nr. 2, S. 199–220.

[Hepp et al., 2005] M. Hepp, F. Leymann, J. Domingue, W. Wahler, and D. Fensel. Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management. In *Proceedings of the IEEE ICEBE* 2005, October 18–20, Beijing, China. Bejing, China, S. 535–540.

[Hepp und Roman, 2007] M. Hepp und D. Roman. An Ontology Framework for Semantic Business Process Management. In A. Oberweis et al. (Eds.): *eOrganisation: Service-, Prozess-, Market-Engineering : 8. Internat. Tagung Wirtschaftsinformatik*, Karlsruhe, 28. Feb. – 2. März; Bd. 1. Karlsruhe : Univ.-Verlag, S. 423–440.

[Kindler, 2006] E. Kindler. On the semantics of EPCs: Resolving the vicious circle. In *Data & Knowledge Engineering* 56, Nr. 1, S. 23–40.

[Koschmider und Ried, 2005] A. Koschmider und D. Ried. Semantische Annotation von Petri-Netzen. In *Workshop für Algorithmen und Werkzeuge für Petrinetze (AWPN'05)*. Humboldt-Univ. zu Berlin, S. 66–71.

[Lautenbacher und Bauer, 2006] F. Lautenbacher und B. Bauer. Semantic Reference- and Business Process Modeling enables an Automatic Synthesis. In K. Hinkelmann et al. (Eds.): *Proceeding of the Workshop on Semantics for Business Process Management*, Budva, Montenegro, June 2006, S. 89–100.

[Lin und Ding, 2005] Y. Lin und H. Ding. Ontology-based Semantic Annotation for Semantic Interoperability of Process Models. In *Proceedings of CIMCA-IAWTIC'06*, Volume 01. Washington, DC, USA : IEEE, S. 162–167.

[Lin und Strasunskas, 2005] Y. Lin und D. Strasunskas. Ontology-based Semantic Annotation of Process Templates for Reuse. In *Proceeding of 10th CAiSE/IFIP8.1/EUNO Internat. Workshop on Evaluation of Modeling Methods in System Analysis and Design (EMMSAD05)*, Porto, Portugal, June 2005, S. 162–167.

[Ly et al., 2009] L.T. Ly, S. Rinderle-Ma, K. Göser, and P. Dadam. On enabling integrated process compliance with semantic constraints in process management systems. In *Information Systems Frontiers*, S. 1–25.

[Mendling, 2009] J. Mendling. Empirical Studies in Process Model Verification. In Jensen, K.; van der Aalst, W. M. P. (Eds.): *Transactions on Petri Nets and Other Models of Concurrency II*. Heidelberg : Springer (LNCS 5460), S. 208–224.

[Ortner, 1997] E. Ortner. *Methodenneutraler Fachentwurf : Zu den Grundlagen einer anwendungsorientierten Informatik.* Stuttgart : Teubner (Reihe Wirtschaftsinformatik).

[Pfeifer 1996] T. Pfeifer. *Qualitätsmanagement*. 2 Aufl. München, Hanser.

[Recker et al., 2009] J. Recker, M. Rosemann, M. Indulska, and P. Green. Business Process Modeling – A Comparative Analysis. In *Journal of the Association for Information Systems* 10, Nr. 4, S. 333–363.

[Rosemann und Schwegmann, 2002] M. Rosemann und A. Schwegmann. Vorbereitung der Prozessmodellierung. In J. Becker et al. (Hrsg.): *Prozessmanagement : Ein Leitfaden zur prozessorientierten Organisationsgestaltung*. 3. Aufl. Berlin : Springer, S. 47–94.

[Rosener, 1990] J. B. Rosener. Ways Women Lead. In *Harvard Business Review* 68, Nr. 6, S. 119–125.

[Rosenkranz, 2009] C. Rosenkranz. *The Quality of Organizations : A Communication-Based Measurement Approach*. Johann Wolfgang Goethe Universität Frankfurt am Main, Fachbereichs Wirtschaftswissenschaften, Dissertation.

[Rospocher et al., 2009] M. Rospocher, C. Ghidini, V. Pammer, L. Serafini, and S. Lindstaedt. Moki: the modelling wiki. In Lange et al. (Eds.): *Proc. of the 4th Semantic Wiki Workshop (SemWiki 2009)*, co-located with ESWC 2009, Hersonissos, Crete, Greece, June 1st, 2009, S. 113-128.

[Sarshar et al., 2006] K. Sarshar, M. Weber, and P. Loos. Einsatz der Informationsmodellierung bei der Einführung betrieblicher Standardsoftware : Eine empirische Untersuchung bei Energieversorgerunternehmen. In *Wirtschaftsinformatik* 48, Nr. 2, S. 120–127.

[Schaffert et al., 2009] S. Schaffert, F. Bry, J. Baumeister, and M. Kiesel. Semantische Wikis. In A. Blumauer und T. Pellegrinio (Eds.), *Social Semantic Web. Web 2.0 – Was nun?* (pp. 245–258). Berlin, Springer.

[Softlab Group, 2007] Softlab Group. IT und Geschäftsabläufe sind getrennte Welten – immer noch. Pressemitteilung zur Studie von meetbiz-research im Auftrag der Softlab Group. München : Softlab Group.

[Studer et al., 1998] R. Studer, V.R. Benjamins, and D. Fensel. Knowledge Engineering: Principles and Methods. In *Data & Knowledge Engineering* 25, Nr. 1–2, S. 161–197.

[Thomas, 2006] O. Thomas. *Management von Referenzmodellen : Entwurf und Realisierung eines Informationssystems zur Entwicklung und Anwendung von Referenzmodellen*. Berlin : Logos (Wirtschaftsinformatik – Theorie und Anwendung; 1).

[Thomas und Fellmann, 2009] O. Thomas und M. Fellmann. Semantische Prozessmodellierung – Konzeption und informationstechnische Unterstützung einer ontologiebasierten Repräsentation von Geschäftsprozessen. In *Wirtschaftsinformatik* 51, Nr. 6, S. 506–518.

[van der Aalst, 1999] W.M.P. van der Aalst. Formalization and verification of event-driven process chains. In I*nformation and Software Technology* 41, Nr. 10, S. 639–650.

[vom Brocke, 2003] J. vom Brocke. *Referenzmodellierung : Gestaltung und Verteilung von Konstruktionsprozessen*. Berlin : Logos (Advances in information systems and management science; 4).

[Wand und Weber, 1995] Y. Wand und R. Weber. On the deep structure of information systems. In *Information Systems Journal* 5, Nr. 3, S. 203–223.

[Weber et al., 2010] I. Weber, J. Hoffmann, and J. Mendling. Beyond soundness: on the verification of semantic business process models. In *Distributed and Parallel Databases* 2010, Nr. 27, S. 271–343.

[Wolf und Harmon, 2010] C. Wolf und P. Harmon. The State of Business Process Management 2010. BPTrends.

Toward an agile knowledge connection of employees with regard to business processes

Michael Leyer
Frankfurt School of Finance & Management
60314, Frankfurt am Main, Germany
m.leyer@fs.de

Nina Claus
Peggy Diggledey GmbH
22609, Hamburg, Germany
nina_claus@web.de

Abstract

Employee knowledge is a valuable and thus very important asset of a company. However, employees are often not aware of existing knowledge within their organization. As a result, wheels are reinvented continuously within organizations and employees spend unnecessary time learning processes in a cumbersome way on their own. To deal with this problem, databases containing explicit knowledge are often built, but they are rarely used – mainly due to the immense effort of keeping them up to date. In this paper we present a new approach aiming at socially connecting employees – an internal social knowledge network. A major novelty is the use of the organization's business processes as a starting point. Employees can connect to each other by indicating their process-related areas of expertise. The aim is to enable sustainable sharing and distribution of knowledge within an organization.

1 Framework for the knowledge connection of employees

We propose an internal social network system that is closely linked to the processes of an organization and does not contain the tacit knowledge itself. The idea is to motivate employees to indicate their areas of expertise and to claim expert status toward other employees:

- *Setting up the process architecture*: The starting point for the introduction of such a social knowledge network lies in the organization's business processes (Figure 1). The processes define the business-related connections between employees. Such a process architecture is provided to employees within the social knowledge system. The processes should be documented on three levels from top down, thus, being specific in terms of the description but general in terms of the activities to be performed within the process.
- *Social knowledge system*: The social knowledge system is the core of the framework. The implementation of the system should take place via the use of software-based tools. Each employee should have easy access to the available but distributed knowledge in the organization.
- *Usage of the social knowledge system*: Employees can reference their relevant areas of expertise, as well as sources for explicit knowledge. Thus, they provide the necessary information individually. The system can easily be used in daily work by everyone and by the management to identify knowledge gaps. Clear rules for communication have to be established to avoid problems. This should cover a rating of knowledge, timespan for knowledge if not used and incentives.
- *Supporting incentive structure*: In order to ensure sustainable application and widespread use of the social network, an incentive system is recommended. A bonus system should be set up, covering the completeness of the profile, answering of inquiries and the rating of colleagues. The bonus system should be part of the personal salary bonus.
- *Technical implementation*: The system is available for the relevant recipients – that is, management and employees – for their use in day-to-day business as a software tool. A personal knowledge page, the knowledge map and a search engine are provided. Depending on existing process- and knowledge-related systems available a technical connection should take place.

An overview of the elements of the framework is provided in Figure 1.

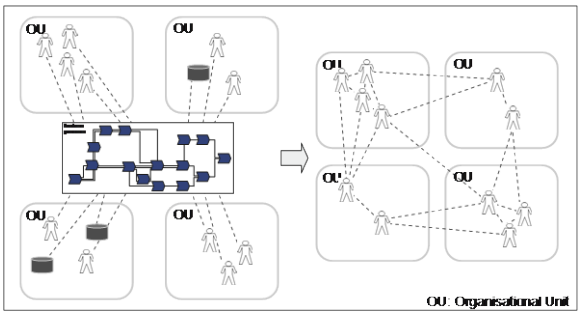


Figure 1. Overview of the social knowledge system

2 Reference of the paper

Michael Leyer and Nina Claus. Toward an agile knowledge connection of employees with regard to business processes. In *Proceedings of the 46th Annual Hawaii International Conference on System Sciences*, pages 3436–3445, Los Alamitos, California, 2013. IEEE.

Using Business Process Management Frameworks to Integrate Knowledge Management Services into Applications

Constantin Jucovschi
Jacobs University Bremen
DE-28757, Bremen, Germany
c.jucovschi@jacobs-university.de

Abstract

In [Dav+12], David et al. present the Semantic Alliance (Sally) — a framework enabling integration of Knowledge Management services into applications typically used by knowledge workers (e.g. spreadsheet programs). While integrating new applications with the Sally framework did not pose serious challenges, adding new Knowledge Management (KM) services was getting ever more challenging with every newly integrated service.

A closer look revealed that compared to other typical tools such as spreadsheet optimizers or rich visualization tools, Knowledge Management services typically have a very restricted set of software objects to which they can be applied. On the other hand, they can be applied in a much richer number of contexts. These requirements call for distinct integration strategies that this paper explores in more detail.

A solution to accommodate these requirements was implemented based on frameworks developed in the Business Process Management research area.

1 Introduction

In their paper [SL04], Stenmark and Lindgren present a collection of best practices for integrating Knowledge Management systems into everyday work. One such best practice is resisting the temptation of introducing new tools to the workplace of knowledge workers. The reason is that new tools undergo a long adoption process — time in which knowledge workers understand how the new tools integrate into their usual workflows. A better solution is to integrate KM tools into applications familiar and widely used by the user (also known as invasive technology [Koh05]). This is known to substantially increase the likelihood of survival of KM services and that the users will eventually discover the benefits of the added tools [SL04].

[Dav+12] presents the Semantic Alliance (Sally), a framework enabling integration of Knowledge Management services into applications typically used by knowledge workers (e.g. spreadsheet programs). The novel idea, coined in the paper as Invasive Design, was to combine invasive technologies with Semantic Illustration [KK09] (linking software objects e.g. spreadsheet cells to ontologies and thus providing fertile ground for KM services). To validate the framework, the authors created an implementation which integrated several KM services into Microsoft Excel and LibreOffice/OpenOffice. The KM services allowed the user to assign ontology concepts to blocks of cells and provided services like definition lookup and semantic navigation.

The initial implementation of the Sally framework proves the feasibility and usefulness of the invasive design paradigm. Subsequent efforts of adding new KM services to the already invaded spreadsheet systems showed that adding new services became ever harder because it was hard to specify when a certain service should be executed and how it should behave in response to events coming from the invaded system. The framework did not show any scalability problems at invading other applications like Autodesk Inventor (CAD System), jEdit (text editor), Planetary (web based framework for active documents). The KM service integration challenges were observed independently of the invaded system. These challenges substantially increased when developing a pricing service that used semantic data from both, a spreadsheet document and a CAD system to provide context-switching functionality from one system to the other [Koh+13]. One reason for this challenge was that services like definition lookup had to function in the same way independently whether the software object for which definition was requested was a CAD object or a spreadsheet cell.

The encountered scalability problems are not consequences of the Invasive Design paradigm. Invading new systems and linking software objects to an ontology did not pose difficulties. It merely provided a fruitful ground for KM services to come together. The real challenge was mashing up KM services together into one coherent system that the knowledge worker is willing to use.

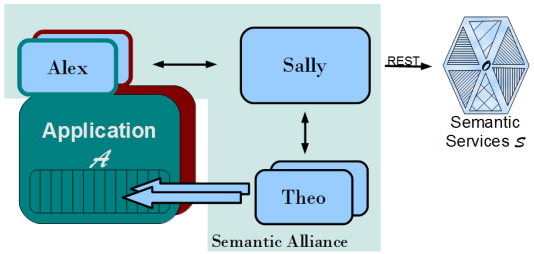
In this paper, I would like to present the new extension mechanism of Semantic Alliance that partially alleviates the aforementioned problems by employing techniques from the field of Business Process Management (BPM). Even with the new extension mechanism, integrating new KM services into Sally might still require adaptations in configuration and even implementation of existing services. The number and difficulty of such adaptations is however reduced.

In the following section, I will shortly describe the main components of the Semantic Alliance framework and introduce important concepts used extensively for the rest of the paper. To give the reader a better idea of the encountered scalability problems, I describe in section 3 two services and show the problems that might occur if integrated into one system. In section 5, I will introduce a metamodel used in the area of Business Process Management to categorize existing process management frameworks. In section 4, I will describe in more detail the observations

that were made during the efforts of mashing up KM tools. These will define the requirements an existing BPM framework would have to provide in order to be used for our purposes. Section 6, presents some detail about our implementation. The paper ends with a conclusion and a short discussion.

2 Sally Components

a)



b)

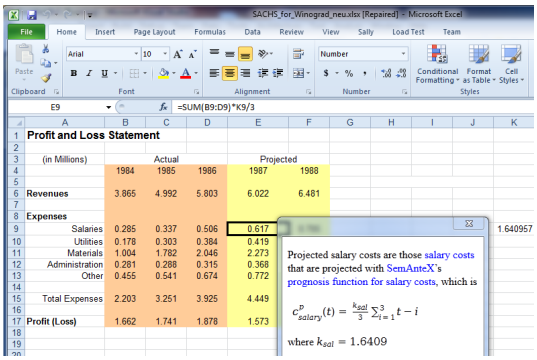


Figure 1: Sally Components while doing definition lookup

To explain the main components of the Sally framework, lets us look at Figure 1b). One can see an instance of Microsoft Excel and a window on top of it showing definition lookup for the cell E9. Microsoft Excel corresponds to Application $\mathcal{A}$ in Figure 1a) and is the system which was invaded. That means that Excel hosts a custom plugin (called an Alex[1]) which connects with the Semantic Ally component (Sally) and sporadically sends it information about the opened documents and user events. Based on the information and events sent by Alex, the Sally component may, at some point, decide that it wants to interact with the user of the invaded system. This is achieved through the Theo screen manager — a standalone program also connected to Sally and it is only job is to create windows anywhere on the screen (just like the definition lookup window) and load a certain web-page inside that window. In contrast to a normal browser window, the JavaScript loaded in a Theo window, can interact back with Sally and hence with Alex. Finally, semantic services like definition lookup, are implemented as standalone services possibly running on other servers and using completely different frameworks.

The original paper [Dav+12] introducing the Semantic Alliance framework discusses in great detail the reasons behind integrating KM services in such a way. It also provides the reader with a good intuition on the type of messages being exchanged between Alex, Sally, Theo and the semantic services.

[1]named after Alexander the Great; one of the mightiest invaders in history

3 Challenges in Mashing Up Knowledge Management Tools

The aim of this section is to make the reader familiar with the type of challenges one encounters when mashing up several Knowledge Management services.

The first service was already introduced in the previous section — the **definition lookup** service. The service works as follows:

1. The user makes Sally aware that she wants to perform definition lookup (e.g. clicking some menu item).
2. She clicks on a cell (e.g. E9) and if that cell has an ontology concept (e.g. "Projected salary costs") assigned to it,
3. Sally creates a Theo window showing the definition (Figure 1b).
4. Any following clicks on spreadsheet cells (e.g. E8), result in the Theo window to be moved to the right of the clicked cell and content of the definition lookup window to be updated.

The second service, called **ontology linking service**, allows the user to assign ontology concepts to blocks of cells. Here is a description of the service.

1. The user makes Sally aware that she wants to perform ontology linking (e.g. clicking some menu item).
2. She clicks on a range of cells (possible just one cell), and if that range has:

3a. no ontology concept assigned to it, then Sally creates a Theo window showing a form that the user needs to complete and so assign an ontology concept to that range of cells. If she selects another range of cells, only one element in the forms gets updated, namely the one showing the range for which the ontology link will be assigned.

3b. has an ontology concept assigned to it, a form with previously saved information is shown to the user allowing editing. If the user does not change the form but selects a new range, Sally closes the current Theo window and goes to step 2, otherwise, updates the range field of the opened form.

Both services have some relatively simple way of responding to events coming from the invaded spreadsheet system. Nevertheless, mashing them up in one system is not trivial. Imagine the user requests ontology linking service, selects a range of cells without a link to an ontology concept and starts filling in the form. In the middle of the task, she feels the need to do definition lookup on some other cell and so she selects it and invokes the definition lookup service.

At this point the Sally component is requested to run two services in parallel and there are several ways it can respond to that. Sally could implement a "one service at a time" policy which would frustrate the user. Sally could also let both services run in parallel and forward all incoming events to each service. That would eventually confuse the user because her action would have consequences in services that for her mental model are not in focus / not active. Hence Sally needs to have an event forwarding strategy which would forward events only to one service — the one for which the event was intended.

The initial implementation of the Semantic Alliance framework used state variables to control which compo-

nents would be notified of incoming events. As the number of services grew, it became very hard to manage such forwarding rules. In particular, to integrate a new service, changes in several related services had to be performed and hence modularity was violated. Additionally, in order to customize this behavior, the whole framework had to be recompiled.

4 Mashing up Semantic Illustration Services

The Semantic Illustration (SI) architecture [KK09] enhances an application with the Interpretation Mapping function IM which assigns each software object a concept in an ontology $\mathcal{O}$. The ontology $\mathcal{O}$ formalizes relationships among concepts and possibly connects them to external ontologies. Note that the ontology $\mathcal{O}$ is not static. As the user changes or opens new documents, the ontology $\mathcal{O}$ gets updated/extended with new concepts and relationships.

Semantic Illustration Services heavily tap into the information provided by the interpretation mapping and hence it is worthwhile to analyze how the properties of the IM function affects the type of interactions Semantic Illustration Services might be able to provide.

4.1 Dynamic Applicability

Semantic Illustration Services usually have very specific requirements that need to be satisfied before they can be applied. For instance the definition lookup service requires that the input software object has a link to an ontology concept which in turn has a definition. A complex service like the pricing service in [Koh+13] requires that the input CAD object cad satisfies

$$\begin{aligned} \exists cell1, cell2. \quad & \text{such that} \\ IM(cell1) &= IM(cad) \wedge \\ IM(cell2) &= \text{"http://economics.org/prices"} \wedge \\ costOf(cell1, cell2) &\in \mathcal{O} \end{aligned}$$

As documents change or new documents are opened, Semantic Illustration services can become applicable or cease to be applicable.

The requirements that need to be satisfied before a Semantic Illustration service can be invoked may become extremely convoluted and should be regarded by a mashup framework as black boxes. The mashup framework should provide a mechanism for the user to get the list of services are applicable in the current context.

4.2 High Reusability

An important observation about the interpretation mapping IM is that it does not depend on any context information. Hence independently of whether spreadsheet cell $C4$ is used as part of a formula or as parameter in a wizard, the $IM(C4)$ is always the same. As a consequence, Semantic Illustration services that solely use information provided by the interpretation mapping are context independent and can be invoked from a multitude of contexts.

The definition lookup service is an example of such a service as it solely uses the IM function to fetch the definition of the software object at hand and display it. Indeed, the user should be able to invoke the definition lookup service for cell $C4$ in many situations e.g. if she: 1) clicks on cell $C4$ inside the spreadsheet document; 2) inspects a reference of cell $C4$ as part of a formula; 3) uses a wizard or plugin to perform operations on cell $C4$. Moreover, it may very well happen that a knowledge worker wants to get a definition lookup on a concept mentioned in the definiens provided by the content produced by the definition lookup service itself.

Due to the multitude of contexts in which a Semantic Illustration service can be applied as well as due to modal windows, it is often unfeasible to provide fixed user interface components for accessing the service. Context menus generally provide a better solution. It is the responsibility of the invading Alex component to reliably detect the software objects the user selects. This, in turn, depends a lot on the architecture of the invaded system.

4.3 Support for Subtasking

Knowledge Management tasks often require the user to perform actions she is not familiar with e.g. creating ontology concepts and making their relationship with other concepts explicit. In such cases, the knowledge worker would be tempted to look how similar tasks were performed before and complete current task by following the example of another one. That naturally introduces subtasking i.e. starting a new task while in the midst of another task.

In the example described in section 3, I have presented the challenges associated with the use case when two services run simultaneously and compete for in the same software object selection event. Such situations are quite common for Semantic Illustration services due to the fact that IM function needs a software object as parameter.

A framework for mashing up Semantic Illustration needs to have a policy for broadcasting events to SI services which does not confuse the user. Moreover, there should be an intuitive way how the knowledge worker can resume an interrupted service.

4.4 Access to Common Resources

KM services often rely on other KM services to achieve their goals. For example, the IM editor, a service linking software objects to ontologies needs to rely on some service that provides CRUD (Create, Read, Update, Delete) operations on ontologies. Our implementation uses the Planetary [Koh12] system for that, but there is no reason why one could not connect use Protege [Pro] or Semantic Media Wiki [Sem].

Similarly, in our experiments invading Web environments, we needed to make sure that services don't use the desktop version of the Theo screen-manager but the Web-based one.

It is desirable that KM services depend on each other because that is the key to reusability. The dependencies should be loose, so that one can easily exchange used implementations based on the context.

5 Dimensions of Business Process Modeling

The area of Business Process Modeling aims at describing actions, agents, resources, and relations among them necessary to complete a certain task. Due to the practical nature of the problem and applicability for business, there is a large body of research, best practices, standards, and implementations available. Most of these frameworks support basic modeling tasks but handle more complex situations in some particular way. This makes it very difficult to decide whether a framework is suitable for particular needs without becoming an expert in it. To partially alleviate this problem [LK06] developed a generic meta-model, i.e., a set of dimensions that can be used to categorize particular frameworks. In this section I provide a short description of

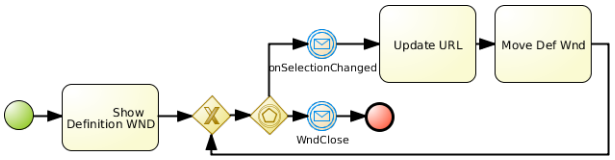


Figure 2: Definition Lookup

these dimensions. In section 5.1, I will evaluate how the mashup requirements from section 4 can be supported by Process Modeling Tools.

Functional Dimension captures the actionable elements of a process and the flows of data relevant to them [CKO92]. It is represented as a directed graph where the nodes represent the actionable elements and the edges represent the type of objects these elements need to exchange. Frameworks are classified in 3 categories, those which: do not represent activities (e.g. Petri Nets); only consider atomic activities; and those which allow subprocesses divisible into other subprocesses and atomic activities.

Behavioral Dimension describes when process elements get executed. This dimension captures loops, branching conditions, decision making, exit criteria etc. Frameworks are analyzed in respect with types of control flows they support e.g. AND splits/joins, XOR splits/joins, N-out-of-M joins etc.

Organizational Dimension captures where and by which agents process actions will be executed. Categorization criteria is based on the types of agents that can be modeled in the framework. For example, is role based execution possible? Can one differentiate between computer services and human workers?

Informational Dimension describes the types of entities that are exchanged among and changed by activities. Such entities include: Events, Database Tables but also Services, Applications.

Context Dimension presents a high-level view on the business process. It describes goals, their measures, deliverables, process owners, process types etc. It captures relations between processes (e.g. support processes); what deliverables are created and what goals these achieve.

5.1 Evaluation of Requirements for SI Service Mashups

The minimum requirements a Business Process Modeling framework needs to support for mashing up SI Services depend on two factors:

- the individual requirements of each SI Service towards the BPM framework.
- requirements to support mashing up SI services.

This section evaluates only the later.

An observation that makes modeling of dynamic applicability, high reusability and subtasking requirements much easier is that they can run in a process separate from the SI services themselves. To illustrate this, let us look at Figure 2 depicting the BPM diagram of the definition lookup service. The diagram does not contain any special actions or events that would allow for dynamic applicability or high reuseability. The later are enabled through the

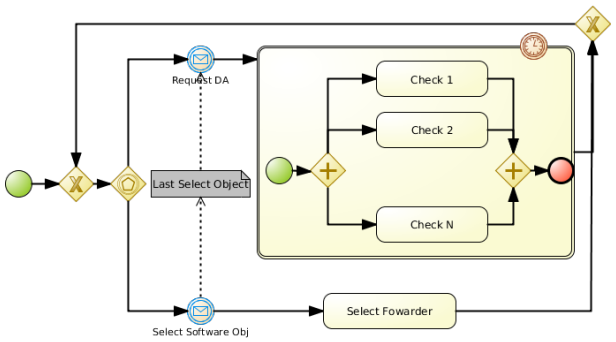


Figure 3: The EventLoop Process

EventLoop process in Figure 3. This process is always active and is the only one which directly receives events about selection events of software object ("Select Software Obj" event) or that the user requested a list of services applicable to the current context ("Request DA"). In case a new software object was selected, it runs a task ("Select Forwarder") which uses some heuristics to decide which service (e.g. definition lookup) should be notified of the event and forward the event to that particular service. This is how "onSelectionChanged" event from Figure 2 gets triggered.

In case the user wants to see what services are available for the current context, the engine starts a subprocess in which it runs in parallel a series of N service applicability checks, optionally have a timeout strategy and a way to create a list of services for which applicability checks succeeded.

The proposed solution gives us the following requirements for BPM frameworks:

- support for subprocesses for the functional dimension
- support for AND joins/splits for the behavioral dimension
- no special requirements for the context and informational dimensions

6 Implementation

To implement the extension mechanism for the Semantic Alliance, I used the jBPM Business Process Management suite [Jbp]. It complied with all the requirements presented in section 4 and also is Java based can so could easily be integrated with the existing Semantic Alliance implementation.

The implemented extension mechanism is based on distributing jar files. On startup, Sally loads all the jar files from a fixed directory and uses Java reflections mechanisms to find all BPMN2 process files. Each BPMN2 process file represents a Knowledge Management service. Each task inside such a process file, specifies an implementation class and method that should be called as soon as the workflow reaches that element. Such BPMN2 files can be created using the jBoss Eclipse plugin or using the web interface of jBoss Drools Guvnor.

7 Conclusions

This paper tackles the problem of mashing up Knowledge Management services in a modular way, in systems typically used by knowledge workers. The aim is to create a framework which could provide plug and play support for KM services. As described in section 4, Knowledge Management services have different integration requirements

into applications than software artifacts such as optimizers or rich visualization tools. In comparison to these tools, Knowledge Management services typically have a very restricted set of software objects to which they can be applied. On the other hand, they can be applied in a much richer number of contexts. Supporting such an integration strategy is the main challenge and contribution of this paper.

The solution presented in the current paper is based on the very mature Business Process Management research area which aims at describing actions, agents, resources and relations among them necessary to complete a certain task. It seemed to provide a natural solution to our challenges and yet the presented solution does not cover all the requirements we identified. Further research is necessary to see if the presented architecture could be improved.

Another research area that could provide some useful insights is the one of ubiquitous computing. This area has similar challenges of making heterogeneous systems work together and respond to user actions in way that makes sense. The author found several goal-based frameworks which have the potential of vastly improving presented architecture.

The Semantic Alliance framework [Dav+12] and the challenges encountered in extending it with new Knowledge Management services, was used both as a tool to understand the root the integration challenges as well as a testbed for developing solutions. An experimental implementation of the extension mechanism was implemented and we are still in the process converting existing KM services to the extension mechanism described in this paper.

8 Acknowledgements

I want to thank Andrea Kohlhase and Michael Kohlhase as well as the anonymous reviewers for their constructive suggestions. This work has been funded by the German Research Council under grant KO-2484-12-1.

References

[CKO92] Bill Curtis, Marc I. Kellner, and Jim Over. "Process modeling". In: *Commun. ACM* 35.9 (Sept. 1992), pp. 75–90. ISSN: 0001-0782. DOI: 10.1145/130994.130998. URL: http://doi.acm.org/10.1145/130994.130998.

[Dav+12] Catalin David et al. "Semantic Alliance: A Framework for Semantic Allies". In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics (CICM) (Bremen, Germany, July 9–14, 2012). Ed. by Johan Jeuring et al. LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012, pp. 49–64. ISBN: 978-3-642-31373-8. URL: http://kwarc.info/kohlhase/papers/mkm12-SAlly.pdf.

[Dav12] Catalin David. "Semantic Alliance Framework: Integrating Documents and Semantic Services". M. Sc. Thesis. Jacobs University Bremen, 2012. URL: https://svn.eecs.jacobs-university.de/svn/eecs/archive/msc-2012/cdavid.pdf.

[Jbp] *jBPM Business Process Management (BPM) Suite from jBoss Community*. URL: http://www.jboss.org/jbpm (visited on 07/15/2013).

[Jeu+12] Johan Jeuring et al., eds. *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics (CICM) (Bremen, Germany, July 9–14, 2012). LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012. ISBN: 978-3-642-31373-8.

[KK09] Andrea Kohlhase and Michael Kohlhase. "Semantic Transparency in User Assistance Systems". In: *Proceedings of the 27th annual ACM international conference on Design of communication (SIGDOC)*. (Bloomington, Indiana, USA, 2009). Ed. by Brad Mehlenbacher et al. ACM Special Interest Group for Design of Communication. New York, NY, USA: ACM Press, 2009, pp. 89–96. DOI: 10.1145/1621995.1622013. URL: http://kwarc.info/kohlhase/papers/sigdoc09-semtrans.pdf.

[Koh+13] Andrea Kohlhase et al. "Full Semantic Transparency". In: *Proceedings of the 14th IFIP TC 14 International Conference on Human-Computer Interaction*. INTERACT'13. in press. Cape Town, South Africa: Springer-Verlag, 2013. URL: http://kwarc.info/kohlhase/papers/Interact2013_FST.pdf.

[Koh05] Andrea Kohlhase. "Overcoming Proprietary Hurdles: CPoint as Invasive Editor". In: *Open Source for Education in Europe: Research and Practise*. Ed. by Fred de Vries et al. Proceedings at http://hdl.handle.net/1820/483. Open Universiteit Nederland. Heerlen, The Netherlands: Open Universiteit Nederland, Nov. 2005, pp. 51–56. URL: http://hdl.handle.net/1820/483.

[Koh12] Michael Kohlhase. "The Planetary Project: Towards eMath3.0". In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics (CICM) (Bremen, Germany, July 9–14, 2012). Ed. by Johan Jeuring et al. LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012, pp. 448–452. ISBN: 978-3-642-31373-8. arXiv: 1206.5048 [cs.DL].

[LK06] Beate List and Birgit Korherr. "An evaluation of conceptual business process modelling languages". In: *Proceedings of the 2006 ACM symposium on Applied computing*. SAC '06. Dijon, France: ACM, 2006, pp. 1532–1539. ISBN: 1-59593-108-2. DOI: 10.1145/1141277.1141633. URL: http://doi.acm.org/10.1145/1141277.1141633.

[Pro] *Protégé*. URL: http://protege.stanford.edu.

[Sem] *Semantic MediaWiki*. URL: http://semantic-mediawiki.org (visited on 03/04/2010).

[SL04] Dick Stenmark and Rikard Lindgren. “Integrating knowledge management systems with everyday work: Design principles leveraging user practice”. In: *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. IEEE. 2004, 9–pp.

Potentialanalyse des prozessorientierten Wissensmanagement für die Baubranche

Sarah Gessinger, Ralph Bergmann
Universität Trier – Wirtschaftsinformatik II
D-54286, Trier, Deutschland
Email: { gessinge | bergmann }@uni-trier.de
www.wi2.uni-trier.de

Zusammenfassung

Während Projektmanagement-Systeme in der Baubranche zur Planung fast jedes Projekts eingesetzt werden, sind Systeme zum prozessorientierten Wissensmanagement in dieser Branche sehr selten vertreten. In diesem Beitrag stellen wir die Ergebnisse einer Potentialanalyse des prozessorientierte Wissensmanagement für die Baubranche vor. Dabei gehen wir auf das Prozessmanagement allgemein ein und zeigen dessen Bedeutung für das prozessorientierte Wissensmanagement. Es werden die einzelnen Schritte der Potentialanalyse erläutert und ein Anwendungsszenario der Nutzungsphase eines Bauprojekts skizziert.

1 Einleitung

Die wachsende Bedeutung von Wissen kann heute als allgemein bekannt vorausgesetzt werden. Wissen ist ein wichtiger Bestandteil in der Gesellschaft und der Wirtschaft. Für Unternehmen ist der richtige Umgang mit Wissen zu einem Erfolgsfaktor geworden, der ihre Wettbewerbsfähigkeit entscheidend beeinflusst [Lehner, 2006]. Ein gezieltes Managen von Wissen kann die Effizienz einer Organisation und den unternehmerischen Erfolg steigern bzw. aufrechterhalten. Wissensmanagement ist auch ein Instrument des Managements, das die Entscheidungsqualität eines Unternehmens verbessern kann, z. B. die Verwendung von Wissen zur langfristigen Unternehmensplanung [Lehner, 2006]. Ein stärkeres Unternehmenswachstum, bessere Abgrenzungsmöglichkeiten gegenüber Konkurrenten, mehr Innovationen, verbesserte Prozesse und eine höhere Kundenzufriedenheit sind die möglichen Vorteile, die sich aus dem Einsatz eines Wissensmanagementsystems ergeben können [Wildemann, 2003].

Prozessorientiertes Wissensmanagement kombiniert Ansätze des Prozessmanagements mit denen des Wissensmanagements. Verschiedene Ansätze des Wissensmanagements beschreiben den Ablauf der Aktivitäten des Wissensmanagements in Form von Prozessen [Probst 2012; Reinmann-Rothmeier/Mandl 1997]. Diese Prozesse werden als Wissensprozesse verstanden, da sie für die Verarbeitung von Wissen verantwortlich sind. Wissensmanagementprozesse beschreiben hingegen die Prozesse, die sich mit der Verwaltung und Pflege von Wissensprozessen befassen. Remus definiert prozessorientiertes Wissensmanagement als „Managementaufgabe, die für die regelmäßige Auswahl, Umsetzung und Evaluation von prozessorientierten WM-Strategien zuständig ist, mit dem Ziel die Wissensverarbeitung in den operativen wissensintensiven Geschäftsprozessen[1] zu unterstützen, zu verbessern und weiterzuentwickeln, um schließlich zur Kernwertschöpfung des Unternehmens beizutragen.“ [Remus, 2002] Bisher findet das prozessorientierte Wissensmanagement Anwendung in projekthaften Prozessen der Automobil- und Anlagenentwicklung [Rose *et al*., 2002] oder in Projekten der Produktentwicklung [Kleiner und Kirsch, 2009; Schnauffer *et al*., 2004]. Auch wurde dessen Einsatz bereits in den Branchen des Anlagen- und Maschinenbaus, Automotive und Elektronik im Rahmen eines Verbundprojekts (ProWis: 2005-2010) der Fraunhofer-Institute untersucht und dessen Einführung begleitet [Orth *et al*., 2009].

Auch die Baubranche ist gekennzeichnet durch komplexe Prozesse und eine Vielzahl an Beteiligten. Änderungen an geplanten Abläufen sind an der Tagesordnung. Wir verfolgen daher die Hypothese, dass der Einsatz von prozessorientiertem Wissensmanagement auch in der Baubranche zu erheblichen Produktivitätsvorteilen führen sollte. Wir erwarten, dass neue Methoden und Werkzeuge des prozessorientierten Wissensmanagements eine automatisierte Ablaufsteuerung für Prozesse im Bauwesen bieten können, die die Agilität von Prozessen einbezieht und technologisch unterstützt. Das Vorkommen von Agilität in Bauprozessen ist ein spezifisches Charakteristikum von vielen Prozessbereichen im Bauwesen. Z. B. sind witterungsbedingte Störungen, Aufgaben, die nicht rückgängig gemacht oder unterbrochen werden können, typische Merkmale, welche einen starken Einfluss auf die Ablaufsteuerung von Bauprozessen haben. Diese Umstände machen eine entsprechende Flexibilität des Systems notwendig, um den Beteiligten eine systematische und softwaretechnische Unterstützung bieten zu können. Der Zeit- und Kostenaufwand könnte sich in Bauprojekten deutlich reduzieren, da korrigierende Maßnahmen rechtzeitig kommuniziert und eingeleitet werden könnten. Im Rahmen einer von uns durchgeführten Analyse von operativen Geschäftsprozessen im Bauwesen konnte festgestellt werden, dass zurzeit keine solche Software zur flexiblen Prozessunterstützung im Einsatz ist.

Der vorliegende Beitrag fokussiert daher auf einer Potenzialanalyse von Prozessen und Workflows im Bauwesen für das prozessorientierte Wissensmanagement. Dabei wird auf Vorarbeiten des Lehrstuhls für Wirtschaftsinformatik II der Universität Trier zu agilen Workflows zurückgegriffen, die im Rahmen mehrerer Projekte entstan-

[1] Eine Definition von wissensintensiven Prozesstypen und eine Klassifikation von wissensintensiven Geschäftsprozessen kann bei Remus 2002, S. 82f nachgelesen werden.

den sind. Hierbei wurde insbesondere als experimentelle Grundlage das CAKE[2] (Collaborative Agile Knowledge Engine) Framework entwickelt [Görg *et al*., 2013]. CAKE ist eine universelle Architektur und ein Werkzeug zur wissensbasierten Unterstützung flexibler, kollaborativer Arbeitsvorgänge. Entsprechend ist die Potentialanalyse vor dem Hintergrund des prozessorientierten Wissensmanagement mit CAKE durchgeführt worden. Hieraus ergeben sich spezielle Anforderungen, insbesondere an die Agilität von Prozessen, welche CAKE einbeziehen und technologisch unterstützen kann.
In den folgenden Abschnitten wird das Prozessmanagement erläutert und CAKE aus der Perspektive des Wissensmanagements skizziert. Die Untersuchung von Bauprozessen hinsichtlich des Potenzials für prozessorientiertes Wissensmanagement wird in Abschnitt 3 aufgezeigt. Im anschließenden Abschnitt wird ein mögliches Anwendungsszenario für den Bereich der Gewährleistung vorgestellt. Abgeschlossen wird der Beitrag mit einer Zusammenfassung der erarbeiteten Aspekte und einem Ausblick.

2 Prozessmanagement in der Baubranche

Der Wandel der Unternehmensstrukturen und -abläufe ist heute geprägt durch einen zentralen Begriff, den Prozessbegriff. Es zeichnet sich eine klare Entwicklung weg von der funktionalen hin zu einer prozessualen Betrachtung ab. Getrieben ist dieser Wandel von den veränderten Wettbewerbsbedingungen, z. B. den schnell veränderlichen Kundenanforderungen, Bedingungen am Markt oder der Zunahme des vorhandenen Wissens [Lehner, 2006]. Unternehmen haben inzwischen erkannt, dass das Prozessmanagement eine wichtige Aufgabe ist und diskutieren dessen Relevanz und Notwendigkeit nicht mehr.

2.1 Prozess- und Workflowmanagement

Gadatsch definiert das Prozessmanagement als einen „zentrale(n) Bestandteil eines integrierten Konzepts für das Geschäftsprozess- und Workflow-Management. Es dient dem Abgleich mit der Unternehmensstrategie, der organisatorischen Gestaltung von Prozessen sowie deren technischer Umsetzung mit geeigneten Kommunikations- und Informationssystemen.“ [Gadatsch, 2010] Das Prozessmanagement erfolgt auf fachlich-konzeptueller Ebene in drei Phasen: der Prozessabgrenzung, der Prozessmodellierung und der Prozessführung. Die Phasen umfassen das Identifizieren, Modellieren, Verbessern und Implementieren von Geschäftsprozessen [Brocke und Rosemann, 2010]. Hammer und Champy umschreiben den Begriff des Geschäftsprozesses als Sammlung von Aktivitäten, für die ein oder mehrere Arten von Inputs benötigt werden und die für den Kunden ein Ergebnis von Wert erzeugen [Hammer und Champy, 1994].

In einem Prozess[3] repräsentieren Tasks eine logische Arbeitseinheit (Aktivität, Aufgabe) und der Kontrollfluss die Reihenfolge in der Tasks in einem Prozess ausgeführt werden sollen. Eine Vielzahl von Beteiligten in verschiedenen Rollen auf unterschiedlichen Ebenen eines Unternehmens kennzeichnet das Prozessmanagement. Im Rahmen der Ressourcen- und Zeitplanung wird zugeordnet von wem, z. B. einem Akteur oder einer Anwendungsfunktion, eine Task ausgeführt werden soll [Leymann und Roller, 2000]. Auf der operativen Ebene erfolgt schließlich im Rahmen des Workflow-Managements die Workflow-Modellierung, -Ausführung und das Workflow-Monitoring. Workflows sind die Automatisierung von Geschäftsprozessen im Ganzen oder in Teilen, wodurch Dokumente, Informationen oder Aufgaben in einer durch Regeln festgelegten Reihenfolge von einem Bearbeiter zu einem Nächsten gereicht werden können [WfMC, 2002]. Die Workflow-Modellierung setzt auf den modellierten Geschäftsprozessen auf und erweitert sie um Spezifikationen, die für die automatisierte Ausführung des Prozesses durch ein Workflow-Management-System (WfMS), genauer gesagt der Workflow-Engine, benötigt werden.

2.2 Nutzung in der Baubranche

Während der durchgeführten Analyse konnte festgestellt werden, dass zurzeit Software zum Prozessmanagement im Bauwesen nur eingeschränkt verfügbar ist. Es existiert u. a. Software zum Projektmanagement, welche rechnergestützte Dokumentenmanagement-Steuerung, Planung und Budget-Tracking ermöglicht. Auch wird Software zum Service-Management angeboten, die Prozessabläufe für den Service und die Wartung von statischen, mechanischen oder anderen Komponenten des Bauwesens automatisiert [BVBS, 2013].

In Abbildung 1 ist als erste Demonstration der Grundidee ein einfacher Workflow im WfMS CAKE aus dem Anwendungsbereich des Bauwesens abgebildet. Der Prozess bildet die Bauabnahme zweier geschlossener Bauleistungsbereiche auf einer Baustelle durch den Bauleiter vor Ort ab. Im unteren Drittel sind die Ein- und Ausgabe-Daten abgebildet, die während der Ausführung des Prozesses entweder benötigt (Eingabe) oder erzeugt (Ausgabe) werden. Workflows dieser Art bilden das prozedurale Wissen eines Bauunternehmens ab und können durch den Fachanwender (z. B. den Projektmanager) modelliert und in entsprechenden Repositories zur Wiederverwendung abgelegt werden.

Im Anschluss an die Modellierung kann dann die Ausführung der modellierten Workflows unter Kontrolle des WfMS erfolgen. Nach dem Starten des Workflows wird die erste Task eines Prozesses aktiviert. Im vorherigen Beispiel wäre dies, die Aufgabe ‚Tagesdaten‘ zu erfassen. Diese Aufgabe erscheint sogleich auf der Worklist (Aufgabenliste) der für die Bearbeitung zugewiesenen Akteure, in diesem Falle der Bauleiter vor Ort. In der Abbildung 2 ist die noch ausstehende Aufgabe auf der Worklist des Bauleiters zu sehen.

Die herkömmliche Workflow-Technologie, wie sie in heutigen Software-Produkten zu finden ist, erlaubt bei der Ausführung keine Änderungen an bereits laufenden Prozessen ohne den Workflow komplett abzubrechen und neu zu starten [Minor *et al*., 2008]. CAKE verfügt hingegen über die Möglichkeit flexibel mit ad-hoc auftretenden Änderungen von bereits geplanten und laufenden Prozessen umzugehen; es können Änderungen an Workflows während der Ausführungsphase vorgenommen werden. Eine solche Technologie wird als agile Workflow-Technologie bezeichnet [Schmalen, 2011; Minor *et al*., 2008; Weber und Wild, 2005; Reichert und Dadam, 1998] und ist ein relativ junges Forschungsgebiet. Die agile Workflow-Engine steuert die Workflow-Ausführung, lässt

[2] Weitergehende Informationen unter http://www.cake.wi2.uni-trier.de/.

[3] Geschäftsprozesses sind Prozesse, die eine betriebswirtschaftliche Ausrichtung aufzeigen.

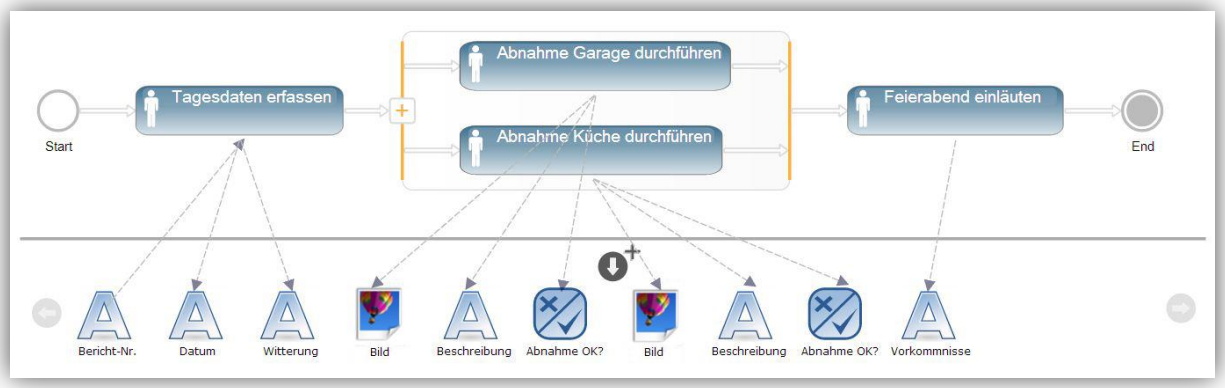


Abbildung 1: Ein einfacher Workflow im CAKE System

Abbildung 2: Worklist im CAKE System

Änderungen währenddessen zu und leitet diese Information an die beteiligen Personen weiter.

Hierbei soll eine geeignete Modellierungsassistenz die verantwortlichen Personen nicht nur bei der Erstellung von Workflows erfahrungsbasiert unterstützen, sondern auch während der Änderung von Workflows. Das Erstellen und das flexible Anpassen von Prozessen soll dabei unter dem bestmöglichen Einsatz von bestehendem Wissen und vorhandenen Erfahrungen erfolgen. Sprich, es sollen Workflows aus früheren, erfolgreichen Projekten wieder berücksichtigt werden, um dabei die Erfordernisse des aktuellen Projekts anstatt standardisierte Prozess-Pools /-Repositories zu verwenden. Möchte ein Architekt beispielweise für einen Kunden einen Bauantrag bei einer zuständigen Behörde stellen, so könnte er auf vergangene, erfolgreich abgeschossene Prozesse der Bauantragsstellung zurückgreifen und sie als Ausgangsbasis zur Modellierung des neuen Bauantragsprozesses verwenden. Die Modellierung würde dabei systemseitig durch eine ähnlichkeitsbasierte Suche unterstützt werden. Die Suche würde dem Architekten dann den besten Prozess (‚best-matching' Case) zur Suchanfrage aus dem Prozess-Repository zur Verfügung stellen [Bergmann & Gil 2012].

Konzepte oder sich im Einsatz befindende Systeme zum prozessorientierten Wissensmanagement auf der Basis eines WfMS, die alle Phasen eines Bauvorhabens[4] unterstützen, sind zurzeit nicht verfügbar. So stellte bereits Mikuláková fest, dass „in der Ausführungsphase des Bauprozesses […] Werkzeuge und Methoden des Wissensmanagements in der Regel nicht eingesetzt [werden]." [Mikuláková, 2010] Zurzeit wird CAKE nicht im Bauwesen eingesetzt, dies ist aber langfristig zur Unterstützung von Bauvorhaben geplant. Die Ausführungen zu den möglichen Einsatzszenarien haben folglich nur exemplarischen Charakter. Wissensbasierte Modelle, welche in den letzten Jahren entwickelt wurden, fokussieren auf die Unterstützung der Bauwerksplanung [Weber *et al.*, 2010]. Dies erfolgt unter der Nutzung von konkreten Daten aus Gebäudemodellen, also auf der Basis von Bauwerksinformation und nicht anhand des Bauablaufs.

3 Potentialanalyse

Im Folgenden wird das Ergebnis der Potentialanalyse des prozessorientierten Wissensmanagement für die Baubranche vorgestellt. Dabei werden zunächst verschiedene Blickwinkel / Sichten auf Bauprozesse aufgezeigt und es erfolgt eine Differenzierung sowie weitergehende Klassifizierung von Bauvorhaben. Die detaillierte Einteilung von Bauvorhaben ist notwendig, um den zu untersuchenden Prozessbereich nachvollziehbar einschränken zu können. Im Anschluss werden die Kriterien erläutert, anhand derer der Prozessbereich zum Aufbau eines prozessorientierten Wissensmanagements identifiziert werden soll. Die Kriterien wurden u. a. auf der Basis eines Experteninterview mit einem mittelständigen Bauunternehmen[5] im

[4] Vgl. Abschnitt 3.1 bzgl. der Phasen eines Bauvorhabens.

[5] Die Experteninterviews wurden am 29.04.2013 mit der Geschäftsleitung, dem Manager Prozess- und Ablauforganisation sowie Mitarbeitern aus dem operativen Geschäft geführt.

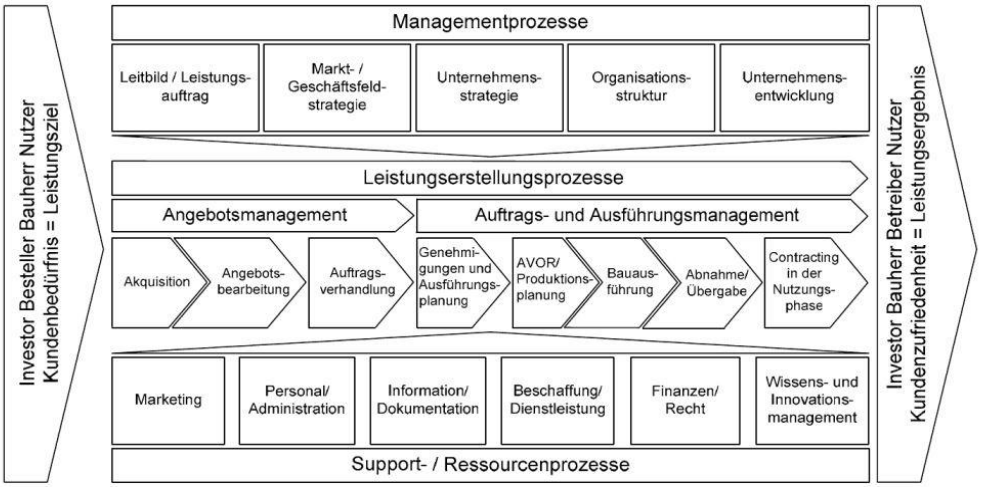


Abbildung 3: Prozessmodell der Wertschöpfungskette eines Bauunternehmens [Girmscheid, 2003]

Bereich des Hochbaus[6] identifiziert und insbesondere vor dem Hintergrund der exemplarischen Umsetzung des prozessorientierten Wissensmanagement mit CAKE ausgewählt. Von ‚klassischen Kriterien'[7] zur Prozessauswahl wie sie im Bereich des Prozessmanagement und der kontinuierlichen Verbesserung verwendet werden [Sieben, 2012], würde in dieser frühen Phase der Analyse abgesehen.[8] Im Abschnitt 3.3 werden die Ergebnisse der Potentialanalyse, gliedert nach den Phasen eines Bauvorhabens (Planung, Bauausführung, Nutzung) vorgestellt. Es werden die als geeignet identifizierten Prozessbereiche aufgezeigt und die Ergebnisse einer ersten Extraktion dargestellt.

3.1 Strukturierung von Geschäftsprozessen im Bauwesen

Die Vielfalt und Komplexität von Bauaufgaben, die mitunter große Anzahl an Beteiligen und nicht zuletzt die Art und Größe eines Bauvorhabens, welche sich von einem Einfamilienhaus bis zum Bau eines Krankenhauses oder Flughafens erstrecken kann, ist für einen Fachfremden schwer zu überblicken. Eine Abgrenzung und Systematisierung von Bauprozessen erleichtert den Zugang zur Prozesslandschaft des Bauwesens.

Verschiedene Sichten auf Bauvorhaben

Die Sicht auf ein Bauvorhaben und die damit verbundenen Bauprozesse kann aus verschiedenen Betrachtungswinkeln erfolgen:

a. Produktorientiert: Es kann das Bauwerk und damit das Bauprojekt im Fokus stehen. Nach Girmscheid ist ein Bauwerk mit seinen Funktionen sowie seinen Planungs-, Ausführungs- und Bewirtschaftungsprozessen als ein komplexes System zu sehen [Girmscheid, 2010]. Durch die Bauwerkserstellungs- und Nutzungsprozesse erfolgt eine systematische Abarbeitung der Aufgaben in den einzelnen Phasen.
b. Prozessorientiert: Aus der Sicht eines Unternehmens gliedern sich Geschäftsprozesse entlang der Wertschöpfungskette in den Primärprozess der Leistungserstellung und den Sekundärprozess der steuernden und unterstützenden Prozesse eines Bauvorhabens. Das Prozessmodell der Wertschöpfungskette eines Bauunternehmens ist in Abbildung 3 dargestellt.

Hierbei ist zu beachten, dass sich die Anbieterprozesse je nach Leistungsanbieter im Bauwesen unterscheiden. Der Bauherr hat zu entscheiden, in welcher der verschiedenen Projektabwicklungsformen er sein Bauvorhaben realisie-

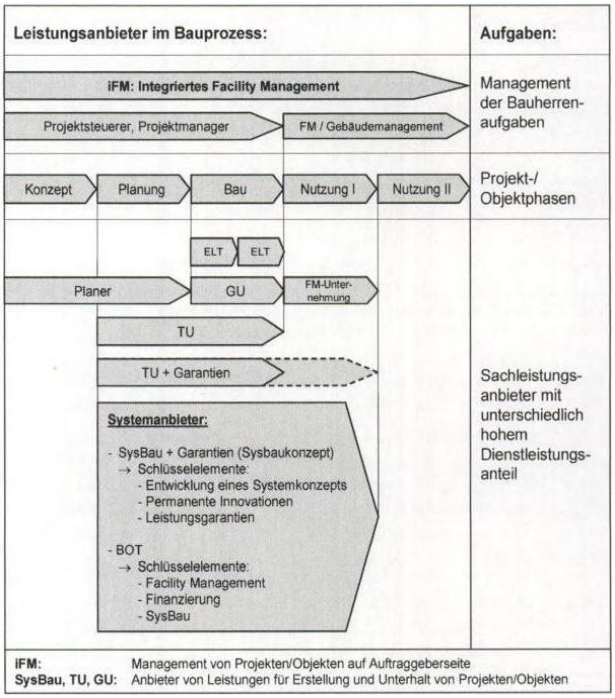


Abbildung 4: Eingliederungssystematik der Leistungsanbieter im Bauwesen [Girmscheid, 2010]

[6] Das Bauunternehmen ist spezialisiert auf den Bau von Ein- und Zweifamilienhäusern.

[7] Nach Sieben sind dies u. a. Prozesse mit kritischen Erfolgsfaktoren, hohen Auswirkungen auf die Kundenzufriedenheit sowie Prozesse die neu und ohne Routine sind.

[8] Eine nähere Betrachtung dieser Kriterien ist erst sinnvoll, wenn der Prozessbereich festgelegt ist und konkrete Informationen vorhanden sind, die eine Beurteilung der Kriterien ermöglichen.

ren möchte. Die Projektabwicklungsformen

- Einzelleistungsträger (ELT),
- Generalunternehmer (GU) und
- Totalunternehmer (TU)

sind die drei am häufigsten auftretenden Formen. In der Abbildung 4 ist eine Eingliederungssystematik für Leistungsanbieter im Bauwesen abgebildet. Die Abwicklungsform Systemanbieter ist die Ausführungen dieses Beitrags unerheblich.
Wie in der Abbildung 4 zu sehen ist, unterscheiden sich die verschiedenen Projektabwicklungsformen in Art und Umfang der erbrachten Leistungen, die für die Erstellung und den Unterhalt von Bauprojekten erbracht werden. Entscheidet sich der Bauherr bspw. für die Form der ELT, so vergibt er alle Aufträge individuell an Einzelplaner und -unternehmer. Die ausführenden Unternehmen tragen keine Verantwortung für das Gesamtprojekt. Wählt der Auftraggeber hingegen einen TU, so ist dieser für die Planung und Erstellung des Bauvorhabens zuständig. Neben diesen Leistungsträgerorganisationen existieren noch Leistungsvermittlerorganisationen, wie General- und Totalübernehmer. Diese treten zwar mit dem Bauherrn in Verbindung, fungieren jedoch als reine Managementorganisation und bieten nur Dienstleistungen und Garantien an. Sämtliche Planungs- und Bauleistungen werden von Subunternehmern erbracht [Girmscheid, 2010].

Differenzierung nach Bauwerken

Neben den beiden beschriebenen Sichten auf ein Bauvorhaben, ist zu beachten, dass Bauwerke grundlegend zwischen Hoch- und Tiefbau differenziert werden. Diese Trennung erfolgt, da sie einen entscheidenden Einfluss auf die Bauprozesse hat und z. T. andere Bauabläufe nach sich zieht. Weiterhin wird nach Trassen-, Kunst- und Untertagebauten unterschieden.

Weitere Klassifizierungen

Nach der Honorarordnung für Architekten und Ingenieure (HOAI) werden Bauvorhaben nach folgende Investitionsarten unterschieden: Hochbauten, Ingenieurbauwerke, Verkehrsanlagen, Anlagenbauten, Altlastensanierung [HOAI, 2009]. Nach Structurae[9] werden Bauwerke nach der **Bauweise / Bauwerktypen** klassifiziert: es wird u. a. unterschieden in Brücken, Gebäude, Hochsee- und Seebauwerke, Türme und Masten. Ebenso kann nach der **Nutzung / Funktion** differenziert werden: Wohnen (Ein-, Mehrfamilienhäuser, Altenwohnheime, …), Ver- und Entsorgung, Transport und Verkehr, Stadtplanung, Industrie und Handel [Sommer, 2009]; nur um einige Kategorien zu nennen. Weitere Klassifikationen könnten u. a. nach Bauträger, Baustil oder Bauverfahren erfolgen. Auch ist es möglich nach der **Bauform** zu unterschieden, also der typologischen Einordnung eines Bauwerks in eine bestimmte Kategorie oder Gruppe. Z. B. die Bauformen eines Gebäudes: in Bungalow und Hochhaus oder Einzel- und Reihenhaus. Eine Einteilung nach der **Baukonstruktion** kann anhand von verschiedenen Kriterien wie nach Bauteilen, nach Material oder auch nach Bauphasen, wie z. B. Rohbau und Innenausbau erfolgen [Sommer, 2009].

[9] Struturae ist eine internationale Datenbank für Ingenieurbauwerke. Siehe http://de.structurae.de/.

3.2 Kriterien für die Auswahl eines Prozessbereichs

Die Auswahl eines Prozessbereichs erfolgt anhand von Kriterien, die einen erfolgversprechenden Ansatz zum Aufbau eines prozessorientierten Wissensmanagements mit CAKE erwarten lassen.

Der Einsatz eines prozessorientierten Wissensmanagements kann die Agilität der Prozesse mit einbeziehen, technologisch unterstützen und so eine automatische Ablaufsteuerung ermöglichen. Die erfahrungsbasierte Wissensmanagement-Komponente gestattet dabei das Erstellen und Anpassen von Bauprozessen unter dem bestmöglichen Einsatz von bestehendem Wissen und vorhandenen Erfahrungen. Das **Auftreten von Agilität** ist also ein wichtiges Kriterium zur Bestimmung des Prozessbereichs.

Das Vorhandensein einer großen **Vielfalt an Prozessen** in einem Prozessbereich ist ebenfalls von Relevanz. Die Wissensbasis (Repository) muss erst mit einer entsprechenden Fülle an Workflows aus vergangenen, erfolgreich abgeschlossen Workflows aus einem Prozessbereich gefüllt werden, bevor die Nutzung einer erfahrungsbasierten Modellierungsassistenz sinnvoll ist. Die Beantwortung einer Suchanfrage seitens eines Projektbeteiligten kann nur Prozesse zurückliefern, die zuvor in das System eingepflegt wurden. Ebenfalls sollten die Prozesse aus einem Prozessbereich eine große Varianz aufweisen. Bei fehlender Varianz würde das Problem der Workflow-Auswahl gar nicht auftreten. Es ist also eine bestimmte Mindestkomplexität notwendig, damit die systematische und technische Unterstützung einen Nutzen bringen kann. Auch die Anzahl der Prozesse ist wichtig. Es wird angenommen, dass eine größere Anzahl und Variation von Bauprozessen aus einem bestimmten Bereich eines Bauvorhabens, die Wahrscheinlichkeit verbessern, dass der gefundene ‚best-matching' Case den Erfordernissen des aktuell zu modellierenden Bauprozesses entspricht und von dem Prozessmodellierenden als hilfreich eingestuft wird.[10]

Auch ist die **Zugänglichkeit zu Prozessinformationen** ein wichtiges Kriterium. Eine unkomplizierte Beschaffung von Bauworkflows erleichtert den ersten Aufbau des Repositories. Ebenso ist die **Perspektive auf ein Bauvorhaben** von Relevanz. Steht das Bauwerk im Vordergrund der Betrachtung, so wird nach dessen Planung vorwiegend auf die Steuerung und Kontrolle des Bauvorhabens nach diversen Gesichtspunkten fokussiert und der Projektstatus u. a. via graphischen Darstellungen (z. B. Gantt-Diagrammen) wiedergegeben. Liegt der Schwerpunkt der Betrachtung auf der eines Bauunternehmens, kann in allen drei Phasen eines Bauprojekts (Planungs-, Bauausführungs- und Nutzungsphase) neben der prozessunterstützenden Modellierung auch die Ausführung der Workflows systematisch und technisch unterstützt werden. Dies ist wichtig, da ein hohes Aufkommen an Agilität insbesondere bei der Ausführung von Bauprozessen erwartet wird.

Ein weiteres Kriterium ist der **Bedarf seitens der Beteiligten** an einer systematischen und technischen Unterstützung des Bauprozesses. Die am Bauprozess Beteiligten sollten eine derartige Unterstützung für sinnvoll erachten und auch anwenden wollen. Auch kann in Betracht

[10] Diese Annahme abstrahiert von dem Umstand, dass die Art der ähnlichkeitsbasierten Suche einen Einfluss auf das Suchergebnis bzw. die Eignung des gefundenen Prozesses zur Modellierung des gegenwärtigen Prozesses hat.

gezogen werden, ob der gewählte Prozessbereich in einem **abgeschlossenen Bereich** liegt: Die Phasen Planung und Bauausführung sind stark miteinander verwoben. Bspw. gilt ein Bauprojekt als abgeschlossen, wenn die Nutzungsphase beginnt.

3.3 Identifikation eines Prozessbereichs und Extraktion von Bauworkflows

Zur Identifikation eines Prozessbereiches, der für das prozessorientierte Wissensmanagement in der Baubranche geeignet ist, müssen die vorhandenen Prozesse im Bauwesen gesichtet und hinsichtlich der oben eingeführten Kriterien untersucht werden. Die Analyse ist der erste Schritt zum Realisieren einer Wissens- und Prozessmanagementunterstützung im Bauwesen. Sie ermöglicht den initialen Aufbau des Repositories, um die erfahrungsbasierte Komponente zum Erstellen und Ändern von Bauworkflows zu realisieren.

Phase der Bauausführung

Bei der ersten Auswahl wurde auf einen Bereich fokussiert, der ein hohes Maß an Flexibilität bei der Abwicklung von Bauprojekten erfordert bzw. ein hohes Aufkommen an unvorhersehbaren Änderungen und nicht mehr rückgängig machbaren Aufgaben aufweist. Der ausgewählte Prozessbereich wurde wie folgt eingeschränkt: es wurden ausschließlich Hochbauten der Bauform Gebäude mit der Nutzung / Funktion, Wohnen (Ein- und Mehrfamilienhäuser), und der Baukonstruktion nach den Bauphasen, Roh- und Innenausbau, aus der Sicht eines Bauunternehmens berücksichtigt, welches als Leistungsträger (TU, GL oder ELT) fungiert. Die Konstruktionsphase, die ein Teilbereich der Bauausführung ist, wurde bewusst gewählt: In dieser Bauphase wird ein starkes ‚Auftreten von Agilität' erwartet, da dort eine Vielzahl an Beteiligten aufeinander trifft und deren Arbeitsschritte abgestimmt sein müssen, auch wenn unvorhersehbare Störungen oder Änderungen im Arbeitsablauf auftreten. Der Bereich des Hochbaus/Wohnen/Ein- und Mehrfamilienhäuser wurde aus der Erwartung heraus gewählt, dass hier eine größere Anzahl an Bauprojekten vorhanden ist, welche mit einer ‚Vielfalt von Prozessen' einhergeht. Eine Einschränkung der Baukonstruktion auf die Bauphasen Roh- und Innenausbau war notwendig, da für eine erste formative Evaluierung der erfahrungsbasierten Komponenten eine bestimmte Anzahl von exemplarischen Bauprozessen notwendig ist, welche aus einem relativen engen Anwendungsbereich stammen. Die ‚Informationsbeschaffung' wurde aufgrund der größeren Anzahl an Bauprojekte in diesem Bereich ebenfalls als unkompliziert eingeschätzt. Die Auswahl der Bauprozesse wurde aus der ‚Sichtweise eines Bauunternehmens' gewählt, da das prozessorientierte Wissensmanagement insbesondere die Workflowausführung unterstützen kann.

Extraktionsergebnis - Bauausführungsphase

Auf der Basis einer Literaturrecherche[11] konnten keine soliden Prozesse aus dem oben beschriebenen, engen Anwendungsbereich zusammengetragen werden. Es war auch nicht möglich, nur einzelne wenige Bauprozesse aus diesem Prozessbereich zu finden. Dokumentationen von erfolgreichen Bauprojekten auf dieser Detailierungsebene und für diesen Ausführungsbereich sind nicht verfügbar, auch nicht für andere Bereiche der Bauausführung. Dies mag an der Heterogenität von Bauprojekten liegen und dem damit einhergehenden Unikatscharakter sowie der Tatsache, dass aufgrund der gegebenen Bearbeitungskomplexität von Bauprozessen in der Ausführung das benötigte Expertenwissen „[…] derzeit stark verteilt, nicht formalisiert und personenabhängig" ist [Mikuláková, 2010]. Auch könnte in dem Experteninterview mit dem mittelständischen Bauunternehmen, von diesem bestätigt werden, dass der Bauablaufprozess in der Ausführung in ihren Unternehmen nicht IT gestützt ist und dies auch mittelfristig nicht geplant ist.

Phase der Planung

Die Phase der Planung eines Bauvorhabens ist von geringerer ‚Agilität' gekennzeichnet als die Phase der Bauausführung, weshalb diese Phase erst in einem weiteren Schritt genauer untersucht wurde. Im Rahmen der Literaturrecherche konnte festgestellt werden, dass die Phase der Bauplanung stärker dokumentiert ist und eine Reihe von Bauprozessen detailliert betrachtet wurden [Mikuláková, 2010; Ansorge, 1998; Scheifele, 1991]. Eine mögliche Erklärung ist, dass diese Phase ein integraler Bestandteil von Projektmanagement-Software ist, die bereits seit vielen Jahren im Bauwesen eingesetzt wird [Bauer, 2004]. In dem bereits erwähnten Experteninterview mit dem mittelständischen Bauunternehmen stellte sich jedoch heraus, dass in der Planungsphase kein flexibler Ansatz zur systemseitigen Unterstützung gewünscht wird. Es wird vielmehr ein fest definierter Rahmen zur Modellierung der Planungsprozesse gefordert. Der Hintergrund dieser Einstellung ist, dass die Unternehmensleitung so eine maximale Kontrolle über den Bereich der Planung gewährleisten kann und einer möglichen Zertifizierung bzgl. des Qualitätsmanagements (Dokumentation von Geschäftsprozessen und Kontrollen) nichts im Wege steht. Es bleibt festzuhalten, dass ein ‚Bedarf seitens der Beteiligten' an einer systematischen und technischen Unterstützung in der Planungsphase nicht gewünscht ist. Von einer weiteren Untersuchung des Prozessbereiches der Planung wurde folglich abgesehen.

Phase der Nutzung

Während des Experteninterviews wurde auch der Bereich der Nutzungsphase elaboriert. Im Rahmen der hier zu leistenden Gewährleistung verfügt das Unternehmen nur über eine rudimentäre, softwaretechnische Unterstützung; es würde eine Entwicklung in diesem Bereich aber sehr begrüßen, womit ein ‚Bedarf seitens der Beteiligten' gegen ist. Die fehlende technische Ablaufunterstützung begründet das Bauunternehmen mit der Heterogenität der Prozesse. Momentan wird jede Anzeige eines möglichen Mangels papiergebunden und ohne eine strukturierte Vorgehensweise bearbeitet.

Die Phase der Gewährleistung verspricht ein hohes Potential an Agilität: unvorhersehbare Ereignisse oder Zustände, die noch nicht berücksichtigtet wurden, können hier auftreten, wie z. B. eine Kombination aus verschiedenen Baumängeln. Auch bietet diese Phase den Vorteil, dass es sich um einen abgeschlossenen Bereich und zwar dem der Nutzungsphase handelt. Bauunternehmen können das entwickelte System zum prozessorientierten Wissensmanagement unabhängig von ihrer im Einsatz befindlichen Systeme zur Planung und Ausführung einführen

[11] Es wurde einschlägige Fachliteratur gesichtet via Online-Fernleihe der Digitalen Bibliotheken sowie sonstigen Internetrecherchen und Recherchen in Literaturverzeichnissen.

und Schritt für Schritt die Wissensbasis (Repository) mit Prozessen aus dem Bereich der Gewährleistung aufbauen. Es wird insbesondere von einer hohen Akzeptanz des Systems ausgegangen, besonders in Unternehmen, die bisher eine rein papiergebundene Bearbeitung ohne entsprechende Prozessunterstützung vornehmen.

Im Zuge der Extraktion von Bauprozessen aus der Nutzungsphase wird im nächsten Abschnitt ein Anwendungsszenario zu diesem Bereich aufgezeigt und ein möglicher Nutzen aus dem Einsatz eines prozessorientierten Wissensmanagements skizziert.

4 Anwendungsszenario

Bei dem gewählten exemplarischen Anwendungsszenario handelt es sich um einen stark simplifizierten Gewährungsleistungsprozess der Nutzungsphase aus Sicht eines Bauunternehmers.

Der Prozess der Gewährleistung wird angestoßen durch einen Brief, Fax oder Dokumenten-Anhang in einer Email, in welcher der Mangel angezeigt wird. Der für diesen Bereich zuständige Sachbearbeiter kann den entsprechenden Prozess starten. Angenommen, dass die Wissensbasis bereits mit wiederverwendbaren Workflows gefüllt ist, kann der Sachbearbeiter z. B. nach dem Begriff ‚Bodenbelag' suchen und einen bereits erfolgreich abgeschlossen Workflow zur Mangelbeseitigung an einem Bodenbelag als Vorlage erhalten. Diese Prozess- / Workflow-Vorlage enthält neben standardisierten Aufgaben, wie der Erfassung aller wichtigen Daten einer Mangelanzeige oder die Überprüfung des Einhaltens der generellen Anzeigefristen, auch Arbeitsschritte, die speziell auf die Beseitigung eines Mangels an einem Bodenbelag abgestimmt sind. Die dem Sachbearbeiter bereitgestellte Prozess- / Workflow-Vorlage enthält jedoch die notwendigen Arbeitsschritte zur Beseitigung eines Schadens an einem Linoleumboden. Mehrere Schweißnähte waren gebrochen. Im aktuellen Fall wurde jedoch ein Schaden an einem Teppichboden gemeldet. Folglich muss von dem Sachbereiter die Vorlage dahingehend angepasst werden, dass der Arbeitsschritt ‚Schweißnähte überarbeiten' gelöscht und durch den Schritt ‚Teppichstück erneuern' ergänzt wird.

Im Zuge eines prozessorientierten Wissensmanagements kann ein Mängelmanagement etabliert werden, welches über die Verwaltung von gemeldeten Mängeln hinausgeht. Es ermöglicht neben einer flexiblen prozessgestützte Bearbeitung und der Einbindung aller Beteiligten auch ein Erstellen und Verändern von Prozessen unter dem bestmöglichen Einsatz von bestehendem Wissen und vorhandenen Erfahrungen. Im vorausgegangen Anwendungsszenario wäre z. B. denkbar, dass der zuständige Sachbearbeiter relativ neu im Unternehmen ist und mit dem generellen Ablauf eines Gewährleistungsprozesses noch nicht vertraut ist. Es könnte sein, dass ihm nicht bewusst ist, dass das Anzeigen eines Mangels innerhalb einer bestimmen Frist erfolgen muss. Folglich könnte er die Reparatur des Teppichbodens veranlassen, obwohl das Bauunternehmen dazu gar nicht mehr verpflichtet ist. Ist die Aufgabe der Fristenüberprüfung jedoch Bestandteil der Prozess- / Workflow-Vorlage kann der Sachbearbeiter darauf aufmerksam werden und eine mögliche Fehlentscheidung, welche mit Kosten für das Unternehmen verbunden wäre, vermeiden. Auch ist eine schnellere Bearbeitung des Vorgangs möglich, da der Sachbearbeiter von den Erfahrungen seiner Kollegen profitieren kann und sich die einzelnen Arbeitsschritte nicht selbst erarbeiten muss, sondern eine Vorlage zu einem ähnlichen Fall verwenden kann.

Ein prozessorientiertes Wissensmanagement lässt erwarten, dass sich der Zeit- und Kostenaufwand eines Unternehmens der Baubranche erheblich reduzieren lässt. Insbesondere die gespeicherten Erfahrungen in einer Wissensbasis, z. B. die effiziente Schadensbehebung eines Gewährleistungsfalles aus einer Kombination von verschiedenen Baumängeln, kann Sachbearbeitern einen wesentlichen Wettbewerbsvorteil gegenüber Konkurrenten bieten. Auch ermöglicht die Bearbeitung von Vorgängen auf Basis von bereits erfolgreich gelösten Prozessen ein Freisetzen von Ressourcen und bietet damit Raum zur Steigerung der Produktivität.

In einem weiteren Forschungsschritt ist der potentielle Nutzen, den ein prozessorientiertes Wissensmanagement in der Phase der Nutzung bieten kann, genauer zu untersuchen: hierzu sind weitere Prozesse aus dem Bereich der Gewährleistung zu identifizieren und zu extrahieren. Die ausgewählten Bauprozesse sind in formale Bauworkflows zu überführen, um eine erste formative Evaluierung vorzunehmen. Dabei ist zu untersuchen, welche Anpassungen für eine Umsetzung in CAKE vorzunehmen sind, um Bauworkflows abbilden zu können und es muss festgelegt werden, anhand welcher Kriterien die empirische Untersuchung durchzuführen ist.

5 Zusammenfassung und Ausblick

In dem vorliegenden Beitrag wurden die Vorgehensweise und die Ergebnisse einer Potentialanalyse des prozessorientiertes Wissensmanagement aufgezeigt. Als Ergebnis kann festgehalten werden, dass kurzfristig die Auswahl der exemplarischen Bauprozesse aus dem Bereich Nutzungsphase / Gewährleistung am Vielversprechendsten (vgl. Abschnitt 4) ist. Auch wird gegenwärtig nicht angenommen, dass in den Phasen der Planung und Bauausführung ein System zum prozessorientierten Wissensmanagement zum Einsatz kommen wird (vgl. Abschnitt 3.3). Dem Bereich der Bauausführung wird aufgrund des hohen Maßes an Flexibilität, welches bei der Abwicklung von Bauprojekten notwendig ist, langfristig ein großes Potential für die Realisierung eines prozessorientierten Wissensmanagements zugesprochen. Zurzeit ist keine ausreichende Prozessdokumentation in diesem Bereich vorhanden, wie in Abschnitt 3 (Extraktionsergebnis - Bauausführungsphase) bereits ausgeführt wurde. Auch wird es als kritisch angesehen, mit der Einführung eines prozessorientierten Wissensmanagement in einem Bereich zu beginnen, der eine so hohe Bearbeitungskomplexität aufweist. So haben die befragten Personen im Experteninterview über eine mögliche systematische und technische Unterstützung in der Phase der Bauausführung noch nicht nachgedacht und sehen auch mittelfristig noch keinen Bedarf für eine Einführung. Die Akzeptanz eines solchen Systems wird als größer eingeschätzt, wenn die Beteiligten damit in einem weniger komplexen Umfeld, wie z. B. der Gewährleistung, zum ersten Mal in Kontakt kommen. Sind die Beteiligten von Bauvorhaben mit dem prozessorientierten Wissensmanagement erst einmal vertraut, wird eine Einführung vielleicht auch schon seitens der Beteiligten gewünscht.

Acknowledgements

This work is part of the WEDA project (Web-based design, reuse, documentation and execution of agile Workflows). WEDA is funded by Stiftung Rheinland-Pfalz für Innovation, grant no. 974.

Literatur

Ansorge, D. (2008). Pfusch am Bau. Band 6: Planung und Bauüberwachung (Fraunhofer IRB Verlag).

Bauer, T. (2004). Kooperation von Projekt- und Workflow-Management-Systemen. Informatik Forschung und Entwicklung *19*, 74–86.

Bergmann, R., und Gil, Y. (2012). Similarity Assessment and Efficient Retrieval of Semantic Workflows. Information Systems, Special Issue on Processoriented Case-Based Reasoning, Forthcoming.

Brocke, J. vom, und Rosemann, M. (2010). Handbook on business process management introduction, methods and information systems (Berlin; London: Springer).

BVBS Bausoftware-Definition – Bundes Verband Bausoftware.

Gadatsch, A. (2005). Grundkurs Geschäftsprozess - Management [Elektronische Ressource] : Methoden und Werkzeuge für die IT-Praxis: Eine Einführung für Studenten und Praktiker (Wiesbaden: Vieweg Teubner Verlag).

Girmscheid, G. (2003). Wettbewerbsvorteile nutzen - Konzepte für Bauunternehmen Markt- und ressourcenbasierte Perspektiven (Bern: hep, Baufachverl.).

Girmscheid, G. (2010). Projektabwicklung in Der Bauwirtschaft: Wege zur Win-Win-Situation für Auftraggeber und Auftragnehmer (Springer DE).

Görg, S., Bergmann, R., Gessinger, S. und Minor, M. (2013). Real-Time Collaboration and Experience Reuse For Cloud-Based Workflow Management Systems. In Proceedings of 15th IEEE Conference on Business Informatics, Workshop on Social Business Process Management, p. to be printed.

Hammer, M., und Chamy, J. (1995). Business reengineering: die Radikalkur für das Unternehmen; (Frankfurt/Main [u.a.: Campus-Verl.).

HOAI (2009), Honorarordnung für Architekten und Ingenieure, http://www.hoai.de/online/HOAI_2009/HOAI_2009.php#anlage3 [Online; abgerufen 24-Jun-2013].

Kleiner, S., und Kirsch, M. (2009), Einführung des prozessorientierten Wissensmanagements in Produktentwicklungsprozessen Praxisbericht aus der Entwicklung von Hochauftriebsysstemen für Verkehrsflugzeuge, KnowTech, Bad Homburg.

Lehner, F. (2006). Wissensmanagement: Grundlagen, Methoden und technische Unterstützung (München [u.a.]: Hanser).

Leymann, F., und Roller, D. (2000). Production workflow: concepts and techniques (Prentice Hall PTR).

Mikuláková, E. (2011). Wissensbasierte Bauablaufplanung mit fallbasiertem Schließen / Eva Mikuláková. Hrsg. von der Professur Informatik im Bauwesen (Weimar: Verl. der Bauhaus-Univ.).

Minor, M., Tartakovski, A., Schmalen, D., und Bergmann, R. (2008). Agile Workflow Technology and Case-Based Change Reuse for Long-Term Processes. International Journal of Intelligent Information Technologies, 4(1), 80–98.

Orth, R., Finke, I., und Voigt, S. (2009). ProWis II – Wissensmanagement im Mittelstand fördern.

Probst, G., Raub, S., und Romhardt, K. (2012). Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen (Wiesbaden: Gabler).

Reichert, M., und Dadam, P. (1998). ADEPT flex—supporting dynamic changes of workflows without losing control. Journal of Intelligent Information Systems *10*, 93–129.

Reinmann-Rothmeier, G., und Mandl, H. (1997). Wissensmanagement: Phänomene - Analyse - Forschung - Bildung.

Remus, U. (2002). Prozessorientiertes Wissensmanagement. Konzepte und Modellierung. Ph.D. Thesis, Universität Regensburg, http://epub.uni-regensburg.de/9925/, [Online; abgerufen 16-Jun-2013].

Rose, T., Fünffinger, M., Knublauch, H., und Rupprecht, C. (2002). Prozessorientiertes Wissensmanagement. KI *16*, 19–24.

Scheifele, D.R. (1991). Bauprojektablauf: Grundlagen und Modelle für eine effiziente Ablaufplanung im Bauwesen (Verlag TÜV Rheinland).

Schmalen, D. (2011). Adaptives Workflow Management - Referenzmodell und Umsetzung. Ph.D. Thesis, Dr. Hut Verlag, 80538 München. ISBN 9783868539905.

Schnauffer, H.-G., Stieler-Lorenz, B., und Peters, S. (2004). Wissen Vernetzen: Wissensmanagement in Der Produktentwicklung (Springer DE).

Sieben, S. (2012). Prozessmanagement und kontinuierliche Verbesserung: mit Trainingsleitfaden (BoD – Books on Demand).

Sommer, H. (2009). Projektmanagement im Hochbau 35 Jahre Innovationen bei Drees & Sommer (Heidelberg; New York: Springer).

Weber, B., und Wild, W. (2005). Towards the agile management of business processes. In Professional Knowledge Management, (Springer), pp. 409–419.

Weber, M., ., Langenhan, C., Roth-Berghofer, T., Liwicki, M., Dengel, A., und Petzold, F. (2010). a.SCatch: Semantic Structure for Architectural Floor Plan Retrieval. In Case-Based Reasoning. Research and Development, I. Bichindaritz, and S. Montani, eds. (Springer Berlin Heidelberg), pp. 510–524.

WfMC (1999). W. M. Coaltion, "Terminology and glossary", http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf, [Online; abgerufen 24-Jun-2013].

Wildemann, H. (2003). Wissensmanagement : ein neuer Erfolgsfaktor für Unternehmen / Horst Wildemann (München: TCW, Transfer-Centrum-Verl.).

A Case Study in Knowledge Acquisition in the Domain of Cataract Surgery

Jochen Reutelshoefer and Joachim Baumeister
denkbares GmbH
Friedrich-Bergius-Ring 15
Würzburg, Germany
<firstname>.<lastname>@denkbares.com

Abstract

This paper discusses the development of an intelligent information system for cataract surgery. The system shall provide a knowledge base combining formal and informal knowledge to assist ophthalmologists in clinical practice. The application scenario requires application site specific adaptation and long-term maintenance of the knowledge base, ideally performed directly by experts. To comply with these requirements we discuss a customized knowledge acquisition environment and report about early experiences.

1 Introduction

In this paper we describe a case study of a knowledge management project within the domain of cataract surgery. A cataract is a cloud dark lens in the human eye, which is appearing quite frequently, especially affecting older people. Usually, a cataract can only be treated operatively by replacing the lens by an artificial one. The medical methods applied today have quite high success rates while being rather efficient in general. With about 20 million surgical intrusions per year, cataract surgery is the most widely applied type of operation applied on humans worldwide.

There are still ongoing research efforts to improve methods optimizing success rates and cost effectiveness. Practical experiences have shown that about 90% of the cases can be considered ordinary cases were a standardized treatment is applied, providing an extremely high success rate at comparably low costs. The remaining cases however, show a considerable high complexity, making the treatment process much more demanding. A suitable treatment has to be determined choosing from a number of surgery methods by incorporating many boundary conditions. Being demanding even for experienced surgeons, these cases often benefit from new methods evolved recently in the field.

The goal of the Wissass project[1], discussed in this paper, is the development of an intelligent information system that is able to effectively support ophthalmologists in cataract surgery. The focus of this paper considers the aspect of knowledge acquisition and maintenance of that system. The captured knowledge includes informal knowledge (e.g., text and figures) as well as formal knowledge to support automated reasoning. This combination of knowledge at different degrees of formalization poses special challenges to knowledge acquisition. For this purpose a special knowledge acquisition tool is designed. We also report about the experiences made during the collaboration with ophthalmologists considering this aspect.

The remainder of this paper is structured as follows: In Section 2 the application scenario is described, outlining how the developed system is going to support the clinical practice of cataract surgery. A customized knowledge acquisition tool for the described scenario is presented in Section 3. A discussion of the current state of the project is given Section 4. The paper presents related work in Section 5 and concludes with a short summary and outlook.

2 Application Scenario

The goal of the Wissass project is to provide an intelligent information system, that is suitable to assist the physicians in practice. Especially for the treatment of the difficult non-standard cases computer-based assistance by expert knowledge would be valuable. Therefore, a knowledge system was designed serving two use cases:

- **Second Opinion System:** A traditional knowledge-based system is employed routinely to run in parallel with the treatment process of each patient. The anamnesis and examination data of each patient is entered into the knowledge-based system, which checks whether there are deviations from the standard case that need to be considered. If so, the system provides the ophthalmologist hints about special issues that need to be considered for the treatment of this patient. It further determines a proper surgery method if appropriate. The system also demands additional examination data for the patient, if it is required to exclude the prevalence of particular complications.
- **Tutoring System:** In an intelligent information system with simple access, ophthalmologists can look up and study a comprehensive up-to-date presentation of the latest knowledge about the domain of cataract surgery. The content is based on standard text book knowledge on the domain being enhanced and updated with new content when new best practices or research results emerge. The tutoring system provides intelligent interactive navigation and is illustrated with multi-media content. It also shall provide means to further research about hints or propositions made by the second opinion system as this only provides a very scarce explanation.

While the latter application scenario allows surgeons to look up particular aspects they are currently interested in, the second opinion system automatically runs in background providing treatment hints on complicated cases.

[1] funded by BMWI (https://www.bmwi.de)

These two major use cases require the combination of formal knowledge, i.e., to generate treatment recommendations, and informal knowledge, such as illustrative contents including text and figures. For the representation of the formal knowledge the concepts of the domain are modeled by an ontology, which is extended by simple derivation rules for the treatment recommendations. Beside a hierarchical organization of the ontology concepts, numerous cross connections between associated concepts are established using a small set of predefined relation types with specific semantics. In that way, a semantic network is established that allows to enable semantic navigation within the tutoring system. Further, these cross connections between concepts can be used to create rules for the second opinion system. Concepts that correspond to values of an anamnestic patient data set (e.g., *eye pressure increased*) can be connected to target concepts using special relations. In surgical practice, if a patient data set matches these source concepts and treatment hint proposing the target concept is generated according to the used relation.

For each concept illustrative content is included describing the role of the concept in the domain. Additional narrative content in text book style is included, being interlinked with relevant domain concepts.

The body of knowledge developed within the scope of this project however is only forming a basic seed of knowledge. At any application site, such as hospitals or doctor's practices, the knowledge base needs to be adapted to the local conditions and requirements. This includes modifications with respect to the available examination equipment and surgery methods as well as the predominant category of patients. Therefore, the ability for performing changes in a simple way at reasonable costs is an important selling factor for the overall product. Consequently, one major challenge of the project is to provide a knowledge acquisition concept that allows for easy adaptation and maintenance of the knowledge. The goal is to enable the clinic personnel to perform minor adaptations of the knowledge on their own. For more complex modifications, easy remote collaboration on the content between a knowledge engineer and local experts should be supported. Hence, beside the formation of a knowledge seed also the design and development of a knowledge acquisition tool, that is fulfilling these requirements, is an important goal of the Wissass project.

3 A Custom Knowledge Acquisition Tool

The developed knowledge acquisition tool is based upon the wiki system KnowWE [1]. For the project specific customization of the tool the meta-engineering approach for document-centered knowledge acquisition has been employed [2]. It allows for smooth and ongoing adaptation of the tool towards the requirements of the project settings. The tool modifications were designed in close cooperation with the medical expert in joint sessions of discussion and assessment. Beside special markup languages for the knowledge formalization, this includes components for navigation, search, visualization, and authoring support.

In document-centered knowledge acquisition the presentation of the knowledge within the tool very much relies on the structure of the documents. While technically not being a characteristic of the tool, we also consider the establishment of a suitable and understandable document structure as part of the meta-engineering process. In the following we describe the knowledge acquisition tool resulting from that customization process.

Figure 1 shows the developed wiki-based knowledge acquisition tool presenting a document describing the domain concept *Augenuntersuchung Befund* (eye examination results) in a particular structure. The structure, which any domain concept of the ontology is/should be described in, is a follows:

1. A custom concept definition markup defines a new concept of the ontology . (A)
2. The label of the concept is defined using the custom markup for concept labels. *(optional)* (B)
3. A list of the sub-concepts of the local concept defines the hierarchical structure of the ontology. Introduced by 'Unterkonzepte:', the comma-separated list markup specifies which concepts are sub-concepts of the local concept of this document. *(optional)* (C)
4. Then, further relations of the local concept within the semantic network can be defined. Therefore, the comma-separated list-based markup with the respective keyword are used. *(optional)* (D)
5. Concluding, the informal description of the concept is defined using normal wiki syntax. (E)

The structure defined above only defines a convention recommended to authors not being enforced by the system.

Navigation & Search Underneath the rendering of the document content, a graph visualization is shown (F). It presents a view on an excerpt of the semantic network displaying the concept described on this document and its neighbors, including relations that are defined in other documents (e.g., association from *Anamnese Patientensituation*). In that way, the user at one glance can get an overview of the concept and its role within the semantic network, also providing instant feedback after editing the document content. Any node can be used to open the corresponding document by click. At the bottom of the left panel the history of recently visited pages is shown (G). The search slot, also located in the left panel (I), provides access to a search mechanism, combining semantic search and full-text search.

Authoring Assistance Figure 2 shows the source text of the document, which is managed by the document-centered knowledge acquisition environment. It can be edited in different ways. Any parts of the contents can freely be edited, using (extended) wiki syntax.

The user can edit the document content in source mode, as shown in Figure 2, or in a section-editing mode which allows to edit any paragraph within the document view as shown in Figure 1.

In the left panel of Figure 1 a hierarchical collection of concepts is shown (H). It resembles a selection of the domain concepts from the ontology that recently were within the focus of the user, i.e, that have been used for editing or appeared on the visited documents. For the editing of the formalized parts of the content, i.e., the comma-separated lists of sub-concepts or other kinds of relations, the system enables drag-and-drop editing. Any concept within the left panel can be dragged onto a list of the document content and will be appended to it in the source text of the document. When a desired concept is currently not present in the left panel, it can be looked up using the search slot above it. The auto-completion functionality allows to select the concept and adds it to the collection of concepts. In

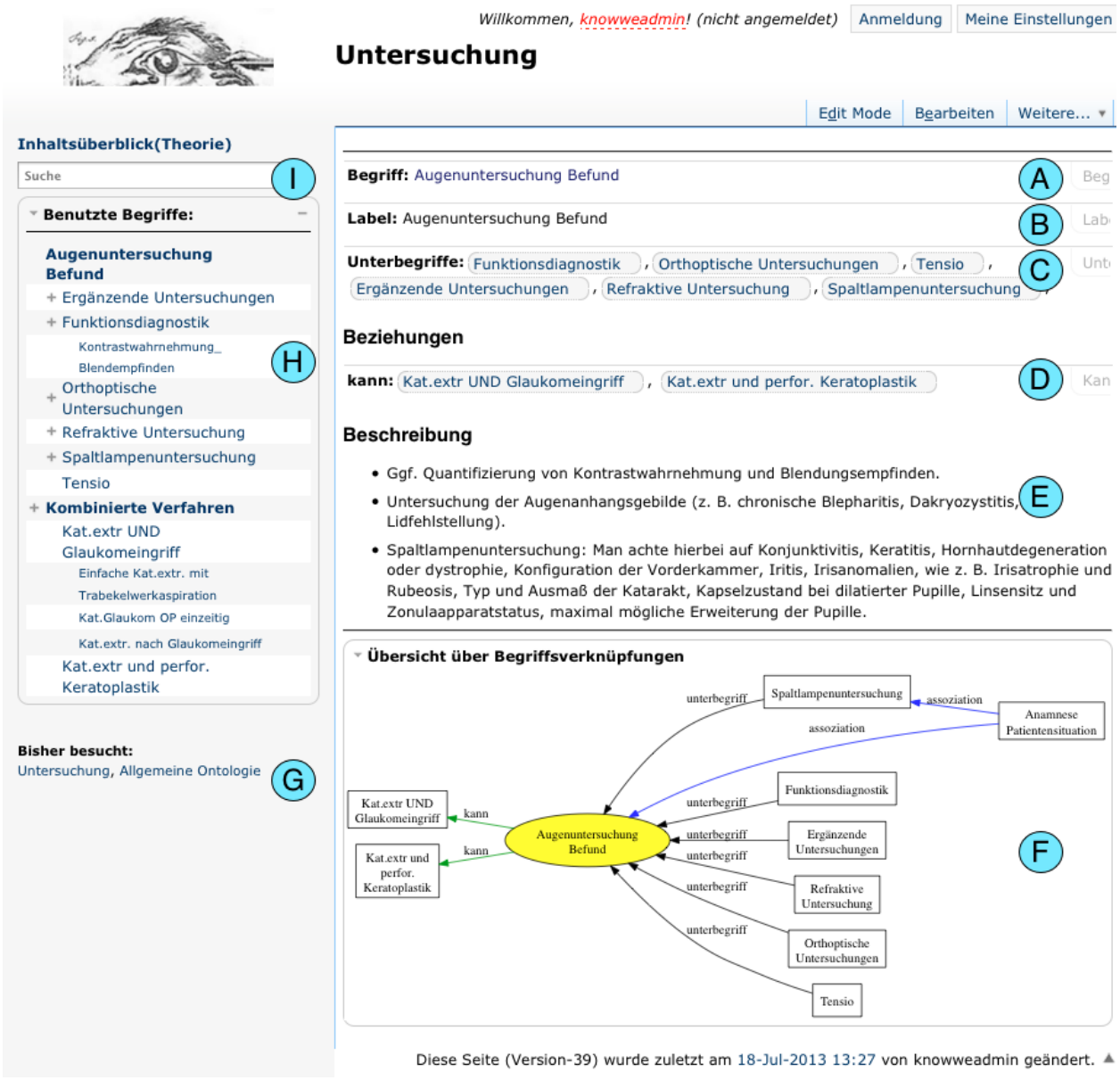


Figure 1: The document describing the concept *Augenuntersuchung Befund* (in German language).

that way, the entire semantic network can easily be edited mostly by using drag-and-drop editing, while the freedom and simplicity of document editing is retained.

4 Discussion

In the current phase of the project the initial body of knowledge to a large extent is captured within the system. Currently, there are more than 320 concepts of the domain contained in the ontology, each being described by a document as discussed in Section 3. There is also a text book chapter about cataract surgery included, where each section is annotated with the relevant domain concepts from the ontology. Further, there are about 200 cross connections defined between concepts. A subset of these relations are used to generate simple rules for treatment recommendations. More complex kinds of rules can be inserted by the use of textual rule syntax [1], which is hardly required by now.

The designed knowledge acquisition tool allows for simple maintenance of the knowledge by a uniform interaction paradigm. This allows to perform minor modifications of the knowledge in a consistent way. The expert supporting the Wissass project is capable to perform most knowledge base editing operations on his own. This includes editing of illustrative knowledge, creation of new concepts, editing of the hierarchy and establishing cross connections between concepts. Currently, we do not have any experiences about knowledge maintenance by independent experts yet. However, we expect the document-centered approach makes it easy for specialists to get involved with the knowledge acquisition activities. The editing of informal content only requires editing of normal wiki content. Further, editing of coma-separated lists supported by drag-and-drop appears to be a simple way to maintain the semantic network. Additionally, the use of a centralized document authoring environment allows for easy (remote) collaboration between experts and knowledge engineers.

For guaranteeing the consistency of the ontology during the development process, automated tests are integrated into the system being executed after editing op-

```
Begriff: Augenuntersuchung Befund

Label: Augenuntersuchung Befund

Unterbegriffe: Funktionsdiagnostik, Orthoptische Untersuchungen,
Tensio, Ergaenzende Untersuchungen, Refraktive Untersuchung,
Spaltlampenuntersuchung,

!! Beziehungen

kann: Kat.extr UND Glaukomeingriff, Kat.extr und perfor. Keratoplastik

!! Beschreibung

* Ggf. Quantifizierung von Kontrastwahrnehmung und Blendungsempfinden.

* Untersuchung der Augenanhangsgebilde (z. B. chronische Blepharitis,
  Dakryozystitis, Lidfehlstellung).
```

Figure 2: The raw text view of the document describing the concept *Augenuntersuchung Befund*.

erations. The following deficiencies are detected: concepts, which are not integrated into the concept hierarchy (orphans); concepts with multiple parent concepts; cycles within the concept hierarchy; concepts with more than 10 sub-concepts (recommending further categorization). The results of these tests can be viewed on a distinct page and are monitored, and in case of need fixed, by the knowledge engineers.

An important aspect to support the maintenance of the semantic network is visualization. As a distinct sub-project to the development of the knowledge acquisition tool special visualization methods are developed. These visualizations will serve the knowledge maintainer as well as the user of the tutoring system to get an overview of the modeled knowledge. However, the visualizations are not within the focus of this paper.

5 Related Work

The customization of knowledge acquisition tools to ease the knowledge authoring task has long tradition. The specification and development of customized tools, based on graphical user interfaces, have been discussed by Musen et. al. [3; 4]. The use of (active) documents for building knowledge bases has been addressed only by few researchers [5; 6]. With the customization of document-centered knowledge acquisition tools however only very little experiences are reported. Even though the use of documents allow for a rather smooth customization process. Further, one major advantage of the document-centered approach is that informal knowledge can be included in a very flexible way. This is, formal knowledge that is defined by the use of markup languages can be intermixed with illustrative contents for documentation or justification. In this aspect, many GUI-based tools show shortcommings. However, the document-centered approach requires to define and to maintain the document structure. The change of this structure during the project can cause considerable refactoring workload.

6 Conclusion

In this paper we discussed the knowledge acquisition concept of the Wissass project. We introduced a customized authoring environment for editing the knowledge base of cataract surgery. The main goal is the design of a tool that allows for simple knowledge maintenance and adaptations at the customer/clinical site. For this purpose we employed the meta-engineering approach for customizing document-centered knowledge acquisition tools.

We reported about early experiences of using the tool in cooperation of an expert ophthalmologist. For a more meaningful evaluation we plan to test the usability of the knowledge authoring tool with independent experts.

7 Acknowledgements

The project is a cooperation of the Karlsruhe Institute of Technology, Germany (KIT) and the denkbares GmbH. It is funded as a ZIM-KOOP[2] project by the German Federal Ministry of Economics and Technology (BMWI). The authors also thank the project executing organisation AiF in Berlin, which is responsible for the allocation of the budget to the research centre and the commercial company.

References

[1] Baumeister, J., Reutelshoefer, J., Puppe, F.: KnowWE: A Semantic Wiki for Knowledge Engineering. Applied Intelligence **35**(3) (2011) 323–344

[2] Reutelshoefer, J., Baumeister, J., Puppe, F.: A Meta-Engineering Approach for Customized Document-centered Knowledge Acquisition. In: Modellierung. Volume 201 of LNI., GI (2012) 187–202

[3] Musen, M.: Automated Generation of Model-Based Knowledge-Acquisition Tools. Pitman Publishing London (1989)

[4] Eriksson, H., Musen, M.: Metatools for knowledge acquisition. IEEE Softw. **10** (May 1993) 23–29

[5] Gaines, B.R., Shaw, M.L.G.: Embedding formal knowledge models in active documents. Commun. ACM **42**(1) (January 1999) 57–64

[6] Molina, M., Blasco, G.: Using electronic documents for knowledge acquisition and model maintenance. In: Knowledge-Based Intelligent Information and Engineering Systems. Volume 2774 of LNCS. Springer Berlin Heidelberg (2003) 1357–1364

[2] http://www.zim-bmwi.de/

Wissens- und Erfahrungsaustausch im Technischen Kundendienst mittels semantischer Wikis: Einsatzmöglichkeiten, Systeme und Erfolgsfaktoren

Michael Fellmann, Sara Schulze, Deniz Özcan

Universität Osnabrück,
Institut für Informationsmanagement und Unternehmensführung,
Katharinenstraße 3, 49069 Osnabrück
{michael.fellmann, deniz.oezcan}@uni-osnabrueck.de,
sara_schulze@yahoo.de

Abstract

Semantische Wikis als Erweiterung herkömmlicher Wiki-Systeme ermöglichen es, Wissen zu strukturieren und mit anderen zu teilen. Gleichzeitig gehen sie über traditionelle Wikis hinaus und bieten eine verbesserte Suche und Analyse der im Wiki enthaltenen Daten. Diese Möglichkeiten können zur Unterstützung des Technischen Kundendienstes (TKD) eingesetzt werden. In dieser Arbeit werden semantische Wikis vorgestellt und auf Einsatzszenarien und Erfolgsfaktoren für die Wissensakquisition im Bereich des Technischen Kundendienstes untersucht. Dabei werden unterschiedliche Einsatzmöglichkeiten dargestellt, ein Systemvergleich präsentiert und Erfolgsfaktoren, die für eine effiziente Nutzung des semantischen Wikis benötigt werden, hergeleitet.

1 Wissen und Erfahrung im Technischen Kundendienst

Wissen und Informationsverarbeitungskapazitäten stellen eine maßgebliche Ressource für Unternehmen dar [Picot, 1989]. Unternehmen müssen heutzutage jegliche Wissensressourcen mobilisieren, um dem Wettbewerbsdruck standhalten zu können. Der Wissenstransfer zwischen Mitarbeitern und unterschiedlichen Abteilungen stellt hierbei einen wichtigen Wettbewerbsfaktor dar, der zu Produktivitäts- und Qualitätssteigerungen führen kann [North, 2011]. Insbesondere im Technischen Kundendienst ist der Servicetechniker eine wichtige Ideen- und Informationsquelle und leistet mit seinem Wissen und den Informationen, die er im Laufe eines Serviceprozesses ansammelt, einen wesentlichen Beitrag zum Wertschöpfungsprozess. Es existieren Produktinformationen, über die ein Servicetechniker vor Ort verfügen muss, ebenso existieren auch Informationen zu Kunden und Märkten, die ebenfalls für den gesamten Wertschöpfungsprozess von Bedeutung sind [Österle und Senger, 2006]. Es gilt daher, Servicewissen zu erheben, zu kommunizieren, seine Anwendbarkeit sowie seine Auswertung zu ermöglichen [Schlicker und Leinenbach, 2010]. Der Einsatz semantischer Wikis unterstützt den TKD im kollaborativen Management dieses Servicewissens. Servicetechniker können auf diese Weise nicht nur Wissen abrufen, sondern auch kommentieren, diskutieren und Erfahrungen austauschen. Die semantische Auszeichnung (Annotation) der Wiki-Inhalte erlaubt dabei eine umfassende Analyse des Wissensschatzes.

1.1 Der Servicetechniker als Ideen- und Informationsquelle

Innerhalb des Technischen Kundendienstes nimmt der Servicetechniker eine entscheidende Rolle ein. Er fungiert als Bindeglied zwischen Herstellern technischer Sachgüter und dem Kunden. Er betreut ausschließlich technische Produkte und hat seinen Aufgabenschwerpunkt vor allem im Zeitraum nach dem Kauf. Im Vergleich zu anderen Unternehmensbereichen fällt die Betreuung eines Kunden und eines Produktes im TKD am zeitintensivsten aus [Harms, 1999]. Durch den intensiven Kundenkontakt ist er meist die erste Anlaufstelle für Kundenwünsche und Beschwerden, die eine entscheidende Rolle im Wettbewerb spielen. Von diesem engen Kundenkontakt, den Erfahrungen sowie den technischen Kenntnissen des Servicemitarbeiters profitieren ebenso andere Unternehmensbereiche. Insbesondere die Bereiche Forschung und Entwicklung (F&E), Produktion, Vertrieb und die Marketingabteilung können das vorhandene Wissen dazu nutzen, die Wettbewerbsfähigkeit des Unternehmens stetig zu verbessern [Harms, 1999; Benkenstein, 2001].

Im Bereich Forschung und Entwicklung werden gezielt Maßnahmen ergriffen um neue technologische Fortschritte zu erzielen und den Kenntnisstand des Unternehmens zu erweitern [Witte, 2007]. Insbesondere während der Entwicklung neuer Produkte kann der TKD wichtige Informationen zu den Kundenwünsche an die F&E-Abteilung weiterleiten [Deuse et al., 2009]. Ihr technisches Know-How sowie eigene Ideen und Änderungsvorschläge können zudem zu Beginn der Entwicklung in das Produkt mit eingebracht werden [Benkenstein, 2001]. Daraus können neue Innovationen oder verbesserte Produkte resultieren [Herrmann et al., 2009]. Insbesondere durch das fachliche Wissen kann der TKD auch für die Erstellung von Bedienungsanleitungen oder Ersatzteilkatalogen verantwortlich sein oder diese unterstützend bearbeiten [Harms, 1999].

Im Rahmen der betrieblichen Produktion ist es durch die Mitarbeiter des TKD möglich, anhand von Informationen die sie vom Kunden erhalten haben, sowie durch Rückmeldungen in Form von Einsatz- und Reparaturberichten, Schwachstellen der Produkte zu erkennen. Produkt- und Fertigungsfehler können so schnell erkannt und

behoben werden. Auf Basis der Erfahrungswerte kann ebenso eine beratende Funktion übernommen und durch Hilfestellungen, Hinweise oder Änderungsvorschläge eine effizientere Produktion und verbesserter Qualität erreicht werden [Harms, 1999].

Auf diese Weise ist es möglich zusätzlich auf neue Produkttypen hinzuweisen, die für den Kunden interessant sein könnten. Der Servicemitarbeiter gibt durch seine Beratungstätigkeit dem Kunden Auskunft darüber, welche Zusatzgeräte zu seiner bestehenden Ausstattung vorteilhaft sind, um eine Maschine optimal nutzen zu können [Kirchgeorg, 1991].

Neben seinen technischen Aufgaben und der Funktion als Ideen- und Informationsquelle für die technischen und vertrieblichen Bereiche eines Unternehmens, stellt der TKD ebenso ein entscheidendes Marketinginstrument dar [Schröder, 1997]. Er unterstützt den Marketing-Mix eines Unternehmens, indem er im Rahmen der Produktpolitik Fachkenntnisse bei der Gestaltung einzelner Produkte und Dienstleistungen und dazugehöriger Serviceleistungen einfließen lässt [Bruhn, 2010]. Ist der TKD durch seine Leistungsfähigkeit und ein positives Image geprägt, kann dies in der Distributionspolitik als verkaufsförderndes Argument genutzt werden [Harms, 1999]. Zusätzlich kann er einen wesentlichen Beitrag zur Marktforschung leisten, indem er die in Kundengesprächen identifizierten Bedürfnisse und Erwartungen der Kunden weiterleitet [Kirchgeorg, 1991].

Mit diesen weitgreifenden Informationen stellt der Servicetechniker eine wichtige interne Ideen- und Informationsquelle für das Unternehmen dar [Benkenstein, 2001], die im Kontext des Product-Service-System von großer Bedeutung ist.

1.2 Wissensbezogene Herausforderungen

Den Nutzen und die Vorteile, die sich für das Unternehmen aus dem Wissen des Servicetechnikers generieren lassen, können mit einigen Problemen einhergehen, die für eine effiziente Wissensakquisition beseitigt werden müssen. So kann es innerhalb der Kommunikation des Servicetechnikers mit anderen Unternehmensbereichen zu einem fehlerhaften oder sogar mangelnden Informationstransfer kommen [Harms, 1999]. Es kann der Fall eintreten, dass aufgrund der immer größer werdenden Anforderungen an die Kundendienstmitarbeiter, diese zeitlich nicht in der Lage sind, neben ihren Kernaufgaben auch andere Unternehmensbereiche mit Informationen zu versorgen [Schlicker und Leinenbach, 2010].

Aus dem steigenden Angebot an immer komplexer werdenden und unterschiedlicheren Varianten von Sachgütern, erhöht sich zusätzlich der Anspruch an die Techniker. Sie müssen ihr Wissen über die jeweiligen Maschinen stetig erweitern, um die mit den Sachgütern verbundenen Kernleistungen schnell und erfolgreich erbringen zu können. Jedoch ist es für Techniker nicht möglich, bei einer großen Variantenvielfalt und Komplexität jedes Produkt bis ins Detail zu kennen. Oftmals benötigen sie zusätzliche Informationen für die Bearbeitung. Da die Informationsversorgung der Techniker diesen Anforderungen jedoch nicht immer gerecht wird, ist es ihnen nicht möglich, sich das kurzfristig benötigte Wissen fristgerecht anzueignen [Walter, 2009; Schlicker und Leinenbach, 2010]. Aus diesem Grunde sind sie oftmals auf das Wissen und die Erfahrungen ihrer Kollegen angewiesen.

Um diese Herausforderungen bewältigen zu können und die Arbeit der Kundendienstmitarbeiter effizienter zu gestalten, stellt die Nutzung eines semantischen Wikis eine Lösung dar. Das benötigte Wissen kann einfacher erhoben, bereitgestellt und geteilt werden, wodurch die produktbegleitende Dienstleistung optimiert werden kann.

2 Wissensakquisition mit semantischen Wikis

Der Informationsaustausch zwischen dem TKD und anderen Unternehmensbereichen ist, wie zuvor dargestellt, mit vorteilhaftem Nutzen für das Unternehmen verbunden. Um das Wissen der Servicemitarbeiter jedoch effizient nutzen zu können, benötigt ein Unternehmen geeignete Instrumente zur Wissenserhebung und Wissensrepräsentation. Das kollektive Wissen bildet die Grundlage für eine effektive organisationale Wissensbasis [Probst et al., 2010]. Die Erhebung von Wissen aus verschiedenen Wissensquellen und die anschließende Umsetzung in eine Wissensbasis wird auch als Wissensakquisition bezeichnet [Curth et al., 1991]. Das Management dieser Wissensbasis findet zuletzt immer häufiger unter der Verwendung von Wikis statt. Durch den Einsatz eines Wikis im TKD wird eine Plattform geschaffen, die ein kooperatives Arbeiten an Texten und Inhalten ermöglicht [Ebersbach und Glaser, 2005]. Ein semantisches Wiki erweitert die Flexibilität des normalen Wikis, indem semantische Annotationen eingesetzt werden, die eine erweiterte maschinelle Verarbeitung der Daten zulassen [Schaffert et al., 2009]. Die technische Infrastruktur ermöglicht es zudem, eine Vielzahl verschiedener Wissensquellen und -nutzer miteinander zu verbinden [Probst et al., 2010].

2.1 Kollaborative Erfassung und Pflege von Wissensbeständen

Bevor ein innerbetrieblicher Wissenstransfer erfolgen kann, müssen zuvor die individuellen und kollektiven Wissensbestände erhoben werden. Im TKD stellen Kundendienstmitarbeiter, Hersteller der technischen Produkte und die dazugehörigen Bedienungsanleitungen mögliche Wissensträger dar. Aber auch Kunden können die Rolle einer Wissensquelle einnehmen.

In der Fachliteratur existieren drei unterschiedliche Ansätze zur Wissensakquisition [Kurbel, 1992]. Neben der *direkten* und *automatischen Wissensakquisition*, die eher darauf ausgelegt sind, die Erweiterung und Modifizierung einer vorhandenen Wissensbasis vorzunehmen, anstatt ihrer Erstellung förderlich zu sein [Haun, 2000], wird die *indirekte Wissensakquisition* als geeignetes Verfahren und Ausgangspunkt für die Wissenserhebung im Rahmen von semantischen Wikis für den TKD angesehen.

Bei der indirekten Wissensakquisition ist der Prozess der Wissenserhebung in drei zentrale Phasen unterteilt [Curth et al., 1991]. Innerhalb der *Erhebungsphase* erfolgt eine reine Datensammlung. Sie stellt die Grundlage für die Wissensanalyse und Wissensrepräsentation dar [Weichert, 2003]. Durch Befragung von Mitarbeitern oder Sammlung relevanter schriftlicher Materialien wird von einem *Knowledge Engineer* Expertisen erstellt und in Wissensprotokollen dokumentiert [Gabriel, 1992].

In der darauffolgenden *Analysephase* werden die erstellten Wissensprotokolle analysiert und interpretiert [Curth et al., 1991], woraus sich die Wissensstruktur und die Vollständigkeit des erhobenen Wissens ableiten las-

sen. Im Rahmen der *Operationalisierungsphase* erfolgt die Formalisierung des analysierten Wissens, sodass eine maschinelle Verarbeitung durch einen Computer möglich ist [Karbach und Linster, 1990]. Dieser Ansatz nutzt einen Wissenseditor, über den die Wissensquelle sein Wissen in einer vorgegebenen Form in das System eingeben kann. Ein semantisches Wiki stellt solch einen Wissenseditor dar und kann als Schnittstelle zwischen Mensch und Maschine gesehen werden [Weichert, 2003]. Es bietet im Vergleich zu herkömmlichen Wikis eine explizite Repräsentation des beinhalteten Wissens [Schaffert et al., 2009], die über Kategorien und semantisch typisierte Links erreicht wird, welche zu einer verbesserten Navigation und Suche genutzt werden können. Zur Strukturierung des Wissens im Wiki kann die in [Fellmann et al. 2012] beschriebene Ontologie herangezogen werden.

2.2 Potenziale semantischer Wikis zur Ergänzung von Servicemanagementsystemen

Da davon auszugehen ist, dass semantische Wikis die im Technischen Kundendienst eingesetzten Servicemanagementsysteme nicht vollständig ersetzen können, wird im Folgenden untersucht, welche Potenziale semantische Wikis zur Ergänzung von Servicemanagementsystemen besitzen. Letztere bündeln die bei der Problemdiagnose/-lösung benötigen Informationen aus diversen IT-Systemen, können jedoch auch als ein eigenständige Systeme realisiert werden. Weitere Eigenschaften von Servicemanagementsystemen werden in [VDMA, 2008] beschrieben. Es werden die gängigen Komponenten von Servicemanagementsystemen betrachtet, die [VDMA, 2008] samt ihrer wesentlichen Funktionen aufgelistet sind. Funktionalitäten von Servicemanagementsystemen, die durch semantische Wikis unterstützt werden können, sind in Tabelle 1 in der rechten Spalte angegeben. Es handelt sich hierbei um Funktionen, die einer Unterstützung durch semantische Wikis besonders zugänglich sind. Dies rührt daher, dass keine komplexen Berechnungen (wie etwa bei der Ressourcenplanung) erforderlich sind, da die Speicherung und der Abruf von Informationen im Vordergrund stehen. Auch Funktionalitäten, die Feedback-Funktionen und integrierte Assistenzfunktionen beinhalten, sind geeignet, durch semantische Wikis unterstützt zu werden (zum Vergleich semantischer Wikis siehe Abschnitt 3).

Betrachtet man Tabelle 1, so fällt auf, dass vor allem der Funktionsbereich der Wissensdatenbank gut durch semantische Wikis unterstützt werden kann. Dies ist plausibel, da Suchmöglichkeiten in Metadaten ein grundlegender Aspekt semantischer Wikis sind, wie auch die semantische Verlinkung von Informationsobjekten. So können Dokumente und Lösungsbeschreibungen mit weiteren Informationsobjekten, wie z.B. Maschinen oder Fehlern verlinkt werden. Der automatische Vorschlag

Tabelle 1: Unterstützungspotenziale semantischer Wikis

Servicemanagementsystemkomponente	Beschreibung	Unterstützungspotenzial durch semantische Wikis
Kontaktmanagement	Feststellung der Bedarfe für Serviceleistungen. Bearbeitung der Serviceleistungen im Helpdesk oder Self-Service.	• Suche in der Lösungsdatenbank • Geführte Problemdiagnose
Auftragsabwicklung	Bearbeitung von ungeplanten wie geplanten Servicevorgängen.	• Suche in der Lösungsdatenbank • Geführte Problemdiagnose • Automatische Generierung von Lösungsvorschlägen
Einsatzplanung	Ressourcenzuordnung und operatives Management der Servicetechniker.	• Zusammenstellung von Information an die Arbeitskräfte
Ausführung und Rückmeldung	Unterstützung der Arbeit vor Ort oder direkte Unterstützung des Kunden durch Teleservice.	• Suche in der Lösungsdatenbank • Geführte Problemdiagnose • Rückmeldung von Fehler, Ursache und Lösung
Fakturierung	Mit der Rechnungsstellung wird der Serviceprozess abgeschlossen.	• Dokumentation von Leistungsarten
Installationsverwaltung	Verfolgung und Aktualisierung der beim Kunden installierten Basis.	• Pflege der Information über die installierten Basis inkl. Änderungs-, Störungs- und Wartungshistorie • Visualisierung der Kundeninstallation
Serviceverträge	Automatische Prüfung von Vertragskonditionen und Monitoring von Service Level Agreements.	• Dokumentation der Vertragstypen
Garantieabwicklung	Abwicklung von Gewährleistungen auch unter Einbeziehung von Wiederverkäufern.	• Registrierung der verwendeten Geräte und Maschinen und Verknüpfung mit Endkundendaten
Servicelogistik	Einbindung der Serviceprozesse in ein Enterprise Resource Planning (ERP)-System.	• Dokumentation der Seriennummern und Herstellernummern
Beschwerdemanagement	Erhöhung der Kundenzufriedenheit durch adäquate Beschwerde- und Retourenprozesse.	• Information zu häufigen Problemen und Rückrufaktionen
Lösungs- und Wissensdatenbank	Die Lösungsdatenbank kann als eigenständige Anwendung aufrufbar sein, in die Helpdesk-Anwendung integriert sein oder sie steht im Internet als Self-Service für Kunden und Partner direkt als Informationsquelle zur Verfügung.	• Suchmöglichkeiten im Volltext und Metadaten der verschiedenen Informationsquellen wie Lösungsbeschreibungen, Maschinendokumentation etc. • Verknüpfung von Dokumenten und Lösungsbeschreibungen mit Geschäftsobjekten (z. B. Maschine, Fehler) • Automatischer Vorschlag von Lösungen • Administration und Pflege der Lösungen und Probleme • Feedbackmöglichkeiten
Servicecontrolling	Umsetzung einer aktuellen Sicht auf die Kundendaten.	• Dokumentation von Kennzahlen zu Serviceprodukten, Serviceverträgen, Serviceeinsätzen
Analysen	Schaffung von Transparenz, Erkennung von Problemen und Trends.	• Auswertung der Nutzung des Wikis erlaubt Rückschlüsse auf Qualitätsprobleme bei Produkten und/oder deren Dokumentation • Auswertung der Kommentare/des Feedbacks im Wiki

von Lösungen wird in neueren Wikis ebenso umgesetzt wie Feedbackmöglichkeiten, etwa durch die Diskussion der Wiki-Inhalte auf Diskussionsseiten oder in den Wikiseiten eingebettete Kommentierungs- und Bewertungsfunktionen.

2.3 Benötigte Funktionsbereiche semantischer Wikis

Die zur Umsetzung der in Tabelle 1 identifizierten Unterstützungspotenziale erforderlichen Funktionalitäten können in fünf Funktionsbereiche eingeteilt werden, die im Folgenden vorgestellt werden.

Um das im semantischen Wiki repräsentierte Wissen zu editieren, wie dies bspw. die Unterstützung einer Feedbackfunktion der Lösungs- und Wissensdatenbank-Komponente erfordert, sind *Editierfunktionalitäten* (EDIT) erforderlich. Zum Abruf der Wissensstrukturen, wie dies etwa bei der Suche in der Lösungsdatenbank erfolgt, sind Funktionen im Bereich des *Browsing und Retrieval* (RETR) erforderlich.

Um mittels eines semantischen Wikis in Form von Dokumenten existierende Informationsquellen, z.B. in Form von Lösungsbeschreibungen und Maschinendokumentationen abzurufen, können diese in das Wiki integriert werden. Ist der hierfür erforderliche Aufwand allerdings zu hoch oder müssen die ursprünglichen Dokumente beibehalten werden, weil Spezialfunktionen wie eingebettete 3D-Darstellungen oder Animationen genutzt werden, so können diese Dokumente im Wiki zumindest referenziert werden. Eine weitere Möglichkeit besteht darin, Dokumente als Attachment einer Wikiseite zu hinterlegen, womit je nach Implementierung des Wikis auch dessen Versionsmanagement für Attachments genutzt werden kann. Der Umfang der benötigen Funktionen im Bereich des *Managements von Dokumenten* (DMGT) hängt letztlich davon ab, wie stark das semantische Wiki die Aufgaben eines Content-Management-Systems übernimmt.

Die Realisierung geführter Problemdiagnosen etwa in Form interaktiver Fehlerbäume oder die automatische Generierung von Lösungsvorschlägen kann allgemein zum Funktionsbereich der *Assistenzfunktionen* (ASFU) zusammengefasst werden. Zur Realisierung eines Austauschs von Wissen zwischen den Akteuren, wie er über die Rückmeldung von Fehler, Ursache und Lösung oder als Feedback in Bezug auf die Inhalte der Lösungs- und Wissensdatenbank auftritt, muss das semantische Wiki *Kollaborationsfunktionen* (KOLL) implementieren.

Die folgende Tabelle 2 zeigt noch einmal die Komponenten von Servicemanagementsystemen aus Tabelle 1 und ergänzt diese um fünf Spalten für die zuvor beschriebenen Funktionsbereiche semantischer Wikis. Durch ein „x“ wird ausgedrückt, dass ein Funktionsbereich des semantischen Wikis erforderlich ist, um die in Tabelle 1 identifizierten Unterstützungspotenziale semantischer Wikis für Servicemanagementsysteme umzusetzen.

Aus der Betrachtung von Tabelle 2 kann geschlossen werden, dass hohe Anforderungen im Sinne der Menge bereitgestellter Funktionalitäten an semantische Wikis im Bereich der Ausführung und Rückmeldung von Servicevorgängen bestehen wie auch zur Unterstützung der Lösungs- und Wissensdatenbank-Komponente von Servicemanagementsystemen.

Tabelle 2: Zuordnung von Funktionsbereichen sem. Wikis zu Servicemanagementsystemkomponenten

	Benötigter Wiki-Funktionsbereich				
Servicemanagementsystemkomponente	EDIT	RETR	DMGT	ASFU	KOLL
Kontaktmanagement		x	x	x	
Auftragsabwicklung		x	x	x	
Einsatzplanung		x			
Ausführung und Rückmeldung	x	x	x	x	x
Fakturierung		x			
Installationsverwaltung	x	x		x	
Serviceverträge		x			
Garantieabwicklung	x	x			
Servicelogistik	x	x			
Beschwerdemanagement		x			
Lösungs- und Wissensdatenbank	x	x	x	x	x
Servicecontrolling		x			
Analysen		x			

3 Funktionaler Vergleich ausgewählter Systeme

Im nachfolgenden werden einige wesentliche Eigenschaften ausgewählter semantischer Wikis dargestellt, um den State-of-the-Art der Entwicklung abzubilden. Die Auswahl der semantischen Wikis erfolgte auf Basis der in der Literatur häufig genannten Systeme, die sowohl in Forschungsprojekten als auch im kommerziellen Bereich vorzufinden sind. Bei den dargestellten semantischen Wikis ist zu beachten, dass einige noch nicht ausgereift sind und sich daher einige Charakteristika noch ändern können.

Knowledge in a Wiki (KiWi) ist ein von der EU gefördertes Wissensmanagement-Projekt, das 2011 seinen Abschluss fand. Durch seine Eigenschaften kann es bei der Erstellung von Decision-Support-Systemen eingesetzt werden [Baumeister et al., 2010]. Im konkreten handelt es sich bei KiWi um eine zur Social-Web-Plattform erweiterte Wiki-Software [Schaffert et al., 2009b]. Zielsetzung ist es, auf Basis eines semantischen Wikis ein System für das Wissensmanagement zu entwickeln, das Unterstützung bei der flexiblen Verwaltung von implizitem Wissen ermöglicht [Salzburg Research, 2011].

KnowWE (Knowledge Wiki Environment) ist ein auf Java-basierendes semantisches Wiki, das Ontologien zur Problemlösung im Zusammenspiel mit formularbasierten Nutzereingaben heranziehen kann. Es kann als webbasiertes Werkzeug zur Gestaltung von Entscheidungsunterstützungssystemen genutzt werden. Zu diesem Zweck bietet es nicht nur die Möglichkeit, Ontologien zu entwickeln, sondern diese auch in Verbindung mit konkreten Ausprägungen zu setzen, beispielsweise durch die Festlegung von Regeln und Fehlermodellen [Baumeister et al., 2010].

Moki (The Modelling Wiki) basiert auf dem Semantic MediaWiki (SMW) und bietet eine Unterstützung bei der Geschäftsmodellierung durch den Einsatz von Wiki-Seiten. Es ermöglicht seinen Nutzern die agile Zusammenarbeit, um ein Geschäftsmodell zu entwerfen. Dafür werden sowohl formale, semi-formale als auch informale Wissensbestände in dem Wiki verwendet. Auf diese Weise wird die Zusammenarbeit und der Wissensaustausch

insbesondere heterogener Gruppen erleichtert, da die Entwicklung eines Geschäftsmodells oftmals unterschiedliche Wissensträger benötigt [Ghidini et al., 2009].

OntoWiki ist eine Software zur Unterstützung von Knowledge-Engineering-Szenarien [Auer et al., 2006]. Sie fungiert als Schnittstelle zur kollaborativen Erstellung und Wartung von Ontologien. Eine umfassende semantische Suche und Navigation sowie die Unterstützung der Versionierung von Metadaten sind weitere Funktionalitäten dieser Software [Schaffert et al., 2009a].

Semantic MediaWiki+ (SMW+) ist ein semantisches Wiki, das ebenfalls auf dem MediaWiki wie auch dem Semantic MediaWiki beruht. Das SMW+ ergänzt die Editier-/Diskussions- und Feedback-Funktionalitäten, die charakteristisch für textbasierte Wikis sind, um grafische Datenvisualisierungen und Möglichkeiten zu gezielten Abfragen des Wissens und eine vielfältige Visualisierung der Informationsabrufe [Greaves, 2012].

Semantic Guide ist ein ontologiebasiertes Ratgebersystem, dass eine Automatisierung von beratungs- und wissensintensiven Prozessen umfasst. Zielgruppen dieses Systems sind vor allem Kompetenzkräfte aus dem industriellen Bereich, die auf Erfahrungs- und Expertenwissen zurückgreifen müssen. Die aktuelle Version (Semantic Guide 6.2) wurde um weitere Funktionalitäten ergänzt, um den Einsatz des mobilen Clients effizienter zu gestalten. Die Erweiterung um eine Agentenfunktion ermöglicht den Zugriff auf externe Systeme und Daten sowie eine automatisierte Fragebeantwortung [Ontoprise, 2011].

Tabelle 3 zeigt den Vergleich der vorgestellten semantischen Wikis anhand der bereits in Tabelle 2 eingeführten fünf funktionalen Bereiche EDIT, RETR, DMGT, ASFU und KOLL. Die Unterstützung eines Kriteriums wird mit den Symbolen ● (umfangreich), ○ (rudimentär) und – (nicht vorhanden oder nicht aus der Dokumentation ersichtlich) angedeutet. Im Bereich EDIT ist für die Einarbeitung in die Wikinutzung, die Unterstützung durch einen WYSIWYG (What You See Is What You Get)-Editor empfehlenswert, der bei der Formatierung von Wikiseiten keine spezielle Syntax erfordert. Weiter sollte das Wiki den Nutzer auch bei Annotationen, also der Erstellung semantisch präziser Aussagen durch den Einsatz typisierter Links, Attribute und Kategorien, unterstützen, beispielsweise durch eine Auto-Vervollständigungsfunktion. Weitere Editierkriterien sind, ob Metadaten und Metadatenschemata, z.B. in Form von Ontologien, importiert werden können, um somit die Inbetriebnahme eines semantischen Wikis zu erleichtern sowie ob Metadaten exportiert werden können, um sie außerhalb des Wikis verwenden zu können. Zur Anpassung des Wikis an die sich ändernde Anforderungen ist es ferner relevant, ob die Metadatenstruktur innerhalb der Wikiseiten geändert werden kann. Zusätzlich sollte eine Versionierung sowohl der Wikiseiten als auch der zu ihnen gehörenden semantischen Metadaten gewährleistet werden.

Im Rahmen des RETR ist – neben einer Volltextsuche zum schnellen Auffinden von Inhalten – das Vorhandensein einer Funktionalität zum facettenbasierten Browsen relevant. Somit können die zur Verfügung stehenden semantischen Metadaten zu einer effektiven und multiperspektivischen Auswahl von Informationsobjekten herangezogen werden. Die im Wiki vorhandenen Werteausprägungen können in diesem Kontext dazu verwendet werden, schrittweise komplexe Filter zur Anzeige einer gewünschten Teilmenge des Modellbestands zu erzeugen.

Tabelle 3: Vergleich semantischer Wikis

	KiWi	KnowWE	Moki	OntoWiki	SMW+	Semantic Guide
Editierfunktionalitäten (EDIT)						
WYSIWYG-Editor	●	●	●	●	●	●
Unterstützung der Annotation	●	●	●	●	●	●
Import von Wissensstrukturen	-	●	●	-	●	●
Export von Wissensstrukturen	○	●	●	-	●	●
Editieren von Wissensstrukturen in den Wikiseiten	●	●	-	●	●	●
Versionierung	●	●	●	●	●	●
Browsing und Retrieval (RETR)						
Facettenbasiertes Browsen	●	-	-	●	●	●
In Seiten eingebettete Abfragen	-	●	-	-	●	-
Nutzung einer Anfragesprache	●	●	-	●	●	-
Unterstützung der Anfragekonstruktion	●	○	-	●	●	-
Anfrage mit Inferenz	○	●	-	-	●	●
Volltextsuche	●	●	●	●	●	●
Management von Dokumenten (DMGT)						
Schnittstelle zu betrieblichen Anwendungssystemen	-	-	-	-	●	●
Import von Dokumenten über Austauschformate	-	-	●	-	●	●
Speicherung von Dokumenten als Attachment	-	●	-	-	●	●
Rechteverwaltung	●	●	-	○	●	●
Assistenzfunktionen (ASFU)						
Inhaltsextraktion aus importierten Daten	●	-	●	-	-	-
Vorschlag von Inhalten / Nutzerführung	●	●	●	●	●	●
Diagnosefunktion	-	●	-	-	-	●
Kollaborationsfunktionen (KOLL)						
Bewertung und Popularität	●	-	-	●	●	●
Diskussionsseiten, Kommentare	●	-	●	●	●	●
Tagging	●	○	-	●	●	●

Zusätzlich kann das Browsing im Informationsbestand durch Übersichtsseiten erleichtert werden, die Informationsobjekte nach bestimmten Kriterien auflisten. Um eine manuelle Pflege derartiger Seiten zu vermeiden, sollte das Wiki über eine Möglichkeit verfügen, Wikiseiten über eingebettete Suchanfragen (Inline Queries) dynamisch zu erzeugen. Diese Suchanfragen werden zur automatisierten Auswertung des Informationsbestands verwendet. Kriterien zur manuellen Suche sind die Unterstützung einer Anfragesprache und das Vorhandensein einer Hilfestellung zur Erstellung korrekter Abfragen, bspw. in Form einer Vorlage oder eines Suchformulars, das die Spezifikation der gesuchten Information erleichtert. Zur Ausschöpfung der mit der Erfassung semantischer Metadaten verbundenen Potenziale im Rahmen von Anfragen ist die Nutzung von Inferenzmaschinen ein wichtiges Kriterium, da somit neue Fakten geschlossen werden können, die nicht explizit (d.h. manuell) im Wiki erfasst wurden, son-

dern zum Anfragezeitpunkt aus den vorhandenen Daten dynamisch abgeleitet werden.

Im Bereich des DMGT ist ein wesentliches Kriterium, inwiefern die Repräsentationen von Dokumenten wie Servicehandbüchern in ein Wiki importiert werden können. Um unterschiedliche Berechtigungen für die Nutzung des Wikis z.B. im Intranet, Extranet oder dem Internet abbilden zu können, sollte das Wiki ergänzend über eine Verwaltung von Zugriffsrechten verfügen.

Durch den Einsatz von Assistenzfunktionen (ASFU) kann die Qualität der Unterstützung gesteigert werden, indem das Wissen aus bereits importierten Daten extrahiert wird und dadurch ohne großen Aufwand angewendet werden kann. Ein impliziter Vorschlag von Inhalten reduziert den Mitwirkungsbedarf der Nutzer und die Verwendung einer Diagnosefunktion kann zusätzlich den Nutzen der Assistenzfunktion erhöhen, da diese somit automatisch am richtigen Punkt startet.

Im Bereich der KOLL wird die Zusammenarbeit der an der Erstellung von Wiki-Inhalten beteiligten Akteure unterstützt. Als relevante Kriterien sind insbesondere Funktionalitäten zum gemeinschaftlichen Indexieren (Tagging), zur Diskussion und zur Qualitätsbeurteilung (Bewertung) von Inhalten zu nennen.

Wie aus Tabelle 3 ersichtlich ist, werden von allen Wikis WYSIWYG-Editoren zu einer leichteren Bearbeitung der Inhalte umgesetzt, die Annotation unterstützt und dem Nutzer werden Vorschläge für den Abruf von Wiki-Inhalten präsentiert. Sehr wenig implementiert werden in Seiten eingebettete Abfragen (Inline Queries).

4 Erfolgsfaktoren für den Einsatz semantischer Wikis

Um den Nutzen eines Wiki-Einsatzes innerhalb eines Unternehmens und insbesondere im Technischen Kundendienst feststellen zu können, sollte der Erfolg, der durch ein semantisches Wiki resultiert, gemessen werden. Es existiert eine Vielzahl an unterschiedlichen Faktoren, die den Erfolg eines semantischen Wikis im TKD beeinflussen. Ein Ansatz zur Erfolgsmessung stellt das Modell von DeLone und McLean (2003) dar. Das 1992 entwickelte multidimensionale Modell ermöglicht anhand von sechs Dimensionen die Erfolgsmessung von Informationssystemen [Reisberger et al., 2008]. Nach einer Bearbeitung des ursprünglichen Modells sind seit 2003 die Erfolgsfaktoren definiert als *Informationsqualität*, *Systemqualität*, *Servicequalität*, *Anwendung*, *Nutzerzufriedenheit* und *Nettonutzen* (Abbildung 1). Zwischen diesen Erfolgsfaktoren bestehen positive, gerichtete Wirkbeziehungen. Um das Modell auf die Gegebenheiten semantischer Wikis anwenden zu können, muss die Betrachtung einiger Dimensionen erweitert werden [Reisberger et al., 2008].
Im Rahmen der *Servicequalität* gilt die Benutzerfreundlichkeit als Grundvoraussetzung für eine erfolgreiche Nutzung semantischer Wikis [Stock und Tochtermann, 2010]. Nur wenn ein semantisches Wiki und seine Funktionen schnell erlernbar und leicht bedienbar für die Nutzer sind, sind diese auch bereit mit dem System zu arbeiten. Das Verfassen und Korrigieren von Inhalten sollte keine große Herausforderung für einen Techniker des TKD darstellen, um nicht den zeitlichen Aufwand und die Komplexität der Anwendung zu erhöhen. Ebenso wichtig ist eine schnelle und einfache semantische Suchabfrage zur Identifikation von Kunden, Geräten und Ersatzteilen.

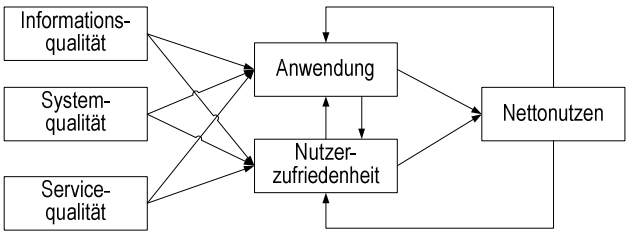


Abbildung 1: Delone/McLean-Modell

Ein weiterer Aspekt der Servicequalität umfasst die Modifikation des Wikis an unternehmensspezifische Anforderungen [Petter et al., 2008]. Der Abruf und die Verarbeitung von Informationen über unterschiedliche Schnittstellen im semantischen Wiki deutet auf eine gelungene Systemintegration hin [Smolnik und Riempp, 2006]. Grundsätzlich gilt, dass sich das Unternehmen für die Einführung eines semantischen Wikis Zeit nehmen sollte, um einen nachhaltigen Aufbau genau zu planen. Eine zeitlich großzügig bemessene Einführungsphase und eine Probelaufzeit ermöglichen eine starke Verankerung des semantischen Wikis im Unternehmen [Stock und Tochtermann, 2010], die bspw. durch die Integration oder enge Verknüpfung des Wikis in das/mit dem bestehenden Servicemanagementsystem erreicht werden kann.

Der Aspekt der *Informationsqualität* hat bei semantischen Wikis einen besonderen Stellenwert, denn erst durch eine breite Auswahl abrufbarer Informationen gewinnt es an Relevanz [Reisberger et al., 2008]. Ist die Anzahl an Artikeln innerhalb eines semantischen Wikis beispielsweise hoch, so werden die Anfragen der Mitarbeiter höchstwahrscheinlich positiv bedient. Bei den bereitgestellten Inhalten im semantischen Wiki ist deren Relevanz und Detaillierungsgrad ebenso von Bedeutung. Es sollten nur Informationen enthalten sein, die für die Arbeit der Techniker von Relevanz sind. Ein ausreichender Detaillierungsgrad wird zum einen durch das ausführliche Ausarbeiten eines Themas erreicht, zum anderen ist bei einem semantischen Wiki die typisierte Verlinkung der einzelnen Informationen zueinander wichtig. Neben dem Verfassen detaillierter Artikel müssen die Servicemitarbeiter auch auf die Richtigkeit der angegebenen Informationen achten [Petter et al., 2008]. Denn korrekte und aktuelle Informationen sind eine wichtige Grundvoraussetzung für das effektive Arbeiten mit einem semantischen Wiki [Reisberger et al., 2008; Stock und Tochtermann, 2010]. Ebenso ist die Benutzerfreundlichkeit von besonderer Bedeutung [DeLone und McLean, 1992], um die Lesbarkeit und Klarheit der Informationen sicherzustellen. Um die Klarheit der Inhalte gewährleisten zu können, könnten entsprechende Schulungen der Mitarbeiter zum verständlichen Verfassen von Beiträgen förderlich sein [Petter et al., 2008] oder Auszeichnungen besonders gelungener Artikel vorgenommen werden [Stock und Tochtermann, 2010].

Innerhalb der *Servicequalität* ist die schnelle Fehlerbehebung eine zentrale Aufgabe. Sind Funktionen des semantischen Wikis nicht fehlerfrei abrufbar, so könnte das die Arbeit mit dem System erschweren oder sogar unmöglich machen. Daher ist es maßgeblich den fehlerfreien Ablauf wieder herzustellen, da andernfalls die freiwillige Nutzung des semantischen Wikis und die Nutzerzufriedenheit negativ beeinflusst werden könnten. Der Einsatz einer Online-Service-Unterstützung kann behilflich sein, Probleme der Kundendienstmitarbeiter bei der Arbeit mit

dem semantischen Wiki zu lösen, beispielsweise durch FAQ-Bereiche oder Foren, in denen die Mitarbeiter des TKD Informationen einholen und sich gegenseitig austauschen können [Reisberger et al., 2008]. Falls Probleme dennoch nicht behoben werden können, sollte eine direkte und schnelle Kontaktmöglichkeit mit der für das semantische Wiki verantwortlichen Person ermöglicht werden [Petter et al., 2008]. Die Kompetenz und Erreichbarkeit kann das Servicepersonal auch dazu befähigen, für die Wartung des Wikis verantwortlich zu sein und das System stetig zu prüfen, um so die Qualität zu steigern.

In die Dimension *Anwendung* spielen alle zuvor genannten Erfolgsdimensionen mit ein. Werden im Rahmen der Messung von System-, Informations- und Servicequalität positive Ergebnisse erzielt, so wirken sich diese ebenfalls positiv auf die Anwendung aus. Dies schlägt sich in der Nutzeranzahl, Nutzungshäufigkeit und Nutzungsdauer nieder, welche entscheidende Faktoren bei der Beurteilung der Systemanwendung sind [Petter et al., 2008]. Von Wichtigkeit ist auch die Anzahl bestehender Artikel [Smolnik und Riempp, 2006]. Bei der Einführung sollten bereits ausreichend viele Beiträge bereitstehen, um das semantische Wiki auf Anhieb als relevante Wissensquelle zu etablieren und dadurch die Akzeptanz der Mitarbeiter zu erreichen [Stock und Tochtermann, 2010].

Auch die *Nutzerzufriedenheit* wird von den zuvor beschriebenen Qualitätsfaktoren beeinflusst. Je nachdem, wie gut das semantische Wiki in den einzelnen Dimensionen abschneidet, wirkt sich dies auf die Zufriedenheit der Nutzer aus. Da die Nutzerzufriedenheit in einer Wechselbeziehung zur Anwendung steht [DeLone und McLean, 2003], bedeutet die steigende Zufriedenheit somit auch eine steigende Nutzeranzahl, Nutzungshäufigkeit und -dauer. Insbesondere die Erwartungen der Techniker sollten erfüllt werden, ebenso wie die an spezifische Bereiche des semantischen Wikis, um die Nutzerzufriedenheit zu gewährleisten [Reisberger et al., 2008].

Der *Nettonutzen* ergibt sich aus der Produktivität des Unternehmens [Petter et al., 2008], worunter verbessertes Mitarbeiterwissen und schnelle Informationsbeschaffung gezählt wird [Reisberger et al., 2008]. Die Wissensteilung hat zusätzlich Auswirkungen auf den Umfang und die Qualität der Ausbildung der Mitarbeiter, da diese zu jeder Zeit die benötigten Informationen einholen können [Hasan und Pfaff, 2006]. Ein weiterer Aspekt für ein erfolgreiches semantisches Wiki stellt zuletzt auch eine offene Firmenkultur dar, die über eine transparente Kommunikation verfügt und diese fördert. Nur so finden die Mitarbeiter des TKD ein Umfeld vor, in dem eine effektive Nutzung des semantischen Wikis möglich ist. Jedoch sollte die Offenheit bei vertraulichen und sensiblen Informationen Grenzen aufweisen [Stock und Tochtermann, 2010]. Datenschutz, Vertrauens- und Kontrollaspekte sind wichtige Bereiche, die bei der Nutzung von Wikis berücksichtigt werden müssen [Kuhlenkamp et al., 2006].

Es ist festzustellen, dass eine Vielzahl unterschiedlicher Faktoren Einfluss auf den Erfolg eines semantischen Wikis im Bereich des TKD haben. Die Wechselbeziehungen einiger Faktoren veranschaulichen, dass ihre Kontrolle von enormer Wichtigkeit ist, um den Erfolg des Einsatzes zu gewährleisten.

5 Fazit und Ausblick

Die Anforderungen an den Servicemitarbeiter steigen mit der wachsenden Bedeutung des Technischen Kundendienstes. Um die Arbeit des TKD noch effizienter zu gestalten, wurde der Beitrag semantischen Wikis untersucht. Es wurde gezeigt, dass in semantischen Wikis ein großes Potential steckt die Arbeit des TKD zu unterstützen. Die Nutzung ermöglicht den Erfahrungsaustausch der Mitarbeiter untereinander und fördert den Informationsfluss der für die Arbeitsprozesse benötigten Daten. Semantische Wikis zeichnen sich insbesondere dadurch als geeignetes Werkzeug aus, da sie in der Lage sind, das Wissen der Mitarbeiter des Technischen Kundendienstes strukturiert und aufbereitet darzustellen.

Zur Realisierung eines Nutzens aus den Potenzialen, die die Nutzung eines Wikis innerhalb des TKD mit sich bringt, sind einige Erfolgsfaktoren zu beachten, die im Rahmen dieses Beitrags vorgestellt wurden. Aus diesem Grund ist bereits vor der Einführung eines semantischen Wikis darauf zu achten, dass die Vielzahl von Faktoren, die einen Einfluss auf den Erfolg des semantischen Wikis haben, kontrolliert und positiv beeinflusst werden. Die weitere Forschung muss sich daher auf diese Faktoren konzentrieren. So könnte im Rahmen zukünftiger Forschung die Gewichtung der Erfolgsfaktoren empirisch bestimmt werden.

Danksagung. Teile dieser Arbeit werden vom BMBF gefördert (Projekt EMOTEC, FKZ 01FL10023).

Literaturverzeichnis

[Auer et al., 2006] S. Auer, S. Dietzold und T. Riechert. OntoWiki – A Tool for Social, Semantic Collaboration. The Semantic Web 2006 - ISWC 2006: Proceedings of the ISWC 2006 (pp. 736–749). Athen, Springer.

[Baumeister et al., 2011] J. Baumeister, J. Reutelshoefer und F. Puppe. KnowWE: a Semantic Wiki for knowledge engineering. Applied Intelligence, 35(3):323–344.

[Benkenstein, 2001] M. Benkenstein. Entscheidungsorientiertes Marketing: Eine Einführung. Wiesbaden, Gabler.

[Bruhn, 2010] M. Bruhn. Marketing: Grundlagen für Studium und Praxis. 10. Auflage. Wiesbaden, Gabler.

[Curth et al., 1991] M. A. Curth, A. Bölscher und B. Raschke. Entwicklung von Expertensystemen. München, Hanser.

[DeLone und McLean, 2003] W. DeLone und E. McLean. Information systems success: the quest for the dependent variable. Information System Research, 3(1):60–95.

[DeLone und McLean, 2003] W. DeLone und E. McLean. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems, 19(4):9–30.

[Deuse et al., 2009] J. Deuse, S. Wischniewski und S. Birkmann. Knowledgebase für die kontinuierliche Innovationsarbeit im Technischen Kundendienst. In T. A. Herrmann, U. Kleinbeck, und C. Ritterskamp (Hrsg), Innovationen an der Schnittstelle zwischen technischer Dienstleistung und Kunde. 2. Auflage. (pp. 156–176). Heidelberg, Physica.

[Ebersbach und Glaser, 2005] A. Ebersbach und M. Glaser. Wiki. Informatik-Spektrum, 28(2):131–135.

[Fellmann et al., 2012] M. Fellmann, P. Reinke, M. Matijacic, M. Schlicker, O. Thomas, M. Nüttgens: Custom-

er-oriented configuration and specification of mobile application systems for improved TCS productivity and empowerment. In Proc. of the XXII. Internat. RESER Conference, September 20–22, Bucharest, Romania, p. 126 ff., 16 pages.

[Gabriel, 1992] R. Gabriel. Wissensbasierte Systeme in der betrieblichen Praxis. London, McGraw-Hill.

[Ghidini et al., 2009] C. Ghidini, B. Kump, S. Lindstaedt, N. Mahbub, V. Pammer, M. Rospocher und L. Serafina. MoKi: The Enterprise Modelling Wiki. Proc. Of ESWC 2009 (pp. 831–835). Heraklion, Springer.

[Greaves, 2012] M. Gabriel. Semantics and the crowd. Journal of Zhejang University SCIENCE C, 13(4):247-249.

[Harms, 1999] V. Harms. Kundendienstmanagement. Dienstleistung, Kundendienst, Servicestrukturen und Serviceprodukte; Aufgabenbereiche und Organisation des technischen Kundendienstes. Herne, Neue Wirtschaftsbriefe.

[Hasan und Pfaff, 2006] H. Hasan und C. C. Pfaff. The Wiki: an environment to revolutionise employees' interaction with corporate knowledge. OzChi 2006: Proc. of the 18th Australia Conference on Computer-Human Interaction, Design, Activities, Artefacts and Environments (pp. 377–380). Sydney.

[Haun, 2000] M. Haun. Wissensbasierte Systeme: Eine praxisorientierte Einführung. Renningen, Expert-Verlag.

[Hermann et al., 2009] T. A. Herrmann, U. Kleinbeck und C. Ritterskamp. Innovationen an der Schnittstelle zwischen technischer Dienstleistung und Kunde 2: Methoden und Strategien. Heidelberg, Physica-Verlag.

[Karbach und Linster, 1990] W. Karbach und M. Linster. Wissensakquisition für Expertensysteme. Techniken, Modelle und Softwarewerkzeuge. München, Hanser.

[Kirchgeorg, 1991] M. Kirchgeorg. Kundendienst und Marketing - Nutzen Sie Synergien im Kundendienst. In H. Budde (Hrsg), Verkaufs- und Markenprofi (pp. 1–34). Landsberg, Moderne Industrie.

[Kuhlenkamp et al. 2006] A. Kuhlenkamp, S. Manouchehri, I. Mergel und U. Winand. Privatsphäre versus Erreichbarkeit bei der Nutzung von Social Software. HMD Praxis der Wirtschaftsinformatik, 43(252):27–35.

[Kurbel, 1992] K. Kurbel. Entwicklung und Einsatz von Expertensystemen: Eine anwendungsorientierte Einführung in wissensbasierte Systeme. 2. Auflage. Berlin, Springer.

[North, 2011] K. North. Wissensorientierte Unternehmensführung - Wertschöpfung durch Wissen. 5. Auflage. Wiesbaden, Gabler.

[Ontoprise, 2011] Ontoprise GmbH. Semantic Guide 6.2. Produktbeschreibung. Karlsruhe – Ontoprise wurde von Semafora übernommen (vgl. www.semafora-systems.com/de/loesungen/semanticguide/)

[Österle und Senger] H. Österle und E. Senger. Innovative Geschäftskonzepte im After Sales Service. Erfolgreich mit After Sales Services. Berlin, Springer.

[Petter et al., 2008] S. Petter, W. DeLone und E. McLean. Measuring information systems success: models, dimensions, measures, and interrelationships. European Journal of Information Systems, 17(3):236–263.

[Picot, 1989] A. Picot. Der Produktionsfaktor Information in der Unternehmensführung. Thexis, 6(4):3–9.

[Probst et al., 2010] G. Probst, S. Raub und K. Romhardt. Wissen managen - Wie Unternehmen ihre wertvollste Ressource optimal nutzen (6th ed.). Wiesbaden, Gabler.

[Reisberger und Smolnik, 2008] T. Reisberger und S. Smolnik. Modell zur Erfolgsmessung von Social-Software-Systemen. In M. Bichler et al. (Hrsg.), Multikonferenz Wirtschaftsinformatik 2008 (pp. 565–577). Berlin, GITO Verlag.

[Salzburg Research, 2011] Salzburg Research. KIWI - Knowledge in a Wiki. http://www.salzburgresearch.at/projekt/kiwi/, abgerufen am 04. März 2013

[Schaffert et al., 2009a] S. Schaffert, F. Bry, J. Baumeister, und M. Kiesel. Semantische Wikis. In A. Blumauer und T. Pellegrini (Hrsg.), Social Semantic Web. Web 2.0 - Was nun? (pp. 245–258). Berlin, Springer.

[Schaffert et al., 2009b] S. Schaffert, J. Eder, S. Grünwald, T. Kurz und M. Radulescu. Kiwi – A Platform for Semantic Social Software. Proc. of the ESWC 2009 (pp. 888–892). Berlin, Springer.

[Schlicker und Leinenbach, 2010] M. Schlicker und S. Leinenbach. Das INTERACTIVE-Serviceportal. In O. Thomas, P. Loos, und M. Nüttgens (Hrsg), Hybride Wertschöpfung: Mobile Anwendungssysteme für effiziente Dienstleistungsprozesse im technischen Kundendienst (pp. 236–258). Berlin, Springer.

[Schröder, 1997] M. Schröder. Informationsverarbeitung im Kundendienst. Wiesbaden, Deutscher Universitäts-Verlag GmbH.

[Smolnik und Riempp, 2006] S. Smolnik und G. Riempp. Nutzenpotentiale, Erfolgsfaktoren und Leistungsindikatoren von Social Software für das organisationale Wissen. HMD Praxis der Wirtschaftsinformatik, 43(252):17–26.

[Stock und Tochtermann, 2010] A. Stock und K. Tochtermann. Wissenstransfer mit Wikis und Weblogs - Fallstudien zum erfolgreichen Einsatz von Web 2.0 in Unternehmen. Wiesbaden, Gabler.

[VDMA, 2008] VDMA Informatik. Entscheidungshilfe zur Einführung von Servicemanagementsystemen. Frankfurt am Main, VDMA Verlag.

[Walter, 2009] P. Walter. Modellierung technischer Kundendienstprozesse des Maschinen- und Anlagenbaus als Bestandteil hybrider Produkte. In O. Thomas und M. Nüttgens (Hrsg), Dienstleistungsmodellierung (pp. 129–145), Berlin, Physica.

[Weichert, 2003] S. Weichert. Der Knowledge Engineering Prozess bei der Entwicklung eines wissensbasierten Konfigurationssystems für die Blaupunkt GmbH. Universität Hildesheim, Fachbereich III – Informations- und Kommunikationswissenschaften.

[Witte, 2007] H. Witte. Allgemeine Betriebswirtschaftslehre: Lebensphasen des Unternehmens und betriebliche Funktionen. München, Oldenbourg Wissenschaftsverlag.

Explanation in Episodic and Continuous Decision Support Systems

Joachim Baumeister[1,2] **and Albrecht Striffler**[2]
[1] University of Würzburg, Institute of Computer Science, Am Hubland, 97076 Würzburg, Germany
[2] denkbares GmbH, Friedrich-Bergius-Ring 15, 97076 Würzburg, Germany
{firstname.lastname}@denkbares.com

Abstract

Advanced decision support systems demand for their episodic and collaborative use in order to solve complex problems. Further, continuous knowledge representations help to build large knowledge spaces. Decision support systems enhanced in these ways, however, require new approaches to explain the derived decisions. In this paper, we propose an explanation approach that is based on the standardized PROV ontology. We discuss its applicability by giving practical examples.

1 Motivation

A new type of decision support systems is emerging in the practical use in industry. *Episodic and continuous decision support systems* are an advanced interpretation of knowledge-based systems. In comparison to classical decision support systems they emphasize the episodic use of the system for finding (complex) decisions and the use of a continuous knowledge representation for representing large knowledge spaces.

- *Episodic decision making:* A (complex) decision is not made during a single session, but the actual decision process is partitioned over time into different *episodes*. Each episode typically covers a different aspect of the decision process and often more than one user is participating in the episodes. In consequence, we face a (collaborative) decision process; here, a complex decision is taken by the aggregation of a collection of sub-decisions. The sub-decisions cover different aspects of the decision and they are derived in different episodes and by possibly different users.
- *Continuous knowledge representation:* In traditional knowledge systems a single knowledge representation is used to build the knowledge base. More complex and larger systems do benefit from the use of hybrid approaches, integrating different representations into one knowledge base. Here, for a single decision/fact different knowledge representations can be *continuously* interweaved. Systems then have to deal with multiple representations during the reasoning process.

For (complex) decision support systems it is necessary to provide helpful explanation for the derived decisions. In the literature [Roth-Berghofer and Richter, 2008] an explanation scenario is described as depicted in Figure 1.

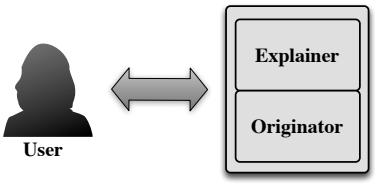


Figure 1: A general explanation scenario.

The user demands for an explanation and interacts with an explanation component. The component consists of the *Explainer* and the *Originator*; in our application scenario the originator is the decision support system, whereas the explainer is the component of the system to generate the explanations to the user.

Transparent explanations improve the general acceptance of users, but can be also used for tutorial and legal purposes by showing the reasons for a particularly derived decision. The commercial application of an episodic and continuous decision support system showed the demand for a new approach in explanation. The following requirements were expressed by the regular users of the system:

- The explanation has to show the temporal development (i.e., the episodes) of a particular decision process.
- All participating users and their competencies have to be integrated in the explanation.
- The explanation has to handle the use of different knowledge representations that were applied for decision making.

This paper presents an extensible explanation approach that meets the requirements stated above. Its main idea is the interpretation of the explanation data as provenance of the entered data and derived decisions but also the decision process itself. As the formal model to represent provenance data the PROV ontology is applied to implement this approach [W3C, 2013a].

The rest of the paper is organized as follows: Section 2 introduces the characteristics of episodic and continuous decision support systems. Section 3 sketches the main ideas of the provenance ontology PROV-O and describes its application within the explanation of decision support systems. Explanation queries can be interpreted as queries to the ontology. In Section 4 typical explanation queries are implemented as SPARQL queries to demonstrate the principal applicability. The paper concludes with a summary of the presented work in Section 5. Also an outlook to future work is given.

2 Episodic and Continuous Decision Support Systems

Before we put more detail on the explanation concept we introduce the concepts of episodic and continuous decision making.

Episodic Decision Making

A complex decision is made not in a "one-shot" session, but multiple users have to contribute their expertise to find a reasonable overall decision. Typically the decision process can be partitioned into a number of aspects that are covered separately. The aggregation of different aspects contributes to an overall decision. Typically, the outcomes of the aspects are represented as (sub-)decisions.

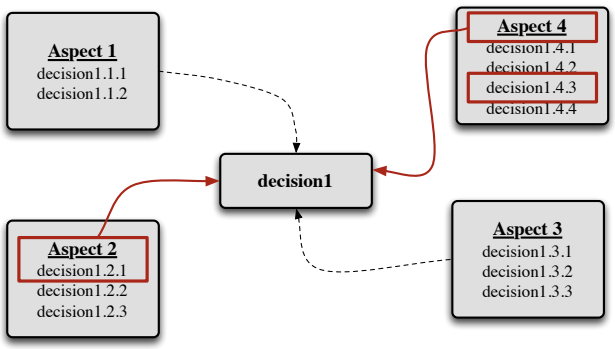


Figure 2: Four aspects are contributing to a final decision `decision1`. Aspects 2 and 4 are responsible for the actual derivation of the decision.

In Figure 2 the aspects 1–4 for the single decision `decision1` are shown. Each aspect itself is represented by a number of questions that need to be answered, so that a decision can be derived. In the example, aspect 2 and aspect 4 actually provide the derived decisions `decision1.2.1` and `decision1.4.3` that support the derivation of `decision1`.

Often not all aspects are covered by a single person but every aspect is handled by a different person or expert group. In medicine, for instance, there exists specialists for the different organs of the human. For a complex evaluation of the patient's physical state, more than one specialist may be consulted. In the technical domain, we see a similar setting: In complex machinery there also exist specialists for the different components of the machine.

Figure 3 shows an exemplary decision process, where the different Users 1–3 are collaborating in a decision process over the time.

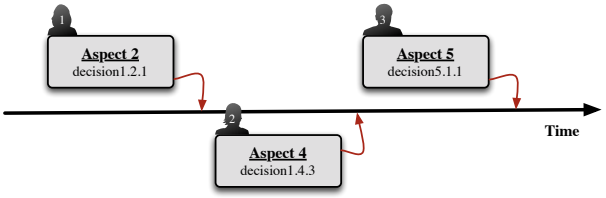


Figure 3: Example decision episodes over time, where different aspects are covered.

The first two decisions `decision1.2.1` and `decision1.4.3` are taken from the previous example shown in Figure 2, whereas User 3 contributes `decision5.1.1` which not relevant for the final derivation of `decision1`.

Use of Continuous Knowledge Representations

In complex domains it often is not possible/reasonable to build the entire knowledge base by a single knowledge representation. More precisely, some parts of the knowledge are preferably not formalized by explicit knowledge representations—such as rules or models. Typical reasons for a hybrid approach are as follows:

- *Uncertain domain knowledge:* Parts of the domain are not well-understood in a technical sense. Here, decisions in practice are often based more on past experiences, evidence, and intuition than on strict domain laws and rules.
- *Bloated domain knowledge:* For some parts of the domain, the explicit representation of the knowledge would be too time-consuming and too complex. For instance, much background knowledge needs to be included, that is required for proper decision making. Here, the expected cost-benefit ratio [Lidwell *et al.*, 2003, p. 56] is low, e.g., because many parts will be rarely used in real-world decisions.
- *Restless domain knowledge:* Especially in technical domains, some parts of the domain knowledge are frequently changing due to technological changes. The explicit representation of these parts would require frequent and costly maintenance. Here, also the cost-benefit of the maintenance vs. the utility of the knowledge needs to be evaluated.

Consequently, mixing different knowledge representations with less formal elements seems to be promising. In the past, the knowledge formalization continuum [Baumeister *et al.*, 2011a] was introduced as a mental model to represent different representations in a single systems.

As a pragmatic reasoning approach, we propose to connect the different representations by a common taxonomy of decisions. That way, the different knowledge elements share the same decision space and thus are able to derive the same set of decisions within one process.

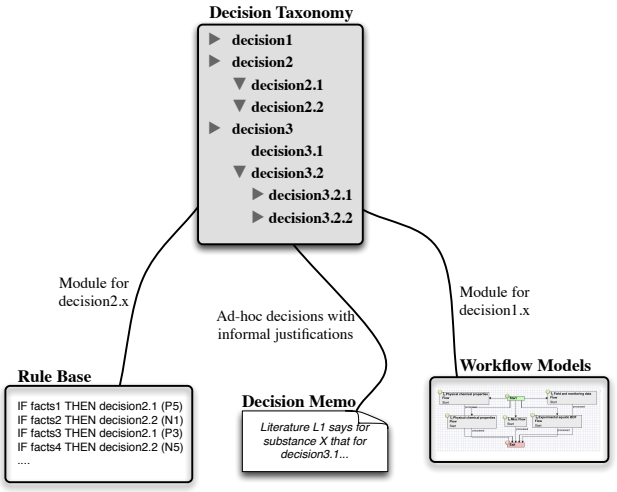


Figure 4: Connecting different knowledge elements by the use of a common decision taxonomy.

In Figure 4, rules, workflow models, and informal decision memos are connected by a taxonomy of decisions.

The combined knowledge base is able to derive the same set of decisions albeit the representations used in reasoning are differing. In literature, formal approaches such as RIF [W3C, 2013b] apply a comparable connection, i.e., decisions are formalized as concepts/instances and rules are defined to derive the existence of the concept/instance.

In a more elaborated approach, the knowledge elements are able to derive decisions in a weighted manner. Here, we propose a score-based approach. Scores have been regularly used as a weighting scheme in knowledge engineering [Puppe, 1998; Miller *et al.*, 1982]. By using scores each knowledge element is not only able to derive a solution categorically, but can attach a score weight to a decision. Every decision provides an account that stores the scoring weights given to the decision during the reasoning process. If a knowledge element "fires", then the corresponding score weight is added to the account of the particular decision. All scoring weights of a single decision are aggregated to a final weight. If the final weight exceeds a predefined threshold, then the decision element is established.

Example: We define a universal set of score weights S = {N1, N2, N3, 0, P1, P2, P3}, where P1,...,P3 are positive score weights and N1,...,N3 are negative score weights. The sum of two equal categories results in the next higher category (e.g. P2 + P2 = P3). A negative and the corresponding positive score weight nullify each other (e.g., N2 + P2 = 0). A decision is established (confirmed), if the aggregation of the collected scoring weights exceeds the score weight P3. In Figure 5, we see that a rule fired a score weight P1 to `decision1`. The decision `decision2` and `decision5` are established, since the aggregation of their score weights exceeds the threshold P3. The decision `decision4` is not established, since a negative weight N3 nullified the positive weight P3.

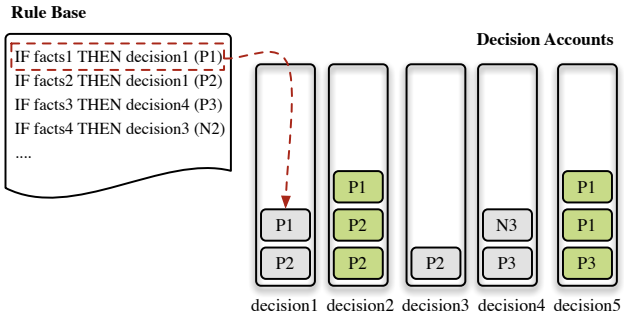


Figure 5: Score accounts of five decisions, and a rule firing a new score weight to the account of `decision1`.

The same decision accounts are also filled by other used knowledge types. For instance, the score weight of an entered decision memo contributes to the account of `decision1`. Further, a traversed workflow model can derive a score weight for `decision2`, which is also added to the decision account.

In this section we described the internal representation of a continuous knowledge representation by the common use of a decision taxonomy. In the following section we introduce an approach to provide explanation capabilities for such a kind of systems.

3 Provenance and Explanation in Decision Support Systems

As motivated above, the process of making a complex decision often involves a number of people contributing to the decision process. Furthermore, the process itself is taking place over a longer period of time. For these reasons, it is very important that the derived decisions are understandable and transparent for all users. These requirements imply the versioning and documentation of decisions and data entries. Changes need to be traceable, as for instance described by [Noy and Musen, 2002; Franconi *et al.*, 2010]. Further also a means of representing the decision process itself is required to be used in an explanation component. Such a component needs to answer (at least) the following questions:

- At which time a particular data was entered and who entered that data?
- Which knowledge elements are responsible for a particular decision?
- What is the history of a particular data and decision?
- Which persons contributed to the process of a particular decision?

We propose the application of the PROV ontology to knowledge elements and the entities of the decision process. The PROV ontology explicitly represents the provenance of entities, i.e., in our case *decisions*, *entered data*, etc. are interpreted as PROV entities. We first give a brief overview of PROV-O and then show its application to decision processes.

3.1 The PROV Ontology in a Nutshell

The PROV ontology [W3C, 2013a] distinguishes three levels of terms defined in the ontology:

1. *Starting Point Terms* to be used to express the basic knowledge about provenance of data.
2. *Expanded Terms* for more expressive definitions of relationships in provenance.
3. *Qualified Terms* integrating the Qualification Pattern [Dodds and Davis, 2012] into the PROV ontology for a very expressive representation of the provenance of data.

In this section we select a helpful subset of "starting points" and "expanded terms" and describe the concepts and relations, that are useful to represent the provenance of decision support systems. In Figure 6 these concepts and relations are depicted.

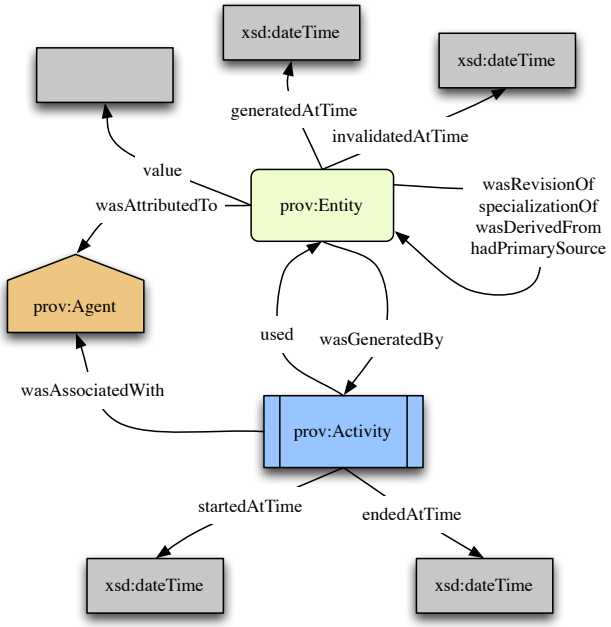


Figure 6: Selected elements of the PROV ontology.

For concepts defined in the PROV ontology the prefix `prov` is used. The three classes `prov:Agent`, `prov:Activity`, and `prov:Entity` are central for describing provenance information. An `prov:Agent` is executing an `prov:Activity` and produces an `prov:Entity`. Consequently, an `prov:Entity` can be attributed to an `prov:Agent` and the `prov:Entity` was generated by a specific `prov:Activity`. An `prov:Activity` is also associated with an `prov:Agent`. In some processes an `prov:Activity` uses an `prov:Entity` for the creation of another `prov:Entity`. An `prov:Activity` has a start and an end time; this is related to the generation time of an `prov:Entity`. When the `prov:Entity` is superseded by a revision (`prov:wasRevisionOf`), then the prov:Entity is invalidated at a specified time. The following properties are also of interest: The property `prov:wasDerivedFrom` states that an instance of `prov:Entity` was transformed into another instance. In decision support systems, the primary source of a specific `prov:Entity` is also of interest (see property `prov:hadPrimarySource`).

In its basic setting we see that the PROV ontology is a suitable starting point for general explanation capabilities in decision support systems. In the following we describe the specific extensions for such systems together with application scenarios.

3.2 The PROV Ontology for Decision Support Systems

When integrating the PROV ontology into decision support systems, we consider a *process-centered provenance*. Here, actions and steps are represented that are used for producing a particular decision.

For the application of the PROV ontology in decision support systems we introduce a number of new concepts sub-classing the known core concepts of PROV. In Figure 7 the most important subclasses are depicted; the prefix `dss` is used for classes introduced for decision support.

At the top of the figure the subclasses of `prov:Entity` are shown: A simple entity can be a decision (`dss:Decision`) or entered data (`dss:FormValue` and `dss:DecisionMemo`). Every decision is associated with a `dss:DecisionAccount` instance, which stores the score weights using `dss:ScoreWeight` instances. The decision account itself is a complex entity, i.e., a `prov:Collection`.

The extension of the concept `prov:Activity` knows two sub-classes: 1) for entering data in memos (`dss:MemoEntry`) and for answering question in forms (`dss:FormEntry`); 2) for the actual derivation of a decision. The latter activity is central for the explanation of different decisions derived during a process.

Two different `prov:Agent` sub-classes are introduced: `dss:TeamMember` to represent users participating in the collaborative decision process and `dss:DomainSpecialist` for building the explicit knowledge base and for giving expert decisions. Further `dss:DSS` represents the actual decision support system.

In a concrete scenario instances of the classes are created storing the provenance information of the decision making process. We demonstrate the concrete use by an example, where instances of the example use the prefix `ex`.

A taxonomy of decisions is typically formalized by narrower/broader relations of the SKOS [W3C, 2009] ontology.

```
ex:decision1 rdf:type dss:Decision;
   skos:narrower ex:decision1.1;
   skos:narrower ex:decision1.2;
   skos:narrower ex:decision1.3.
ex:decision2 rdf:type dss:Decision;
   skos:narrower ex:decision2.1;
   skos:narrower ex:decision2.1.
```

In the example, the decisions `ex:decision1` and `ex:decision2` are defined together with more specific sub-decisions. A particular decision can be inferred in different ways: 1) The use of explicit inference knowledge used by a decision support system. 2) The creation of a decision memo by a user. We examine both alternatives in the following.

Explicit Inference of Decisions

The explicit reference and the ontological representation, respectively, is exemplified by the activity of a firing rule. The instances are graphically depicted in Figure 7-(4).

The user `ex:teamMemberMM` enters a `ex:formValue1` entity during the activity `ex:formEntry1`. The entity `ex:formValue1` is responsible for deriving a specific score weight `ex:scoreWeight1`. The derivation of the score weight is represented by an instance of `dss:DecisionDerivation`, i.e., `ex:decisionDerivation1`, which itself uses a rule included in the rule base `ex:ruleBase1` The activity `ex:decisionDerivation1` added the score weight to the defined score account `ex:decisionAccount1`, which itself is linked to the corresponding decision `ex:decision1`.

Inference by Decision Memos

The reasoning and representation of decisions taken by decision memos is very similar to the approach described above for rules. Actually, only the instances of `dss:DecisionDerivation` and `dss:Data` are connected differently to the score weight of the decision. In Figure 7-(4) a dotted box is depicted at the right. The box

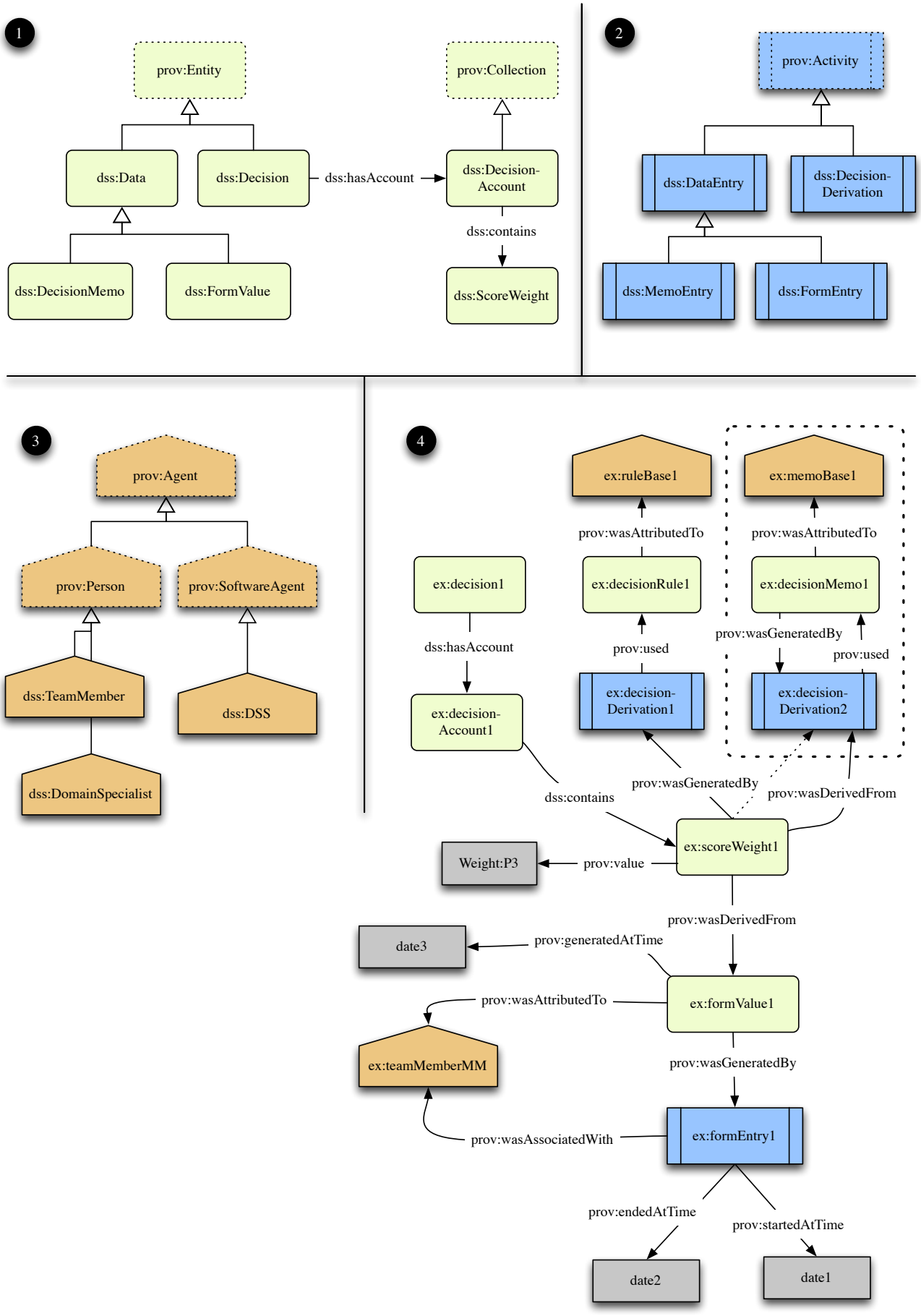


Figure 7: (1) Subclasses and relations `prov:Entity`; (2) Specific subclasses of `prov:Activity` and (3) specific subclasses of `prov:Agent`; (4) Instances and relations created when a decision rules fires for a concrete decision `ex:decision1`. The dotted boxes show the alternative use case, where a decision memo is responsible for deriving the decision.

shows the alternative instances when entering a decision memo.

The user `ex:teamMemberMM` enters the decision memo `ex:decisionMemo1` during the activity `ex:memoEntry1`. In the decision memo the user also enters a score weight for `ex:decision1`. Consequently, the decision memo derives the `ex:scoreWeight1` that was generated by the instance `ex:decisionDerivation2`. This instance is connected with the actual `ex:decisionMemo1` stored in the data base for memos (`ex:memoBase1`).

4 Querying for Explanation

By using an ontology representation of the decision process, the justification of a particular decision can simply be queried. Moreover, ad-hoc explanations can easily be constructed by new queries. When representing the ontology as RDF triples, the standard query language SPARQL [W3C, 2013c] can be used. Of course, an intuitive visualization of the query results needs to be defined, but this is typically up to the application front-end of the decision support system.

In the following, we exemplify the explanation capabilities of the presented PROV extension by defining the SPARQL queries for the questions posed at the beginning of Section 3. It is worth noticing that the definition of SPARQL queries is up to the administrators of the system. For end-users the results of these queries should be presented in a user-friedly manner.

For demonstration purposes we implemented an extended version of the example depicted in Figure 7 in the knowledge modelling environment KnowWE [Baumeister *et al.*, 2011b; 2012].

At which time was a particular data entered and who entered the data?

The following SPARQL query lists all entries and the persons involved in creating a corresponding entry. Additionally the generation time of the entry is shown.

```
SELECT ?entry ?person ?time
WHERE {
  ?entry
     prov:wasGeneratedBy ?activity.
  ?activity
     prov:wasAssociatedWith ?person.
  ?entry
     prov:generatedAtTime ?time.
}
```

Figure 8 shows the results of the SPARQL query above: The three entities `decisionMemo2`, `formValue1`, and `formValue3` are listed with their creators and creation date.

entry	person	time
decisionMemo2	teamMember1	2013-07-01T16:06:00
formValue3	teamMember3	2013-06-02T16:10:00
formValue1	teamMember1	2013-07-01T16:10:00

Figure 8: Entities and persons involved in the creation of the entities.

The query can be further constrained to a specific data entry. Then this query is similar to the last query of this section.

Which knowledge elements are responsible for a particular decision?

The following SPARQL query inspects the connected nodes of the decision instance `ex:decision1` in order to check for derived values. The `FILTER NOT EXISTS` extension guarantees that only valid entities are shown.

```
SELECT ?givenValue ?byKnowledge
WHERE {
  ex:decision1
     dss:hasAccount ?account.
  ?account
     dss:contains ?weights.
  ?weights
     prov:value ?givenValue.
  ?weights
     prov:wasGeneratedBy ?activity.
  ?activity
     prov:used ?byKnowledge.

  FILTER NOT EXISTS {
    ?weights
    prov:invalidatedAtTime
    ?invalidated. }
}
```

Figure 9 shows the results of the SPARQL query above: Derived values are shown together with the knowledge elements—decision memos and rules–that are responsible for the existence of the values.

givenValue	byKnowledge
P1	decisionRule3
P3	decisionMemo2

Figure 9: Valid values derived together with the acting knowledge element.

What is the history of a particular data and decision (including involved persons)?

For a given decision `ex:decision1` the following SPARQL query identifies all values that were given to that decision. For each value, also the used knowledge element and the acting person is retrieved. Also the validity of the value is printed; values with empty invalidated column are currently valid.

Figure 10 shows the results of the query: We see that `ex:decision1` retrieved three values, whereas value P2 is already invalidated.

value	usingKnowledge	byPerson	generated	invalidated
P1	decisionRule3	teamMember3	2013-06-02T16:10:00	
P2	decisionRule1	teamMember1	2013-07-01T16:10:00	2013-07-01T12:00:00
P3	decisionMemo2	teamMember1	2013-07-01T16:06:00	

Figure 10: History of values for the given decision `ex:decision1`.

```
SELECT ?value ?usingKnowledge ?byPerson
       ?generated ?invalidated
WHERE {
  ex:decision1
     dss:hasAccount ?account.
  ?account
     dss:contains ?data.
  ?data
     prov:value ?value.
  ?data
     prov:wasDerivedFrom ?entity.
  ?data
     prov:wasGeneratedBy ?activity.
  ?activity
     prov:used ?usingKnowledge.
  ?entity
     prov:wasAttributedTo ?byPerson.
  ?entity
     prov:generatedAtTime
        ?generated.

  OPTIONAL {
     ?data prov:invalidatedAtTime
     ?invalidated. }
}
```

5 Conclusions

We conclude the paper with a brief discussion and an outlook to the future work.

5.1 Discussion

Advanced decision support systems allow for the distributed and episodic handling of complex decision problems. They handle large knowledge spaces by mixing different knowledge representations with informal decision justifications. When implemented in a distributed setting, the transparent justification of derived decisions is of prime importance. In this paper we introduced an explanation approach of continuous knowledge representations that is based on the PROV ontology. We described how an ontology representation of the decision process and the derived decisions can be used to generate transparent explanations.

In the literature the related ontology models can be found: [Evangelou *et al.*, 2005] describe an ontology to support collaborative decision-making. They propose the model KAD (Knowledge-Argument-Decision) to facilitate exchange between decision makers and their argumentation. The KAD ontology model defines the three main classes `discussionParticipant`, `coreEntity`, and `coreProcess`, where their semantics is related or can be aligned to the starting point terms of the PROV ontology. Here, more focus is set on supporting the argumentation and discussion between decision makers. [Kornyshova and Deneckère, 2010] also propose an ontology for decision making. The *decision making ontology* (DMO) tries to support IS engineers in their decision making during an information systems project. The proposed ontology is evaluated by instantiating it to a requirements engineering process. The ontology is very elaborated and could be connected with the PROV ontology. For our purposes (continuous knowledge representation and episodic use) the extensions described in Figure 7 (1) need to be also made.

5.2 Future Work

At the current state, explanations are based on SPARQL queries. Albeit a very general approach, the construction of such queries can be cumbersome for standard users. For this reason we aim to define a simplified language to define explanation queries quickly in an intuitive manner. In Section 4 we demonstrated the access to typical explanation queries by SPARQL expressions. Although the shown results include all relevant information needed for an explanation, the presentation is likely to be not very intuitive. Therefore we are planning to investigate ontology visualization approaches [Fluit *et al.*, 2002; Katifori *et al.*, 2007] to render the results of the explanation query in a more user-friendly manner.

As the next practical step we are planning to implement and evaluate the proposed ontology and explanation capabilities for a decision support system, that is already in use. The KnowSEC system supports the decision work for chemical safety within a unit of the Federal Environment Agency in Germany (Umweltbundesamt). At the current state, the systems manages more than 42.000 sub-decisions for more than 11.000 chemical substances; many of the decisions were automatically derived. We refer to [Baumeister *et al.*, 2013] for more details.

References

[Baumeister *et al.*, 2011a] Joachim Baumeister, Jochen Reutelshoefer, and Frank Puppe. Engineering intelligent systems on the knowledge formalization continuum. *International Journal of Applied Mathematics and Computer Science (AMCS)*, 21(1), 2011.

[Baumeister *et al.*, 2011b] Joachim Baumeister, Jochen Reutelshoefer, and Frank Puppe. KnowWE: A semantic wiki for knowledge engineering. *Applied Intelligence*, 35(3):323–344, 2011.

[Baumeister *et al.*, 2012] Joachim Baumeister, Jochen Reutelshoefer, Volker Belli, Albrecht Striffler, Reinhard Hatko, and Markus Friedrich. KnowWE - a wiki for knowledge base development. In *The 8th Workshop on Knowledge Engineering and Software Engineering (KESE2012)*, http://ceur-ws.org/Vol-949/kese8-05_04.pdf, 2012.

[Baumeister *et al.*, 2013] Joachim Baumeister, Albrecht Striffler, Marc Brandt, and Michael Neumann. Towards continuous knowledge representations in episodic and collaborative decision making. In *The 9th Workshop on Knowledge Engineering and Software Engineering (KESE2013)*, 2013.

[Dodds and Davis, 2012] Leigh Dodds and Ian Davis. *Linked Data Patterns Linked Data Patterns Linked Data Pattern*. http://patterns.dataincubator.org/book, 2012.

[Evangelou *et al.*, 2005] Christina Evangelou, Nikos Karacapilidis, and Omar Abou Khaled. Interweaving knowledge management, argumentation and decision making in a collaborative setting: the kad ontology model. *International Journal of Knowledge and Learning*, 1(1):130 – 145, 2005.

[Fluit *et al.*, 2002] Christiaan Fluit, Marta Sabou, and Frank van Harmelen. Supporting User Tasks through Visualisation of Light-weight Ontologies. In *Handbook on Ontologies in Information Systems*, pages 415–432. Springer, Berlin, 2002.

[Franconi *et al.*, 2010] E Franconi, T Meyer, and I. Varzinczak. Semantic diff as the basis for knowledge base versioning. In *13th International Workshop on Non-Monotonic Reasoning (NMR)*, pages 7–14, 2010.

[Katifori *et al.*, 2007] Akrivi Katifori, Constantin Halatsis, George Lepouras, Costas Vassilakis, and Eugenia Giannopoulou. Ontology visualization methods - a survey. *ACM Comput. Surv.*, 39(4), November 2007.

[Kornyshova and Deneckère, 2010] Elena Kornyshova and Rébecca Deneckère. Decision-making ontology for information system engineering. In Jeffrey Parsons, Motoshi Saeki, Peretz Shoval, Carson C. Woo, and Yair Wand, editors, *ER*, volume 6412 of *Lecture Notes in Computer Science*, pages 104–117. Springer, 2010.

[Lidwell *et al.*, 2003] William Lidwell, Kritina Holden, and Jill Butler. *Universal Principles of Design*. Rockport Publishers, October 2003.

[Miller *et al.*, 1982] Randolph A. Miller, Harry E. Pople, and J. Myers. INTERNIST-1, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. *New England Journal of Medicine*, 307:468–476, 1982.

[Noy and Musen, 2002] Natalya F. Noy and Mark A. Musen. PromptDiff: a fixed-point algorithm for comparing ontology versions. In *In 18th National Conference On Artificial Intelligence (AAAI-2002*, pages 744–750, 2002.

[Puppe, 1998] Frank Puppe. Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-Kit D3. *International Journal of Human-Computer Studies*, 49:627–649, 1998.

[Roth-Berghofer and Richter, 2008] Thomas R. Roth-Berghofer and Michael M. Richter. On explanation. *Künstliche Intelligenz*, 22(2):5–7, May 2008.

[W3C, 2009] W3C. SKOS Simple Knowledge Organization System reference: http://www.w3.org/tr/skos-reference, August 2009.

[W3C, 2013a] W3C. PROV-O: The PROV Ontology: http://www.w3.org/tr/prov-o/, April 2013.

[W3C, 2013b] W3C. RIF-Core Recommendation: http://www.w3.org/tr/rif-core/, February 2013.

[W3C, 2013c] W3C. SPARQL 1.1 recommendation: http://www.w3.org/tr/sparql11-query/, March 2013.

Case-Based Forecasts Enhancing Decision Support for Capacity-Planning in Higher Education

Daniel Pöppelmann
University of Osnabrueck
D-49069, Osnabrueck, Germany
dpoeppel@uos.de

Abstract

Academic capacity planning is a knowledge-intensive process that has to be based upon forecasts of course demand. Forecasts have to take into account each student's current course achievements, prospective future course selections, time constraints as well as a wide range of different rules for graduation. This paper presents an innovative concept for forecasting of course enrolments serving as demand-figures in academic capacity planning processes and fulfilling information needs of decision makers on various levels in German higher education. Adaptability to a wide range of different study programs is ensured by employing a refined case-based reasoning approach. The case-base is dynamically interpreted with regard to stored cases' problem descriptions and solutions. Moreover the structure of cases is heterogeneous depending on the students' course achievements. Furthermore the methodology of case-based reasoning is enhanced by including a rule-based reasoning component as well as a web-based component for the adaptation of proposed solutions. The results of the case-based reasoning processes are loaded to a star-schema to support capacity planning by a data-driven decision support system. The concept is evaluated in terms of correctness of retrieval as well as accuracy of forecasts by contrasting its results with those of a simple regression forecast using real student data.

1 Introduction

1.1 Motivation and Problem Statement

Dramatic changes have taken place in the German Higher Education environment in the recent decade. The Bologna process led to a high increase in the number of different study programs with heterogeneous, complex curricula. Institutions are intensively competing with each other for enrolments and are confronted with a scarcity of monetary and non-monetary resources (Alt & Auth, 2010; U. Hansen, Henning-Thurau, & Langer, 2000; Küpper, 2010). Additionally, the total number of enrolments constantly rises and universities are confronted with very heterogeneous groups of demanding students (Löwer, 2012). As a consequence universities need to offer a demand-oriented portfolio of advisory services and education (Rieger, Haarmann, Höckmann, & Lüttecke, 2009). With regard to the scarcity of resources it is indispensable to ensure and increase efficiency and effectiveness within institutions especially in the core process of teaching. To enable efficient decision making different groups within the university have to be supplied with information regarding this process on different levels of aggregation. Top management, i.e. the presiding committee, is concerned with achievement-oriented allocation of resources among the university's faculties (Reichwald, 1998). Thus it needs information on capacity utilization (Nusselein, 2002), i.e. for example the number of exams to be taken or student enrolments, on the level of different faculties. Information has to be highly aggregated for decision makers within this group (Postert, 2001). Middle management, i.e. deans and deans of study affairs, need to allocate monetary and non-monetary resources to the faculty's chairs and are coping with satisfying student demand for specific classes by assigning workloads to lecturers and eliminating overlapping of lectures (T. F. Burgess, 1996; Reichwald, 1998). Hence they need information on capacity utilization on the level of chairs, lecturers or single classes and the respective classrooms' sizes and utilization. The maximum aggregation level for this group of decision makers will be a single department (Postert, 2001). Lower management, i.e. professors, is directly concerned with teaching. In order to assign assistants to the right courses as well as to adapt course-contents to the skills and interests of their students they need information on the number and type, e.g. repeaters, of students in each of their courses (Rieger et al., 2009). The aggregation level of information will be relatively low for this group of decision makers (Postert, 2001). To ensure competitive advantages students should be considered as a relevant group of decision makers and be proactively supplied with information (Rieger et al., 2009), too. Information for this group of decision makers is on the lowest level of aggregation, i.e. a single student, and has a primary focus on planning the course of studies (Postert, 2001). In summary, universities and their decision makers are faced with an environment that resembles that of private businesses with international competition (Rieger et al., 2009) and information needs are heterogeneous in terms of aggregation levels across the groups of decision makers.

As a means of efficient and effective resource utilization private businesses employ the method of capacity planning (Oden, Langenwalter, & Lucier, 1993; Slack, 2010). For applying this method knowledge of current or preferably forecasted demand figures is necessary (Schonberger & Knod, 1991; Slack, 2010). Unfortunately, demand figures within the academic environment, i.e. future course enrolments, tend to be fraught with high uncertainty as they are strongly influenced by students'

individual choices e.g. the postponing of enrolment for courses or the selection of majors. Additionally, the increasing complexity of curricula and graduation requirements as well as the growing heterogeneity of student groups makes it hardly possible to estimate future demands. Thus information needs of decision makers involved in capacity planning –as described above – can hardly be satisfied. Therefore capacity planning processes within institutions are difficult to implement and rarely found in the German Higher Education Area.

1.2 Research objective and contribution

In terms of (March & Smith, 1995) the paper presents an innovative concept, namely for forecasting future course enrolments serving as demand-figures in academic capacity planning processes and fulfilling information needs of decision makers on various levels in higher education. The underlying methodology is case-based reasoning (CBR) since it is perfectly suited for weak theory domains for which deep causal models can hardly be derived (Cunningham, 1998) – as it is the case in the domain at hand. In contrast to previous approaches students' individual choices are explicitly considered. Moreover the concept ensures adaptability to various different programs of study with little knowledge engineering effort. The concept is derived and evaluated by employing the method of prototyping (Wilde & Hess, 2006) and includes some major revisions of the CBR methodology that were necessary to fit it to the higher education domain. Evaluation is done with real student data for an undergraduate program at a medium sized German university. Addressed decision makers are on multiple levels of a university, namely students, faculty, deans and top management. Practitioners as well as researchers in the fields of higher education management and artificial intelligence, especially the area of CBR, are addressed in the paper.

With regards to the field of higher education management literature offers some concepts that explicitly target at supporting decision makers with forecasted demand figures. Forecasts are mainly derived by Markovian analyses (Bessent & Bessent, 1980; Kassicieh & Nowak, 1986), network simulation models (R. R. Burgess, 1970) or failure rate-based mathematical models (Deniz, Uyguroglu, & Yavuzr, 2002). These approaches focus on mathematical modeling for single programs of study only and none of them explicitly considers students' individual choices. With influencing factors being considered as parameters of a mathematical model these approaches are static in nature and hardly adaptable to changing environments. Applying these approaches to today's complex and constantly changing curricula would lead to a prohibitive increase in the models' complexity. The concept presented in this paper will overcome these drawbacks. Due to the application of a refined CBR approach the effort for knowledge engineering can be kept low (Watson & Marir, 1994). Moreover the approach is easily adaptable to a wide range of different programs of study. Thus the paper contributes to the field of higher education management by offering an improved, innovative approach to support decision makers on various levels.

Regarding the area of artificial intelligence a contribution is made to the field of CBR: The CBR methodology focuses on using specific knowledge from past experiences for solving new problems (Aamodt & Plaza, 1994). Experiences are stored within a case-base with cases typically containing at least a problem-description and a solution (Cunningham, 1998; Watson, 1997, 2003). An important distinction has to be made between homogeneous and heterogeneous case-bases. Within homogeneous case-bases all cases share the same structure and the same classes of attributes whereas heterogeneous ones are characterized by cases differing from each other in terms of structure and attributes (Watson, 2003). Heterogeneous case-bases implicate difficulties especially in the retrieve and reuse phases of a CBR cycle as attributes cannot be unambiguously assigned to either description or solution (Abou Assali, Lenne, & Debray, 2009; Lopez De Mantaras et al., 2005). The refined CBR approach presented in this paper offers an innovative solution to this problem by introducing a dynamic splitting point for differentiating description and solution attributes. Moreover the CBR cycle proposed by (Aamodt & Plaza, 1994) is enhanced - amongst others – by:

- a strategy for automatically generating new cases from the case-base thus enabling efficient solving of huge amounts of new cases,
- a rule-based component interacting with an ontology for making sure forecasts are in line with graduation requirements,
- a new phase supporting temporarily independent multi-user revision,
- an interface to a data-driven decision support system that supplies decision makers with derived predictive information.

1.3 Research Methodology

In the research project resulting in this paper the principles of design science research were applied. Design science research aims at improving the environment by introducing innovative artifacts (Hevner, 2007). According to (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) the design science research process consists of six activities altogether. The problem addressed by the paper at hand was identified (activity 1) due to deficiencies within a practical setting as well as a literature review aiming at the assessment of approaches to forecast demand figures regarding their applicability in the setting at hand. The problem's relevance is presented in 1.1. A literature review was conducted to identify previous approaches to demand forecasting in higher education. Due to space limitations only a short overview of the deficiencies of existing approaches identified is given in section 1.2. Based on these shortcomings the objectives for a new solution were inferred (activity 2) as presented in 1.2. Employing the method of prototyping (Wilde & Hess, 2006) design and development (activity 3) resulted in the artifact of an innovative concept for forecasting course enrolments and support for decision makers on multiple levels of a university. The concept is thoroughly described in section 2. The developed prototype facilitates the concept's demonstration (activity 4). It was implemented and tested within a real academic setting, forecasting demand

figures for an undergraduate program at a medium sized German university as a proof of concept. Forecasts were statistically evaluated (activity 5) and contrasted with the results of a forecast derived by simple regression. First evaluation results are shown in section 3. With this paper research is made available for the research community (activity 6). Communication with practitioners was achieved by presenting outcomes at one of the major forums on data warehousing in higher education in the beginning of 2012.

2 Conceptual Approach and Implementation

2.1 Conceptual Overview

The designed and prototypically implemented concept consists of a compound decision support system comprising two components. The first one, named CBR component, at its core includes a workflow aiming at forecasting students' individual course-enrolments for one or more upcoming semesters that is based on extensions and refinements of the CBR-cycle introduced by (Aamodt & Plaza, 1994) and that is thus called refined CBR cycle. Forecasts are stored within a database which is used as an interface for the second component, named data-driven component. This second component targets at supplying decision makers on various levels within a university with derived predictive data. A high-level overview of the concept's architecture is given in figure 1:

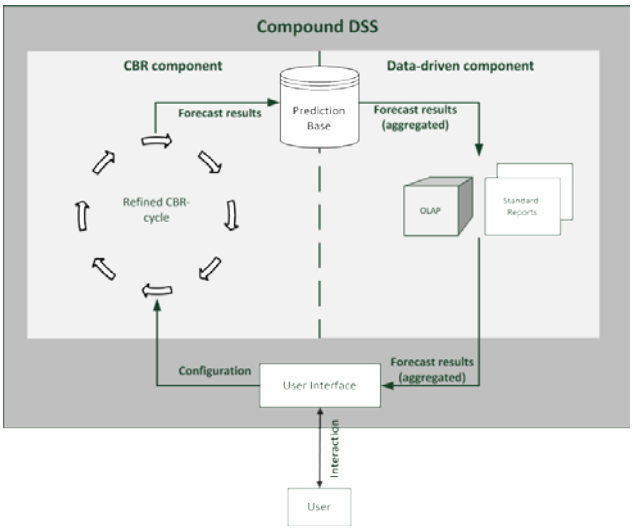


Fig. 1. Conceptual overview

2.2 Refined case-based reasoning cycle forecasting individual enrolments

To fit the CBR methodology to the domain of forecasting enrolments in higher education several refinements had to be made to the CBR cycle introduced by (Aamodt & Plaza, 1994). From a high level point-of-view the concept aims at forecasting students' individual course enrolments by reusing past enrolments of similar students. Figure 2 presents an overview of the concept of the refined CBR cycle.

The refined CBR cycle consists of the seven phases Initialize, Retrieve, Reuse, Repeat, Revise 1, Save, and Revise 2. For the development of the prototypical application the jColibri2 CBR framework (Recio García, 2008) was used in its version 2.1 as it provides predefined components that are especially useful for object-oriented case representation, persistence of cases in relational databases as well as predefined (local) similarity functions.

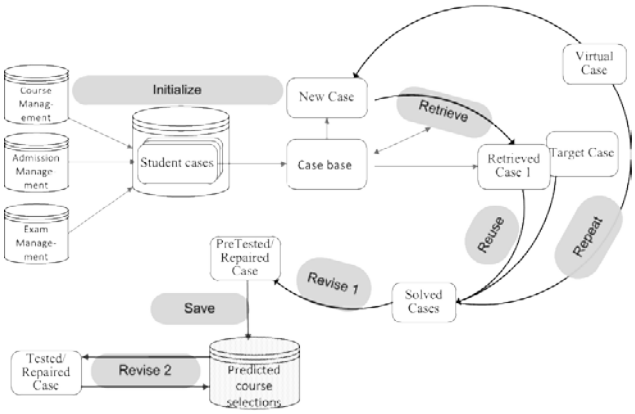


Fig. 2. Component 1: Refined case-based reasoning cycle

During the first phase **Initialize** student data is extracted from operational systems, transformed to a case model and loaded to an Oracle database. At runtime case data is loaded to an in-memory case-base. This may be seen as pre-processing not belonging to the core of a CBR-cycle. However supplying the refined CBR-cycle with fresh data from operational systems at runtime is a key to deriving correct forecasts. Thus building cases from raw data is explicitly included in the cycle. The concept comprises a case model regarding each student as one case containing personal attributes, e.g. age, gender and a-levels grade, as well as attributes representing the student's previously taken courses, e.g. each semester's gpa and exams written. Each case consists of about 30 attributes altogether that were selected pragmatically from the set of attributes available from a campus management system to keep the effort for knowledge engineering low. As the number and kinds of taken programs, courses as well as the study semesters courses were taken in differ from student to student a flexible case-representation is needed. Thus representation follows the principals of object-orientation allowing for cases with different structures (Bergmann & Stahl, 1998). The Initialize phase results in a case-base being heterogeneous (Watson, 2003) with regard to the amounts of instances of classes representing previously taken courses, semesters and programs of study. Cases only contain a description this far as – resulting from the cases' heterogeneity - a distinction of description and solution attributes can only be made with regard to a new case. Figure 3 shows two exemplary cases highlighting the heterogeneous structure – case one represents a student who finished one semester, case two represents a student who finished two semesters already, resulting in multiple instances of the class `StudySemester`.

In contrast to the traditional CBR methodology new cases stem from the case-base itself. Employing rules and utilizing domain knowledge from an ontology all cases within the case base are checked regarding the represented student's progress within a study program. If a case represents a student who did not finish a study program yet it is treated as a new case. All new cases identified are iteratively solved against the remaining cases within the case-

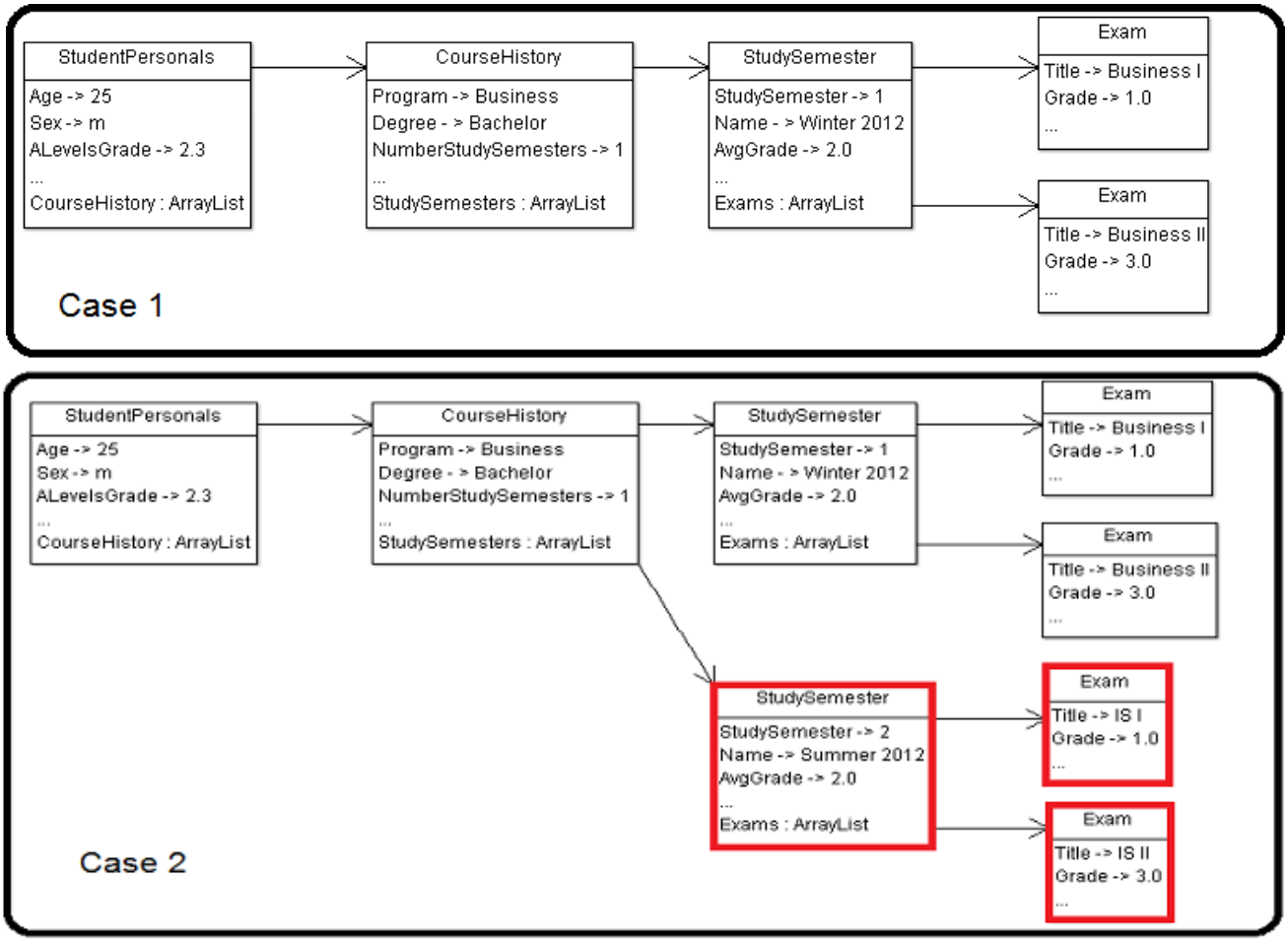


Fig. 3. Case Representation

base. This way solving of hundreds or even thousands of student cases can be achieved automatically.

The second phase **Retrieve** is enhanced in order to overcome retrieval problems which are mainly invoked by the varying structure of cases within the heterogeneous case-base including cases that contain a description only. For this purpose a so-called dynamic splitting point (dsp) is introduced to align the structure of all cases within the case-base with the structure of one specific new case. The dsp represents the amount of study semesters in the highest program being studied by a new case. When trying to retrieve similar cases all cases within the case-base are first reorganized according to the dsp, i.e. attributes are assigned dynamically to description and solution: All cases' `StudySemester` objects being associated with a `CourseHistory` object having the same values for their Program and Degree attributes as the new case are identified. Those having a value for their StudySemester attribute lower than the highest semester of the new case (<dsp) are assigned to belong to the case's description. All `StudySemester` objects having a higher value for their StudySemester attribute (>dsp) are assigned to belong to the case's solution. With regard to the exemplary cases in figure 3 case 1 might be a new case and case 2 might be part of the case base. During retrieval case 2 is aligned with case 1's structure: According to the dynamic splitting point – StudySemester 1 in Program Business with Degree Bachelor – the instances of classes with a light border belong to case 2's description and those with a bold, red border make up its solution. After the alignment of all cases within the case base the case with the highest object similarity (Wess, 1995) to the new case is retrieved by a k-NN retrieval algorithm.

The third phase **Reuse** employs a transformational reuse as according to (Aamodt & Plaza, 1994). The only transformation made to the solution of the retrieved case is to project semester numbers and descriptions according to the highest semester of the new case. With regard to the example in figure 3 the `StudySemester` object would be transformed so that its Name attribute is Summer 2013 instead of Summer 2012. The transformed solution makes up an initial forecast of future course selections of the student represented by a new case.

Cases available for retrieval from the case-base are heterogeneous with regard to the amount of represented study semesters. Thus a retrieved case might offer a solution, i.e. forecast, of one semester only – as it is the case in the example described above. The **Repeat** phase is an optional one that aims at extending the amount of future semesters for which course selections are forecasted. The solved case – consisting of the initial new case's description and the solution reused from the retrieved case – is transformed into a virtual case – consisting of an extended description only. Running the second and third phases of the cycle again with the virtual case being treated as a new case the amount of solution semesters can be significantly extended as now the dsp will be higher than in the first iteration of the cycle and thus only higher semesters can

be assigned to case solutions. The content of this phase could also be interpreted as being part of the reuse phase. It was designed as an additional phase to emphasize the iteration of previous phases based on a newly created virtual case that exceeds the steps typically carried out in the reuse phase.

First experiments with an initial version of the prototype demonstrated that it is unlikely that the solved case generated by the first four phases is in line with the specific examination regulations of the program the represented student is enrolled in. This is due to the fact that real experiences of only partially similar students are combined. Thus the fifth phase **Revise 1** employs transformation adaptation (Lopez De Mantaras et al., 2005) in order to alter solution objects. Therefore, a rule-based reasoning component is integrated using the solved case as facts and domain knowledge encoded both in an ontology and action rules (Herbst, Knolmayer, Myrach, & Schlesinger, 1994). The ontology represents knowledge on single courses and their feasible or mandatory use in different programs as well as course alternatives or prerequisites. Action rules are used to enforce the examination regulations, i.e. alter solutions derived by the first four phases with respect to ontology information. The result of phase five is a pretested case, i.e. a forecast of future course selections of the student represented by the new case that are approvable with regard to examination regulations.

Within the following sixth phase **Save** the solution is serialized to a forecasted course selections database that is independent from the case-base (see figure 2) making it available for further processing and analyses. As phases one through six are executed for all new cases identified within the case-base, this database will contain solutions, i.e. forecasted course selections, of all students that are likely to continue their studies for at least one upcoming semester.

The phases described work on the premise that similar students behave in a similar way – they select the same courses – in due consideration of their examination regulations. Students' individual choices are thus included only implicitly. In order to explicitly consider individual choices and also supply students with decision support regarding their planning processes an optional seventh phase, a web-based revise phase called **Revise 2**, is introduced. Based on the pretested solutions stored within the forecasted course selections database each single student is presented his or her potential future course selections as well as a list of alternative courses. The web-application enables students to alter the proposition or simply approve it. Additional information on the student's achievements is given by tooltips, e.g. indicating the student's major, missing obligatory courses or hints on the area where to best write a final thesis in. Altered solutions are automatically checked for alignment with examination regulations by rerunning phase 5 and stored within the database afterwards. This way not only students benefit from decision support but forecasts may also be significantly improved. The feedback given by individual students in this phase might be used to improve the knowledge of the system. E.g. the retrieval phase might be enhanced by obtaining students' feedback and using it for learning similarity measures as suggested by (Stahl, 2004). This might be subject to further research but is out of scope of this paper.

There is no retain phase as suggested by the classical CBR cycle. Instead a new case-base with fresh data of real students is initialized each time forecasts are to be made. Learning is thus achieved based on real-world data only. Cases already serialized within the Oracle database from a previous forecast, e.g. one semester ago, will be updated by the study achievements the represented students made up to the point of time of the forecast. For freshmen students, i.e. students not yet represented by a case in the serialized case-base, a new case will be constructed during initialization. Thus the case-base grows with each forecast to be made.

2.3 Aggregation of forecasted course-enrolments to support capacity-planning processes of various decision makers

An approach to enhance a university's stakeholders ability to make decisions in capacity planning processes is to combine the component of the refined CBR cycle with a data-driven decision system in terms of (Power, 2008). Data-driven decision support systems can be based on data warehouse systems and often include production reports, alerts and ad hoc data retrieval (Power, 2008). As a first step towards such a system data is extracted from the database holding forecasted course selections (see figures 1 and 2), transformed and loaded to the star schema shown in figure 5 once the phases of the case-base component have been run.

As discussed in 1.1 stakeholders of a university need information on different levels of aggregation. Based on the star schema design production reports can be created to fulfill all identified stakeholders' needs. In addition to standard reporting OLAP functionality is included and dashboards can be created for different decision makers. Top management's information needs are satisfied by reporting the key figure ExamCount by Dimension LecturerGroup or Lecturer respectively, filtered by chairs of single faculties. A report for deans might show ExamCount by Semester and Chair, deans of study affairs can be supplied with a report of ExamCount by Program and Exam. For eliminating overlapping of lectures reporting ExamCount by combinations of Exams per Semester is possible. Lecturers can be supported with detailed reports on the students they are likely to cope with in future, e.g. the ExamCount by Exam and Program filtered on Attempt_No>=2 (number of students repeating their course).

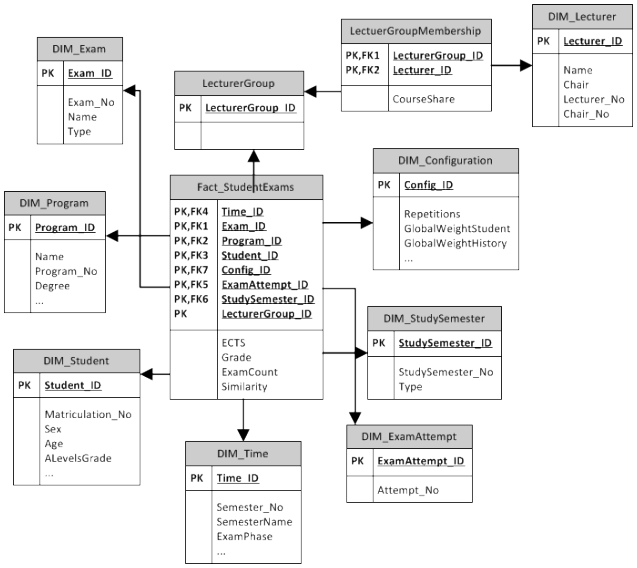


Fig. 4. Star Schema

3 Evaluation

This paper presents research conducted by a design science approach. Evaluation is a critical task in design science research and needs to demonstrate utility, quality and efficacy of the designed artifact (Hevner, March, Park, & Ram, 2004). As (Hevner et al., 2004) state artifacts can e.g. be evaluated in terms of accuracy, reliability, usability and fit with the organization. The artifact presented in this paper can be classified as innovative in a particular way as it is an alternative concept for forecasting course enrolments, utilizing a different methodology than previous approaches. As innovative artifacts need time to be accepted in the real world (Frank, 2006) evaluation results regarding fit with the organization and usability cannot be presented by the time of writing this paper.

Thus the paper focuses on an evaluation of correctness of the Retrieve phase as well as forecasting accuracy. Concerning accuracy of forecast results the impact of the phase Revise 2 is neglected as empirical evidence on the influences of confronting students with forecasted information on their actual future course selections is still missing. For evaluation purposes the prototypical implementation of the concept at a medium sized German university is utilized, the case-base comprises 1306 cases representing students and alumni of an undergraduate business program.

Correctness was assessed – as suggested by (Althoff, 1997) - by taking a copy of a case as new case and having the system solve it by retrieving and adapting a case from the case-base. Solving the new case is regarded as success if the system finds the case it was copied from (original case) as the best match. Altogether 50 cases were copied from the case-base and solved by running the phases Retrieve through Revise 1 of the refined CBR-cycle. The process terminated successfully in 100% of times thus it can be concluded that the retrieval task is performing in a correct way.

Forecasting accuracy was evaluated on the aggregation level of single courses by contrasting real enrollment figures for the winter semester 2012/2013 of three third-semester and eight fifth-semester courses of an undergraduate business program with forecasting results (forecast-horizon = 1 Semester) generated by the refined CBR-cycle and those generated by employing a simple linear regression. Input data for the regression is a five-year time-series of course-enrollment numbers (cases in the case-base cover the same period of time and the same program/students). Forecast accuracy of both methods is measured by three standard error measures, namely the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the Symmetric Mean Absolute Percentage Error (SMAPE). Further the difference between RMSE and MAE was calculated. The results are summarized in table 1.

Table 1: Forecast Errors of Linear Regression vs. CBR

	MAE	RMSE	RMSE-MAE	SMAPE
Linear Regression	35,018	42,3944	7,375	54,930
CBR	15,761	20,423	4,661	20,841

As the table shows the prototype employing the refined CBR-cycle performs better on all chosen error measures. The MAE indicates that on average forecasts provided by the refined CBR-cycle are more than twice as close to real values as those provided by the linear regression. To assess the magnitude of the errors the RMSE was calculated. It indicates that the magnitude of errors is about twice as high with the linear regression as with the refined CBR-cycle. For both, the linear regression and the refined CBR-cycle the difference between RMSE and MAE is rather low which indicates that the variance in the errors is rather low for both forecasting methods. The SMAPE is used as percentage error since it is applicable when observations contain near-zero values (Hyndman & Koehler, 2006) which is the case for some of the considered courses. Again the refined CBR-cycle scores better than the linear regression. To summarize evaluation results give supportive evidence that accurate forecasts can be derived by employing the concept presented in this paper.

4 Summary and Outlook

An innovative concept for forecasting course enrolments, serving as demand figures in academic capacity planning, has been presented. The application of a refined CBR approach ensures flexibility in terms of adaptation to different programs of study and provides the opportunity to include students' individual choices. Support for decision makers on various levels of a university is provided by embedding the refined CBR component with a data-driven decision support system. First evaluation results demonstrate correctness of case retrieval and accuracy of forecasts derived by a prototypical implementation of the concept. Further research will have to focus on the evaluation of usability and fit with the organization especially regarding effects of the confrontation of students with forecasted information on their progression within a study program. Moreover comparing the results derived be the refined CBR-cycle with those of a standard CBR-application appears to be a further interesting step in evaluation.

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, *7*(1), 39–59.

Abou Assali, A., Lenne, D., & Debray, B. (2009). Case Retrieval in Ontology-Based CBR Systems. In B. Mertsching, M. Hund, & Z. Aziz (Eds.), *KI 2009 Proceedings of the 32nd annual German conference on Advances in artificial Intelligence* (Vol. 5803, pp. 564–571). Springer Berlin Heidelberg.

Alt, R., & Auth, G. (2010). Campus-Management-System. *Wirtschaftsinformatik*, *52*(3), 185–188.

Althoff, K. D. (1997). Evaluating case-based reasoning systems: The Inreca case study. *Postdoctoral thesis (Habilitationsschrift), University of Kaiserslautern*.

Bergmann, R., & Stahl, A. (1998). Similarity measures for object-oriented case representations. In B. Smyth & P. Cunningham (Eds.), *Advances in Case-Based Reasoning* (Vol. 1488, pp. 25–36). Springer Berlin Heidelberg.

Bessent, E. W., & Bessent, A. M. (1980). Student Flow in a University Department: Results of a Markov Analysis. *Interfaces*, *10*(2), 52–59.

Burgess, R. R. (1970). *GERT Models of the University.* Virginia Polytechnical Institute.

Burgess, T. F. (1996). Planning the academic's workload: different approaches to allocating work to university academics. *Higher Education*, *32*(1), 63–75.

Cunningham, P. (1998). CBR: Strengths and weaknesses. In A. Pasqual del Pobil, J. Mira, & M. Ali (Eds.), *Tasks and Methods in Applied Artificial Intelligence* (Vol. 1416, pp. 517–524). Springer Berlin Heidelberg.

Deniz, D. Z., Uyguroglu, M., & Yavuzr, H. (2002). Departmental Workload Administration Using Group Forecasting in Universities. In *The 2012 ICEE Conference Proceedings on CD-ROM*.

Frank, U. (2006). *Towards a pluralistic conception of research methods in information systems research* (Research Report No. 7). Duisburg-Essen: Universität Duisburg-Essen. Retrieved from http://hdl.handle.net/10419/58156

Hansen, U., Henning-Thurau, T., & Langer, M. F. (2000). Qualitätsmanagement von Hochschulen: FACULTY-Q als Erweiterung von TEACH-Q. *Die Betriebswirtschaft*, *60*(1), 23–38.

Herbst, H., Knolmayer, G., Myrach, T., & Schlesinger, M. (1994). The specification of business rules: A comparison of selected methodologies. In *Methods and Associated Tools for the Information System Life Cycle* (pp. 29–46). Presented at the IFIP Working Group 8.1 Conference CRIS 94, Amsterdam: Elsevier.

Hevner, A. R. (2007). The three cycle view of design science research. *Scandinavian Journal of Information Systems*, *19*(2), 87–92.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Mis Quarterly*, *28*(1), 75–105.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Kassicieh, S. K., & Nowak, J. W. (1986). Decision support systems in academic planning: Important considerations and issues. *Information Processing & Management*, *22*(5), 395–403.

Küpper, H.-U. (2010). Steuerungsinstrumente für Hochschulen. In R. von Lüde (Ed.), *Neue Wege der Hochschulgovernance: Symposium zum Hochschulmanagement an der Fakultät Wirtschafts- und Sozialwissenschaften der Universität Hamburg* (pp. 61–90). Hamburg: Hamburg University Press.

Lopez De Mantaras, R., Mcsherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Watson, I. (2005). Retrieval, Reuse, Revision and Retention in Case-Based Reasoning. *The Knowledge Engineering Review*, *20*(03), 215–240.

Löwer, W. (2012). Ein Jahrzehnt Hochschulreform in Deutschland. *Zeitschrift für Hochschulrecht, Hochschulmanagement und Hochschulpolitik: zfhr*, *11*(1), 7–23.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, *15*(4), 251–266.

Nusselein, M. (2002). Empirische Erkenntnisse einer Informationsbedarfsanalyse an bayerischen Hochschulen. *Beiträge zur Hochschulforschung*, *24*(1), 100–114.

Oden, H. W., Langenwalter, G. A., & Lucier, R. A. (1993). *Handbook of material & capacity requirements planning*. New York: McGraw-Hill.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77.

Postert, S. (2001). *Gestaltungspotentiale eines MSS-gestützten Hochschulmanagements am Beispiel der Universität Osnabrück.* Diss. Universität Osnabrück, Osnabrück.

Power, D. J. (2008). Understanding Data-Driven Decision Support Systems. *Information Systems Management*, *25*(2), 149–154.

Recio García, J. (2008). *jCOLIBRI: A multi-level platform for building and generating CBR systems* (PhD Thesis (condensed version)). Universidad Complutense de Madrid.

Reichwald, R. (1998). Universitätsstrukturen und Führungsmechanismen für die Universität der Zukunft. In H.-U. Küpper & E. Sinz (Eds.), *Gestaltungskonzepte für Hochschulen: Effizienz, Effektivität, Evolution* (pp. 237–258). Stuttgart: Schäffer-Poeschel.

Rieger, B., Haarmann, T., Höckmann, E., & Lüttecke, S. (2009). Data Warehouse basierte Entscheidungsunterstützung für das Campus-Management von Hochschulen. In R. Hansen, D. Karagiannis, & H.-G. Fill (Eds.), *Wirtschaftsinformatik Proceedings 2009* (Vol. 2, pp. 493–502). Presented at the Business Services: Konzepte, Technologien, Anwendungen: 9. Internationale Tagung Wirtschaftsinformatik, Wien: Österreichische Computer Gesellschaft.

Schonberger, R., & Knod, E. M. (1991). *Operations management : improving customer service*. Homewood, IL: Irwin.

Slack, N. (2010). *Operations Management* (6th ed.). Harlow, England; New York: Pearson Education.

Stahl, A. (2004). *Learning of Knowledge-Intensive Similarity Measures in Case-Based Reasoning.* Dissertation.de.

Watson, I. (1997). *Applying Case-Based Reasoning: Techniques for Enterprise Systems* (1st ed.). Morgan Kaufmann.

Watson, I. (2003). *Applying knowledge management : techniques for building corporate memories*. Morgan Kaufmann San Francisco, Calif.

Watson, I., & Marir, F. (1994). Case-Based Reasoning: A Review. *The Knowledge Engineering Review*, *9*(04), 327–354.

Wess, S. (1995). *Fallbasiertes Problemlösen in wissensbasierten Systemen zur Entscheidungsunterstützung und Diagnostik : Grundlagen, Systeme und Anwendungen* (Diss.). Universität Kaiserslautern, Kaiserslautern.

Wilde, T., & Hess, T. (2006). Methodenspektrum der Wirtschaftsinformatik: Überblick und Portfoliobildung. *Arbeitspapiere des Instituts für Wirtschaftsinformatik und Neue Medien, LMU München*, *2*.

Explanation-Aware Maintenance of Distributed Case-Based Reasoning Systems Thesis Proposal

Pascal Reuss[1] **and Klaus-Dieter Althoff**[2]

[1]University of Hildesheim, Institute of Computer Science
[2]Competence Center Case-Based Reasoning, German Research Center for Artificial Intelligence, Kaiserslautern
reusspa@uni-hildesheim.de, klaus-dieter.althoff@dfki.de

Abstract

This paper is a thesis proposal and describes the idea of an integrated, explanation-aware maintenance approach for distributed case-based reasoning systems. We describe the related work and the foundations our own research will base on and the derived research goals for the dissertation. We describe in detail the integration of several existing approaches with our ideas to develop a methodology and maintenance strategy for distributed case-based reasoning systems.

1 Introduction

Developing and implementing knowledge-based systems, especially case-based reasoning (CBR) systems, has been explored for several years. Today we have methodologies to define and implement knowledge-based or especially case-based reasoning systems. For our research we will first focus on CBR systems and later trying to generalize the approach. For maintaining a single case-based reasoning system there are also several approaches that deal with maintaining the case base, the similarity or the adaptation knowledge. In general all the knowledge sources belonging to a knowledge-based system have to be considered, too. A knowledge source in this context is a software agent with access to a CBR system. These knowledge sources have dependencies between each other that have to be taken into account for a maintenace approach for distributed CBR systems. This thesis proposal describes the idea of an integrated, semi-automatic maintenance approach for distributed CBR systems with explanation capabilities on maintenance.

In Section 2 related work and foundations are described which are underlying our ideas. Section 3 presents the research goals based on these ideas. Section 4 further describes what we mean by explanation-aware maintenance. At last, Section 5 summarizes the paper and gives a short outlook on the next steps.

1.1 Example Domain and Application

In this subsection we describe the application *docQuery* which is settled in the travel medicine domain. The travel medicine domain and the *docQuery* application are used below to illustrate the approach presented in this paper. This application is an mulit-agent-system that uses several knowledge sources to find a solution to a user-given query. The query can contain information like travel destination, planned arrival, age, activities and chronical diseases. Based on this information, the necessary knowledge source, in this case software agents with CBR systems, are identified and requested. The *docQuery* application contains seven different CBR systems that represents several sub-domains of the travel medicine domain like regions, diseases, medication or activities [Bach, 2012][Reuss, 2012].

2 Related Work and Research Basics

This section contains related work that is used as foundations for the ideas in this paper. In his PhD thesis Carsten Tautz developed a methodology for experience management systems called DISER [Tautz, 2000]. This methodology does not consider maintenance in detail, but the methodology can be used as a basis for an integrated maintenance approach. Based on DISER, Markus Nick developed DILLEBIS [Nick, 2005]. This methodology focuses on maintenance that is based on user feedback to identify necessary maintenance actions. Both methodologies are not specifically designed for CBR sytems like INRECA [Bergmann *et al.*, 2003][Althoff and Weis, 1996] is. The latter is focusing on developing CBR applications, but does not explicitly consider maintenance tasks. A methodology for CBR systems originating from the INRECA context and focusing on the maintenance task is Roth-Berghofer's SIAM methodology [Roth-Berghofer, 2003]. This methodology extends the CBR cycle from Aamodt und Plaza [Aamodt and Plaza, 1994] with two steps called *Review* and *Restore*. These steps contain tasks for evaluating and maintaining a CBR system. However, the methodology does not consider distributed CBR systems.

The SEASALT architecture [Bach, 2012][Reichle *et al.*, 2011] provides a general framework for developing distributed knowledge-based systems. The architecture is domain-independent and modular and allows the creation of customized systems. The so-called Knowledge Line is responsible for managing the different knowledge sources. A software agent coordinates the queries and answers to and from the knowledge sources [Reichle-Schmehl, 2008][Bach *et al.*, 2008]. An approach for automating the selection of the needed knowledge sources using CBR was presented by Reuss [Reuss, 2012]. Our future research will base on the SEASALT architecture as a method for developing distributed knowledge-based systems. Additionally we will focus on the process of selection approach knowledge sources for maintenance purposes.

Althoff, Hanft and Schaaf developed the idea of a Case Factory to evaluate and maintain case bases [Althoff *et al.*, 2006]. The Case Factory is based on the Experience Factory approach from software engineering. It consists of several software agents for different tasks like evaluating

incoherence or modifying the case base. A central idea of the Case Factory is that an agent not only learns from his individual experience, but also learns from the experience of other agents. The idea of the Case Factory is integrated into the Knowledge Line of the SEASALT architecture. Each knowledge source, in this context CBR systems, has its own Case Factory that is responsible for maintaining the dedicated knowledge. Based on these ideas the Case Factory can be extended to maintain not the case base, but maintaining the vocabulary, the similarity measures and the adaptation knowledge, too.

The methodology to be developed within this thesis shall not only focus on CBR systems and maintenance tasks, but also considers the characteristics of distributed knowledge-based systems (e.g. the SEASALT architecture) and explanation capabilities [Roth-Berghofer *et al.*, 2005].

In addition to these methodologies, there are many different maintenance approaches for CBR systems, which should be taken into account. Ferrario and Smyth described an approach for collaborative maintenance of a case base. The feedback of several users is evaluated and an appropriate maintenance action derived [Ferrario and Smyth, 2000][Ferrario and Smyth, 2001]. Other authors like Ioannis Iglezakis [Iglezakis, 2001][Iglezakis and Roth-Berghofer, 2000], Racine and Yang [Racine and Yang, 1997][Racine and Yang, 1998][Yang and Zhu, 2001], Smyth, Keane and McKenna [Smyth, 1998][Smyth and Keane, 1995][Smyth and McKenna, 2001] or David Wilson[Wilson, 2001] described different approaches to maintain a case base. Armin Stahl describes the learning of feature weights [Stahl, 2001] and Patterson et al showed a strategy to maintain the similarity of a CBR system[Patterson *et al.*, 2000]. All these approaches are set up to maintain a single CBR system, but neither consider the use of multiple CBR systems nor the dependencies between these single systems. The maintenance approach to be developed within this thesis will consider such dependencies and will be able to combine single maintenance actions to an integrated maintenance strategy for distributed CBR systems.

3 Research Goals

Based on Section 2, four major research goals are formulated. All major research goals are split into several minor goals and tasks. The first goal is to develop a methodology, being able to define, implement and maintain distributed CBR systems based on the SEASALT architecture. This methodology contains all necessary tasks that lead to a functional multi-agent-system, as defined in the SEASALT architecture, extended with tasks covering the maintenance of CBR systems and the explanation awareness of the knowledge maintenance. The ideas underlying this methodology are described in more detail in Section 4.

The second research goal is to extend the concept of the Case Factory approach. The maintenance strategy has to be extended to cover the dependencies between several different homogenous or heterogeneous knowledge sources for cross-system maintenance and to improve the maintenance of a single CBR system. Of course for maintenance all four knowledge containers should be considered. Again a more detailed description can be found in Section 4.

The third research goal is to integrate maintenance explanation capabilities into a mulit-agent-system with distributed CBR systems. These explanations capabilities have to be considered when developing the methodology. The explanation process has to be integrated into the design and implementation of a system. For more detailed information on the explanation process see Section 4. The second and the third goal could be seen as minor goals of the first reseach goal, because of the dependencies between these goals. We decided to treat them as seperate goals, because these goals will take a large part of the research and can be divided in minor goals, too.

The fourth major research goal is to empirically evaluate the developed maintenance strategy, methodology and improved Case Factory approach. Therefore we will integrate the explanation-aware maintenance approach into the *docQuery* application and in an industrial multi-agent decisionsupport system. With the integration of the approach into *docQuery* we will show, that our methodology can be used on existing systems to extend those systems with explanation-aware maintenance capabilities. While building the new industrial multi-agent-system we will show, that our methodology can be used to build up a new multi-agent-system with explanation-aware maintenance capabilities. On both systems experts can evaluate the maintenance suggestions and related explanations.

4 Explanation-Aware Maintenance

This section describes the current status of our the ideas and goals of explanation-aware maintenance in more detail. With explanation-aware maintenance we mean an approach that enables a multi-agent-system to suggest maintenance actions to keep the knowledge in this system correct and consistent and provides explanations for the suggested maintenance actions. The suggestions can be presented to a knowledge engineer. With the related explanations the knowledge engineer should be able to understand why the multi-agent-system gives the specific suggestions. This way the knowledge engineer can make a quicker selection of the maintenance actions that should be executed and he will be able to identify potential problems in the knowledge of the multi-agent-sytem if an explanation is not comprehensible.

There are two main ideas behind this thesis proposal. The first idea is that maintenance of knowledge cannot be done per knowledge source only, but the dependencies between the knowledge sources have to be taken into account for an integrated maintenance strategy. An example for a dependency between CBR systems is the change of the vocabulary. If both systems have the same or partially the same vocabulary, a change in one system may cause a change in the other system for consistency reasons. Another example is removing one or more cases from a case base. Cases in other CBR systems could depend on one of the removed cases, so they may become inconsistent (to some degree). The system should suggest an appropriate maintenance action like removing the depending cases to keep the system's correctness/consistency.

For example, the *docQuery* application has a CBR system for regions and a CBR system for infectious diseases. Both systems have partly the same vocabulary for region names. For the region CBR system the names are part of the solution of the case structure and for the disease CBR system the names are part of the problem description of the case structure. These CBR systems have a dependency between each other. When a certain region is retrieved from the region CBR system the related infectious diseases can be retrieved from the disease CBR system using the name of the region. If the name of one region changes in the vocabulary of the region CBR system, an inconsitency be-

tween this CBR systems will occur, because the retrieved region has no match in the disease CBR system. The related infectious diseases cannot be retreived anymore. To keep the CBR systems consistent a change of the vocabulary of the disease CBR system should be suggested. Another example is the CBR system for medication. If one medicament must not be applied for a disease any longer, the system has to check if the medicament can still be applied for another disease. If the medication is not longer used the case representing the specific medicament can be deleted, ohterwise the case has to be adapted to keep the system's correctness.

The Case Factory approach and the SEASALT architecture support distributed knowledge sources in a multi-agent-system. A Case Factory can support the maintenance of the case base of a single CBR system. The original approach has to be extended to support the maintenance of the other three knowledge containers, namely vocabulary, similarity and adaptation knowledge. The original approach contains several software agents to monitor the case-base and one agent to do the necessary maintenance actions. To support all knowledge containers some more agents are needed to monitor these containers and the maintenance tasks should be split into several agents. An own maintenance agent per knowledge container is needed to support parallel maintenance of the knowledge containers. Additionally a supervising agent is required to coordinate the maintenance actions. This coordination agent is also responsible for the communication between the multiple Case Factories. Figure 1 gives an overview of an extended Case Factory.

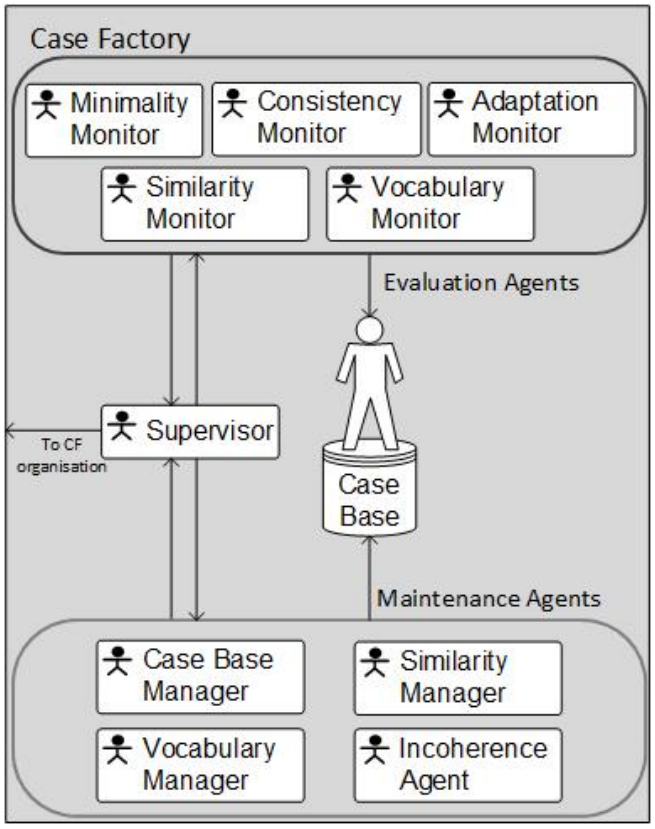


Figure 1: Extended Case Factory

Another idea is the combination of the SIAM approach and the Case Factory approach. SIAM extends the so-called *4R* cycle with two steps to a *6R* cycle. The *4R* cycle consists of four steps, called *Retrieve*, *Reuse*, *Revise* and *Retain*. In the *Retain* step a given problem is mapped to the case structure of the CBR system and the most similar case(s) is retrieved from the casebase. The *Reuse* step adapts the solution of the retrieved cases to the given problem. This adapted solutions are checked by an expert or a user of the CBR system in the *Revise* step. The result of the check can be stored in the specific case. In the last step *Retain* the retrieved, adapted and revised case can be added to the casebase. This way the CBR system can learn [Aamodt and Plaza, 1994].

The steps *Review* and *Restore* are not part of the original *4R* cycle. These steps support the monitoring and maintenance of single CBR systems [Roth-Berghofer, 2003]. Each step consists of three tasks. The steps could be mapped to a Case Factory with software agents for the defined tasks. The *Review* task contains the steps *Assess*, *Monitor*, and *Notify*. Each step could be assigned to an agent in the Case Factory. An agent responsible for the *Assess* task evaluates the knowledge of a CBR system. Another agent compares the evaluation result with the available constraints or thresholds (*Monitor*). The constraints and thresholds are defined by the knowledge engineer and stored in a so-called Maintenance Map. This Map will be described in more detail later in this section. The notifying agent decides if and whom to inform that a maintenance action is necessary. This agent sends a message to an agent at a high-level Case Factory organization. This organization is also described in more detail later. On the higher level the notifications are collected and used to create a maintenance plan.

For an example we take the CBR systems for disease and medication. An assess agent evaluates the casebase of the medication CBR systems and finds a case for a medicament that is not dedicated to a disease. These result is passed to the monitoring agent. The Maintenance Map contains the constraint that every medicament has to be dedicated to a disease. The information about the constraint violation is passed to the notify agent, which sends this information and a request for a maintenace action to a coordination agent ouside the Case Factory of the medication CBR system.

Three additional agents are responsible for the tasks of the *Restore* step. The single tasks to perform are *Suggest*, *Select* and *Modify*. The suggestions for maintenance actions are made in the high-level Case Factory organization, because the suggestions depend on the maintenance plan. The suggested actions are sent to an agent in the relevant Case Factory and this agent will select the appropriate maintenance action. It is possible, that more than one maintenance action is selected. The agent responsible for modifying the knowledge gets a message with the selected maintenance actions and executes the modification on the knowledge containers or notifys the knowledge engineer if the action needs additional input. When the modification is done or an error occurs, the agent sends a notification to the high-level Case Factory organization.

A high-level Case Factory organization is needed to control the integrated maintenance between these Case Factories. Therefore several software agents have to supervise the communication and the adherence of high-level maintenance goals. Additionally, an agent collects the suggested maintenance actions from multiple Case Factories, while another agent combines the individual maintenance actions to a maintenance plan. A Case Factory can suggest more

than one maintenance action. The planning agent is also responsible for checking constraints or solving conflicts between the individual maintenance actions. A suggested action from one Case Factory can trigger a necessary maintenance action of another Case Factory, based on the dependencies between the knowledge sources. So these actions have to be integrated into the maintenance plan, too. The relevant maintenance actions are passed back by the collector agent to the relevant Case Factories.

Based on the previous example the information about the constraint violation is passed to a coordination agent in the Case Factory organization. This coordination agent checks the possibilities to repair the constraint violation. Based on the knowledge in the Maintenance Map two suggestions to repair the problems are found. The first suggestion is to delete the case to keep only cases in the casebase that could be retrieved. The second suggestion is to adapt the case and add the dedicated diseases to the problem description. Both suggestions are send back to the Case Factory of the medication CBR system. If the selection agent can decide on his own the appropriate maintenance action based on the information in the Maintenance Map, then the agent selects an action. In our example we will assume that the Maintenance Map contains the information that a case should be adapted if possible before deleting it. The selection agent will notify the knowledge engineer that additional information is needed to dedicate the medicament to a disease. The knowledge engineer can select the disease that should be dedicated to the medicament. The modify agent takes the information an adapts the problem description of the case. At last the agent sends a success or error message to the Case Factory organization. Figure 2 shows a Case Factory organization with example agents.

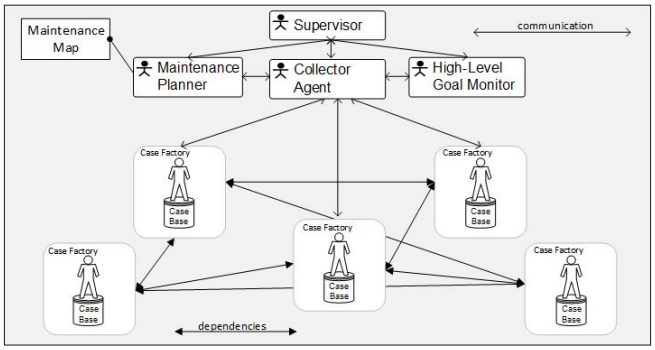


Figure 2: Case Factory organization

To coordinate the maintenance actions for the single CBR systems a Maintenance Map is introduced. This Maintenance Map is based on the Knowledge Map from Davenport and Prusak [Davenport and Prusak, 2000], that was adapted to multi-agent-systems from Bach et al. [Bach *et al.*, 2008]. In contrast to a Knowledge Map the Maintenance Map is a bidirectional graph. The vertices represent the knowledge sources in a distributed knowledge-based system and the edges represent the dependencies between the single sources. The weights of the edges could be used to describe the importance of the dependency. Additionally, the maintenance goal for a single knowledge source could be stored as well as the integrated maintenance goal for the overall system. For every maintenance goal the metrics for the empirical evaluation and the thresholds and constraints could be defined. Another idea is to store the preferred maintenance action for every knowledge source in the Maintenance Map. By this it will be possible for the relevant agent to decide quickly which maintenance action to choose. The Maintenance Map could be in XML or RDF format to share the knowledge between different systems in an easy way. The Maintenance Map should be defined and updated by the knowledge engineer.

To define and implement a distributed CBR system with explanation-aware maintenance capability a methodology is needed that contains tasks to acquire the necessary knowledge, maintenance goals, and actions to realize an integrated maintenance strategy. The SEASALT architecture implies tasks to develop a multi-agent-system with distributed knowledge-based systems (e.g. CBR systems). These tasks could be derived and organized in a methodology. For example the use of software agents implies the definition of an agent model which describes the roles and responsibilities of the agents. The resulting methodology has to be extended with tasks for maintenance and explanations.

Methodologies like DISER, DILLEBIS and INRECA contain tasks that may be integrated in the new methodology. The methodology to be developed shall be applicable to multi-agent-systems with different single CBR systems as knowledge sources. While the focus is set on using CBR systems as knowledge sources, the methodology shall contain several template tasks, that could be replaced by tasks for different knowledge-based systems. These templates have to be included in a way that substituting a template with a concrete instantiation will not affect the other tasks. Our research will focus on a methodology that can be used to define and implement a multi-agent-system with distributed CBR systems. From this methodology it could be possible to identify task that could generalized with templates to support the definition of other knowledge-based systems. An example task affecting maintenance is shown in Figure 3.

Task: define maintenance goal for Case Factory organisation

Maintenance Goal: keep overall solution correct/complete
Needed Knowledge: expert feedback, consistency rules
Metrics:-
Trigger for Evaluation: getting feedback to solution
Input: overall solution
Output: evaluation results, maintenance suggestion and explanation

Figure 3: Example task to define an overall maintenance goal

The second idea is that a CBR system should be able to explain why a maintenance action is suggested. This explanation will support the knowledge engineer's decision which maintenance action should be executed. To give a CBR system explanation capabilities a lot of knowledge is necessary. The introduced idea of an integrated maintenance strategy focuses on explanations for maintenance. The underlying research assumption here is that the minimal knowledge necessary for the explanation of the maintenance actions is the same knowledge that is necessary for the CBR system to suggest a maintenance action. It fol-

lows, that the minimal knowledge for explanations already exists in the system, if the system is able to (reasonably) suggest maintenance actions. One challenge is to identify and extract the needed part of the knowledge to formulate the explanations for a given maintenance action. For a single explanation of a maintenance action not the whole knowledge of the CBR system is needed. Another challenge is to identify and combine the knowledge of several CBR systems to explain maintenance suggestions that are necessary to keep the overall sytem correct and consistent.

Knowledge that can be used for explanations are logging information, rules, evaluation results, metrics, thresholds, etc. Additionally, knowledge can be extracted from social media like expert forums or external data sources. A scenario for extracting knowledge from sources outside the multi-agent-system can be the prohibition of a medicamant. This information can be extracted from the webside of the European Medicines Agency. Monitoring this website the multi-agent-sytem can react on new information and suggest a maintenance action for deleting or adapting a case in the medication CBR system. The same information that triggers the maintenance suggestion can be used to explain the suggestion. Such knowledge extracting has already been carried out in the scope of the SEASALT architecture [Bach, 2012] [Bach *et al.*, 2010] and can be plugged into our methodology.

5 Conclusion and Outlook

This paper describes the ideas of an integrated explanation-aware maintenance approach for distributed (knowledge-based) systems. This includes the development of an integrated maintenance strategy and the definition of a methodology with tasks that consider maintenance of distributed systems and explanations. The successful implementation of distributed knowledge-based systems in research and industrial environments to evaluate the maintenance approach and the methodology is a part of the idea, too.

The next steps are to define the goals of an integrated maintenance strategy and the extended Case Factory concept. Based on these definitions the implied tasks from the SEASALT architecture (e.g. define agent model, define knowledge acquisition, etc.) can be derived. Some of these implied tasks already exist in DISER, DILLEBIS or INRECA, other tasks like defining an agent model have to be formalized. With these task a first version of the target methodology can be set up. This methodology will be used to define and implement an explanation-aware maintenance extension for the *docQuery* application. The experience collected during this task will be used to refine and extend the methodology. The second version of the methodolgy will be used to define the mentioned industrial multi-agent-system. Both systems will be continuoisly evluated with the help of domain experts and knowledge engineers. The results will be used to improve the methodology, the maintenace strategy and the implemented multi-agent-systems.

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enrice Plaza. Case-based reasoning: Foundational issues, methodological variations and system approaches. *AI Communications*, 7(1):39–59, 1994.

[Althoff and Weis, 1996] Klaus-Dieter Althoff and Karl-Heinz Weis. An evaluation of the inreca cbr system. *Proceedings of the 9th German Workshop on Machine Learning*, 1:6–11, 1996.

[Althoff *et al.*, 2006] Klaus-Dieter Althoff, Alexandre Hanft, and Martin Schaaf. Case factory: Maintaining experience to learn. In *Proceedings of the 8th European conference on Advances in Case-Based Reasoning*, pages 429–442, 2006.

[Bach *et al.*, 2008] Kerstin Bach, Meike Reichle, Alexander Reichle-Schmehl, and Klaus-Dieter Althoff. Implementing a coordination agent for modularised case bases. In *Proceedings of the 13th UK Workschop on Case-Based Reasoning*, pages 1–12, 2008.

[Bach *et al.*, 2010] Kerstin Bach, Christian Sauer, and Klaus-Dieter Althoff. Deriving case base vocabulary from web community data. In *ICCBR-2010 Workshop Proceedings: Workshop on Reasoning From Experience On The Web*, pages 111–120, 2010.

[Bach, 2012] Kerstin Bach. *Knowledge Acquisition for Case-Based Reasoning Systems.* PhD thesis, University of Hildesheim, 2012.

[Bergmann *et al.*, 2003] Ralph Bergmann, Klaus-Dieter Althoff, Sean Breen, Mehmet Gker, Michel Manago, Ralf Traphner, and Stefan Wess. *Developing Industrial Case-Based Reasoning Applications: The INRECA Methodology*. Springer Verlag Berlin, 2003.

[Davenport and Prusak, 2000] Thomas H. Davenport and Laurence Prusak. *Working Knowledge: How Organizations Manage What they Know*. Havard Business School Press, 2000.

[Ferrario and Smyth, 2000] Maria Angela Ferrario and Barry Smyth. A user-driven distributed maintenance strategy for large-scale case-based reasoning systems. In *ECAI Workshop Notes*, pages 55–63, 2000.

[Ferrario and Smyth, 2001] Maria Angela Ferrario and Barry Smyth. Distributing case-based maintenance: The collaborative maintenance approach. *Computational Intelligence*, 17(2):315–330, 2001.

[Iglezakis and Roth-Berghofer, 2000] Ioannis Iglezakis and Thomas Roth-Berghofer. A survey regarding the central role of the case base for maintenance in case-based reasoning. In *ECAI Workshop Notes*, pages 22–28, 2000.

[Iglezakis, 2001] Ioannis Iglezakis. The conflict graph for maintaining case-based reasoning systems. In *Case-Based Reasoning Research and Development: Proceedings of the Fourth International Conference on Case-Based Reasoning*, pages 263–275, 2001.

[Nick, 2005] Markus Nick. *Experience Maintenance through Closed-Loop Feedback*. PhD thesis, University of Kaiserslautern, Computer Science Department, AG Software Engineering, 2005.

[Patterson *et al.*, 2000] David Patterson, Sarabjot Anand, and John Hughes. A knowledge light approach to similarity maintenance for improving case-base competence. In *ECAI Workshop Notes*, pages 65–78, 2000.

[Racine and Yang, 1997] Kristi Racine and Qiang Yang. Maintaining unstructured case bases. In *Case-Based Reasoning Research and Development*, pages 553–564, 1997.

[Racine and Yang, 1998] Kristi Racine and Qiang Yang. Redundancy and inconsistency detection in large and

semi-structured case bases. In *IEEE Transactions on Knowledge and Data Engineering*, 1998.

[Reichle *et al.*, 2011] Meike Reichle, Kerstin Bach, and Klaus-Dieter Althoff. Knowledge engineering within the application independent architecture seasalt. *Joachim Baumeister, Grzegorz J. Nalepa (Hrsg.). International Journal of Knowledge Engineering and Data Mining (IJKEDM)*, 1:202–215, 2011.

[Reichle-Schmehl, 2008] Alexander Reichle-Schmehl. Design and implementation of a software agent to coordinate the dynamic retrieval on distributed, heterogeneous case bases (in german). Bachelorarbeit, University of Hildesheim, 2008.

[Reuss, 2012] Pascal Reuss. Concept and implementation of a knowledge line - retrieval strategies for modularized, homogeneous topic agents within a multi-agent-system (in german). Master's thesis, University of Hildesheim, 2012.

[Roth-Berghofer *et al.*, 2005] Thomas Roth-Berghofer, S. Schulz, and A. Woody (Eds). Explanation-aware computing (exact 2005), technichal report fs-05-04, washington d.c., usa. Technical report, AAAI Press, 2005.

[Roth-Berghofer, 2003] Thomas Roth-Berghofer. *Knowledge maintenance of case-based reasoning systems. The SIAM methodology.* Akademische Verlagsgesellschaft Aka GmbH, 2003.

[Smyth and Keane, 1995] Barry Smyth and Mark Keane. Remembering to forget: A competence-preserving case deletion policy for case-based reasoning systems. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 377–382, 1995.

[Smyth and McKenna, 2001] Barry Smyth and Elizabeth McKenna. Competence models and the maintenance problem. *Computational Intelligence*, 17(2):235–249, 2001.

[Smyth, 1998] Barry Smyth. Case-based maintenance. In *Proceedings of the 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 1998.

[Stahl, 2001] Armin Stahl. Learning feature weights from case order feedback. In *Case-Based Reasoning Research and Development: Proceedings of the Fourth International Conference on Case-Based Reasoning*, 2001.

[Tautz, 2000] Carsten Tautz. *Customizing Software Engineering Experience Management Systems to Organizational Needs*. PhD thesis, University of Kaiserslautern, 2000.

[Wilson, 2001] David Wilson. *Case-Based Maintenance: The Husbandry of Experience*. PhD thesis, Faculty of the University Graduate School, Department of Computer Science, University of Indiana, 2001.

[Yang and Zhu, 2001] Qiang Yang and Jun Zhu. A case-addition policy for case-base maintenance. *Computational Intelligence*, 17(2):250–262, 2001.

An ontology-based Recommender System to Support Nursing Education and Training

Marjan Khobreh, Fazel Ansari, Mareike Dornhöfer and Madjid Fathi
Institute of Knowledge Based Systems, University of Siegen
Hölderlinstrasse 3 D-57068 Siegen, Germany
{marjan.khobreh@, fazel.ansari@, m.dornhoefer@, fathi@informatik.}uni-siegen.de

Abstract

The need of healthcare organization on highly knowledgeable and qualified human recourses to guarantee quality of performance is indispensable. A desired performance level is tailored with obtaining competences and job knowledge, as the major influential factors. This is especially critical due to high rate of changes in knowledge domains and technological infrastructure over time i.e. before or within employment of job holders and applicants. Therefore, applicants as well as employees and practitioners are also dealing with upgrading their level of job knowledge and qualifications.

Adaptive Medical Profession Assessor (Acronym: Med-Assess) as a European funded project, proposes a knowledge based system for assessment of the competences and job knowledge of the applicant/employee to perform a certain job role in the domain of healthcare i.e. nursing and care giving to neuro-patients. In this context, recommendation of learning materials is an integral part. It subjects to the required training due to the lack of competence(s) for performing a specific nursing task(s). This paper presents the system architecture of Med-Assess, and discusses how the applied semantics i.e. ontologies and rules are developed. It especially presents the background in nursing education and training, and conceptually presents the design of the recommender components.

1 Introduction

With no doubt "humans are generally considered the most valuable resource in any organizations" [Lianga *et al.*, 2013; Hesketh and Fleetwood, 2006]. Therefore one of the vital requirements of organizations is to recruit competent applicants for announced job vacancies. Organizations examine and interview the applicants to find the most qualified one for a job vacancy. However, in most job fields, knowledge requirements are associated within the today's market with a moderately high rate of change over time [Pilz, 2012]. The causes are, for example, development of new technologies like smart phones or cloud computing, development of new know-how, or integration of organizational processes. Such changes raise the needs of individuals and organizations to regularly improve the level of job related knowledge, and to obtain competences continually before and within employment.

The importance of job specific knowledge is drastically increasing especially in the field of personnel selection [Mol *et al*., 2013]. In this way "general mental ability (or intelligence) is the single best predictor of job performance, regardless of job type" [Mol *et al*., 2013], [Schmidt, 2009], [Schmidt and Hunter, 2004]. Despite empirical evidences such as [European Commission, 2012a] and [Salgado *et al*., 2003], the associated rationales are not consistently figured out, and thus, considered as a progressive research topic in the fields of psychology [Mol *et al*., 2013]. The anticipation is in turn that "job knowledge therewith appears to be a more proximal predictor of job performance than general mental ability" [Mol *et al*., 2013]. In this paper, we consider the term of job knowledge to refer to domain specific knowledge required for obtaining (hard and soft) competences towards performing nursing jobs with desired qualities defined by the healthcare organizations. In this context, the stakeholder's points of view are classified in three categories.

The **first** is an organizational perspective. Organizations classify individuals in the two groups:
(1) employees of the organization as human resource and,
(2) applicants or potential employees, who apply for an announced job vacancy.
Obviously, the main concern of organizations is to recruit the qualified applicants, and simultaneously improve the professional levels of their employees to sustain and guarantee quality of performance, influence on the effectiveness of processes and deliver excellent outcomes.

The **second** is the applicant perspective. An applicant needs to possess certain knowledge to improve the level of his/her competences, based on the market requirements. The underlying reason is a hard competition on the job market towards being hired at a desired job with fair income and good reputation. In turn, employees concern with keeping their positions or getting promoted to higher levels of the organization.

The **third** perspective is in regard of educational institutes. They should provide teaching and training schemes (e.g. curriculum, courses, e-learning) for students, job applicants and practitioners. The training should also address the market needs subject to (further) vocational education and training [Pilz, 2012].

Blancero *et al*. [Blancero *et al*., 1996] and Parry [Parry, 1998] defined competences as knowledge, skills and attitude to perform a job. In the European Qualifications

Framework (EQF) for lifelong learning, the term *Competence* is defined as "the proven ability to use knowledge, skills and personal, social/methodological abilities, in work or study situations and in professional and personal development" [European Commission, 2008]. However, as job descriptions are changed over time, the competences cannot remain solid [Glosson and Schrock, 1985], [Pilz, 2012]. The realization of competences may differ based on organizations' requirements. Moreover, learning expresses the rate of transmitting and absorbing knowledge, which depends on cognitive abilities [Wu & Lee, 2007]. Weinstein and Underwood [Weinstein and Underwood, 1985] considered four main learning strategies, namely (1) information processing strategies, (2) affective learning strategies, (3) typical reading strategies, and (4) metacognitive strategies. In this sense, "typical reading" is a learning concept that addresses the development of the learning process through reading, learning materials, doing exercises and preparing for attending associated tests. There are meaningful relationships between learning materials, competences and job roles in each domain. The concept of Adaptive Medical Profession Assessor (Med-Assess) is developed for bridging these three aspects in nursing education [Mol *et al*., 2013]. One of the Med-Assess objectives are identifying the required competences to perform nursing tasks [Mol *et al*., 2013]. In addition, Med-Assess is used to assess and evaluate the competences of applicants (i.e. novice or experienced nurses), based on multiple-choice tests on domain knowledge and general mental ability [Mol *et al*., 2013]. Finally the system recommends respected learning materials e.g. courses, workshops, e-learning courses, textbooks. In this framework, a Recommender System (RS) is being developed, as a sub-system of Med-Assess, to suggest learning material(s), in case the test candidate lacks competence(s) in (a) certain task(s) [Mol *et al*., 2013]. Further details about the general concept of Med-Assess are discussed in [Mol *et al*., 2013].

Considering the given introduction, this paper consists of 6 sections. Section 2 discusses the background of Med-Assess. The Med-Assess system architecture and its components are presented in section 3. Section 4 and 5, respectively, describe the concept and related methodologies of RS, and related aspects of ontology engineering and the execution of rules. Finally, section 6 concludes the paper and indicates the future research steps.

2 Background

As mentioned one of the primary objectives of Med-Assess is to recommend learning material(s), if applications refer to lack of competence(s) to perform certain task(s). In the following sub-sections, firstly, the nursing job is discussed to clarify which nursing competences enable nurses to perform their tasks in a high quality. Secondly, the nursing education and training is described to elucidate what learning materials are available in different categories, based on the job classifications. Both sub-sections focus on market requirements in Germany. However, the geographical transfer of the project findings and product is not limited to Germany; thereby the entire European health sector is considered.

2.1 Nursing job role

"Deutscher Pflegerat (DPR)" (German care council) and Krankenpflegegesetz (KrPflG) (nursing act), as a special administrative law in the scope of the Federal Republic of Germany, noted the personal responsibility tasks of professional nurses as follows [KrPflG, 2003][1], [DPR, 2004][2]:

- Determine the need for care, planning, organization, conduction and documentation of care.
- Evaluation of care, protection and development of care quality.
- Advice, guidance and support of caregivers.
- Initiation of life-sustaining emergency measures until a doctor arrives.

In addition, the tasks which should be performed as assistance are [DPR, 2004], [KrPflG, 2003]:

- Independent implementation of medical interventions that were prescribed by a doctor,
- Provision of medical diagnosis, treatment or rehabilitation,
- Action in crisis and disaster situations.

Nurses should establish multidisciplinary solutions to health problems and work together with other professionals in hospitals [DPR, 2004].

Williams *et al*. [William *et al*., 2009] identified nursing tasks in four different categories based upon the daily workload.
(1) **Direct care:** defined as "all activities involving direct interaction between the nurse and patient/family" [William *et al*., 2009], namely, communication, medication, nutrition and fluid intake, elimination, personal care, positioning or turning, escorting patients, assisting other health professionals, routine checks, vital signs, collecting specimens, nursing procedures [William *et al*., 2009].
(2) **Indirect care:** defined as all activities related to the patients but execute away from the patient, namely, charting/form completion, reports, communication, meeting preparation [William *et al*., 2009].
(3) **Unit-related:** defined as all activities for handling the unit/ward namely, housekeeping, clerical errands, communication, and maintenance [William *et al*., 2009].
(4) **Personal time:** defined as all activities that lead to nurses' well-being, namely, education/training and meal break [William *et al*., 2009].

Furthermore Nursing Interventions Classification (NIC) provided a taxonomy to represent nursing constructs [Bulechek *et al*., 2013]. The NIC consists of 7 domains and 554 interventions. The 7 domains are: *Basic, Complex, Behavioral, Safety, Family, Health System, and Community* [Bulechek *et al*., 2013]. In the framework of Med-Assess, NIC is considered as one of the validation sources. Other methods are interviewing with nursing supervisors, educators and physicians in cooperation with the clinical partner of the project.

[1] Translated and adopted by the authors **according to Section 2 " Vocational Training",§ 3" training target", paragraph (1)**
[2] Translated and adopted by the authors **according to §2 "Tasks" paragraph (3)**

2.2 Nursing competence through education and training

In the literature on healthcare, competence is often used to only describe knowledge that enables practitioners to perform a particular task [Schroeter, 2008]. However, competence is more than knowledge [Norman, 1985]. It consists of understanding of various knowledge merging skills to have capability, and abilities to perform the clinical, technical and communication tasks, and also to solve problems successfully [Schroeter, 2008].

Obtaining the required competences can be integrated into curricular coursework [Rudolph, 1999]. Practitioners learn and practice certain cognitive results such as concepts, significations, principles, strategies, problem solving and having reversibility, (re) construction and improvement characteristics [Neacşu, 2011].

According to article 31 (6) of the European parliament and of the council on the recognition of professional qualifications: "Training for nurses responsible for general care shall provide an assurance that the person in question has acquired the following knowledge and skills" [European Parliament and Council, 2005]:

(a) Adequate knowledge of the sciences of structure, physiological functions and behavior of healthy and sick persons.

(b) Sufficient knowledge of the nature and ethics of the profession.

(c) Adequate clinical experience.

(d) "The ability to participate in the practical training of health personnel and experience of working with such personnel".

(e) Experience of team working with other professions in the hospital.

As listed in this treaty[3], the training program shall consist of theoretical instruction as well as clinical instruction [European Parliament and Council, 2005]. Theoretical instruction includes [European Parliament and Council, 2005]: (a) **Nursing**: nature and ethics of the profession, general principles of health and nursing, general and specialized medicine, general and specialized surgery, child care and pediatrics, maternity care, mental health and psychiatry, care of the old and geriatrics [European Parliament and Council, 2005]. (b) **Basic science**: anatomy and physiology, pathology bacteriology, virology and parasitology, biophysics, biochemistry and radiology, dietetics, hygiene. (c) **Social science**: sociology, psychology, principles of administration, principles of teaching, social and health legislation, legal aspects of nursing.

Furthermore, the "Ausbildungs- und Prüfungsverordnung für die Berufe in der Krankenpflege (KrPflAPrV)" as a vocational training and examination regulation of occupations in nursing in Germany, expresses two training parts[4] [KrPflAPrV, 2003]: (a) **Practical**: internal medicine, geriatrics, surgery, gynecology, neurology, birthing, newborn Care, psychiatry, pediatrics and ambulant care. (b) **Theoretical**: nursing and health science, natural science and medicine, human and social sciences law, politics and business.

KrPflAPrV [KrPflAPrV, 2003] mentioned that to assess the competences, skills and knowledge of the practitioners, they should take the national examination which is in written, oral and practical form[5]. The results of written and oral exams are graded[6] in 6 degrees: (1) **Very good**: (a nurse) who is particular competent (2) **Good**: who is fully competent (3) **Satisfactory**: who is generally competent (4) **Sufficient**: who has deficiencies in her/his competence, (5) **Poor**: who is not competent, however, it is possible to recognize that the necessary knowledge exists and the deficiencies can be solved in the foreseeable future (6) **Unsatisfactory**: who has not even the basics of the competences, the deficiencies are more than could be resolved in the foreseeable future.

In addition, many states of Germany defined their individual curriculum frameworks for training and educating nursing students in nursing schools. In general, while the curriculum frameworks are the same; they have some different point of views. Examples are [Oelke *et al.*, 1998] and [Müller-Klepper (Ed.), 2005].

3 Med-Assess System Architecture

The main functionality of Med-Assess is to provide adaptive tests for clinics, hospitals and nursing schools to assess the competences of their applicants, employees and practitioners. The aim is to provide recommendations of learning materials for further training and education. In this context, Med-Assess incorporates the core elements of the Onto-HR solution [OntoHR, 2009]. Onto-HR is a former European Project, where two of the Med-Assess project partners took part. It focuses on the assessment of IT-specialists. Med-Assess project is categorized as a transfer of innovation project [European Commision, 2012b], which indicates that a given innovative solution (i.e. Onto-HR) is customized and transferred to another domain or country while enhancing its functionalities and performance. Therefore, the transfer of Med-Assess is twofold, first from the IT to the medical sector, and second from the Netherlands and Hungary to Germany. Eventhough Med-Assess is developed based on Onto-HR, it differs in certain characteristics. The major improvements are the adaptivity of the solution, the focus on recommending based on the assessment results, and offering a decision component in the field of personnel selection. Compared to other existing technologies like KnowWE (Knowledge Wiki Environment) that "enables domain specialists and experienced users to build knowledge–based consultation systems collaboratively on the web" [Baumeister et al., 2007],or Experience Factories that aimed at "capitalization and reuse of life cycle experience and products" [Basili,2009], Med-Assess system focuses on the assessment of job knowledge and compensating the lacks through recommendation of further education and training.

The system architecture of Med-Assess is illustrated in Figure 1. It consists of the three modules *Adaptive Testing*, *Recommender* and *User Interface.*

The *User Interface* is developed in order to supply the user with a graphical interface for performing the tests and gathering user information. It ultimately transfers the test results, recommended learning materials and applicant ranking to the user.

[3] **ANNEX V, Point 5.2.1**

[4] Translated and adopted by the authors **according to § 1 (1)**

[5] Translated and adopted by the authors **according § 3 "State Examination"**

[6] Translated and adopted by the authors **according § 7**

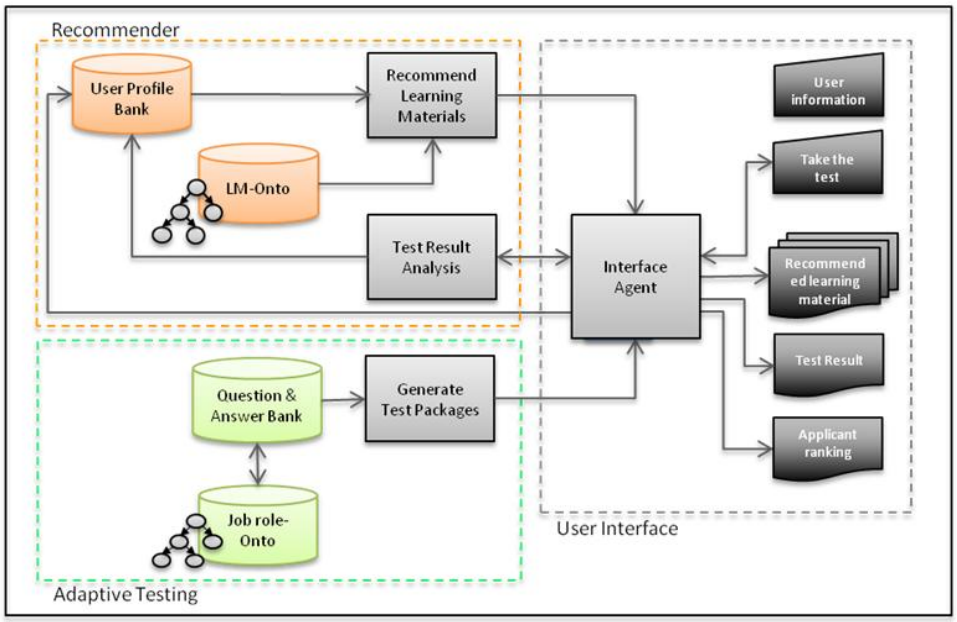


Figure 1 Med-Assess System Architecture

This component is connected to both, *Adaptive Testing* and *Recommender*, to receive the test package from *Adaptive Testing* and recommended learning materials as well as test results from *Recommender*. The *User Interface* also delivers the test package, which contains the questions as well as user information for *Recommender*.

The *Adaptive Testing* consists of a Job-role ontology (Job role-Onto), *Test Bank* and *Test Generator*. The Job role-Onto formalizes and represents all the nursing tasks and activities. Ontology is a means to structure and represent knowledge about a domain in a formal way [Guarino, 2009]. Ontology is discussed more in section 5.

To assess the required competence to perform a certain task, a group of questions is employed and stored in the *Test Bank*. In this context, the tests are classified into different groups based on their level of difficulty. The *Test Generator* provides the different test packages in the range of difficulties, refers to the profile of the user and considers especially the user's job experience and professional level. The Med-Assess *Recommender* is discussed in section 4.

4 Recommender System

"Recommendation systems are software tools and techniques providing suggestions for items to be of use to a user" [Ricci *et al*., 2011]. RS refers to a kind of Information System (IS) which analyzes "*User's Need*", collects the "*Items*", and suggests them to the "*Users*" [Ricci, 2011], [Klahold, 2009]. Recommendation techniques are made out based on knowledge source [Burke, 2007]. These knowledge sources can be fed by "the knowledge of other users' preferences" or "ontological or inferential knowledge about the domain, added by a human knowledge engineer" [Burke, 2007]. Burke [2007] distinguished six types of recommendation approaches (see Table 1).

Med-Assess utilizes the combination of content and knowledge-based recommendation approach i.e. hybrid recommender system. In particular, items are learning materials, which are gathered based on the analysis of the user's level of competence. Therefore, at first the needs analysis should be applied to identify the requirements for recommendation. Here the Items include domain knowledge to clarify how they meet the "User's Needs" [Ricci et al., 2011]. In knowledge-based recommender systems, the Users' Needs (based on the user profiling) are mapped to Items through involving the associated domain expert(s) (e.g. physicians, nursing educators, or nursing supervisors). The quality of the recommended items by content-based and knowledge-based recommender depends on the quality of the entered data in the system by knowledge engineers [Burke and Ramezani, 2011]. A knowledge-based recommender needs not only what features are associated with what items, but also an ontology over the item features to allow the system to reason about the relationship between the features [Burke and Ramezani, 2011].

Table 1 Types of Recommendation Approaches, Adopted by the authors from [Ricci *et al*., 2011].

Recommendation approach	Description
Content-based	The system recommends items refer to the user's likes and dislike based on product features.
Collaborative	The system provides recommendation refers to "users with similar tastes liked in the past".
Demographic	The system generates recommendations based on rating of users in those niches.
Community-based	The system recommends item with regard to the preferences of the user's friends.
Knowledge-based	The system does not gather user ratings. The system provides recommendation refers to specific domain knowledge about "how certain item features meet user needs and preferences".
Hybrid recommendation	The system is developed based on the combination of the above mentioned techniques.

As shown in Figure 1, Med-Assess *Recommender* module contains *Recommend Learning Materials*, *Test Result Analysis*, *Learning Material Ontology (LM-Onto)* and *User Profile Bank*. The *Test Result Analysis* receives the test package which is answered by the user. The incorrect answers show the lack of competence(s) of the respected task(s). In this term the rate of the incorrect or correct answers is considered as the level of competence of the practitioners to perform each task. This result is used as "User's Needs" and stored as information in the *User Profile Bank*. Additionally *Recommender* contains *LM-Onto* in order to formalize and represent the LM domain as "Items". To create semantic recommend, Med-Assess utilized ontology. "Ontologies are now used routinely in recommender systems" [Middleton *et al*., 2009]. In LM-Onto, the knowledge domain of "how to perform the nursing tasks" has been formalized in a hierarchically structure and it can be used as a basis for a knowledge base.With reference to the lack of competence(s) to perform (a) specific task(s), the Recommended Learning Material feature will recommend the appropriate LM(s) to practitioners via the User Interface.

5 Med-Assess Ontologies framework

As mentioned earlier, Med-Assess deploys ontologies as a knowledge representation method to establish the semantics e.g. between learning materials in *LM-Onto* and job profiles in *Job role-Onto*. In the implementation, an in-house software solution, providing features for ontology and test bank creation is applied. As shown in Figure 2, the ontology engineering of Med-Assess consists of three stages.

Stage 1 is the modeling of the nursing processes. The inputs of this stage are nursing literature studies (as partially discussed in sub-section 2.1), and knowledge acquisitions via interviews with the nurses, educators and physicians. The output of this stage is nursing master list of tasks and nursing process which indicates the sequential relation of the activities.

Stage 2 is transforming nursing tasks and processes to build the Job role Ontology (Job role-Onto). In this stage the nursing tasks and sub-tasks are formalized in a hierarchically structure. The output of this stage is Job role-Onto.

Stage 3 contains the modeling and development of Learning Materials Ontology (LM-Onto) refers to the nursing literature studies as partially discussed in sub-section 2.2, and interviews with the domain experts especially consulting with nursing schools. In LM-Onto the know-how to perform nursing tasks are formalized.

The methodology to develop *LM-Onto* refers to the "Ontology engineering methodology" which is provided by [Sure *et al*. 2009]. This process consists of five main steps [Sure *et al*., 2009];

(1) **Feasibility study**: to identify problems/ opportunities.

(2) **Kickoff**: to clarify what this ontology should support, what the valuable knowledge sources are to build a semi-formal ontology.

(3) **Refinement**: to formalize a refined semi-ontology into target ontology and to create a prototype.

(4) **Evaluation**: to evaluate technology, users, and ontology to ready for the roll-out into a productive system,

(5) **Application and evolution**: to apply the ontology and manage evolution and maintenance. Here, this point should be highlighted that "an ideal ontology is one whose models exactly coincide with the intended ones" [Guarino *et al*., 2009].

As the project of Med-Assess is still ongoing, not all aforementioned steps of creating *Job role-Onto* and *LM-Onto* have been established yet. The first steps in creating these ontologies have been done in the form of literature studies about nursing tasks and modeling of work processes and the rest is planned to be accomplished within work packages.

The concept of "Ontology" in general is part of the "Semantic Web" [Berners-Lee *et al.*, 2001], a structure, which according to his inventors "will open up the knowledge and workings of humankind to meaningful analysis by software agents, providing a new class of tools by which we can live, work and learn together" [Berners-Lee et al., 2001]. Med-Assess fulfills the aspect of *working together* with modeling the *Job role-Onto*, which formalizes nursing tasks i.e. creating a common understanding of the daily activities, commotions and requirements e.g.

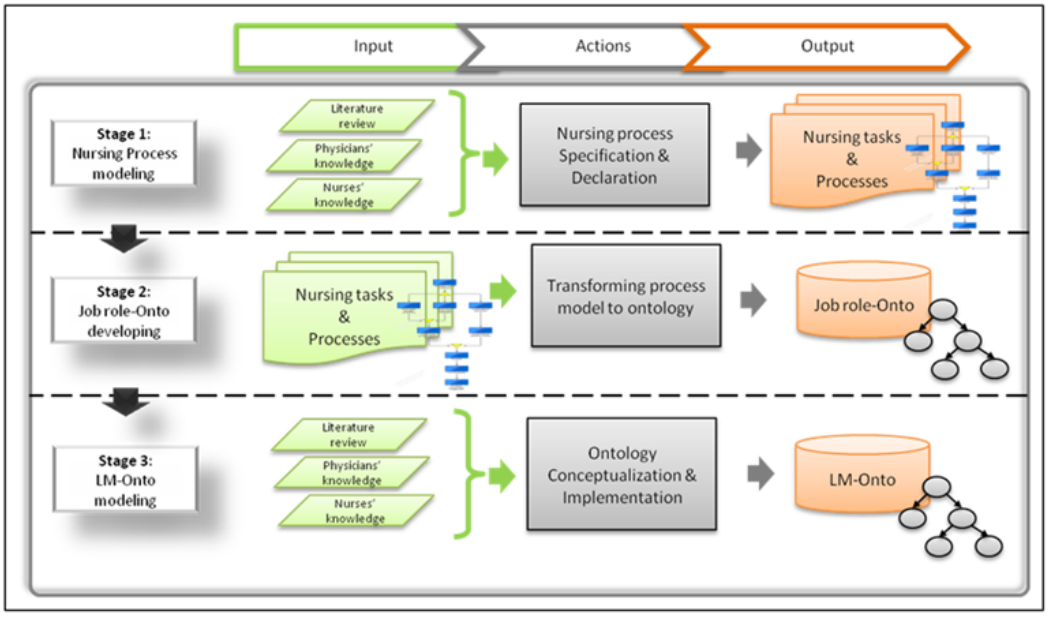


Figure 2 Med-Assess Ontologies Framwork

of a basic and a neuroscience nurse. *LM-Onto* meets the *learning aspect*, as this ontology represents learning material for the aforementioned job roles.

For the recommendation of corresponding learning material from *LM-Onto* to a test candidate, after acquiring his/her *Test Result Analysis*, an inference mechanism is required. Therefore, rules will be applied for establishing the reasoning processes. In addition to the project plan a research study was done by [Demuth, 2013] about the general interaction of ontologies and rules in the context of a medical scenario. This study [Demuth, 2013] contains a scenario, where exemplarily a competence ontology and the related rules have been created and executed, e.g. to filter the nurses, who need training in a specific topic, which is related to their area of work[7]:

```
worksIn(?nurse, ?area) ∩ lacksCompe-
tenceIn (?nurse, ?topic) ∩ topicRela-
tedToArea(?topic, ?area) → need-
sTraining(?nurse, ?topic)
```

Where the variables `?nurse`, `?area` and `?topic` may be a concrete test candidate, who works in the area of neurology and lacks knowledge in the topic of indirect care and as a conclusion is recommended for training. While this is a simple rule, this case study delivered first input results for the structures and associated rules, needed for building up *Job role-Onto*, *LM-Onto* and the executing of the recommendation of fitting learning material. Indeed, verifying such an approach requires several steps, like defining all rules and procedures to merge, aggregate or breakdown all rules in cooperation with the domain experts and knowledge engineers.

6 Conclusion and future research

This paper discusses the role of education and training in the development and improvement of nursing competence. It provides a contribution to (further) vocational education and training in the health sector. In particular, the paper holds the concept of Med-Assess and presents the design of RS as a component/sub-system in the framework of Med-Assess. In fact, Med-Assess is a gateway to bridging a synergistic approach using human resource, experience management and knowledge management methodologies to support nursing education and training. However, there are limitations in the concept and domain of application that should be addressed properly through the progress of the project and within the future research.

In particular, Med-Assess does not provide learning materials, instead it only recommends them. The reason for this approach is due to the lack of existing structured online learning systems in the field of neuroscience nursing. Therefore it is essentially important to establish in-house workshops and develop learning materials (e.g. text books) using online learning technologies and a combination of text-based and multi-media materials for learning and education. In this way, licensing of text-based materials is a major challenge which should be considered and handled through communication with the copyright holders and publishers.

Moreover, the recommendation result of Med-Assess does not reflect a kind of certificate yet. Integration of certification in the framework of Med-Assess might encourage the users to eagerly take part in the tests.

For implementation and evaluation of the performance, a pilot test with a number of test candidates (i.e. minimum of 200 candidates) will take place.If necessary, the solution will be adapted according to the feedback.

In the domain of application, one of the major challenges is the autonomy of nurses. For example, diagnosis is a task that should be fulfilled by the contribution of physicians and nurses. In this way, the authority of nurses is quite limited in Germany, while they have more freedom in other European countries or the United States. This issue should be fully considered in the development of the system and incorporating of learning materials.

Med-Assess also has an influence on the decision process of superiors in medical institutions. In fact, the secondary objective of the Med-Assess is to support superiors on integration of foreign job applicants (e.g. Chinese nurses in the German health sector).

Additionally Med-Assess has direct influence on continuous improvement of nursing performance through regular evaluation of the nurses. Improving nursing performance is directly reflected in doctor-nurse and nurse-patient communications as well as customer satisfaction in hospitals and clinics. Dealing with neuroscience patients and their relatives, this issue is very important and can affect the entire treatment process.

Acknowledgments

The research presented in this paper has been funded with support from the European Commission under the grant no. DE/12/LLP-LdV/TOI/147557. This communication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

The authors would like to thank the members of Med-Assess consortium, Amsterdam Business School of the University of Amsterdam, Beta Klinik GmbH a private international clinic in Bonn, and Corvinno Technology Transfer Center Nonprofit Public Ltd. in Budapest.

References

[Basili, 2009] Victor R. Basili, Gianluigi Caldiera and H. Dieter Rombach. The Experience Factory, 2009, http://agde.cs.uni-kl.de/teaching/qmss/ws2009/material/excercise/experienceFactory.pdf [last visit 30.08.2013]

[Baumeister *et al.*, 2007] Joachim, Baumeister, Jochen, Reutelshoefer and Frank, Puppe. KnowWE: community-based knowledge capture with knowledge wikis . In *Proceedings of the 4th international conference on Knowledge capture*, pages189–190, ACM, New York, NY, USA, 2007.

[Berners-Lee *et al.*, 2001] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284 (5): 34–43, 2001.

[Blancero *et al.*, 1996] Donna Blancero, John Boroski, and Lee Dyer. Transforming human resource organiza-

[7] The rule is adopted from the original source [Demuth, 2013] to address the specific example in the framework of this paper.

tions: a field study of future competency requirements. *Human Resource Management*, 35:383–404, 1996.

[Bulechek *et al*., 2013] Gloria M. Bulechek, Howard K.. Butcher, Joanne M. Dochterman and Cheryl Wagner (Eds.). *Nursing interventions classification (NIC)*, 6th Ed., Elsevier, St. Louis, Missouri, 2013.

[Burke and Ramezani, 2011] Robin Burke and Maryam Ramezani. Matching Recommendation Technologies and Domains .In: F. Ricci et al. (eds.), Recommender Systems Handbook. Springer Science+Business Media, LLC: 1- 35, 2011.

[Burke, 2007] Robin Burke. Hybrid web recommender systems. *The Adaptive Web*, Springer Berlin/Heidelberg, Germany, 377–408, 2007.

[Demuth, 2013] F. Demuth. Analyse und Interaktion von Ontologien und Rules im Bereich Semantic Web und praktische Umsetzung eines medizinischen Beispielszenarios, Bachelor Project Work in the context of Med-Assess, *Institute of Knowledge Based Systems, University of Siegen*, 2013.

[DPR, 2004] Deutscher Pflegerat e. V. Rahmen – Berufsordnung für professionell Pflegende. Deutscher Pflegerat, Berlin, Germany, 2004, http://www.deutscher-pflegerat.de/dpr.nsf/3F6CE4D95D84F8EDC12572B9003A1EF2/$File/Rahmenberufsordnung.pdf [last visit 04.07.2013].

[European Commission, 2008] European Commission, Recommendation of the European Parliament and of the Council of 23 April 2008 on the establishment of the European Qualifications Framework for lifelong learning. *Official Journal C 111*, 6.5.2008, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=oj:c:2008:111:0001:0007:en:pdf [last visit 04.07.2013].

[European Commission, 2012a] European Commission, Online survey on scientific information in the digital age, 2012, http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf [last visit 04.07.2013].

[European Commission, 2012b] European Commission, Lifelong Learning Programme, Transfer of Innovation call 2012, ADAM database entry: http://www.adam-europe.eu/prj/9974/project_9974_en.pdf.

[European Parliament and Council, 2005] European Parliament and Council. DIRECTIVE 2005/36/EC, The European parliament and of the council on the recognition of professional qualifications, *Official Journal of the European Union*: 40-42, September 2005, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2005:255:0022:0142:en:PDF [last visit 04.07.2013].

[Glosson and Schrock, 1985] Linda R. Glosson, Jay R. Schrock. Competencies Needed for Articulation Among and Between Post-secondary Vocational Food Technology Programs in Texas. *Texas Education Agency*, Austin, 1985, https://repositories.tdl.org/ttu-ir/bitstream/handle/2346/9847/31295003969770.pdf [last visit 04.07.2013].

[Guarino *et al*., 2009] Nicola Guarino, Daniel Oberle and Steffen Staab. What Is an Ontology? S. Staab and R. Studer (Eds.). *Handbook on Ontologies*, International Handbooks on Information Systems, Springer Berlin/Heidelberg, Germany, 1-17, 2009.

[Hesketh and Fleetwood, 2006] Anthony Hesketh and Steve Fleetwood. Beyond measuring the human resource management-organizational performance link: applying critical realist meta-theory. *Organization*, 13: 677–700, 2006.

[Klahold, 2009] André Klahold, *Empfehlungssysteme: Recommender Systems - Grundlagen, Konzepte und Lösungen: Grundlagen, Konzepte und Systeme*, Vieweg+Teubner Verlag, 2009.

[KrPflG, 2003] Krankenpflegegesetz. Gesetz über die Berufe in der Krankenpflege, 2003, http://www.gesetze-im-internet.de/bundesrecht/krpflg_2004/gesamt.pdf [last visit 04.07.2013].

[KrPflAPrV, 2003] Ausbildungs- und Prüfungsverordnung für die Berufe in der Krankenpflege, November 2003, http://www.gesetze-im-internet.de/bundesrecht/krpflg_2004/gesamt.pdf

[last visit 04.07.2013].

[Lianga *et al*., 2013] Chiung-Ju Lianga, Ying-Li Linb and Hsiu-Feng Huang. Effect of core competence on organizational performance in an airport shopping center. *Air Transport Management*, 31:23-26, August 2013

[Med-Assess, 2012] Official homepage of Med-Assess Project, www.med-assess.eu

[Middleton *et al*., 2009] Stuart E. Middleton, David De Roure, and Nigel R. Shadbolt. Ontology-Based Recommender Systems. S. Staab and R. Studer (eds.). *Handbook on Ontologies*, International Handbooks on Information Systems, Springer, Berlin/Heidelberg, Germany, 20, 779-7962009.

[Mol et al., 2013] S. T. Mol, G. Kismihok, F. Ansari and M. Dornhöfer. Integrating Knowledge Management in the context of Evidence Based Learning: Two concept models for facilitating the assessment and acquisition of job knowledge. M. Fathi (ed.) *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*. Springer, Germany, 29-45, 2013.

[Müller-Klepper (ed.), 2005] Petra Müller-Klepper (ed.). Rahmenlehrplan für die Krankenpflegeberufe. *Hessisches Sozialministerium Referat Öffentlichkeitsarbeit*, Wiesbaden, Germany, November 2005, http://www.examensfragen.de/Rahmenlehrplan_Hessen.pdf [last visit 04.07.2013].

[Neacşu, 2011] Ioan Neacşu. Academic learning – a generative model of professional competence. Impact on teaching. *Procedia - Social and Behavioral Sciences*, 11: 230–234, 2011.

[Norman, 1985] GR. Norman. Assessing Clinical Competence. Springer : 330-341, New York, USA, 1985.

[Oelke *et al*., 2003] Uta Oelke, Gertrud Hundenborn, Cornelia Kühn, Richtlinie für die Ausbildung in der Gesundheits- und Krankenpflege sowie in der Gesundheitsund Kinderkrankenpflege. *Ausbildungsrichtlinie für die staatlich anerkannten Kranken-und Kinderkrankenpflegeschulen in NRW.*

Ministeriums für Gesundheit, Soziales, Frauen und Familie des Landes Nordrhein-Westfalen, Germany, revised version, 2003, http://www.mais.nrw.de/08_PDF/002/ausbildungsrichtlinien-krankenpflege-nrw.pdf [last visit 04.07.2013].

[OntoHR, 2009] Official homepage of OntoHR Project, www.ontohr.eu

[Parry, 1998] Scott B. Parry. Just what is a competency? (why should you care?), *Training* , 58–64, 1998.

[Pilz, 2012] Pilz, Matthias (Ed.), *The Future of Vocational Education and Training in a Changing World*, Springer (VS) Verlag für Sozialwissenschaften, 2012.

[Ricci et al., 2011] Francesco Ricci, Lior Rokach and Bracha Shapira. Introduction to Recommender Systems. In: F. Ricci et al. (eds.), Recommender Systems Handbook. Springer Science+Business Media, LLC: 1-35, 2011.

[Rudolph, 1999] Rodney D. Rudolph. Desirable competencies of hospitality graduates in year 2007. Ph.D. Thesis. *University of Cornell, USA*, 1999.

[Salgado *et al*., 2003] Jesús. F. Salgado, Neil Anderson, Silvia Moscoso, Cristina Bertua, Filip de Fruyt, and Jean Pierre. Rolland. A Meta-Analytic Study of General Mental Ability Validity for Different Occupations in the European Community, *Journal of Applied Psychology*,88(6): 1068-1081 ,2003.

[Schmidt, 2009] Frank L. Schmidt. Select on Intelligence, E. A. Locke (Ed.) *The Blackwell Handbook of Principles of Organizational Behavior*, Blackwell Publishing, 3-14, 2009.

[Schmidt and Hunter, 2004] Frank. L. Schmidt and John. E Hunter. General Mental Ability in the World of Work: Occupational Attainment and Job Performance, *Journal of Personality and Social Psychology*, 86 (1): 162-173, 2004.

[Schroeter, 2008] Kathryn Schroeter. Competence Literature Review. *Competency & Credentialing Institute*, 2008, http://www.cc-institute.org/docs/default-document-library/2011/10/19/competence_lit_review.pdf [last visit 04.07.2013].

[Sure *et al*., 2009] York Sure, Steffen Staab, and Rudi Studer. Ontology Engineering Methodology. S. Staab and R. Studer (eds.). *Handbook on Ontologies - International Handbooks on Information Systems*, Springer, Berlin/Heidelberg, Germany, 135-152, 2009.

[Williams, 2009] Heather Williams, Ruth Harris and Lynne Turner-Stokes. Work sampling: a quantitative analysis of nursing activity in a neuro-rehabilitation setting. *ADVANCED NURSING*, Blackwell Publishing Ltd: 2097-2107, May, 2009.

[Weinstein and Underwood, 1985] Claire E. Weinstein, Vicki L. Underwood. Learning strategies: The how of learning J.W. Segal, S.F. Chiman, R. Glaser (eds.), *Thinking and learning skills, relating instruction to researc*h, LEA, Hillsdale, 241-258, 1985.

[Wu and Lee, 2007] Shu-Chen Wu and Yong-Yin Lee. The development of effective learning scale for college students. Education and Psychology, 30 (4): 173–197, 2007.

Forschungsansatz für eine wissensbasierte Wirkungsanalyse im Business Intelligence

Axel Benjamins
Universität Osnabrück,
Institut für Informationsmanagement und Unternehmensführung,
Katharinenstraße 3, 49069 Osnabrück
axel.benjamins@uni-osnabrueck.de

Abstract

Die Möglichkeit, computergestützte Analysen zur Unterstützung von Entscheidungsprozessen im Unternehmen einzusetzen, wird aktuell unter dem Begriff Business Intelligence zusammengefasst und stellt einen bedeutenden Wettbewerbsvorteil dar. Ein wichtiger Bestandteil dieser Analysen sind Wirkungsanalysen zum Identifizieren von Wirkungszusammenhängen zwischen Daten. Dadurch können Ursachen für Wirkungen erkannt und bei Bedarf in den Entscheidungsprozessen berücksichtigt werden. Zur Durchführung von Wirkungsanalysen wird allerdings ein umfangreiches Datenverständnis benötigt. Da sich dies bei steigenden Datenmengen zu einer immer schwerer werdenden Aufgabe entwickelt, wird im vorliegenden Beitrag ein Konzept zur Unterstützung von Wirkungsanalysen vorgestellt. Bestandteile des Konzepts sind eine homogene Wissensbasis über Wirkungszusammenhänge, welche Wissen aus unterschiedlichen Quellen beinhaltet, und einer an die menschliche Vorgehensweise angelehnte Verarbeitung des Wissens.

1 Ausgangssituation und Problemstellung

Die Analyse von Daten gilt für Unternehmen als entscheidender Wettbewerbsvorteil (vgl. Davenport & Harris, 2007, S. 48), da die Ergebnisse zur effektiveren Entscheidungsfindung eingesetzt werden können (vgl. Davenport, 2006, S. 99; vgl. Vercellis, 2009, S. 3). Ein wichtiger Bestandteil dieser Analysen ist die Identifikation von Wirkungszusammenhängen in einer konkreten Entscheidungssituation (vgl. Cyert, Simon, & Trow, 1956, S. 237; vgl. Mintzberg, Raisinghani, & Theoret, 1976, S. 274). Bei einer Wirkungsanalyse handelt es sich um ein entscheidungsunterstützendes Verfahren, das zum besseren Verständnis von relevanten Wirkungsketten beiträgt und aus technologischer Sicht in den Bereich des Business Intelligence eingeordnet wird (vgl. Gluchowski, 2001, S. 6). Durch diese Analysen können die potenziellen Konsequenzen einer Entscheidung besser vorhergesehen werden.

Die Anforderungen an Analysten zur Durchführung von erfolgversprechenden Analysen, setzen sich aus dem Geschäftsverständnis, der Methodenkenntnis sowie dem Datenverständnis zusammen (vgl. Laursen & Thorlund, 2010, S. 98–101). Vor allem der Aufbau eines umfangreichen Datenverständnisses, also beispielsweise die Strukturen oder die Integration von Daten, stellt eine wichtige Grundlage von Analysen dar (vgl. Davenport, Harris, & Morison, 2010, S. 23). Bei wachsenden Datenmengen und einer größeren Vielfalt an Datenquellen (vgl. Minelli & Dhiraj, 2013, S. 10) entwickelt sich der Aufbau aber zu einer immer schwerer werdenden Aufgabe.

Die Nachfrage nach Anwendern, die dieses Datenverständnis aufbringen und zudem ausreichende analytische Fähigkeiten besitzen, wird in Zukunft weiter steigen und vor dem Hintergrund des aktuell prognostizierten Fachkräftemangels im Hinblick auf Analysen (vgl. Chen, Chiang, & Storey, 2012, S. 1185) mittelfristig nicht zu decken sein. Dieser drohende Mangel an geeigneten analytischen Fähigkeiten gefährdet die Erlangung oder Sicherung von Wettbewerbsvorteilen. Umso wichtiger ist es, allen Entscheidungsträgern die Möglichkeit zu eröffnen, computergestützte Wirkungsanalysen im Rahmen einer Entscheidungssituation durchzuführen.

2 Zielsetzung

Zur Formulierung eines Forschungsziels werden im Folgenden wichtige Komponenten einer wissensbasierten Wirkungsanalyse hergeleitet.

Ein Entscheidungsprozess wird nach Simon (1977) in die Phasen Intelligence, Design und Choice unterteilt[1]. Die Intelligence-Phase dient zur Identifikation des Problems und dessen Ursachen, in der Design-Phase werden mögliche Lösungsalternativen erarbeitet und in der Choice-Phase wird die eigentliche Selektion einer Lösung durchgeführt (vgl. Simon, 1977, S. 40 f.). Mintzberg, Raisinghani & Theoret (1976) haben eine weitere Unterteilung der Intelligence Phase in die Erkennung von Entscheidungsbedarf und die Diagnose vorgenommen, wobei letztere das Identifizieren und Verstehen von Wirkungszusammenhängen in einer Entscheidungssituation umfasst (vgl. Mintzberg u. a., 1976, S. 252 f.).

Diese Wirkungszusammenhänge nehmen durch eine komplexer werdende Umwelt einen immer höheren Stellenwert in Entscheidungssituationen ein (vgl. Morton, 1971, S. 32; vgl. Sabherwal & Becerra-Fernandez, 2011,

[1] Simon (1977) definierte nachträglich noch eine vierte Phase Review, welche u.a. von Gorry & Morton (1971) nicht aufgenommen wurde und für diesen Beitrag keinen weiteren Einfluss hat.

S. 9). Eine weit verbreitete Methode zur Dokumentation von Ursachen für eine Wirkung stammt aus dem Total Quality Control (TQC) von Ishikawa zur Korrektur von Qualitätsmängeln (vgl. Andersen & Fagerhaug, 2006, S. 12; vgl. Ishikawa, 1972, S. 19 ff.). Durch ein Ursache-Wirkungs-Diagramm[2] lassen sich die Ursachen für eine beobachtete Wirkung aufzeigen. Es wird hierbei keine Bewertung darüber vorgenommen, wie groß der Einfluss einer Ursache in einer konkreten Situation auf die Wirkung war (vgl. Ishikawa, 1972, S. 20). Ein solches Diagramm besitzt einen gewissen Lerneffekt für alle an der Erstellung Beteiligten (vgl. Ishikawa, 1972, S. 26). Im nachfolgenden Konzept sollen neben der Identifikation, die Validierung und Bewertung von Wirkungsbeziehungen im Fokus stehen, um dem Anwender einen Eindruck der Größe des Einflusses von Wirkungen zu vermitteln.

Bei der menschlichen Vorgehensweise zum Lösen von Problemen werden Wirkungszusammenhängen immer in einem konkreten Problemkontext identifiziert und validiert (vgl. Newell & Simon, 1972, S. 809). Nachdem das Problem verstanden wurde, muss dazu ein sogenannter Problemraum konstruiert oder, falls bereits in ähnlicher Form vorhanden, das Problem in einen bestehenden Problemraum übertragen werden (vgl. Newell & Simon, 1972, S. 809). Ein Problemraum besteht dabei aus den Elementen eines Problems, den Operatoren zur Bearbeitung sowie dem für eine konkrete Problemsituation relevanten Wissen (vgl. Newell & Simon, 1972, S. 810). Es wird nicht das gesamte Wissen in einer Problemsituation angewendet, sondern primär das mit dem Problem in Beziehung stehende Wissen zur Lösungsfindung herangezogen. Diese Vorgehensweise und die Eingrenzung des zu verwendenden Wissens sollen in das nachfolgende Konzept mit aufgenommen werden, sodass die Analyse eines Problems über Wirkungszusammenhänge immer im Kontext einer Entscheidungssituation durchgeführt wird.

Die Möglichkeit Probleme zu lösen und die Qualität einer Lösung basieren zu einem großen Teil auf Wissen (vgl. Newell & Simon, 1972, S. 814). Dabei geschieht die Auswahl von relevantem Wissen in einer Problemsituation immer auf Basis des gesamten verfügbaren Wissens. Eine möglichst große Ansammlung von Wissen stellt also einen wichtigen Faktor zur Identifikation von Wirkungszusammenhängen dar, da sich bei einer größeren Wissensmenge die Wahrscheinlichkeit erhöht, dass die einflussreichsten Wirkungszusammenhänge in der Wissensmenge vorhanden sind und gefunden werden können. Dies führt zu der Anforderung, dass die Integration von Wissen aus möglichst vielen Quellen sowie die Erweiterbarkeit der Wissensstrukturen im nachfolgenden Konzept zum Aufbau einer möglichst umfangreichen Wissenssammlung eine elementare Rolle einnehmen sollen.

Das folgende Forschungsziel basiert auf den diskutierten Anforderungen und fasst diese zusammen.

> Forschungsziel: Die Unterstützung der *Identifikation und Validierung* von *Wirkungszusammenhängen* in der Intelligence-Phase von Entscheidungsprozessen im *Kontext einer Entscheidungssituation* mithilfe eines *wissensbasierten* Systems.

[2] Das Ursache-Wirkungs-Diagramm wird auch als Ishikawa- oder als Fischgräten-Diagramm bezeichnet (vgl. Ishikawa, 1985, S. 64).

3 Verwandte Arbeiten

Erste Forschungsarbeiten hinsichtlich Wirkungsanalysen im Business Intelligence existieren von Caron (2013). Dabei wurde die Identifikation, Erklärung und Behebung von Ausnahmewerten in multidimensionalen Analysen untersucht. Ein Ausnahmewert ist als eine signifikante Abweichung von dem erwarteten Wert definiert und führt oftmals zur Identifikation eines Problems oder zu neuem Unternehmenspotenzial (vgl. Caron, 2013, S. 9). Durch den Aufbau eines statistischen Modells werden diese Werte innerhalb eines On-Line Analytical Processing (OLAP)-Würfels durch Daten im gleichen Würfel zu erklären versucht (vgl. Caron & Daniels, 2008, S. 889). Dies setzt jedoch die Annahme voraus, dass eine Erklärung eines Ausnahmewertes innerhalb des vorab abgegrenzten multidimensionalen Datenbereichs vorhanden ist.

Da davon ausgegangen wird, dass eine Wirkungsanalyse immer in einer konkreten Entscheidungssituation durchgeführt werden muss und somit eine auf die jeweilige Situation angepasste Menge von Daten benötigt, ist das Konzept von Caron (2013) zur Erreichung des Forschungsziels nicht anwendbar.

4 Forschungsansatz für eine wissensbasierte Wirkungsanalyse im Business Intelligence

Zur Durchführung einer Wirkungsanalyse müssen zuerst potenzielle Wirkungszusammenhänge identifiziert und anschließend auf ihren Einfluss hin überprüft werden (vgl. Mintzberg u. a., 1976, S. 253). Die Identifikation ist ein wissensintensiver Prozess auf Basis des Daten- und Geschäftsverständnisses (vgl. Kapitel 1). Die Überprüfung der Wirkungszusammenhänge wird durch die Anwendung von statistischen Methoden zur Bestimmung der Korrelation durchgeführt und beruht auf der Methodenkenntnis (vgl. Kapitel 1).

Zur Erreichung des Zieles wird im Folgenden untersucht, wie Wissen über Wirkungszusammenhänge zu deren Identifikation und Überprüfung nutzbar gemacht werden kann. Aufgrund der in Kapitel 2 analysierten Anforderungen wird hierzu ein wissensbasiertes System verwendet, dessen Struktur durch eine Trennung zwischen der Wissensrepräsentation und Wissensverarbeitung geprägt ist (vgl. Beierle & Kern-Isberner, 2008, S. 11; vgl. Ertel, 2009, S. 15)

Der vorliegende Forschungsansatz nimmt Bezug auf diese Trennung und diskutiert die Konzeption einer wissensbasierten Wirkungsanalyse. Zuerst wird hierzu das bereits vorhandene Wissen über Wirkungszusammenhänge in Unternehmen sowie die mögliche Speicherung dieses Wissens erläutert und im Anschluss der Ablauf der Wissensverarbeitung vorgestellt.

4.1 Wissensrepräsentation

Durch die Wissensrepräsentation wird Wissen aus unterschiedlichen Quellen in einer homogenen Struktur zusammengefasst. Dieses Wissen wird in einer Wissensbasis der Wissensverarbeitung bereitgestellt und eine Verknüpfung mit den ursprünglichen Daten hergestellt.

Im Unternehmen existieren diverse Quellen mit Wissen über potenzielle Wirkungszusammenhänge. Im Rahmen eines Business-Intelligence-Systems sind dies bspw. OLAP-Würfel oder Prozesse zur Extraktion, zur Transformation und zum Laden (ETL). Zusätzlich kann Wissen

aus Modellen von Decision Support Systems (DSS) ausgelesen werden, welche eine quantitative Sichtweise auf Wirkungszusammenhänge bieten. Neben diesem strukturierten Wissen besitzen Experten umfangreiches Wissen über Wirkungszusammenhänge. Wird die Annahme zugrunde gelegt, dass Wissen über Wirkungszusammenhänge auch in unternehmensexternen Quellen existiert, kann unter anderem auch das World Wide Web als ein Lieferant für Wissen dienen.

Das in der Wissensbasis vereinheitlicht vorliegende Wissen kann zur systematischen Entdeckung von Wirkungszusammenhängen bereitgestellt werden. Zusätzlich werden Verweise auf die jeweiligen Daten im Data Warehouse zur statistischen Validierung von Wirkungszusammenhängen abgelegt. Die Integration von Wissen aus unterschiedlichen Quellen ist in Abbildung 1 dargestellt, wobei die Auswahl von Quellsystemen erweiterbar ist.

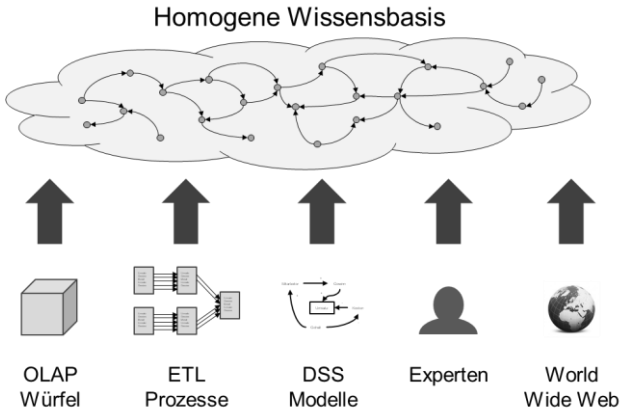


Abbildung 1: Homogene Wissensbasis zur wissensbasierten Entdeckung von Wirkungszusammenhängen

Zur Erstellung der Wissensbasis soll für den Einsatzzweck einer wissensbasierten Wirkungsanalyse möglichst hilfreiches Wissen aus den Quellen, falls möglich automatisiert, extrahiert werden. Das Wissen liegt in unterschiedlicher Form und Qualität vor und muss zur Integration in die homogene Wissensbasis transformiert werden. Diese Vorgehensweise ist an die Extraktion, die Transformation und das Laden beim Aufbau eines Data Warehouse angelehnt (vgl. Kemper, Baars, & Mehanna, 2010, S. 25). Die Übertragung des Wissens in die Wissensbasis wird durch die Repräsentationsform der Wissensbasis sowie das zu extrahierende Wissen aus den unterschiedlichen Quellen beeinflusst. Eine Ausgestaltung des Konzepts kann an dieser Stelle daher noch nicht erfolgen. Exemplarisch werden die abgebildeten Quellen kurz vorgestellt und das potenzielle Wissen über Wirkungszusammenhänge erläutert.

OLAP-Würfel

Zur Durchführung von dynamischen Analysen im multidimensionalen Datenraum werden Daten im Unternehmen in mehrdimensionaler Form in einem OLAP-Würfel aufbereitet (vgl. Gluchowski, Gabriel, & Dittmar, 2008, S. 143 f.). Die Daten werden in betriebswirtschaftliche Kennzahlen und Dimensionen eingeteilt und durch Dimensionshierarchien weiter strukturiert. Zusätzlich kann jeder OLAP-Würfel Metainformationen, bspw. über den Ersteller, den Einsatzzweck oder die Aktualität der Daten besitzen.

Zum Einsatz in einer Wirkungsanalyse sind die Struktur eines OLAP-Würfels und die Metainformationen relevant. Durch die Struktur können Aussagen darüber getroffen werden, welche Kennzahlen in welchen Dimensionen auftreten und dadurch einen potenziellen Zusammenhang haben. Zusätzlich bieten Hierarchien innerhalb einer Dimension die Möglichkeit, die semantischen Zusammenhänge der Daten besser zu verstehen. Hinsichtlich der Metainformationen ist der Einsatzzweck eines OLAP-Würfels besonders interessant, da er Hinweise auf potenzielle Wirkungszusammenhänge geben kann.

ETL-Prozesse

Ein Data Warehouse dient der Bereitstellung von themenorientierten, integrierten, beständigen und zeitorientierten Daten zur Entscheidungsunterstützung von Führungskräften (vgl. Inmon, 2002, S. 31). Der Aspekt der Integration wird durch die Aufbereitung der Daten aus Vorsystemen durch ETL-Prozesse ermöglicht, um einen konsistenten Datenbestand im Data Warehouse aufzubauen.

Wissen über potenzielle Wirkungszusammenhänge kann vor allem aus den Transformationsschritten von Daten aus operativen Vorsystemen zum Data Warehouse entnommen werden. Die Transformation unterteilt sich in die Teilprozesse der Filterung, der Harmonisierung, der Aggregation und der Anreicherung (vgl. Kemper u. a., 2010, S. 28). Aus allen genannten Schritten kann Wissen über potenzielle Wirkungszusammenhänge entnommen werden (vgl. Kemper u. a., 2010, S. 28–38):

- Filterung:
  Zusammenhänge in den Daten aus Transformationsregeln zur Bereinigung der Daten auf semantischer Ebene mithilfe von Plausibilitätskontrollen oder Mustererkennungsverfahren.
- Harmonisierung:
  Zusammenhänge von Bezeichnungen, wie bspw. Synonyme und Homonyme, die zum Aufbau eines konsistenten Datenbestands verwendet werden.
- Aggregation:
  Hierarchische Zusammenhänge in den Daten aus Definitionen und Berechnungsregeln von Hierarchiestufen.
- Anreicherung:
  Zusammenhänge von Daten, die zur Berechnung neuer Kennzahlen genutzt werden.

DSS Modelle

Durch den Einsatz eines repräsentativen oder abstrakten Modells kann die Analyse eines komplexen Problems in der Realität unterstützt werden (vgl. Turban, Sharda, & Delen, 2011, S. 44). Dabei kann eine Optimierung alle Faktoren darauf ausrichten, die in der aktuellen Situation optimale Lösung zu finden oder eine Simulation die Realität durch definierte Formeln zur Beobachtung von Auswirkungen definierter Alternativen fortschreiben (vgl. Power, 2002, S. 169 & 171).

Zur Erreichung dieser beiden Ziele müssen die Einflüsse zwischen den Faktoren möglichst exakt quantifiziert werden. Diese quantifizierten Einflüsse können in Kombination mit der Struktur von OLAP-Würfeln und ETL-Prozessen helfen, eine quantifizierte Sicht auf Wirkungszusammenhänge zu erhalten.

Experten

Zur Lösung von neu auftretenden Problemen oder Problemen ohne eindeutige Lösung besitzen Experten überdurchschnittlich gute Fähigkeiten (vgl. Friedrich & Stumptner, 1990, S. 14). Besonders bei der Bearbeitung von bisher undokumentierten Problemen können Experten ihr Wissen einsetzen, um die Auswirkungen einer Entscheidung zu evaluieren. Dieses Wissen basiert unter anderem auf Erfahrungen aus ähnlichen Situationen, auf dem Allgemeinwissen und auf der Intuition (vgl. Friedrich & Stumptner, 1990, S. 14).

Bei einem Vergleich der bisher vorgestellten Quellen nimmt das Wissen von Experten eine besondere Position ein. Es handelt sich um Wissen, welches noch nicht expliziert wurde. Da durch das Expertenwissen Wirkungszusammenhänge zwischen unterschiedlichen Quellen hergestellt werden können, stellt es einen zentralen Bestandteil zum Aufbau der Wissensbasis dar. Zudem ermöglicht ein Experte durch die Korrektur von Schlussfolgerungen des Systems eine stetige Verbesserung der Wissensbasis und führt langfristig zu besseren Ergebnissen.

World Wide Web

Ausgehend von der Annahme, dass nicht alle relevanten Daten zur Analyse von Wirkungszusammenhängen innerhalb eines Unternehmens vorhanden sind, besteht die Notwendigkeit, ebenfalls externe Quellen einbinden zu können. Dies eröffnet die Möglichkeit externe Einflussfaktoren bei der Analyse von Wirkungszusammenhängen zu berücksichtigen.

Als eine der größten unternehmensexternen Informationsquellen gilt das World Wide Web (vgl. Hitzler, Krötzsch, Rudolph, & Sure, 2008, S. 9) und durch Entwicklungen im Bereich des Semantic Webs können Informationen auch in maschinenlesbarer Form bereitgestellt werden (vgl. Hitzler u. a., 2008, S. 11). Um eine homogene Wissensbasis zu erhalten, müssen Quellen aus dem World Wide Web denselben Bedingungen wie unternehmensinterne Quellen genügen. Neben Metainformationen zur Quelle muss eine strukturierte Bereitstellung von Daten eines eingebundenen Faktors sichergestellt sein.

4.2 Wissensverarbeitung

Das gespeicherte Wissen in der Wissensbasis wird durch die Wissensverarbeitung auf eine konkrete Problemsituation angewendet, um relevante Wirkungszusammenhänge zu identifizieren. Die Vorgehensweise der Verarbeitung ist an die menschliche Vorgehensweise angelehnt (vgl. Kapitel 2). In Form einer Hypothese werden Mutmaßungen über potenzielle Wirkungszusammenhänge im Kontext einer konkreten Problemsituation angestellt (Hypothesengenerierung), welche im Anschluss auf Richtigkeit überprüft werden (Hypothesenverifizierung) (vgl. Gluchowski u. a., 2008, S. 143).

Zu Beginn erfolgt die Initialisierung, welche eine Problemsituation in die Wissensbasis einordnet. Aufbauend auf dieser Einordnung wird die Erkundung der Wissensbasis im Kontext der übertragenen Problemsituation angestoßen, um potenzielle Wirkungszusammenhänge zu identifizieren. Abschließend werden die gefundenen Wirkungszusammenhänge bewertet. Die drei Phasen sind in Abbildung 2 ersichtlich und werden nachfolgend einzeln und in ihrem Ablauf erläutert.

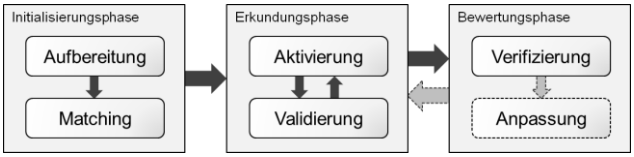


Abbildung 2: Ablauf der Wissensverarbeitung

Durch einen modularen Systemaufbau soll es möglich sein, diese Komponenten gezielt auszutauschen. So könnte beispielsweise eine andere Strategie zur Aktivierung in der Erkundungsphase verwendet oder eine Problemsituation aus einem anderen System aufbereitet werden. Dies erhöht die Flexibilität und Anpassbarkeit des Forschungsansatzes.

Initialisierungsphase

Zur Anwendung des Wissens aus der Wissensbasis in einer Problemsituation muss diese zu Beginn in die Wissensbasis übertragen werden. Dies geschieht durch die automatisierte Aufbereitung der Problemsituation in einem Modell und dem Matching dieses Modells in die Wissensbasis durch das System.

Die Aufbereitung dient der Erstellung eines strukturierten Modells auf Basis einer Problemsituation. Das Modell unterliegt dabei der Bedingung, dass es mit der Wissensbasis kompatibel sein muss oder durch das Matching in die Wissensbasis eingeordnet werden kann. Die Problemsituation kann in der Realität in verschiedenen Formen oder Systemen vorliegen. Die Überführung in ein einheitliches Modell führt dazu, dass aus unterschiedlichen Systemen eine Wirkungsanalyse initiiert werden kann. Die Abbildung 3 stellt beispielhaft die Überführung einer Problemsituation aus einem OLAP-Würfel in ein Modell dar. Der OLAP-Würfel besteht aus den Dimensionen Zeit (Dim1), Artikel (Dim2), Region (Dim3) und der Kennzahl Umsatz. In der Abbildung tritt zu einem bestimmten Zeitpunkt, bei einem bestimmten Artikel und in einer bestimmten Region ein Umsatzeinbruch (Problem) im Vergleich zum Vorjahr auf.

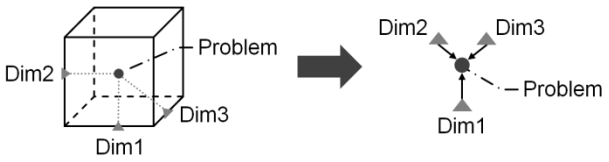


Abbildung 3: Aufbereitung einer Problemsituation aus einem OLAP-Würfel in ein Modell

Durch das Matching wird das zuvor erstellte Modell in die Wissensbasis eingeordnet. Auf Basis der Faktoren aus dem Model werden die dazu passenden Faktoren in der Wissensbasis gesucht. Die Beziehungen zwischen den Faktoren im Modell werden nur auf Existenz in der Wissensbasis überprüft. Ein exakter Abgleich der Beziehungen ist nicht erforderlich, da davon ausgegangen wird, dass dies zu keinem besseren Matching führt und in der Wissensbasis die Beziehungen eine höhere Qualität aufweisen. So könnten beispielsweise mehrere semantische Beziehungen zwischen zwei Faktoren oder anders gerichtete Beziehungen als im Modell existieren.

Das Ergebnis des Matching ist die Einordnung der Faktoren in die Wissensbasis, mit dem Ziel darauf aufbauend weitere Zusammenhänge zu finden. Der Abgleich des Modells mit der Wissensbasis ist in Abbildung 4 dargestellt.

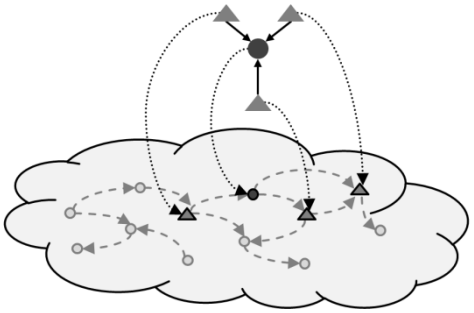


Abbildung 4: Matching eines Modells in die Wissensbasis

Erkundungsphase

Nach Abbildung der Problemsituation in die Wissensbasis erfolgt das Auslesen von relevanten Wirkungszusammenhängen. Dies geschieht durch die Aktivierung von anknüpfenden Faktoren an die Faktoren der Problemsituation und die Validierung von potenziell wirkenden Faktoren auf Basis bestehender Daten.

Durch die Aktivierung werden Faktoren, die direkt mit dem aktivierten Faktor zusammenhängen, zur weiteren Analyse von Wirkungszusammenhängen herangezogen. Es werden alle Faktoren der Problemsituation als Ausgangslage verwendet und die verknüpften Faktoren aktiviert (vgl. Abbildung 5). Dies geschieht durch die Verwendung der unterschiedlichen Verbindungen zwischen den Faktoren. Die aktivierten Faktoren sind unbestätigte Wirkungszusammenhänge in einer konkreten Problemsituation und werden im Rahmen der Validierung überprüft.

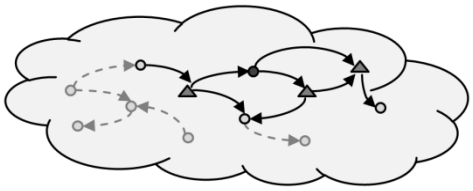


Abbildung 5: Aktivierung anknüpfender Faktoren in der Wissensbasis

Zur Bestätigung der gefundenen Wirkungszusammenhänge werden die einzelnen Faktoren für die konkrete Problemsituation validiert. Schrittweise werden die Daten von jeweils zwei miteinander verbundenen Faktoren aus dem Data Warehouse ausgelesen und die Korrelation mithilfe statistischer Methoden überprüft. Die berechnete Korrelation stellt in der konkreten Problemsituation die Gewichtung der Verbindung dar.

Nach Abschluss der Validierung wird die Erkundungsphase auf Basis der neuen Ausgangslage wiederholt. Die Gewichtung der Verbindungen kann dazu genutzt werden, um Wirkungszusammenhänge mit einer hohen Korrelation von zwei Faktoren auszuwählen. Dadurch kann eine gezielte Aktivierung entlang erfolgversprechender Faktoren erfolgen. Die Wiederholungen enden, wenn keine signifikante Korrelation mehr festgestellt werden kann.

Bewertungsphase

In der Bewertungsphase kann der Anwender die Ergebnisse aus der Erkundungsphase beurteilen. Dies geschieht durch eine Verifizierung und, falls notwendig, eine Anpassung der Ergebnisse.

In der Verifizierung werden die vom System gefundenen und validierten Wirkungszusammenhänge durch den Anwender überprüft. In Abhängigkeit vom Erfahrungsgrad des Anwenders sowie der konkreten Problemsituation, wird ein Abgleich mit der Realität durchgeführt. Dieser Abgleich kann bspw. darin bestehen, dass ein versierter Anwender sein Expertenwissen einsetzt, um zu beurteilen, wie realistisch die Ergebnisse sind.

Wenn der Anwender durch Expertenwissen oder durch Anwendung der Ergebnisse weitere relevante Wirkungszusammenhänge findet, können die ursprünglichen Ergebnisse angepasst und in die Wissensbasis zurückgeschrieben werden. Falls hierzu eine Verbindung komplett fehlt, kann die Wissensbasis durch Unterstützung des Anwenders erweitert werden. Alternativ können bestehende Verbindungen in der Wissensbasis durch den Anwender aufgehoben werden.

Da die Wissensbasis durch die Anpassung verändert wurde, kann die Erkundungsphase erneut gestartet werden. Eine Initialisierung ist nicht notwendig, da die Problemsituation bereits mit der Wissensbasis abgeglichen wurde und nur Beziehungen zwischen Faktoren verändert wurden.

5 Ausblick

Der vorgestellte Forschungsansatz dient als erster Grundstein für weitere Arbeiten. Die Wissensbasis sowie die einzelnen Verarbeitungsschritte müssen ausgestaltet und detailliert definiert werden. Damit die einzelnen Verarbeitungsschritte modular ausgetauscht werden können, ist es notwendig, klare Schnittstellen zu definieren. Vor allem die Ein- und Ausgabewerte müssen festgelegt werden, sodass das Gesamtkonzept weiterhin funktionieren kann. Die noch zu wählende Wissensrepräsentationsform beeinflusst zudem sehr stark die darauf aufbauende Inferenzmaschine zur Wissensverarbeitung.

Zur Verfeinerung des vorgestellten Konzepts kann eine Klassifizierung von Wirkungszusammenhängen durchgeführt werden, um ggf. eine effizientere Nutzung der Wissensrepräsentation und -verarbeitung zu ermöglichen. Je nach Klassenzugehörigkeit der Wirkungszusammenhänge könnten die Aktivierung und Validierung während der Erkundungsphase auf die jeweilige Klasse abgestimmt oder sogar andere Module verwendet werden.

Eine Anforderung an das Konzept ist die dynamische Erweiterbarkeit der Wissensbasis. Neben der vom Anwender gesteuerten Erweiterung wäre auch eine teilautomatisierte Erweiterung denkbar. Das System schlägt automatisiert neue Verbindungen vor und lässt diese vom Anwender bestätigen. Die Vorschläge könnten durch Berechnung der Korrelation von einem Begriff zu bisher nicht zusammenhängenden Begriffen im Kontext einer Analyse in der Wissensbasis erstellt werden. Zusätzlich könnte die Häufigkeit von erfolgreich genutzten Verbindungen als Gewichtung in die Identifikation von Wirkungszusammenhängen einfließen. Dadurch würde bestimmten Verbindungen ein höheres Potenzial als anderen Verbindungen unterstellt. Hierbei stellen auch das systematische Vergessen oder auch Verblassen von Verbindungen innerhalb der Wissensbasis einen interessanten Forschungsbereich dar.

Abschließend muss eine Evaluation der praktischen Anwendbarkeit des Konzepts auf Basis einer prototypischen Implementierung erfolgen. Im Rahmen dieser Evaluation könnte die Abhängigkeit des Identifizierens von Wirkungszusammenhängen vom Umfang der Wissensbasis untersucht und ggf. Aussagen über eine kritische Größe der Wissensbasis gemacht werden.

Literaturverzeichnis

Andersen, B., & Fagerhaug, T. (2006). Root Cause Analysis - Simplified Tools and Techniques. Milwaukee, Wisconsin: ASQ Quality Press.

Beierle, C., & Kern-Isberner, G. (2008). Methoden wissensbasierter Systeme - Grundlagen, Algorithmen, Anwendungen (4., verbesserte Auflage.). Wiesbaden: Vieweg + Teubner.

Caron, E. A. M. (2013). Explanation of Exceptional Values in Multi-dimensional Business Databases. University Rotterdam, Rotterdam.

Caron, E. A. M., & Daniels, H. A. M. (2008). Explanation of exceptional values in multi-dimensional business databases. European Journal of Operational Research, 188(3), 884–897.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165–1188.

Cyert, R. M., Simon, H. A., & Trow, D. B. (1956). Observation of a business decision. The Journal of Business, 29(4), 237–248.

Davenport, T. H. (2006). Competing on analytics. Harvard Business Review, 84(1), 98–107.

Davenport, T. H., & Harris, J. G. (2007). Competing on Analytics: The New Science of Winning. Boston, Massachusetts: Harvard Business School Press.

Davenport, T. H., Harris, J. G., & Morison, A. R. (2010). Analytics at Work: Smarter Decisions, Better Results. Boston: Harvard Business School Press.

Ertel, W. (2009). Grundkurs Künstliche Intelligenz - Eine praxisorientierte Einführung (2., überarbeitete Auflage.). Wiesbaden: Vieweg + Teubner.

Friedrich, G., & Stumptner, M. (1990). Einführung. In G. Gottlob, T. Frühwirth, & W. Horn (Hrsg.), Expertensysteme (S. 1–19). Wien: Springer Verlag.

Gluchowski, P. (2001). Business Intelligence - Konzepte, Technologien und Einsatzbereiche. HMD - Praxis der Wirtschaftsinformatik, (222), 5–15.

Gluchowski, P., Gabriel, R., & Dittmar, C. (2008). Management-Support-Systeme und Business intelligence : computergestützte Informationssysteme für Fach- und Führungskräfte (2., vollständig überarbeitete Auflage.). Berlin; Heidelberg: Springer.

Gorry, G. A., & Morton, M. S. S. (1971). A Framework for Management Information Systems. Sloan Management Review, 13(1), 55–70.

Hitzler, P., Krötzsch, M., Rudolph, S., & Sure, Y. (2008). Semantic Web. Berlin, Heidelberg: Springer Verlag.

Inmon, W. H. (2002). Building the Data Warehouse (Third Edition.). New York, New York: John Wiley & Sons.

Ishikawa, K. (1972). Guide to Quality Control. Tokyo, Japan: Asian Productivity Organization.

Ishikawa, K. (1985). What Is Total Quality Control? The Japanese Way. Englewood Cliffs, New Jersey: Prentice Hall.

Kemper, H.-G., Baars, H., & Mehanna, W. (2010). Business Intelligence - Grundlagen und praktische Anwendungen: Eine Einführung in die IT-basierte Managementunterstützung (3., überarbeitete und erweiterte Auflage.). Wiesbaden: Vieweg + Teubner Verlag.

Laursen, G., & Thorlund, J. (2010). Business Analytics for Managers: Taking Business Intelligence Beyond Reporting. Hoboken, New Jersey: John Wiley & Sons.

Minelli, M., & Dhiraj, A. (2013). Big data, big analytics: emerging business intelligence and analytic trends for today's businesses. Hoboken, New Jersey: John Wiley & Sons.

Mintzberg, H., Raisinghani, D., & Theoret, A. (1976). The structure of „unstructured" decision processes. Administrative science quarterly, 21(2), 246–275.

Morton, M. S. S. (1971). Management Decision Systems - Computer-Based Support for Decision Making (Second Printing.). Boston, Massachusetts: Division of Research, Graduate School of Business Administration, Harvard University.

Newell, A., & Simon, H. A. (1972). Human problem solving (6th printing.). Englewood Cliffs, New Jersey: Prentice-Hall.

Power, D. J. (2002). Decision Support Systems: Concepts and Resources for Managers. Westport, Connecticut: Greenwood Publishing Group.

Sabherwal, R., & Becerra-Fernandez, I. (2011). Business Intelligence: Practices, Technologies and Management. Hoboken, New Jersey: John Wiley & Sons.

Simon, H. A. (1977). The new science of management decision (Revised Edition.). Englewood Cliffs, New Jersey: Prentice-Hall.

Turban, E., Sharda, R., & Delen, D. (2011). Decision support and business intelligence systems. (J. E. Aronson, Hrsg.) (9th Edition.). Upper Saddle River, New Jersey: Pearson Education.

Vercellis, C. (2009). Business intelligence: Data Mining and Optimization for Decision Making. Chichester, United Kingdom: John Wiley & Sons.

Ein adaptiver Ansatz zum Ingest großer Bestände audiovisueller Medien unter heterogenen Anforderungen

Robert Herms, Robert Manthey, Marc Ritter, Maximilian Eibl
Technische Universität Chemnitz
Professur Medieninformatik
D-09107, Chemnitz, Germany
{robert.herms,robert.manthey,marc.ritter,maximilian.eibl}@informatik.tu-chemnitz.de

Zusammenfassung

Adaptive Methoden zur Digitalisierung und Transkodierung von audiovisuellen Medien stellen eine Grundvoraussetzung dar, um den heterogenen Qualitätsanforderungen verschiedener Kooperationspartner gerecht zu werden. Für diese Anwendungsszenarien stellt dieser Beitrag eine flexible Transkodierlösung vor, die in Kombination mit dem Middleware-Framework *Imtecs* als Einspiellösung fungiert, wobei eine etablierte auf großen Datenmengen eruierte Workflow-Prozesskette näher beleuchtet sowie hinsichtlich eines repräsentativen Einsatzszenarios evaluiert wird.

1 Einleitung

In der Produktion, bei Dienstleistern und in Archiven stellt die Handhabung und Verwaltung audiovisueller Medien erhebliche Anforderungen an die Verarbeitungs- und Speicherkapazitäten. Zum einen muss die notwendige Technik und Software vorgehalten und durch qualifiziertes Personal gewartet werden, zum anderen sind gesetzliche Vorgaben einzuhalten und die betreffenden Medien dementsprechend zu handhaben. Zusätzlich dazu stellen der Nutzungszeitraum von mehreren Jahren bis Jahrzehnten sowie die Heterogenität der Workflowanforderungen bei den Anwendern weitere erhebliche Herausforderungen dar. Der Workflow bei Dienstleistern, welche beispielsweise VHS-Kassetten digitalisieren, sollte kostengünstig und schnell den als Ingest bezeichneten Einspielvorgang unterstützen, die Mediendaten zwischenspeichern und geeignet an die Kunden weiterleiten. Demgegenüber liegt der Schwerpunkt bei Archiven im Bereich der qualitativ hochwertigen, verlustfreien und sicheren Speicherung der Mediendaten. [Götzer *et al.*, 2008]

Bedingt durch diese Heterogenität erfordert ein generischer Ansatz eines einheitlichen Workflows eine größtmögliche Flexibilität bei der Verwaltung der daran beteiligten Komponenten und deren Verbindungen zueinander. Dies erfordert beispielsweise sowohl die Anzahl, Typen und Standorte der am Ingest beteiligten Komponenten an veränderte Kostenanforderung anpassen zu können als auch die Qualität und Speicherorte der Daten. Flexible und rekonfigurierbare Verbindungen müssen zwischen vorhandenen Komponenten hinzugefügt, modifiziert oder entfernt werden. Durch diese Möglichkeiten können gegebene Workflows adaptiert und bedarfsgerecht angepasst werden.

2 Archivierungsbedarf audiovisueller Medien

In Europa existieren über 28 Millionen Stunden an audiovisuellem Material, welches vom Verfall bedroht ist und für die zukünftige Nutzung mittels Digitalisierung sowie Annotation erschlossen werden sollte. Dabei existierten bereits im Jahr 2007 ca. 200 Exabyte (1 Exabyte = 10^6 TB) an analogem Video- und Filmmaterial auf alternden Originalträgern. Um diese massive Menge an Daten zu retten, werden Workflows zur Archivierung von digitalem audiovisuellen Material benötigt, welche sowohl automatisierbar als auch finanzierbar sind. [Herla *et al.*, 2010]

Qualitätsstufe	Transferrate (Mbit/s)	Speicherplatzbedarf (GB/h)
Browsen	0.128	0.058
Vorschau	1.5	0.680
Broadcast	4	1.8
Produktion	50	23
Unkomprimiert	270	121.5

Tabelle 1: Eigenschaften unterschiedlicher Qualitätsstufen (nach: [Mauthe and Thomas, 2004])

Existierende Workflowmanagementlösungen für die Verwaltung von Geschäfts- und Verwaltungsabläufen wie kommerzielle Produkte *MS BizTalk*[1] oder *Oracle BPEL Process Manager*[2] unterstützen einerseits nur bestimmte Betriebs- und Datenbanksysteme. Andererseits entstehen gleichzeitig größere finanzielle Belastungen sowie Hersteller- und Technologieabhängigkeiten. Diese Einschränkungen und Bindungen stellen ihrerseits eine erhebliche Behinderung für notwendige Daten-, Format- und Technologiemigrationen und die darauf aufbauende Archivierung dar. [Götzer *et al.*, 2008; Neuroth *et al.*, 2010] Open-Source-Lösungen wie *Activiti*[3] oder *jBPM*[4] reduzieren den Umfang dieser Nachteile durch weitgehende Herstellerunabhängigkeit und geringere Kosten bei Beschaffung und Betrieb. Allerdings ist deren notwendige Anpassung an die speziellen Bedingungen des Ingests, wozu auch das Einspielen großer Datenmengen von bis zu 50 Mbit/s (Tabelle 1), das Sammeln der technischen Randbedingungen unterschiedlichster Hardware, deren Steuerung und die Variabilität des gesamten Einspielworkflows gehören, nur begrenzt realisierbar, so dass eine qualifizierte und für diesen Anwendungsfall spezialisierte Lösung entwickelt werden musste.

[1] http://www.microsoft.com/en-us/biztalk/default.aspx, 10.07.2013
[2] http://www.oracle.com/technetwork/middleware/bpel/overview/index.html, 10.07.2013
[3] http://www.activiti.org, 10.07.2013
[4] http://www.jboss.org/jbpm, 10.07.2013

Datenträger	Kapazität in GB	Datentransferrate in Mbit/sec	Zugriffszeit in ms	mittlerer Preis in EUR/GB
CD	0.8	70 (52x)	50	0.1
Mikrofilm	5	–	–	2
MOD	17	5 (1x)	60	5
DVD	19	80 (8x)	65	0.2
BD	50	288 (8x)	65	0.5
SSD	128	30 – 50	0.2	10
LTO	800	60 – 120	–	0.1 – 0.25
HDD	1500	50 – 100	10	0.5

Tabelle 2: Eigenschaften verschiedener Speichermedien (nach: [Neuroth *et al.*, 2010])

Die kundengegebenen, wirtschaftlichen und rechtlichen Bedingungen beeinflussen die Art der Speicherung der vorhandenen Mediendaten und führen beispielsweise bei Videodaten zu erheblichen Unterschieden in der Größe der Daten und resultieren in den dafür optimierten Speichersystemen (Tabelle 2). Dazu gehören konventionelle Speichermedien von CD über DVD und Bluray (BD) bis hin zu LTO (Linear Tape Open). Letztere bieten gegenüber Festplattenmedien (HDD) Eigenschaften für längerfristige und kostengünstige Speicherung größerer Datenmengen unter der Bedingung, dass auf die Daten nicht allzu oft zugegriffen werden muss, da die Zugriffszeiten des Wiederherstellungsvorgangs ähnlich wie bei Mikrofilmen langwieriger sind.

Weiterhin fallen für die Archivierung in der Regel noch Aufwendungen und Kosten für Personal an, welche durch den Offlinespeicher reduziert werden können. [Klaproth, 2013]

Im nachfolgend vorgestellten Ansatz wurde die notwendige Flexibilität unter Verwendung von Festplatten als Zwischenspeicher und LTO-Laufwerken als Archivierungsmedium exemplarisch umgesetzt und erprobt.

3 Systemarchitektur

Die Gesamtarchitektur für den adaptiven Ansatz zum Ingest audiovisueller Medien setzt sich aus dem Framework *Imtecs* und einer flexiblen Transkodiereinheit zusammen. Beide Lösungen werden in den folgenden Abschnitten hinsichtlich ihrer Funktionalität und Einsatzmöglichkeiten näher erläutert sowie in Kombination zur Realisierung eines adaptiven Ingest-Workflows vorgestellt.

3.1 Imtecs-Framework

Das eigens entwickelte Framework *Imtecs* (Ingest middleware including extraction of metadata from technical constraints) ist eine Open-Source-Lösung für das automatisierte Einspielen audiovisueller Medien in serverbasierte Systeme bei gleichzeitiger Dokumentation jedes Einspieljobs in Form von Metadaten über die technischen Rahmenbedingungen des Einspielens, wozu auch verwendete Hardware zählt.

Die Architektur des *Imtecs*-Framework ist in Abbildung 1 dargestellt. *Imtecs* ist grundsätzlich eine Middleware zur Steuerung der für das Einspielen notwendigen Hardware- und Softwarekomponenten. Die Konfiguration eines Ingest-Workflows erfolgt durch die entsprechende Beschreibung (Workflow Description). Hierfür wurde das XML-basierte *Ingest Workflow Description* Format (IWD) entwickelt, in dem alle anzusteuernden Geräte sowie Dienste und deren Beziehungen zueinander beschrieben werden. [Manthey *et al.*, 2013]

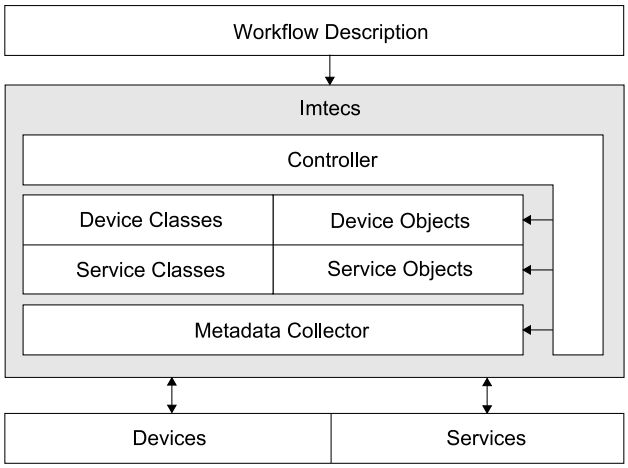


Abbildung 1: Architektur des Imtecs-Frameworks [Manthey *et al.*, 2013]

Dieses Format wird durch Imtecs geladen und initiiert den Workflow. Spezifische Medien wie z.B. VHS erfordern eine entsprechende Workflow-Beschreibung, da nur bestimmte Hardware angesprochen wird. Um auf diese Gegebenheit flexibel einzugehen, können mehrere IWDs im Vorfeld angelegt oder bereits existierende adaptiert werden. Um neue Workflows anzulegen werden generische Workflow-Templates bereitgestellt, um die Konfiguration zu vereinfachen.

In *Imtecs* werden Geräte als *Device Objects* und Dienste als *Service Objects* repräsentiert, wobei entsprechende *Classes* Zusammenfassungen von mindestens einem Object sind. Der *Controller* steuert den Ablauf des Workflows und koordiniert dabei die zeitlichen Abläufe der teilnehmenden Geräte und Dienste, so dass verschiedene audiovisuelle Medien erfolgreich eingespielt werden können. Jedes einzuspielende Medium kann als eine Art Job betrachtet werden, wobei der *Metadata Collector* für jeden Job automatisch die zuvor angegeben formalen Metadaten festhält sowie die Metadaten von den beteiligten Geräten und Diensten sammelt, welche am Einspielprozess beteiligt sind. Weiterhin sind diese Daten in Form einer Wissensbasis bei der Adaption sämtlicher nachfolgender Workflowelemente und -abläufe nutzbar. [Herms *et al.*, 2012]

Das *Imtecs*-Framework fand bereits Anwendung in frühen Testphasen zur Archivierung audiovisueller Medien, um einerseits frühzeitig die Probleme der automatisierten Digitalisierung zu identifizieren und andererseits das aufkommende Datenvolumen verschiedener Qualitätsstufen abschätzen zu können. Eine weitere Testphase befasste sich in der Kombination mit dem Analyse-Framework AMOPA (Automated Moving Picture Annotator) [Knauf *et al.*, 2011; Ritter and Eibl, 2011; 2009], zur automatisierten inhaltsbasierten Annotation der eingespielten Medien [Ritter *et al.*, 2013]. Aktuell wurden die eingesetzten Technologien hinsichtlich eines Langzeittests von vier Wochen im unterbrechungsfreien Betrieb rund um die Uhr sowie bei der Verarbeitung großer Datenmengen evaluiert, wobei 500 Stunden Videomaterial erfolgreich erfasst, gespeichert und analysiert werden konnten.

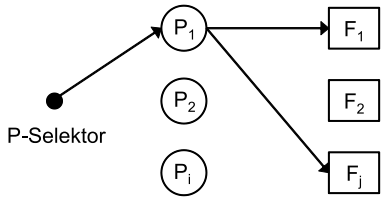


Abbildung 2: Flexibles Transkodieren durch Selektion von Kooperationspartner P_i zugeordneten Formaten F_j

3.2 Flexibles Transkodieren

Die im Rahmen dieser Arbeit entwickelte Transkodier-Lösung nutzt FFmpeg[5] für das Transkodieren von audiovisuellen Medien unter der Berücksichtigung der Formatauswahl und den Qualitätsanforderungen eines Kooperationspartners, dem diese Dienstleistung zur Verfügung gestellt wird. Wie in Abbildung 2 zu sehen ist, werden die von jedem Kooperationspartner P_i präferierten und daran angepassten Formate F_j zugewiesen.

Diese Informationen müssen in irgendeiner Form, beispielsweise in einer Datenbank, hinterlegt werden, damit das Transkodier-System bei einer Eingangsgröße P_i, welche aus den formalen Metadaten von *Imtecs* oder einem anderen System entstammen, die entsprechenden Zielformate für die Transkodierung laden kann. Die Datenbankeinträge sollten nach Möglichkeit modifizierbar sein, da sich die Anforderungen eines Kooperationspartners zu einem späteren Zeitpunkt ändern können.

Diese Vorgehensweise ermöglicht einerseits auf bestimmte Codec- und Containerformate einzugehen, andererseits spezielle Parameter wie die Bitrate aufgrund von mangelnder Speicherkapazität zu justieren. Da zudem jedem Partner mehrere Formate zuordenbar sind, ergibt sich eine auf die individuellen Bedürfnisse zugeschnittene Konfiguration, durch deren Abbildung die Anzahl und Art der durchzuführenden Transkodierungen eindeutig festgelegt ist.

3.3 Prozesskette des Gesamtsystems

Aufgrund der Vielzahl potenzieller Kooperationspartner mit jeweils eigenen Eingangs- und Zielformaten sowie individuellen Anforderungen an Kostenprofile und darauf aufbauende Speichermengen, ergibt sich eine schwer vorhersehbare Menge zu realisierender Prozessabläufe für den Gesamtprozess. Dieser Umfang stellt seinerseits ein beträchtliches Problem in Bezug auf Wartbarkeit und Kosteneffizienz dar, weshalb eine Vereinfachung basierend auf den Prinzipien des vorherigen Abschnitts erweitert und auf den Gesamtprozess angewendet und in Abbildung 3 dargestellt wird.

Der Gesamtprozess erfordert eine Wissensbasis über die einzuspielenden Medien als auch die Zielformate eines Kooperationspartners. Hierzu wird zu Beginn eine Anforderungsanalyse vorgenommen, die ausschlaggebend für die Automatisierung der folgenden Prozesse und die Ergebnisse der Kooperationspartner sind.

Die Anforderungsanalyse befasst sich insbesondere mit der Ermittlung der Art und Menge der einzuspielenden Medien, deren Zielformate und Qualität. Hierbei findet zudem eine Aufschlüsselung der Aufwendungen für die aufzubringenden Speicherkapazitäten statt; eine Anpassung des Vorhabens ist dabei unter Umständen abzuwägen. Das Resultat

[5]http://www.ffmpeg.org, 10.07.2013

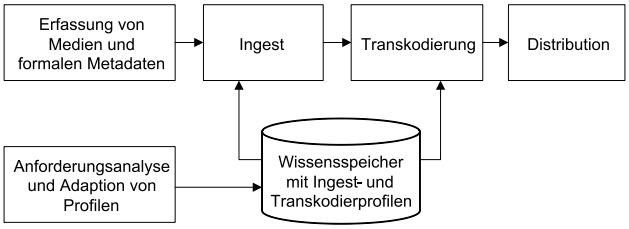


Abbildung 3: Prozesskette des adaptiven Workflows mit anforderungsabhängigen Profilen für den Einspiel- und Transkodiervorgang

der Analyse sind Profile, die sich einerseits auf das Einspielen beziehen, damit der Ingest-Workflow auf die Eingangsformate justiert ist. Andererseits werden Transkodierprofile angelegt, welche jeweils die Zielformate eines Kooperationspartners beinhalten. Ändert sich das Vorhaben eines Kooperationspartners, so können sowohl Ingest- als auch Transkodierprofile adaptiert oder weitere Profile hinzugefügt werden. Ingestprofile liegen in Form von IWD-Formaten vor, die Transkodierprofile sind Datenbankeinträge.

Noch vor dem Ingest werden die einzuspielenden Medien sowie Informationen über diese und den Kooperationspartner in Form formaler Metadaten erfasst und an *Imtecs* übergeben. Um den Workflow automatisiert konfigurieren zu können, wird daraufhin das entsprechende Ingestprofil (IWD) anhand der formalen Metadaten aus dem Wissensspeicher geladen. Anschließend startet der Ingest-Workflow, wobei zunächst qualitativ weitgehend verlustfreie Videodaten erstellt werden, die den Input für den nachfolgenden Transkodierprozess bilden. Zusätzlich werden die über den Einspielprozess sowie die zu Beginn festgelegten formalen Metadaten abgespeichert.

Der nächste Prozess ist die Transkodierung. Aus den Metadaten ist ersichtlich von welchem Kooperationspartner die einzuspielenden Medien stammen. Diese Kenngröße wird als Basis verwendet, um automatisiert das entsprechende Transkodierprofil aus dem Wissensspeicher zu laden und die Transkodierung der zuvor eingespielten Videodaten in die Zielformate durchzuführen. Die Ergebnisse werden in Ausgangsverzeichnisse abgelegt und stehen zur weiteren Verarbeitung zur Verfügung. Die Datensätze der Kooperationspartner sind somit für die Speicherung in Datenbanken oder auf LTO-Bändern zur Archivierung vorbereitet. Die resultierten Videoformate und Metadaten können über den Distributionsschritt zum Kooperationspartner in Abhängigkeit der Größe der aufkommenden Daten und entsprechenden Aufwendungen individuell übermittelt werden.

4 Evaluation

In diesem Abschnitt wird ein repräsentatives Szenario für den adaptiven Ansatz zum Ingest von audiovisuellen Medien unter heterogenen Anforderungen eines Kooperationspartners aufgestellt. Ziel dieser Evaluation ist die Auswertung des gesamten Ingest-Zeitraumes, der aufkommenden Speichergröße je Qualitätsanforderung und einer Kostenabschätzung der sich hieraus ergebenden Speichergröße, welche dabei aufgebracht werden muss.

Das Szenario umfasst den Ingest von insgesamt 444 S-VHS-Kassetten, deren Bandlänge zwischen 30 und 240 Minuten variieren und die inhaltlich sowohl Nachrichtenbeiträge, Sportsendungen, Dokumentationen und Werbe-

Format	F1 (Analyse)	F2 (Preview)	F3 (Prod.)	F4 (Archiv)
Container-format	MP4	MPEG	AVI	MXF (OP1a)
Video-codec	h.264	h.262	DV	IMX 50
Auflösung	720×576			
Chroma YUV	4:2:0			4:2:2
Bildrate	25 fps			
Bitrate (kBit/s)	1.998	8.750	28.812	50.000
Audio-codec	AAC	MP3	PCM S16 LE	
Abtastrate	48 kHz			
Kanäle	2			
Bitrate (kBit/s)	160	128	1.536	12.300

Tabelle 3: Anforderungskatalog der eingespielte S-VHS-Kassetten für verschiedene Anwendungsfälle von Analyse und Preview über Produktion bis hin zur dauerhaften Archivierung.

beiträge zeigen. Aus dem Anforderungskatalog des Kooperationspartners (Tabelle 3) ergeben sich die vier Zielformate F_1, F_2, F_3 und F_4, die in unterschiedlichen Szenarien Einsatz finden.

Um den speziellen Anforderungen des Inhaltes an die Analyse Rechnung zu tragen, wurde das dafür optimierte Format F_1 zur Verwendung durch das Analyse-Frameworks AMOPA definiert. Das Format F_2 stellt eine erheblich größere Version für den Preview-Prozess in der Redaktion des Kooperationspartners dar und wurde auf die dort eingesetzte Software ausgerichtet. Für die weitere Produktion steht F_3 als qualitativ hochwertigstes und größtes Einsatzformat zur Verfügung. Zur Archivierung und als Sicherheitskopie ist das Format F_4 vorgesehen.

4.1 Testaufbau und Durchführung

Für die Umsetzung des Szenarios wurde das zuvor vorgestellte System basierend auf *Imtecs* und der flexiblen Transkodierung mit Hilfe entsprechender Hardware eingesetzt. Der Testaufbau für dieses Szenario umfasste zwei professionelle S-VHS-Player, ein selbstkonstruiertes Magazin mit einer Kapazität von zehn S-VHS-Kassetten sowie einen Laderoboter basierend auf der Lego-NXT-Technologie, zwei Analog-Digital-Wandler und zwei Hardware-Netzwerk-Enkoder. Die S-VHS-Kassetten wurden, damit die Digitalisate zu einem späteren Zeitpunkt eindeutig identifizierbar sind, jeweils mit einem eigenen QR-Code markiert, der eine zu ISAN[6] analoge Identifizierungsnummer codiert. Sobald der Einspielprozess einer Kassette startet, wird dieser Code ausgelesen und den resultierenden Zielformaten zugeordnet. Da die Kapazität des Kassettenmagazins auf zehn Einheiten beschränkt ist, musste dieses fünf mal innerhalb einer Woche gefüllt werden.

[6]http://www.isan.org, 14.07.2013

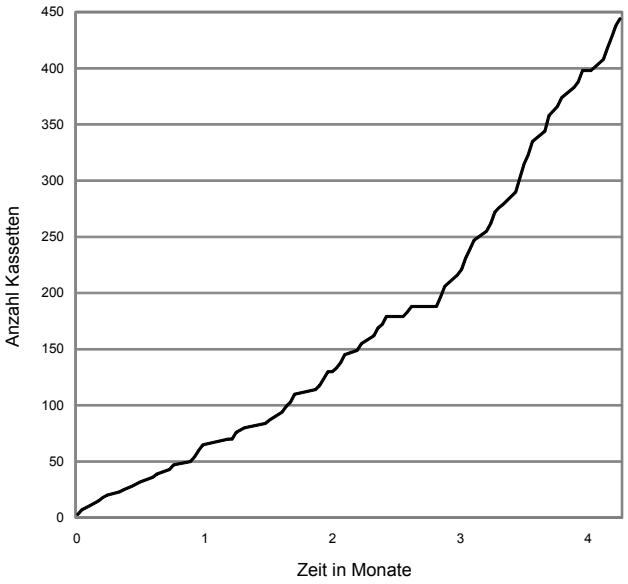


Abbildung 4: Übersicht über den Zeitraum in Monaten, in dem 444 S-VHS-Kassetten digitalisiert und transkodiert wurden.

In Abbildung 4 wird illustriert, wieviel S-VHS-Kassetten über den Zeitraum von 130 Tagen eingespielt wurden. Dabei lässt sich nach zweieinhalb Monaten eine signifikante Wende verzeichnen. Bis zu diesem Zeitpunkt wurden im Schnitt 66 Kassetten pro Monat eingespielt, in der nachfolgenden Zeit konnte eine Menge von 150 Kassetten pro Monat gezählt werden. Dieser massive Umbruch ist auf effizienteres Management des temporären Speichers zurückzuführen, was zu weniger Unterbrechungen des Einspielens führte. Vor der LTO-Abspeicherung wurde dann eine automatische Verteilung der Daten auf verschiedene Zwischenspeicher vorgenommen, weshalb seitdem nahezu unterbrechungsfrei eingespielt werden konnte.

4.2 Speicher- und Kostenabschätzung

Für einen Kooperationspartner sind die aus dem Einspielszenario resultierenden Speichergrößen und die damit verbunden Kosten von besonderer Bedeutung. Tabelle 4 gibt einen Überblick über die Formate, welche zuvor identifiziert wurden, und die entsprechende Gesamtspeichergröße für das Szenario mit einer Gesamtdauer von 1.450 Stunden. Ebenfalls wird eine Abschätzung der Speichergröße für eine Stunde Videomaterial getroffen.

Die Speicherkapazität, die in diesem Anwendungsfall für das Archivformat (F_4) zu erbringen ist, beläuft sich auf 40,4 TB und wird in der Regel Near- oder Offline gespeichert, wobei LTO-5-Bänder mit einer (unkomprimierten) Kapazität von 1,5 TB verwendet werden. Aus Sicht eines Archivars ist dementsprechend die Frage nach den Kosten interessant. Die Kosten der Gesamtspeicherkapazität eines spezifischen digitalen Videoformates K_i wird bestimmt durch die Gesamtspeichergröße des Videoformates s_i, die Größe der Kapazität des zu verwendenden Speichermediums g und eines Kostenfaktors k für das Speichermedium:

$$K_i = \left\lceil \frac{s_i}{g} \right\rceil \cdot k \qquad (1)$$

Beispielhaft lassen sich die anfallenden Archivierungskosten aus einem Speicherbedarf von $s_i = 40,4$ TB, einer

Format	Speicher, 1 h	Speicher, 1.450 h
F_1	150 MB	217,5 GB
F_2	1 GB	1,4 TB
F_3	14 GB	20,3 TB
F_4	28,5 GB	40,4 TB

Tabelle 4: Übersicht über Formate und Speichergrößen je Stunde (links) und für das gesamte Szenario (rechts).

Speicherkapazität von $g = 1,5$ TB und einem fiktiven Kostenfaktor von $k = 50$ EUR ermitteln. Allein für das Szenario der Langzeitarchivierung (F_4) ergeben sich somit die Kosten von $K_4 = 1.350$ EUR.

Basierend auf dem hierbei verwendeten Archivierungsformat und dem gegebenen Kostenfaktor k für ein LTO-5-Band, lässt sich ein Kostenpunkt abschätzen. Da der Datenaustausch aller Formate zwischen Dienstleister und Kooperationspartner durch LTO-5-Bänder vorgenommen wird, ergeben sich die Gesamtkosten K_G aus der Summe aller Teilkosten:

$$K_G = \sum_{i=1}^{n} K_i \tag{2}$$

Die gesamte Speicherkapazität aller Formate, die in diesem Anwendungsszenario und über alle Qualitatsstufen zu erbringen war, umfasst 62,3 TB. Dies entspricht 42 LTO-Bänder und unter Einbezug des obigen Kostenfaktors die Gesamtkosten $K_G = 2.100$ EUR. Mit Hilfe des Kostenmodells und des damit in Relation stehenden Speicherbedarfs kann somit auch der Rückschluss auf Qualität und Anzahl verschiedener Zielformate ermöglicht werden.

5 Zusammenfassung und Ausblick

In Archiven mit audiovisuellen Inhalten sind Medien zum Teil unerschlossen und in analoger Form vorzufinden. Abhilfe schaffen hierbei Archivierungs-Workflows, um die Medien in serverbasierte Landschaften zu überführen indem sie digitalisiert und annotiert werden. Der kombinierte Einsatz der beiden Komponenten *Imtecs* und einer flexiblen Transkodierlösung ermöglicht für bestimmte Qualitätsanforderungen einen flexiblen Ansatz für die Digitalisierung von audiovisuellen Medien und die Transformation in bestimmte Zielformate. Mittels *Imtecs* erfolgt die automatisierte Steuerung von Hardware- und Softwarekomponenten für das parallele Einspielen der Medien, wobei der Workflow als solches konfigurierbar ist. Die in der Prozesskette folgende Transkodiereinheit ermöglicht die Selektion der Zielformate, welche im Vorfeld einem bestimmten Kooperationspartner zugeordnet werden. Somit können wiederkehrende Aufgaben im Wechsel und ohne erneutes Konfigurieren effizient abgearbeitet werden.

Das Gesamtsystem wurde anhand eines Einsatzszenarios evaluiert, bei dem die Anforderungen eines Kooperationspartners identifiziert und das Zielvorhaben angestrebt wurde. Das Szenario umfasst den Ingest von 444 S-VHS-Videokassetten und vier unterschiedliche filebasierte Zielformate. Die Machbarkeit des Vorhabens konnte nach etwas mehr als vier Monaten nachgewiesen werden. Der Ingest ist weiter optimierbar, wenn die Anzahl von Videoplayern höher skaliert ist und das Speichermanagement bezüglich Schreib- und Lesegeschwindigkeit sowie Kapazität angepasst wird.

Aus Sicht eines Archivars sind die entstehenden Kosten für das aufkommende Speichervolumen und der Speicher-Verwaltungsaufwand von besonderer Bedeutung. Zur Orientierung konnte anhand des in dieser Arbeit untersuchten Szenarios eine Aussage über die Kosten im Zusammenhang mit der ermittelten notwendigen Gesamtspeichergröße getroffen werden. Ebenso ist es möglich, Qualitätsstufen und Anzahl der Zielformate in Abhängigkeit der Kosten umzusetzen.

Zukünftige Erweiterungen umfassen einerseits die flexible Integration in andere Produktions- und Archivworkflows in Form von cloud-basierten Systemen. Andererseits wird das System um zusätzliche Dienste zur automatischen und manuellen inhaltlichen Annotation ergänzt, um den Archivierungsaufwand geeignet zu reduzieren und Möglichkeiten für späteres Retrieval zu integrieren.

Danksagung

Diese Arbeit wurde teilweise im Projekt *ValidAX - Validation der Frameworks AMOPA & XTRIEVAL* (Projekt VIP0044) erstellt, das vom *Bundesministrium für Bildung und Forschung* (BMBF) gefördert wird.

Literatur

[Götzer *et al.*, 2008] Klaus Götzer, Ralf Schmale, Berthold Maier, and Torsten Komke. *Dokumentenmanagement - Informationen im Unternehmen effizient nutzen*. dpunkt.verlag, Heidelberg, 2008. 4. Auflage, 378 S.

[Herla *et al.*, 2010] Siegfried Herla, Christian Fey, Birgit Schröter, and Gerhard Stoll. Multimediale Archive der nächsten Generation. In *Fernseh- und Kinotechnik (FKT) 1-2/2010*, 2010. S. 37–41.

[Herms *et al.*, 2012] Robert Herms, Robert Manthey, and Maximilian Eibl. Framework für Ingest mit Annotation technischer Randbedingungen. In *Lernen, Wissen, Adaption; Workshop Information Retrieval 2012 der GI-Fachgruppe Information Retrieval*, Dortmund, 2012. S. 2.

[Klaproth, 2013] Frank Klaproth. DP4lib – Kostenmodell für einen LZA-Dienst, 2013. 28 S., http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-Kostenmodell_eines_LZA-Dienstes_v1.0.pdf, 01.07.2013.

[Knauf *et al.*, 2011] Robert Knauf, Jens Kürsten, Albrecht Kurze, Marc Ritter, Arne Berger, Stephan Heinich, and Maximilian Eibl. Produce. Annotate. Archive. Repurpose – Accelerating the Composition and Metadata Accumulation of TV Content. In *Proceedings of the 2011 ACM International Workshop on Automated media analysis and production for novel TV services (AIEMPro)*, 2011. S. 31–36.

[Manthey *et al.*, 2013] Robert Manthey, Robert Herms, Marc Ritter, Michael Storz, and Maximilian Eibl. A Support Framework for Automated Video and Multimedia Workflows for Production and Archive. In *Proceedings of Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business, 15th International Conference HCII 2013*, Las Vegas, NV, USA, 2013. Springer. S. 336–341.

[Mauthe and Thomas, 2004] Andreas Mauthe and Peter Thomas. *Professional Content Management Systems - Handling Digital Media Assets*. John Wiley & Sons, 2004. 316 S.

[Neuroth *et al.*, 2010] H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, and K. Huth (Hrsg.). *nestor Handbuch - Eine kleine Enzyklopädie der digitalen Langzeitarchivierung Version 2.3*. Verlag Werner Hülsbusch, 2010. 634 S.

[Ritter and Eibl, 2009] Marc Ritter and Maximilian Eibl. Visualizing steps for shot detection. In *Lernen, Wissen, Adaption; Workshop Information Retrieval 2009 der GI-Fachgruppe Information Retrieval*, Darmstadt, 2009. S.98–100.

[Ritter and Eibl, 2011] Marc Ritter and Maximilian Eibl. An Extensible Tool for the Annotation of Videos Using Segmentation and Tracking. In *Proceedings of Design, User Experience, and Usability. Theory, Methods, Tools and Practice - First International Conference, DUXU, Held as Part of HCI International 2011*, Orlando, FL, USA, 2011. S. 295–304.

[Ritter *et al.*, 2013] Marc Ritter, Robert Herms, Robert Manthey, and Maximilian Eibl. Ein ganzheitlicher Ansatz zur Digitalisierung und Extraktion von Metadaten in Videoarchiven. In *Proceedings des 13. Internationalen Symposiums für Informationswissenschaft (ISI 2013)*, Potsdam, 2013. S. 362–371.

Funktionsumfang und Eignung von XML-Datenbanken für Multimedia- und Metadaten

Tom Neumerkel, Robert Manthey
Technische Universität Chemnitz
Fakultät für Informatik
D-09111, Chemnitz, Deutschland
{tomn,mrob}@hrz.tu-chemnitz.de

Zusammenfassung

Diese Arbeit untersucht die Eignung verschiedener Extensible Markup Language (XML)-Datenbanken für die Ablage und Verwaltung von Multimedia- und Metadaten anhand ihrer für diesen Einsatzzweck relevanten und benötigten Eigenschaften sowie durchgeführter Benchmarks.

1 Einleitung

Im Rahmen des Projektes *ValidAX* werden verschiedene Werkzeuge für Analyse, Annotation, Archivierung und Recherche von Audio- und Video-Material[1] entwickelt, welche meist in der betriebssystemunabhängigen Programmiersprache Java implementiert sind und die Handhabung und weitere Verwendung dieser Daten vereinfachen sollen. Hierzu werden Informationen durch Verfahren wie z.B. der Sprach- und Texterkennung aus den Daten extrahiert und als Metadaten in XML-Form in Dateien gespeichert, wie in Quelltext 1 beispielhaft zu sehen, wobei Struktur, Größe, Variabilität und Komplexität dieser Metadaten durch die multimedialen Quelldaten sowie die zu ihrer Analyse eingesetzten Verfahren bestimmt sind.

Aufgrund dessen steigt der Handhabungsaufwand dieser Dateien und die Möglichkeiten damit verbundener Fehler bei größerer Anzahl erheblich an und behindern sowohl einen mehr als prototypischen Einsatz der Werkzeuge, als auch eine Möglichkeit zur Lösung der Problematik durch den Einsatz von Relational Database Management Systems (RDBMSs)[2].

2 Projektspezifische Anforderungen an die Softwarelösung

Bedingt durch dieses Anwendungsszenario scheint eine Datenbanksoftware eine geeignete Lösung darzustellen. Die daran gestellten Anforderungen werden im folgenden Kapitel herausgearbeitet

Essentiell ist die Eigenschaft XML-Daten verwalten zu können, dies umfasst vor allem die folgenden Punkte:

- Nutzerschnittstelle[3] zur Verwaltung der Datenbank an sich
- Nutzerschnittstelle um vorhandene XML-Daten in das System importieren, die Datenbank sichern und wiederherstellen zu können
- Nutzung einer standardisierten Anfragesprache zur einfacheren Handhabung
- Möglichst granulare Rechteverwaltung für einen Mehrbenutzerbetrieb
- Volltextindex, um die Suche zu beschleunigen
- Transaktionsfähigkeit[4] zur Vermeidung störender gegenseitiger Beeinflussungen bei mehreren gleichzeitigen Abfragen
- Sperrmöglichkeiten[5], um Teile des Datenbestandes während der Bearbeitung vor Zugriffen durch andere Nutzer zu sperren

Die Datensätze bestehen aus den Metadaten in XML-Form und Audio- und Videodateien in Formaten wie *mp3*, *mp4* und (unkomprimiertem) *avi* sowie Zusatzdateien wie *iso*. Daher ist die Ablage der Audio- und Videodateien formatunabhängig notwendig. Für verschiedenste Analyse-, Annotations-, Recherche-, Weiterverarbeitungs- und auch Archivierungszwecke müssen aus dem selben Quelldatensatz unterschiedliche Versionen erzeugt, vorgehalten sowie verwaltet werden, woraus weitere wichtige Anforderungen an die Handhabung dieser Daten folgen. Speziell vor dem Hintergrund der damit einhergehenden erheblichen Datenmengen, deren Verteilung zu Zwecken des Lastausgleichs sowie zur Distribution und Weiterverarbeitung ist ein effizientes Management größerer Dateimengen im Bereich von etlichen Gigabyte bis mehreren Terabyte je Datei vorteilhaft, entsprechend den Anforderungen an „Big Data“.

Die Nutzung der verwalteten Daten erfolgt durch andere Programme welche zum einen Metadaten und Dateiversionen für interne Analyse- und Verarbeitungsschritte abfragen und verwalten lassen, zum anderen um Suchanfragen von Nutzern beantworten zu können.

Um die Analyseverfahren zu skalieren und die Werkzeuge verteilen zu können, sowie eine effiziente Resourcennutzung zu ermöglichen, erscheint ein Client-Server-Betrieb vorteilhaft. Sehr resourchenintensive, latenzkritische Werkzeugteile könnten eine direkte Integration in diese erzwingen, sodass ein eingebetteter (embedded) Betrieb der Datenbank möglich sein sollte. Weiterhin erforderlich ist der Einsatz unter verschiedenen Betriebssystemen wie Windows Server, Linux/Unix und Mac OS, sowie Schnittstellen zur Anbindung existierender und zukünftig wahrscheinlicher Werkzeuge mittels Programmiersprachen wie Java-, C#- und PHP: Hypertext Preprocessor (PHP).

1. z.B. im (Jahres-)Programm eines Fernsehsenders
2. Zur Erklärung von RDBMSs siehe [12]
3. Nutzerschnittstelle steht stellvertretend für Command-line interface (CLI) und optional Graphical User Interface (GUI)
4. Modelliert erfolgt die Abarbeitung mehrerer Transaktionen nacheinander, so dass keine Beeinflussung untereinander stattfindet.
5. „*Lock*“ genannt

```
<speech_recognition>
 <file name="TV-20110327-1719-0301.webm.h264.mp4" videoLength="102000">
  <speaker_recogniton>
    <!-- Angaben zum Sprecher -->
  </speaker_recogniton>
  <segment id="0" speaker_id="0" startTime="0" endTime="24000">
   <recognitionEvent id="0" confidence="0.5455173" duration="22710" startTime="60">
    <word startTime="1229" duration="251" confidence="0.6199378">V.</word>
    <word startTime="1481" duration="251" confidence="0.124626793">ein</word>
    <word startTime="1733" duration="352" confidence="0.7838184">hundert</word>
    <word startTime="2086" duration="755" confidence="0.783242166">Sekunden</word>
    <word startTime="2953" duration="574" confidence="0.6497807">und</word>
    <word startTime="3558" duration="705" confidence="0.30492723">Landtagswahl</word>
    <!-- mehr Worte ;) -->
   </recognitionEvent>
   <!-- mehr Events -->
  </segment>
  <segment id="1" speaker_id="0" startTime="24000" endTime="49000">
    <!-- mehr Events -->
  </segment>
    <!-- mehr Segmente -->
 </file>
</speech_recognition>
```

Quelltext 1: Durch Spracherkennung erzeugte XML-Struktur

Um die Handhabung zu vereinfachen sollte in nur einer Anfragesprache nur eine Schnittstelle zur Datenquelle angesprochen werden müssen und die Fähigkeit zum Zugriff auf beliebige Teilstücke von Mediendaten bieten.

3 Anfrage- und Update-Sprachen

Die genannte Vielfalt der bisher bei den Werkzeugen des Projektes verwendeten verschiedenen XML-Konstrukte führt zu Handhabungsproblemen und -fehlern und soll durch die nachfolgend vorgestellte einheitlichere Anfragemethodik weitestgehend beseitigt werden.

Bei XML-Datenbanken gibt es zwar zur Verwaltung der Datenbank keine einheitliche Sprache wie Structured Query Language (SQL) bei RDBMS, aber zum Anlegen, Bearbeiten und Löschen von Datenbanken bzw. der Verwaltung von Berechtigungen dienen SQL-ähnliche Konstrukte. Zur Abfrage und Manipulation von Daten haben sich die Standards XML Path Language (XPath) [22], XML Query Language (XQuery) [23] bzw. XQuery Update Facility (XQUF) [25] des World Wide Web Consortium (W3C) durchgesetzt.

Für alle nachfolgend untersuchten *nativen*[6] XML-Datenbanken haben sich XPath und XQuery als Anfragesprachen etabliert. Beide wurden vom W3C definiert um den Zugriff auf und Umgang mit XML-Daten zu vereinheitlichen. „Die primäre Aufgabe von XPath ist die Adressierung von Knoten innerhalb einer XML-Struktur.“ [22] Zur Formulierung der Anfragen wird XQuery, meist dem For, Let, Where, Order by, Return (FLWOR)-Prinzip folgend verwendet, wie in Quelltext 2 gezeigt.

XQUF erweitert XQuery um die Möglichkeit Daten zu verändern. Es lassen sich so Knoten innerhalb der Struktur einfügen, löschen, verändern und umbenennen, wie in Quelltext 3 dargestellt.

4 Vorstellung einzelner Softwareprodukte

Im folgenden werden die einzelnen betrachteten Softwareprodukte auf ihre Eigenschaften hin untersucht und eine Vorauswahl für die weiteren detailierteren Eignungsanalysen getroffen.

```
<ul>{
for $x in speech_recognition//word
where contains( $x ,"Landtagswahl")
return
  <li>{data($x/ancestor::*[last()]
       /file/@name)}
      -- @ {data($x/../../@startTime)
            + data($x/../@startTime)
            + data($x/@startTime)}
  </li>
}</ul>
```

Quelltext 2: XQuery zur Abfrage des Dateipfades und der Position innerhalb der Video-Datei, anhand von Suchbegriffen. Die Ausgabe erfolgt als HTML-Fragment.

```
for $r in collection('/db/tagesschau')
 /speech_recognition[
   starts-with( ./file/@name, 'TV-')]
return update value $r/file/@name
  with substring($r/file/@name, 4)
```

Quelltext 3: XQuery mit XQUF zum Ändern des Dateinamens der Mediendatei in allen Dokumenten der Collection

6. *Nativ* bezeichnet XML-Datenbanken, die speziell für die Verarbeitung von XML entwickelt wurden, wohingegen *xml-enabled* für Datenbanken Verwendung findet, die um Funktionen zur XML-Verarbeitung erweitert wurden

APACHE XINDICE ist eine der ersten nativen XML-Datenbanken und stammt aus dem Jahr 2002. Da seit dem letzten Release im Dezember 2007 keine Weiterentwicklung stattfand wurde das gesamte Projekt im August 2011 in Apache Attic[7] verschoben und hat damit das Ende seines Lebenszyklus erreicht.

BASEX entstand als Projekt in der Arbeitsgruppe Datenbanken und Informationssysteme des Fachbereichs Informatik und Informationswissenschaft an der Universität Konstanz. Diese Arbeitsgruppe beteiligt sich auch heute an der Weiterentwicklung (vgl. [7] u. [16]). Der Einstieg in die Arbeit mit BaseX wird durch das mitgelieferte GUI erleichtert. Damit lassen sich sofort erste visuelle Erkundungen der Datenbank durchführen und XQuery-Anfragen durchführen. Das Projekt bemüht sich, W3C Standards umzusetzen bzw. diesen eng zu folgen (vgl. [2]). Zusätzliche Funktionen werden durch XQuery Module bereitgestellt, z.B. Funktionalitäten um mit ZIP-Archiven umzugehen oder Volltext-Operationen (vgl. [4]).

BERKELEYDB XML ist eine embedded Datenbank, die heute von Oracle betreut wird. Sie ist dual-lizenziert (kommerziell & open-source) [14]. BerkeleyDB XML baut auf BerkeleyDB auf. Damit kann sie Funktionen von dieser, wie „Replikation zur Hochverfügbarkeit“ [15], nutzen. Die letzte Version (2.5.16) wurde am 22. Dezember 2009 veröffentlicht. Eine Nutzerverwaltung innerhalb der Datenbank ist nicht vorgesehen. Die XML-Daten werden in sogenannten „Containern“ zusammen mit ihren Meta-Informationen, z.B. Indizes, gespeichert. Optional ist eine Validierung der XML-Daten gegen ein Schema[8] möglich (vgl. [11]). Zur Wartung und Verwaltung von Containern ist zusätzlich ein Kommandozeilen-Werkzeug enthalten, das die gleichen Operationen wie die Bibliotheken unterstützt.

EXIST-DB wurde als native XML-Datenbank entwickelt. In der aktuellen Version 2.0 wurde sie zu einer „all-in-one Lösung für die Anwendungsentwicklung“ [6] ausgebaut. Die Verwaltungsoberfläche und die Online-Dokumentation stellen selbst eine Demonstration der Einsatzmöglichkeiten dar. Die Datenbank kann unter anderem als eigenständiger Server oder embedded betrieben werden (vgl. [5]). Die Installation und der Betrieb kann innerhalb eines unprivilegierten Nutzerkontos erfolgen. eXide, ein XQuery-Integrated Development Environment (IDE), kann als Modul in der Weboberfläche genutzt werden. Diese bietet die Möglichkeit, direkt neue Apps[9] entwickeln zu können, die sich dann zu Paketen bündeln und in andere Installationen übertragen lassen. Dokumente können in Collections und Sub-Collections zusammengefasst werden. Locking funktioniert auf Dokumentenebene. Nutzerberechtigungen sind denen in Unix-Dateisystemen nachempfunden. Wartungsaufgaben werden entweder über die Weboberfläche, Apache Ant [1] oder über ein mitgeliefertes Werkzeug, dem „Java Admin Client“ (GUI und CLI) erledigt (vgl. [5]). Wie bei BaseX gibt es auch bei eXist-db XQuery-Module, die den Funktionsumfang erweitern (vgl. [5]). Volltextindex und Caches sorgen auch hier für bessere Leistungen, Indizes müssen manuell angelegt werden.

IBM DB2 ist ein kommerzelles RDBMS der Firma IBM, welches in Version 1 1983 erschien und für diese Arbeit ab Version 9 (2006) interessant ist, da diese seither als hybride Datenbank sowohl mit relationalen als auch mit XML-Daten umgehen kann. IBM nennt die Erweiterung „pureXML“ (vgl. [3]). DB2 speichert die XML-Daten innerhalb von Tabellenspalten. Der Zugriff auf diese ist durch erweiterte XQuery-Funktionen wie *db2-fn:xmlcolumn* und *db2-fn:sqlquery* möglich. Den (kostenlosen) Einstieg bildet „DB2 Express-C“. Diese Edition bietet alle Kernfunktionen (außer Skalierung) zum Betrieb von DB2 auf einem Server in kleinen Umgebungen. Damit entwickelte Anwendungen sollen sich ohne Änderungen mit skalierbaren Editionen nutzen lassen. (vgl. [9])

MONETDB bezeichnet sich selbst als „The column-store pioneer“, die erste open-source Version (MonetDB 4) wurde 2004 veröffentlicht (vgl. [8]). Spaltenspeicher (engl. column-store) bezieht sich hierbei auf die interne Datenorganisation, die nicht, wie bei vielen anderen RDBMS üblich, zeilenorientiert sondern spaltenorientiert erfolgt. Wie bei anderen RDBMS sind Abfragen in SQL zu formulieren. Um Abfrageergebnisse als XML auszugeben bzw. XML-Daten zu importieren sollte bevorzugt „SQL/XML“[10] verwendet werden (vgl. [8]). Somit zählt MonetDB zu den xml-enabled Datenbanken. Eine XQuery-Erweiterung ist verfügbar. Damit wäre MonetDB äußerlich als native XML-Datenbank ansprechbar. Die Entwicklung dieser Erweiterung ist aber seit Mai 2011 „eingefroren“ und in der Dokumentation als „veraltet“ gekennzeichnet, womit eine zukünftige Verwendung für XML-Daten fraglich erscheint (vgl. [8]).

SEDNA, eine freie native XML-Datenbank, entstand als Projekt des „Instituts für Systemprogrammierung“ [10] an der Russischen Akademie der Wissenschaften. Im Gegensatz zu BaseX und eXist-db nutzt Sedna nicht XQUF sondern eine ähnliche Syntax, die auf die Diplomarbeit von Patrick Lehti [13] zurückgeht. Konsistente Backups der Datenbank können während des normalen Betriebs erfolgen (vgl. [19]). Optional lassen sich Nutzer und Berechtigungen je Datenbank aktivieren. Hervorzuheben sind einzelne Berechtigungen für *INSERT*, *DELETE* und *RENAME*, die eine feinere Unterscheidung zulassen als ein *WRITE* bei anderen Produkten. Angewandt werden diese auf Dokumente und Collections. Zur Verwaltung der Datenbank stehen weitere Rechte zur Verfügung (vgl. [19]).

WEBMETHODS TAMINO XML SERVER (nachfolgend kurz Tamino) ist ein proprietäres Produkt der Software AG. Viele Informationen lassen sich über Tamino nicht zusammentragen, da keine Dokumentation ohne Registrierung zugänglich ist. Die zuverlässigste Quelle bildet noch das Datenblatt [21]. Demnach wird für alle Vorgänge rund um ein XML-Dokument XQuery genutzt, ob die Erweiterung XQUF zum Einsatz kommt bleibt offen. Ebenfalls werden dort Fähigkeiten zur Replikation und Hochverfügbarkeit genannt.

5 Eignungsanalyse

Nachfolgend werden die Produkte aus Abschnitt 4 miteinander verglichen. Einen schnellen Überblick der Merkmale bietet Tabelle 1.

Die Anforderung „Big Data“ (Abschnitt 2) ist nur durch IBM DB2 ansatzweise erfüllbar. DB2 bietet zumindest den Konfigurationsparameter „*GetDataLobNoTotal*“. Dessen Wert gibt an, wie viele Bytes eines (Spalten-)Datums an den Client übertragen werden. Der Client kann dann

7. Apache Attic ist ein Archiv zur Verwaltung von Projekte der Apache Foundation, die keine Verwalter bzw. Mitwirkende haben.
8. XML-Schema ist ein W3C-Standard, er dient zur Definition des Aufbaus von XML-Dateien.
9. Modische Kurzbezeichnung für Anwendung
10. SQL/XML ist ISO/IEC Standard (ISO/IEC 9075-14:2011)

weitere „Einheiten“ dieser Größe nachladen. Eine Möglichkeit, eine bestimmte Anzahl Bytes einer Datei (eines Binär-Datums) ab einer bestimmten Position zu laden bietet keines der betrachteten Produkte, sodass diese Anforderung in den weiteren Betrachtungen nicht weiter aufgeführt wird und auf andere Weise gelöst werden muss.

Möglichkeiten der Skalierung über die Grenzen eines Servers hinaus, bieten vor allem die kommerziellen Produkte (IBM DB2 und Tamino). monetDB sieht den Betrieb mehrerer Server innerhalb einer Gruppe vor, die dazu dient, Daten zu verteilen oder zu replizieren. Die Dokumentation dazu ist allerdings äußerst knapp gehalten (vgl. [8]). Etwas ausführlicher wird die Replikation bei eXist-db beschrieben (vgl. [5]).

Die Nutzer- und Berechtigungsverwaltung der Datenbanken ist unterschiedlich ausgeprägt. BerkeleyDB XML verwaltet selbst keine Rechte oder Nutzer. BaseX kennt Nutzer, die eine Datenbank lesen oder auch ändern können bzw. das Recht besitzen neue anzulegen. eXist-db, IBM DB2, monetDB, Sedna, Tamino ermöglichen die Gruppierung von Nutzern [11]. Berechtigungen lassen sich dann auf die Datenbank, eine Collection oder einzelne Dokumente anwenden. eXist-db, IBM DB2 und Tamino können Nutzer auch gegen ein externes System (wie zum Beispiel LDAP oder Kerberos) authentifizieren

Mehrere Anfragen können alle Produkte sicher in einer Transaktion kapseln. Ein explizites Sperren von Dokumenten unterstützen nur eXist-db und IBM DB2.

Um sich von anderen Produkten abzugrenzen bieten sie jeweils zusätzliche Funktionen: eXist-db und Tamino können Dokumente automatisch versionieren, dadurch bleiben ältere Revisionen der Dokumente weiterhin erreichbar, Änderungen am Datenbestand lassen sich nachvollziehen und falls notwendig rückgängig machen. Bei BaseX und eXist-db ist für den Zugriff auf die Dokumente auch Web Distributed Authoring and Versioning (WebDAV)[12] bzw. ein RESTful-API[13] nutzbar.

6 Praxistest

BaseX, eXist-db und Sedna bieten Werkzeuge um ohne weitere Programmierung Daten importieren und Abfragen ausführen zu können. Außerdem nutzen sie mit XQuery for Java (XQJ) [24] ein einheitliches Application Programming Interface (API) für die Integration in Java-Anwendungen. Somit lasen sich diese Produkte gegeneinander austauschen. BerkeleyDB XML, IBM DB2, MonetDB, und Tamino gehen hier eigene Wege. Da die Installation und Konfiguration von IBM DB2 zu umfangreich für den Rahmen dieser Untersuchung ist, der Fortbestand des XQuery-Frontend von MonetDB in Zukunft fraglich erscheint und eine Testversion von Tamino sich nur nach vorheriger Registrierung beschaffen lässt, werden diese nachfolgend nicht weiter betrachtet.

Für die Praxistests wurden die Produkte innerhalb einer Virtuelle Maschine (VM) innerhalb von Oracle VirtualBox [20] installiert, da es selbst auf allen genannten Plattformen aus Abschnitt 2 nutzbar ist. Außerdem kann die VM einfach ex- und importiert werden, was den Transfer auf andere physische Hardware vereinfacht. Innerhalb der virtuellen Umgebung diente die Linux Distribution Ubuntu in Version 12.04 Long Term Support (LTS) als Betriebssystem, da alle getesteten Produkte auf diesem installierbar waren und keine Lizenzverletzungen beim Transfer der VM zu anderen Personen zu befürchten sind.

Die physische Grundlage bildete ein Rechner mit Core™ i5-Prozessor von Intel® (i5-3320M) mit 16GB Arbeitsspeicher und 250GB Solid-state Drive (SSD). Der VM standen 8GB Arbeitsspeicher und 20GB Speicherplatz exklusiv zur Verfügung. Die Installation von BaseX, eXist-db und Sedna verlief problemlos: BaseX über die Paketverwaltung von Ubuntu, eXist-db und sedna über den jeweiligen Installer.

BerkeleyDB XML steht nur als Quelltext zur Verfügung, dessen kompilieren weder auf dem genannten Test-System noch auch auf einem anderen Linux-System möglich war. Als mögliche Ursache hierfür wird die fehlende Weiterentwicklung und Anpassung an aktuelle Betriebssystem und Compiler-Versionen vermutet. Somit wird BerkeleyDB XML nicht weiter betrachtet.

Sedna disqualifizierte sich, da das mitgelieferte Werkzeug zur Verwaltung der Datenbank (im Release 3.5.161) unbenutzbar ist: Es konnten keine Daten importiert werden, dokumentierte Befehle wurden zurückgewiesen.

6.1 Benchmarks

Für die Durchführung von Benchmarks blieben BaseX und eXist-db übrig. Um später Suchanfragen auszuführen, wurden zuerst die Testdaten (Spracherkennung aus 1008 Sendungen der „Tagesschau in 100 Sekunden“) in die beiden Systeme importiert, um die folgenden Anfragen auf diese anzuwenden. Die Abfrage aus Quelltext 2 dient als Beispiel für typische Recherchen. Eine weitere Abfrage, Quelltext 4, kann als Grundlage weiterer Auswertungen dienen, sie zählt die Häufigkeiten der erkannten Wörter innerhalb eines bestimmten Monats. Quelltext 3 dient zum Vergleich bei Änderungen. Alle Anfragen wurden je auf 10 identische Collections ausgeführt.

Um äußere Einflüsse zu vermeiden wurden das Host-System, innerhalb dessen die VM ausgeführt wird, und die VM selbst vor Beginn der Benchmarks neu gebootet und Cron-Jobs (sowie weitere Dienste der Benutzer-Oberfläche, z.B. Update-Checks) deaktiviert. Auch zwischen den Benchmarks von BaseX und eXist-db erfolgte ein Neustart der VM.

Das Diagramm in Abbildung 1 zeigt die durchschnittliche Ausführungszeit der jeweiligen Aktion anschaulich im Vergleich. Die Tabellen 2 und 3 zeigen die Messergebnisse im Detail.

6.2 Auswertung

Der Import der Daten und explizite Indexaufbau erfolgt bei BaseX unauffällig. eXist-db benötigt beim mehrmaligen Import der selben XML-Daten in die selbe Datenbank (Collection) zusätzliche Zeit (siehe Tabelle 3, Spalte Import^b), da sie die Dokumente nicht hinzufügt, sondern ersetzt. Die Daten für die Indexierung bleiben unverändert, der Indexaufbau erfolgt ab dem dritten Lauf (siehe Tabelle 3, Spalte Index^b) etwas schneller.

Quelltext 3 wird bei mehrmaliger Ausführung auf den selben Datenbestand (siehe Tabelle 3, Spalte Q 3^b) von beiden Systemen ab dem zweiten Lauf optimiert. Bei eXist-db

11. Nutzer werden Rollen, Schemata oder Gruppen zugeordnet
12. WebDAV ist ein Protokoll, um per HTTP auf Dokumente zuzugreifen. Für weitere Informationen siehe [17].
13. Representational state transfer (REST) beschreibt den Zugriff auf Ressourcen (hier (Sub-)Collections und Dokumente) über HTTP-Methoden. Adressiert werden diese Ressourcen über einen Uniform Resource Locator (URL). Für weitere Informationen siehe [18].

	BaseX	BerkeleyDB XML	eXist-db	IBM DB2	MonetDB	Sedna	Tamino
Implementiert in	Java	C++	Java	?	C/C++	C/C++	?
Server	✔	×	✔	✔	✔	✔	✔
Embedded	✔	✔	✔	×	×	×	×
Typ	native	native[e]	native	enabled	enabled[f]	native	native
Open Source	✔	dual	✔	×	✔	✔	×
Nutzerverwaltung	✔, R/W je DB	×	✔	✔	✔	✔	✔
Locking	Je DB[a]	×	Datei, Collection	Zeile	×	×	?
Transaktionen	✔	✔	(✔)[g]	✔	(✔)[h]	✔	✔
Backup & Restore	✔	✔	✔	✔	✔	✔	✔
Volltextindex	✔	✔	✔	✔	✔	✔	✔
Plattform	multi[b]	siehe Bindings	multi[b]	z/OS, multi	multi	multi	multi
Bindings bzw. Clients	Java (XML:DB[i] u. XQJ), C#, PHP[c]	C++, C#, Java, PHP, Perl, Python	Java (XML:DB u. XQJ)	C, C++, COBOL, Java (JDBC u. eigener Client), C#, .NET, PHP, Perl	Java (JDBC), Perl, Python, ODBC, MAPI (PHP)	Java (XML:DB u. XQJ), C, C#, PHP, Python, Ruby, Perl, Delphi	Java (XML:DB), JScript, ActiveX, Perl, .NET
Cluster	×	×	×	✔	✔	×	✔
Replikation	×	✔	✔	✔	✔	×	✔
Big Data	×[d]	×	×[d]	✔	×	×	?
WebDAV	✔	×	✔	×	×	×	×
RESTful API	✔	×	✔	×	✔	×	×
Versionierung	×	×	✔	×	×	×	✔

a Einschränkungen siehe auch: `http://docs.basex.org/wiki/Transaction_Management`
b Da in Java implementiert, ist das Produkt auf allen Systemen lauffähig, für die es eine Java Vitual Machine gibt, z.B. Windows, Linux, Mac OS
c Nur „Socket-Wrapper“, keine Namensraum-Unterstützung
d Kann nur mit kleinen Dateien umgehen, werden im Dateisystem des Servers gespeichert, unterliegen dessen Beschränkungen
e Nutzt BerkeleyDB zur Speicherung der Daten
f siehe Informationen im Abschnitt 4
g Wird nur in Beispielen gezeigt, auf Atomicity, Consistency, Isolation and Durability (ACID)-Konformität wird nicht hingewiesen
h Eingeschränkt, wird durch ein Modul realisiert, nur im SQL-Frontend verfügbar
i Nur embedded nutzbar

Tabelle 1: Die Tabelle zeigt die Unterstützung der Anforderungen durch die Produkte.

Lauf	Import[a]	Import[b]	Index[a]	Index[b]	Q 2[a]	Q 2[b]	Q 4[a]	Q 4[b]	Q 3[a]	Q 3[b]
1	4,296	2,400	1,506	1,132	0,201	0,153	28,234	28,550	0,077	0,083
2	2,833	2,450	1,125	1,123	0,107	0,140	27,848	28,605	0,071	0,007
3	2,518	2,201	1,068	1,164	0,105	0,144	28,277	29,510	0,053	0,011
4	2,416	2,082	1,074	1,134	0,097	0,149	37,804	29,102	0,056	0,016
5	2,385	2,118	1,234	1,201	0,098	0,136	28,477	28,679	0,053	0,009
6	2,629	2,204	1,086	1,147	0,090	0,146	28,675	28,310	0,053	0,008
7	2,267	2,290	1,134	1,122	0,109	0,135	28,684	28,566	0,055	0,010
8	2,452	2,223	1,097	1,138	0,103	0,156	29,190	28,194	0,047	0,018
9	2,371	2,287	1,101	1,156	0,092	0,139	27,731	28,131	0,055	0,011
10	2,410	2,310	1,088	1,126	0,109	0,156	27,817	28,317	0,076	0,009
Durchschnitt	2,658	2,257	1,151	1,144	0,111	0,145	29,274	28,596	0,060	0,018
Median	2,434	2,255	1,099	1,136	0,104	0,145	28,377	28,558	0,055	0,011

a Ausführung der Anfrage auf 10 identische Collections
b Mehrmalige Ausführung auf die selbe Collection

Tabelle 2: Die Tabelle zeigt die detaillierte Ausführungsdauer der Anfragen (Import, Indexeraufbau, Suche aus Quelltext 2, Statistik aus Quelltext 4 und Änderungen aus Quelltext 3) in Sekunden an BaseX.

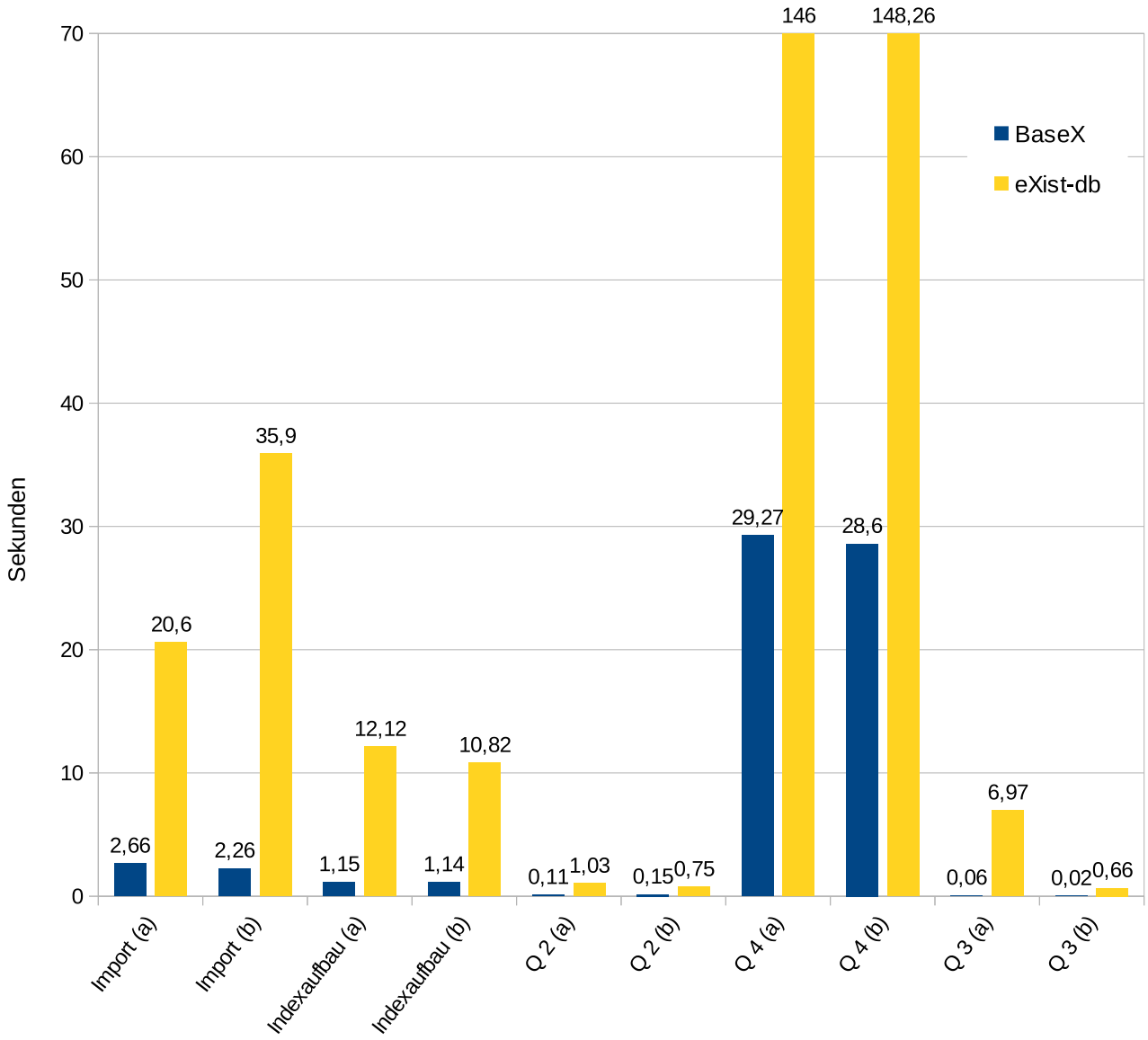


Abbildung 1: Das Diagramm zeigt die Zeit, die für die Ausführung der Abfragen Import, Indexeraufbau, Suche aus Quelltext 2, Statistik aus Quelltext 4 und Änderungen aus Quelltext 3 (a: Ausführung der Anfrage auf 10 identische Collections, b: Mehrmalige Ausführung auf die selbe Collection) benötigt wurde.

Lauf	Import[a]	Import[b]	Index[a]	Index[b]	Q 2[a]	Q 2[b]	Q 4[a]	Q 4[b]	Q 3[a]	Q 3[b]
1	19	21	12,215	12,562	1,687	1,537	146,67	148,33	8,793	6,438
2	21	30	12,155	11,144	0,680	0,632	144,74	147,33	7,952	0,019
3	20	36	12,447	10,451	0,710	0,628	145,25	145,72	6,428	0,022
4	19	27	12,881	10,842	1,453	0,636	146,71	149,59	6,858	0,025
5	21	38	12,621	10,309	0,711	0,758	145,51	148,06	6,793	0,021
6	21	46	11,265	10,842	0,697	0,689	145,11	146,76	6,728	0,024
7	21	34	11,757	10,748	1,459	0,685	145,64	146,68	6,400	0,021
8	20	41	11,781	10,643	0,691	0,622	146,03	147,17	6,665	0,025
9	21	38	11,882	10,390	1,539	0,657	145,36	156,26	6,461	0,019
10	23	48	12,233	10,25	0,719	0,656	149,02	146,69	6,617	0,021
Durchschnitt	21	36	12,123	10,818	1,034	0,750	146,00	148,26	6,970	0,664
Median	21	37	12,185	10,696	0,715	0,657	145,58	147,25	6,697	0,022

a Ausführung der Anfrage auf 10 identische Collections
b Mehrmalige Ausführung auf die selbe Collection

Tabelle 3: Die Tabelle zeigt die detaillierte Ausführungsdauer der Anfragen (Import, Indexeraufbau, Suche aus Quelltext 2, Statistik aus Quelltext 4 und Änderungen aus Quelltext 3) in Sekunden an eXist-db.

```
<ul>{
let $r := collection('/db/tagesschau')
 //speech_recognition[
 starts-with(./file/@name, 'TV-201208')]
for $word in distinct-values($r//word)
let $c := count($r//word[. = $word])
order by $c descending, $word
return <li>{$word} -- {$c}</li>
}</ul>
```

Quelltext 4: XQuery, der alle Begriffe und deren Häufigkeit innerhalb eines bestimmten Monats auflistet (hier August 2012)

hält dies für alle weiteren Durchläufe an, bei BaseX hingegen nur für ein bis zwei.

eXist-db hebt sich durch Versionierung, grundlegende Replikationsmöglichkeiten, Abdeckung der Anforderungen und nicht zuletzt durch eine gute Dokumentation von den anderen Produkten ab. BaseX optimiert die Anfragen vor der Ausführung und ist im Test stets performanter als eXist-db. Vor allem die Anfrage aus Quelltext 4 zeigt, dass die Optimierung der Anfragen, ob durch den Nutzer oder die Datenbank, sehr wichtig ist.

7 Zusammenfassung und Ausblick

Gegenüber eines RDBMS bieten XML-Datenbanken einen großen Gewinn an Flexibilität, da auch nicht exakt identisch strukturierte Daten in einer Collection abgelegt und gleichzeitig durchsucht werden können. XML-Datenbanken eignen sich gut zur Ablage von Metadaten. Um als Archivsystem auch die Multimedia-Daten selbst aufzunehmen, fehlen ihnen allerdings die nötige Funktionalitäten wie z.B. das Ausliefern von Teilstücke von Datei. Außerdem ist die Verwaltung von Berechtigungen, wenn vorhanden, oft auf das Mindeste beschränkt. eXist-db bietet ein gutes Gesamtpaket für erste Entwicklungen. Sie könnte eine gute Grundlage für Webanwendungen zur gemeinsamen Arbeit mit den gespeicherten Informationen sein.

Als Open-Source-Projekte könnten beide an spezielle Anforderungen angepasst werden. Denkbar ist vor allem die Ergänzung der fehlenden Funktionalitäten für Multimedia-Daten.

Die XML-Datenbank könnte sich aus Sicht des Clients auch hinter einer Abstraktionsschicht, einem „Content-Server“ verbergen. Diese Schicht spricht dann zur Recherche und Ablage von Meta-Informationen die Datenbank an, integriert zur Auslieferung der Audio- und Video-Daten aber weitere Komponenten. Wenn diese Schicht mehr Anforderungen übernimmt, ist BaseX eine schlankere und schnellere Alternative zu eXist-db.

Literatur

[1] *Apache Ant Homepage*. Englisch. 2013. URL: http://ant.apache.org/.

[2] *BaseX Homepage*. Englisch. Dez. 2012. URL: http://basex.org/.

[3] R.F. Chong, X. Wang und M. Dang. *Understanding DB2 learning visually with examples*. 2nd ed. IBM Press Pearson plc, 2008. ISBN: 9780768681772. URL: http://proquest.safaribooksonline.com/9780768681772.

[4] *documentation of BaseX*. Englisch. Dez. 2012. URL: http://docs.basex.org/wiki.

[5] *documentation of eXist-db (2.0.x branch)*. Englisch. eXist-db. 2012. URL: http://exist-db.org/exist/apps/doc/documentation.xml.

[6] *eXist-db Homepage*. Englisch. eXist-db. 2012. URL: http://exist-db.org/.

[7] *Homepage der Arbeitsgruppe Datenbanken und Informationssysteme*. Englisch. Arbeitsgruppe Datenbanken und Informationssysteme. 2013. URL: http://dbis.uni-konstanz.de/.

[8] *Homepage of monetDB*. Englisch. MonetDB. 2012. URL: http://www.monetdb.org.

[9] *IBM DB2 Version 10.1 Information Center*. IBM Corporation. 2012. URL: http://pic.dhe.ibm.com/infocenter/db2luw/v10r1/index.jsp.

[10] *Institute for System Programming Homepage*. Englisch. 2013. URL: http://www.ispras.ru/en/.

[11] *Introduction to Berkeley DB XML*. 2009. URL: http://docs.oracle.com/cd/E17276_01/html/intro_xml/BerkeleyDBXML-Intro.pdf.

[12] Georg Lausen. *Datenbanken: Grundlagen und XML-Technologien*. 1. Aufl. München: Elsevier Spektrum Akademischer Verlag, 2005, 281 S. ISBN: 3827414881. URL: http://swbplus.bsz-bw.de/bsz116707615cov.htm.

[13] Patrick Lehti. *Design and Implementation of a Data Manipulation Processor for an XML Query Language*. Aug. 2001. URL: http://www.lehti.de/beruf/diplomarbeit.pdf.

[14] *Oracle Berkeley DB Licensing Information*. Englisch. Oracle. 2013. URL: http://www.oracle.com/technetwork/products/berkeleydb/downloads/licensing-098979.html.

[15] *Oracle Berkeley DB Technology Network Product Page*. Englisch. Oracle. 2013. URL: http://www.oracle.com/technetwork/products/berkeleydb/overview/index.html.

[16] Dimitar Popov. *Advanced Storage Structures for Native XML Databases*. Aug. 2012. URL: http://files.basex.org/publications/Popov%20[2012],%20Advanced%20Storage%20Structures%20for%20Native%20XML%20Databases.pdf.

[17] *RFC4918 - HTTP Extensions for Web Distributed Authoring and Versioning (WebDAV)*. Englisch. IETF. Juni 2007. URL: http://tools.ietf.org/html/rfc4918.

[18] L. Richardson und S. Ruby. *RESTful Web Services*. O'Reilly, 2008. URL: http://books.google.de/books?id=XUaErakHsoAC.

[19] *Sedna Administration Guide*. Englisch. 2012. URL: http://www.sedna.org/adminguide/AdminGuide.html.

[20] *VirtualBox Homepage*. Oracle. 2013. URL: https://www.virtualbox.org/.

[21] *webMethods Tamino XML Server Datenblatt*. deutsch. Software AG. Nov. 2012. URL: http://www.softwareag.com/corporate/images/SAG_TaminoXML_FS_Nov12_Web_tcm16-71285.pdf.

[22] *XML Path Language (XPath) 2.0 (Second Edition)*. Englisch. W3C. Jan. 2011. URL: http://www.w3.org/TR/xpath20/.

[23] *XQuery 3.0: An XML Query Language: W3C Candidate Recommendation 08 January 2013*. Englisch. W3C. Jan. 2013. URL: http://www.w3.org/TR/xquery-30/.

[24] *XQuery API for Java Homepage*. Englisch. URL: http://xqj.net/.

[25] *XQuery Update Facility 1.0: W3C Recommendation 17 March 2011*. Englisch. W3C. März 2011. URL: http://www.w3.org/TR/xquery-update-10/.

Exploration of Spreadsheet Formulae with Fency

Andrea Kohlhase and **Alexandru Toader**
Jacobs University Bremen
D-28717 Bremen, Germany
a.kohlhase and a.toader @jacobs-university.de

Abstract

Spreadsheets are well-known to be frequently-used but error-prone communication devices. They are useful since they are active (e.g., automatic computation), provide a cognitive notation system drawing on visualizing values, meanings and relations at the same time (enabled by labeled, color-coded grids), and provide easy-to-use domain-specific operations (e.g., computational functions). The latter, in particular, is enabled by the text-style formula format in spreadsheets, in which variables are replaced by cell references. For simply-structured formulae this works very well. To keep the formulae simple, computations are modularized into subformulae and as such distributed over and beyond the spreadsheet. This makes the provenance (tree) of spreadsheet values difficult to understand – a probable cause for the high error rate in spreadsheets.

To explore and navigate the subformulae involved in the computation of a cell value we present the subformula explorer "Fency", a tree-based, explorative interface: Whenever a user clicks on a cell its formula becomes the root of a cell-dependency graph. Each child node displays the formula of a cell (or range) reference used in the parent formula. Moreover, each node represents a direct link to the respective cell (or range), so that it can be used for formula navigation as well.

1 Introduction

What is a mathematical formula? According to Wikipedia, in mathematics it is "*an entity constructed using the symbols and formation rules of a given logical language*". Even though there are multiple mathematical communities of practice which use a partly different set of symbols and slightly varying formation rules, there is a common understanding how to encode several information levels into formulae by extending the linear form of text.

On the one hand, this construction of a formula, O'HALLORAN calls a "*grammatical strategy for encoding meaning efficiently [. . . which is achieved . . .] through spatial and positional notation in a form that is not found in language.*" [O'H05, p. 112]. In Fig. 1 we can see some common typographical line elements. The spatial information needed to characterize the form of a typical English

Figure 1: Typographical Line Elements[1]

text can be characterized via these line elements. But very often formulae need more space.

Accommodating our running example in Fig. 3, the equation

$$\sigma_4 = \frac{1}{3}\sum_{j=4}^{7}\delta_{4j}^2 \tag{1}$$

with variables σ_4 and δ_{4j} represents the simple formula used in cell [B4].

Here, if we take a closer look (Fig. 2), we realize that the equation transcends the ascender and descender height with respect to the typographical baseline of the used font quite a bit. If we look closely, we also realize right away that not only specific spatial and positional notation is used, the common font type is also broken, there are, for example, greek letters. For mathematicians these are not unexpected and hardly something to think about since they have internalized the notational naming convention within formulae, that is the relation between fonts and functional status of objects. This common mathematical practice of authoring and interpreting formulae evolved over centuries and proved to be effective and efficient for mathematicians.

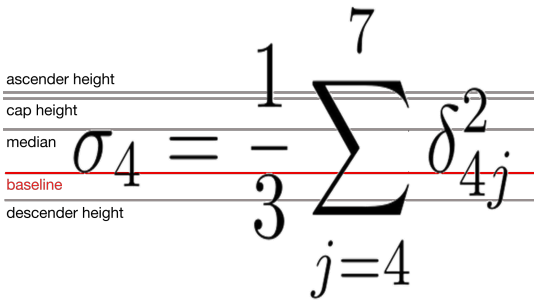


Figure 2: Equation (1) with Typography

On the other hand, in a spreadsheet there are also mathematical formulae. We can, for instance, reformulate Equation(1) as a computational formula in a spreadsheet like this:

$$[\text{B4}] = 1/3 * \texttt{SUMSQ}(D4 : G4) \tag{2}$$

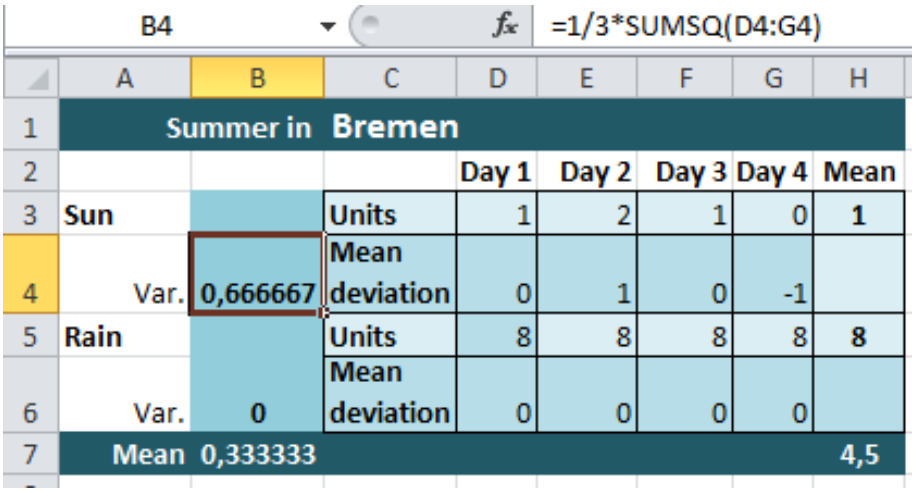


Figure 3: The Spreadsheet "Summer in Bremen"

The differences between the different representations is obviously vast. In this paper we use the example given in Sect. 2 as a running example. In particular we discuss the differences in Sect. 3 to motivate the design of our (sub)formula explorer "Fency" described in Sect. 4. We consider related work in Sect. 5 and conclude in Sect. 6 with an outlook on further work.

2 Running Example "Summer in Bremen"

Let us suppose that we want to describe the summer in Bremen statistically. Real-world distributions are typically not fully known, e.g. the rain could stop for 5 minutes when the observer went to the coffee bar to get some more coffee. In this case, the variance of the whole distribution is *estimated* by computing the **variance of a sample** of n observations drawn suitably randomly from the whole sample space according to Equation (3) where $x_1, \ldots, x_k$ represent the measurements and $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ their **arithmetic mean**.

$$\sigma = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 \qquad (3)$$

In the spreadsheet seen in Fig. 3, observed half-an-hour periods of full sunshine resp. rain in Bremen, i.e., the measurements, on four days in June are noted in ranges [D3:G3] resp. [D5:G5]. The difference $x_k - \bar{x}$ is called the **mean deviation of** x_k. The mean deviation of those measurements can be found in ranges [D4:G4] resp. [D6:G6]. The sample variance for sunshine in Bremen, for example, in cell [B4] is calculated from the mean deviation according to Equation (3) with the spreadsheet formula in Equation (2). Finally, the arithmetic mean of the sample variances is presented in cell [B7].

We use this example throughout the paper as running example.

3 Readability of Spreadsheet Formulae

In general, the set of symbols used in spreadseet formulae consists of given functions like `SUM`, individual macro extensions, numbers, and cell references like [B4] (in A1 referencing style referring to the cell in column B and row 4) or [R4C2] (in R1C1 referencing style pointing to the same cell). In `MS Excel`'10, for example, the set of symbols enlists 339 functions and $2^{20} \times 256$ cell references per worksheet. An essential component of spreadsheet players is their computational foundation: they can compute values from formulae, that is, they can simplify formulae to values. It is important to note that – even though it acts like a programming language – "*the formula language itself is entirely textual*" [Nar93, p. 49].

The formation rules are rather simple: concatenate the ingredients into a string of ASCII characters. From the user perspective NARDI points out that authoring and understanding formulae "*the user must master only two concepts: cells as variables and functions as relations between variables*" [Nar93, p. 42]. This is suspected to be the underlying reason for spreadsheets being the world's most used programming environment: the task of writing formulae (program scripts) is transformed into the task of writing text in a well-understood domain language consisting of typically 3-5 [SP88], at most 10 [Nar93, p. 43] and potentially – in `MS Excel` e.g. – 339 functions. It is rather interesting that the formula language hasn't changed at all since the very first appearance of spreadsheet applications, therefore we can call it a successful formula language.

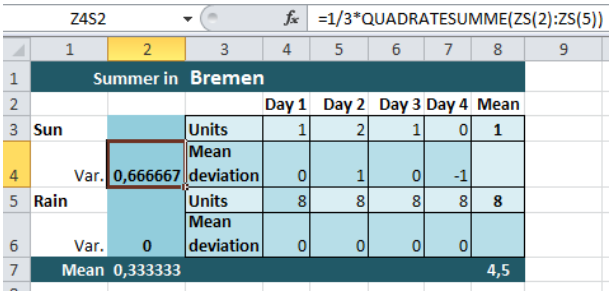


Figure 4: German R1C1 Notation of Equation (4)

Note that the ease of writing down spreadsheet formulae comes at the cost of reading them. For a simple formula, there is no problem in interpreting this linear notation of a formula – if the reader is very familar with the used naming convention for cells in spreadsheets.

The confusion begins if the spreadsheet author used the rather uncommon R1C1 referencing style, e.g., for Equation (2):

$$\text{[R4C2]} = 1/3 * \texttt{SUMSQ}(RC(2) : RC(5)) \qquad (4)$$

Here, the cell referencing is relative to the cell that will con-

tain the calculated value, e.g. $RC(2) = R(0)C(2)$ refers to [D4] (with D=B+2,4=4+0).

It gets even more confusing if the spreadsheet author used e.g. the German `MS Excel` version with R1C1 referencing style (where "Z(eile)" stands for "R(ow)", "S(palte)" replaces "C(olumn)", and "`SUMSQ`" translates to "`QUADRATESUMME`") as in Fig. 4.

Besides this specific representation format knowledge, the reader might also get easily overwhelmed if the formula is complex. As readers are typically experts in their specific fields, but laymen in spreadsheet technology, this is in analogy to command line interfaces which work very well for simple commands used by laymen or for complex commands used by power users. Therefore, one explicit aim of a spreadsheet author has to be the optimal reduction of complex formulae.

This can be done via *modularization*, in particular by collapsing parts of formulae into variables by using these parts as autonomous formulae to calculate different cell values.

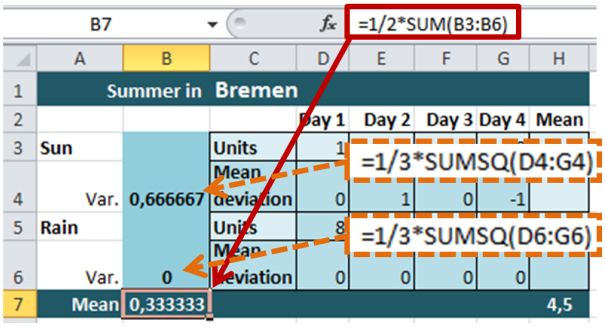


Figure 5: Modularization: Mean and Variance in [B7]

In Fig. 5 we can see a version of Fig. 4, this time in the more common A1 referencing style. The mean of all cell values in range [B3:B6] is calculated in [B7].

Moreover, the cells in the ranges [D4:G4] and [D6:G6] contain formulae of the kind (as shown in Fig. 6):

$$[\text{D4}] = \texttt{SUM}(D3; -\$H\$3) \tag{5}$$

Thus, in [B7] as seen in Fig. 5 we have the recursively resolved equations as shown in Fig. 9.

Note that even though the underlying formulae are one of the most simple ones, already the concatenated formula turns out to look rather complex to grasp. The reason consists of the fact that the cell references in Equation (10) can still be resolved easily by a reader, but the cell references in Equation (11) are more distributed and thus much harder to follow. HERMANS ET AL. report that nested formulae are hard to understand for end-users, which was also speculated in [Bre08]. "*We conclude that users find it difficult to work with long calculation chains*" [HPD12, p. 10]. Somewhat surprisingly they continue that this difficulty "*does not influence their perceived understanding of the formula or their ability to explain it*" [HPD12, p. 10]. A closer read reveals that their users are spreadsheet professionals, thus spreadsheet authors that not only do have the background knowledge for the specific spreatsheet at hand, they also know of the data architecture they created. They do not need to understand the concrete formula any longer as they trust in the underlying (hopefully) sound architecture.

As it is well-known that human short-term memory is rather limited (7 +/- 2 items can be kept in short term memory at any given time), the modularization of formulae is not an option, but rather a requirement for authoring readable spreadsheets. It is obvious that this modularization enables at the same time a high error rate with errors that are hard to debug.

The formula explorer Fency is based on the idea that the cell references can be automatically resolved into a cell-independent format e.g. presentation MathML [Aus+10] with variables that have mnemonic names, that is, names that hint at their meaning. For example, it is a quasi-standard to index a set of data points by a counter variable in $\{i, j, k, l, m, n\}$, to assign the name $\bar{y}$ to the mean of data points y_k, to name variances σ, and to name differences δ. Now look at Equations (10) to (12) in common mathematical notation:

$$0,333333 = \bar{\sigma} \tag{6}$$

$$= \frac{1}{2}\sum_{i=3}^{6}\sigma_i \tag{7}$$

$$= \frac{1}{2}\sum_{i=3}^{6}\left(\frac{1}{3}\sum_{j=4}^{7}\delta_{ij}^2\right) \tag{8}$$

$$= \frac{1}{6}\sum_{i=3}^{6}\sum_{j=4}^{7}(x_{ij} - \bar{x_i})^2 \tag{9}$$

Note that typically a reader familiar with math notation will have noticed at the latest in Equation (7), that there is something strange going on with the mean being a sum of 4 numbers divided by the normalizing term 2. Looking at Fig. 5 we notice why the effect is correct, but the formula isn't. Therefore, math notation might also help to discover semantic errors in formulae.

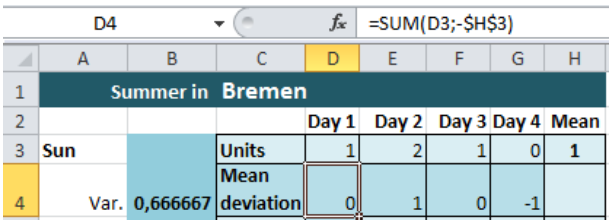


Figure 6: Modularization: Deviation in [D4]

The modularization can be kept, if we visualize the formula dependencies in form of a graph, where every node contains information about a formula.

4 The (Sub)Formula Explorer Fency

To keep spreadsheet formulae simple, computations are modularized into subformulae and as such distributed over and beyond the spreadsheet. Even though the modularization simplifies the formula itself, it resolves in a very complex provenance (tree) of spreadsheet values. The basic idea of Fency consists in an interactive visualization of the modularization of a formula. To explore and navigate the subformulae involved in the computation of a cell value we developed a semantically supported, tree-based, explorative interface: Whenever a user clicks on a cell its formula becomes the root of a "**formula graph**", i.e., a graph with cell/range nodes and cell/range-dependency edges. Each child node displays the formula of a cell (or range) reference used in the parent formula.

For example, in Fig. 7 we can see an entire formula graph developed after the user clicked cell [B7]. This

$$
\begin{aligned}
0,333333 &= 1/2 * \texttt{SUM}(B3:B6) && (10)\\
&= 1/2 * \texttt{SUM}(1/3 * \texttt{SUMSQ}(D4, G4) : 1/3 * \texttt{SUMSQ}(D6, G6)) && (11)\\
&= 1/2 * \texttt{SUM}(1/3 * \texttt{SUMSQ}(\texttt{SUM}(D3; -\$H\$3), \texttt{SUM}(G3; -\$H\$3))\\
&\quad : 1/3 * \texttt{SUMSQ}(\texttt{SUM}(D5; -\$H\$5), \texttt{SUM}(G5; -\$H\$5))) && (12)
\end{aligned}
$$

Figure 9: Recursively Solving Equations for cell [B7]

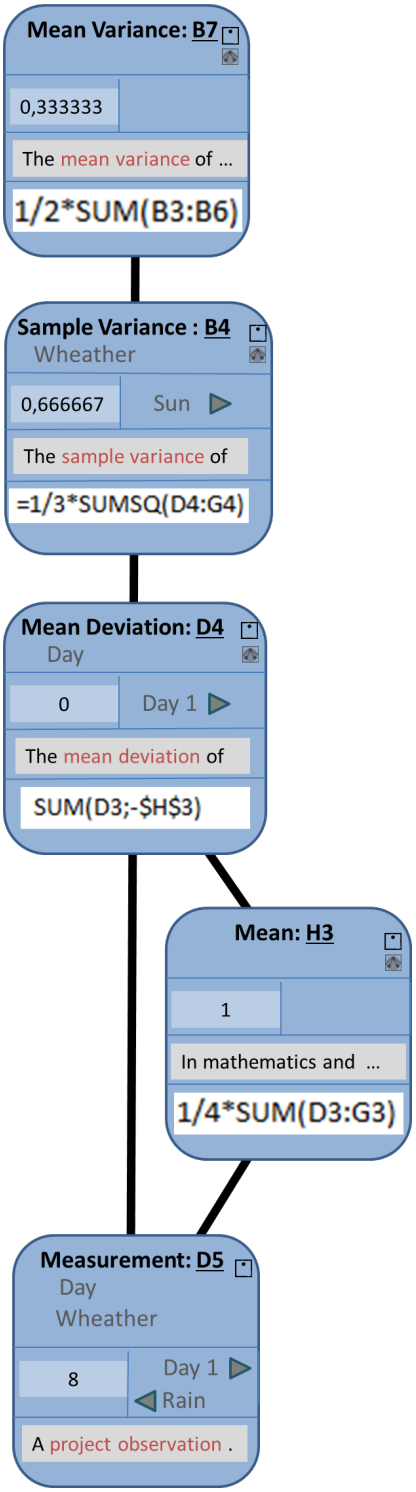


Figure 7: The Expanded Formula Tree in Cell [B7] (with Spreadsheet Formulae)

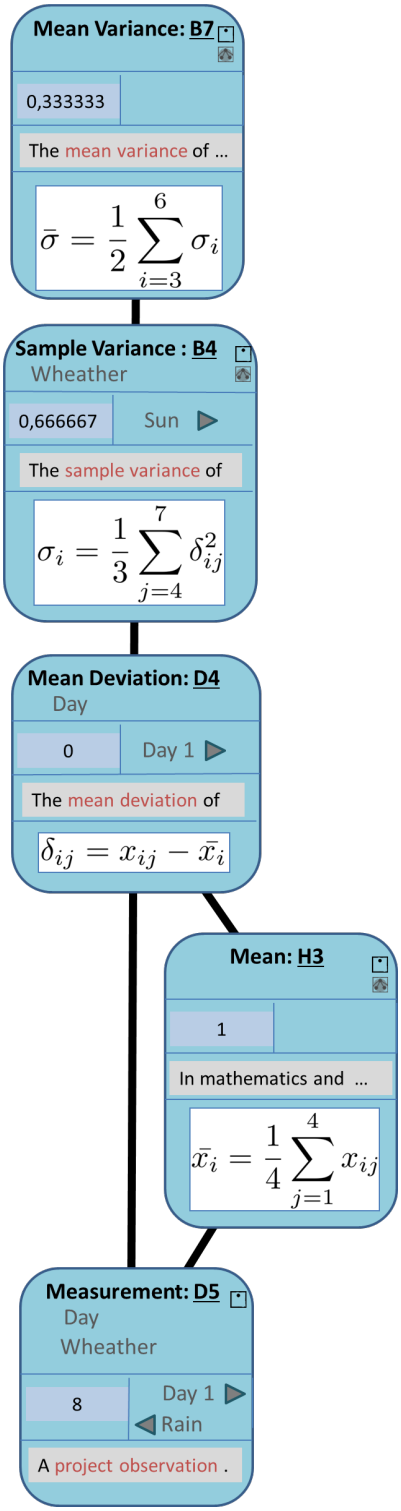


Figure 8: The Expanded Formula Tree in Cell [B7] (with Math Formulae)

cell contains the formula $1/2 * \texttt{SUM}(B3 : B6)$, that is Equation (10). The values in the cells in the cell range [B3:B6] are computed by equivalents of the formula $1/3 * \texttt{SUMSQ}(D4, G4)$ taken from cell [B4][2]. With Fency, if the user clicked cell [B7], the root node as in Fig. 10 would be created and the cell-dependency of the underlying formula on range [B3:B6] would give rise to a child node representing it in the formula graph. If the user wanted to see the child node of this, then she could click the expand button on the upper right and a node for the functional block in range [D4:G4] would appear.

On a more technical note, the formula explorer Fency is a semantic service integrated into the open source Semantic Alliance Framework [Dav+12]. This framework allows to superimpose semantic services over an existing (and possibly proprietary) application provided that it gives open-API access to user events. Elements in the application are connected to according concepts in structured background ontologies, which, for instance, contain a representation of the respective domain and some instance specific information. Semantic services can draw on the ontology information to offer intelligent services, which are offered to the user via the Semantic Alliance framework in local, but application independent windows. For the most common spreadsheet applications `MS Excel` and `LibreOffice` there are already existing Semantic Alliance APIs.

Fency offers more than a tree-based visualization of the (sub-)formulae in a spreadsheet. In a nutshell, every node of the formula graph consists of a list of elements:

- The **title** expressing the underlying meaning of a cell value or a range of values,
- a **link** to the corresponding cell/range in the spreadsheet,
- the **dependencies** this cell/range depends on,
- its data **value**,
- an **explanation** of its meaning,
- the spreadsheet **formula** (or its equivalent math formula), and
- **iterators** to move through the cells with their resp. values of a range.

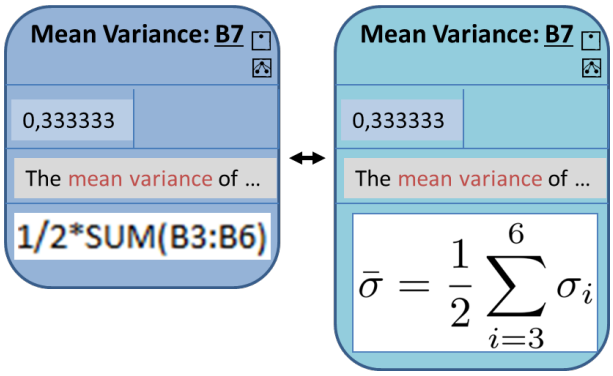


Figure 10: Node Variants in Cell [B7]

Let us have a closer look, for example, at a node like the left one in Fig. 10. The cell [B7] is associated with the ontology concept "mean variance". The title of this concept followed by the cell reference "B7" itself is used as a title for the node. The underline of the cell reference indicates that it represents a link to this cell. On the upper right-hand side we can see a collapse and an expand button, which collapses or expands the formula graph respectively if clicked. The cell value of cell [B7] is $0,333333$ and is shown in the node as well. In the grey box the beginning of the explanation of the concept "mean variance" given in the ontology is visible. Hovering over the grey box will trigger the expansion of it, so that the entire definition will be visible (see an example in Fig. 11). By using the JOBAD framework[JOBAD], the user can even interact with the information items within this explanation: If other concepts are referenced in this definition (indicated by blue font usage), a click will open another window with the according concept definition. This way, a user can explore the background ontology and comprehend the meaning of the formula much deeper. The lower part of the node contains the formula, here the formula for [B7], if existent; see an empty formula example in Fig. 11. The hovering effect kicks in here as well, in particular, if the formula exceeds a certain size, the entire formula will only be visible while hovering over the formula box.

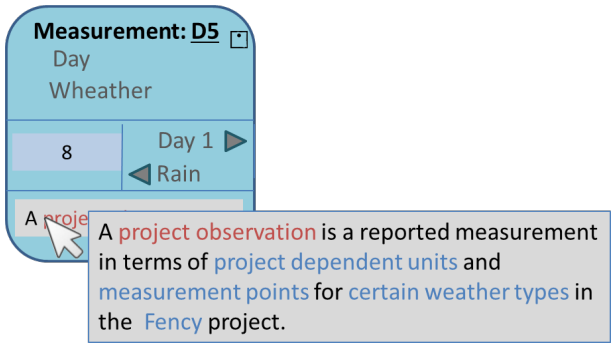


Figure 11: Expansion on Hover over Definition Box

Cell [B7] itself is not part of a functional block, but e.g. cell [D5] is. As the value in [B7] depends down in the formula tree on the value in this cell, we can find the node for [D5] as the last one in the formula graph in Fig. 7 or more conveniently in Fig. 11. This functional block covers the observed and summarized data. Each measurement depends on which day it was taken and what wheather condition is reported, in other words the measurement functional block depends on the day functional block [D3:G3] and the wheather functional block [[A3], [A5]]. This dependency is noted in the node directly under the title (in grey font). Moreover, we can see that cell [D5] contains the value for "Day 1" and "Rain". The triangular buttons allow a user to skim through the values in the respective functional blocks, and navigate to the respective spreadsheet cells via the link "D5" right after the title. This feature allows the user to easily navigate through related information items while abstracting away from the concrete structure. If any of the information items presented above are missing, the UI of the node adapts.

In a future prototype, if the user double clicks on the formula in a node, then the spreadsheet formula is converted into a math formula using MathML (see right node in Fig. 10). The option of presenting both variants seems sensible as a switch of formats should always be easily reversible to avoid confusion. The ontology concept "mean variance" includes knowledge about the symbol notation $\bar{\delta}$. Moreover, as the range [B3:B6] is associated with the concept "sample variance" with its symbol notation δ, a parser should be able to figure the math formula as seen in the

[2]Ranges used as cell references in formulae are typically **functional blocks**, i.e., cell ranges that have the same *functional* content, see [KK13] for more details.

right node in Fig. 10. To give a taste of the potential of this conversion, we include Fig. 8. Another idea, we want to pursue shortly is that the user can even edit the formula and push the changes back to the spreadsheet.

5 Related Work

The visualization of data-flows within spreadsheets is not a new idea. In MS Excel itself there is a *tracing tool* that visualises precedents and dependents of a selected cell. The visualization breaks if the dependencies are beyond the worksheet or even more so beyond the workbook.

In [CKR01] the authors studied the comprehension factor of formulae visualized in distinct ways. They frame formula understanding in terms of the reader's cognitive load and thus as a visual memory problem. They find that the "*ideal organization is the simple tree. It is the easiest to chunk. In the simple tree the surface organization of the formula tree is in harmony with its deep structure.*" [CKR01, p. 487].

KANKUZI and AYALEW presented in [KA08] a graph-based visualization of spreadsheets. Based on a Markov Clustering algorithm they generate a data-flow graph which visualizes cell cluster dependencies in an extra window aside the spreadsheet application window and provides semantic navigation similar to the one presented in Fency. Instead of using functional blocks, i.e., sets of cells that belong together semantically, these authors use statistical clustering. Even though this probably provides a similar grouping effect, the spreadsheet reader won't know why the cells are grouped. With Fency we cannot only offer the reader this reason, i.e., the semantic relating concept, we also allow the reader to dig into the definition of this concept.

In [Raj+00] a tree representation for formulae is suggested according to predominant Software Engineering techniques. In particular, a formula is divided into a structure tree containing operators and functions and an arguments tree containing cell addresses and constants. This tree visualization of a formula is suggested to be done when authoring a spreadsheet, whereas Fency is a tool that supports reading a spreadsheet. In [JMS06] a tool for generating formulae in several formats (possibly spreadsheet format) is presented. Again, the sole focus is given to the developper or author of formulae, nothing is said about the enhanced readibility or comprehendibility of a formula.

http://www.spreadsheetstudio.com/ offers another type of formula explorer. The modularity of MS Excel formulae is made use of as is in Fency. This formula explorer offers a modal pop-up window that presents the formula of the selected cell. The formula is automatically segmented into sensible parts like cells, ranges, function plus function parameters, constants etc. If the user hovers over the formula shown then the corresponding value is presented. If a segment corresponding to a cell or range is left-clicked, then the formula of that MS Excel object is shown as before. Thus, this formula explorer allows a similar navigation thru a formula via its subformulae. Moreover, the MS Excel cursor also moves to the MS Excel object selected in the formula window.

ASUNCION suggests in [Asu11] to capture the provenance of cell values by unobtrusively document their history and to make this set of data available for later querying. This kind of provenance capture certainly is appealing because of its automation facility, but the provenance is not stored on a semantic level. Thus, the author has to recognize data to be able to interpret the provenance correctly. Otherwise this kind of data handling seems to be very tedious.

6 Conclusion and Further Work

In this paper we have presented Fency, a (sub)formula explorer for spreadsheets, that allows readers to deeper understand what formulae, which concrete calculated values, what underlying concepts are spread how and where over the document.

We hope that Fency will prove to be a useful service, especially as we are planning to extend its capability towards a light formula resp. concept editor, that allows to *update* existing formulae resp. ontology items. Even though the cell values are shown in the resp. formula nodes, we believe that the provenance of cell values is still not enough covered. The graph structure gives a hint where the data originally come from, but very often outside data bases are used for data input of spreadsheets. In particular, the spreadsheet author is typically a data architect. For him the primitives are data resources. Therefore, a set of new information objects could be introduced to spreadsheets. If they were present, then Fency could visualize it as well, to obtain a formula visualization that not only keeps all relevant information in one place, it also uses the notation that is most efficient.

Acknowledgements We want to thank the anonymous reviewers for their constructive suggestions. This work has been funded by the German Research Council under grant KO-2484-12-1.

References

[Asu11] Hazeline U. Asuncion. "In Situ Data Provenance Capture in Spreadsheets". In: *eScience*. IEEE Computer Society, 2011, pp. 240–247. ISBN: 978-1-4577-2163-2.

[Aus+10] Ron Ausbrooks et al. *Mathematical Markup Language (MathML) Version 3.0*. W3C Recommendation. World Wide Web Consortium (W3C), 2010. URL: http://www.w3.org/TR/MathML3.

[Bre08] Andrej Bregar. "Complexity Metrics for Spreadsheet Models". In: *CoRR* abs/0802.3895 (2008).

[CKR01] David Chadwick, Brian Knight, and Kamalasen Rajalingham. "Quality Control in Spreadsheets: A Visual Approach using Color Codings to Reduce Errors in Formulae". In: *Software Quality Journal* 9.2 (2001), pp. 133–143.

[Dav+12] Catalin David et al. "Semantic Alliance: A Framework for Semantic Allies". In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics (CICM) (Bremen, Germany, July 9–14, 2012). Ed. by Johan Jeuring et al. LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012, pp. 49–64. ISBN: 978-3-642-31373-8. URL: http://kwarc.info/kohlhase/papers/mkm12-SAlly.pdf.

[HPD12] Felienne Hermans, Martin Pinzger, and Arie van Deursen. “Measuring Spreadsheet Formula Understandability”. In: *CoRR* abs/1209.3517 (2012).

[JMS06] Sven Jörges, Tiziana Margaria, and Bernhard Steffen. “FormulaBuilder: a tool for graph-based modelling and generation of formulae”. In: *Proceedings of the 28th international conference on Software engineering*. ICSE ’06. Shanghai, China: ACM, 2006, pp. 815–818. ISBN: 1-59593-375-1.

[JOBAD] *JOBAD Framework – JavaScript API for OMDoc-based active documents*. URL: `http://jobad.omdoc.org` (visited on 02/18/2012).

[KA08] Bennett Kankuzi and Yirsaw Ayalew. “An end-user oriented graph-based visualization for spreadsheets”. In: *Proceedings of the 4th international workshop on End-user software engineering*. WEUSE ’08. Leipzig, Germany: ACM, 2008, pp. 86–90. ISBN: 978-1-60558-034-0.

[KK13] Andrea Kohlhase and Michael Kohlhase. “Spreadsheets with a Semantic Layer”. In: *Electronic Communications of the EASST: Specification, Transformation, Navigation – Special Issue dedicated to Bernd Krieg-Brückner on the Occasion of his 60th Birthday* (2013). Ed. by Till Mossakowski, Markus Roggenbach, and Lutz Schröder. in press. URL: `Http://kwarc.info/kohlhase/papers/easst11.pdf`.

[Nar93] Bonnie A. Nardi. *A Small Matter of Programming: Perspectives on End User Computing*. 1st. Cambridge, MA, USA: MIT Press, 1993. ISBN: 0262140535.

[O’H05] K.L. O’Halloran. *Mathematical discourse: language, symbolism and visual images*. Continuum, 2005. ISBN: 9780826468574. URL: `http://books.google.com/books?id=5LsAJaBRKRcC`.

[Raj+00] K. Rajalingham et al. “Quality control in spreadsheets: a software engineering-based approach to spreadsheet development”. In: *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*. 2000, 10 pp. vol.1.

[SP88] Jorma Sajaniemi and Jari Pekkanen. “An empirical analysis of spreadsheet calculation”. In: *Softw. Pract. Exper.* 18.6 (June 1988), pp. 583–596. ISSN: 0038-0644.

An approach to visualize ophthalmic ontologies

Christian Henninger, Constantin Rieder, Klaus Peter Scherer
Karlsruhe Institute of Technology
Institute for Applied Computer Science
D-76344, Eggenstein-Leopoldshafen, Germany
<firstname.lastname>@kit.edu

Abstract

In this position paper we introduce an approach of visualizing the ontology describing ophthalmic knowledge. The data consists of semantic information provided by a knowledge based system. This ontology is developed using the Semantic Wiki KnowWE and the visualizations are developed as a plugin. The visualization aims to satisfy the needs of ophthalmological experts and trainees as well as knowledge engineers. The visual exploring assists all users to obtain a general overview of the knowledge and detects specific characteristics. Therefore different approaches of visualizations are shown and evaluated.

1 Introduction

In this paper, we describe the work in a project on visualization of ontologies. This project is part of the "Wissass" Project[1]. The aim of the project is to assist the physicians working in the ophthalmology. In this field, the cataract surgery is the most common procedure. This operation is performed very often. That's why there is very much knowledge available. On the other hand, various cases are known in which a special treatment is required to obtain the bests results possible. In "Wissass", we develop a knowledge-based tool to provide the knowledge to the experts in special cases. Another goal of the system is to help teaching young physicians the knowledge as a tutor system. Based on positive feedback of former projects, for instance [Dieng-Kuntz et al, 2006], we used an ontology to represent the knowledge. In order to do so we based our work on the Semantic Wiki KnowWE. Our aim is to provide a visualization of the knowledge to the user. As a part of the research towards a "knowledge formalization continuum", the "Wisskont" sub-project focuses on fusion of the process of Knowledge Engineering and productive work with the system.

Besides the decision support aspects and the tutoring capabilities, visualization methods can also help during Knowledge Engineering. Visualizing helps the developers to get a quick overview of the knowledge and to spot interesting or even malicious parts of the knowledge base. It also helps the expert to see whether the database is complete. Due to its easy use it helps the user during the Knowledge engineering process to expand the knowledge and alter parts which are out to date. To provide a simple possibility to visualize the knowledge during the productive workflow, the extra use of generic ontology visualization tools is difficult.

Our work is also influenced by the small iterations between the reviews of the involved physician, whose suggestions are tried to be added contemporary. This paper describes the state of implementations and variety of representations we evaluated.

This paper is divided in the following sections:

In section 2 we describe the ontology we developed to describe knowledge in the ophthalmic domain, visualized by the described plugin. In section 3 the different visualization approaches are comparatively discussed. In section 4 we report on the review results of the sessions with ophthalmologist experts. The reviews checked whether the chosen representations are easily understandable by the experts and whether the work and representation are worth the effort. Section 5 discusses related work and gives a brief conclusion.

2 Ontology for cataract surgery

The application will be used essentially by physicians in an ophthalmological domain, especially in the cataract surgery. Like most areas of medical knowledge, the knowledge in this domain is very extensive. The knowledge is represented by concepts connected to other concepts by relations. When we outline all concepts from the knowledge base and their relations between each other the result is a net. The major task for us is to represent the allocated knowledge and accordingly to simplify the information maintenance and retrieval for the end-user by using visualization techniques.

Transferring the knowledge from the predecessor system "Visu-XPS"

At the beginning of the "Wissass" Project, the ophthalmologic knowledge stored in the software developed in the previous project was transferred to the wiki system "KnowWE" shown in figure 1[KnowWE, 2013]. The screenshot presents a concept page from the actual system. The predecessor system "Visu-XPS" was a standalone application developed in Java. It only supported one type on semantic information: *"associated with"*. Therefore, the presentation was limited to a directed graph, which could be simply traversed.

[1] The Wissass project is founded by Zentrales Innovationsprogramm Mittelstand (ZIM) from 2012 - 2014

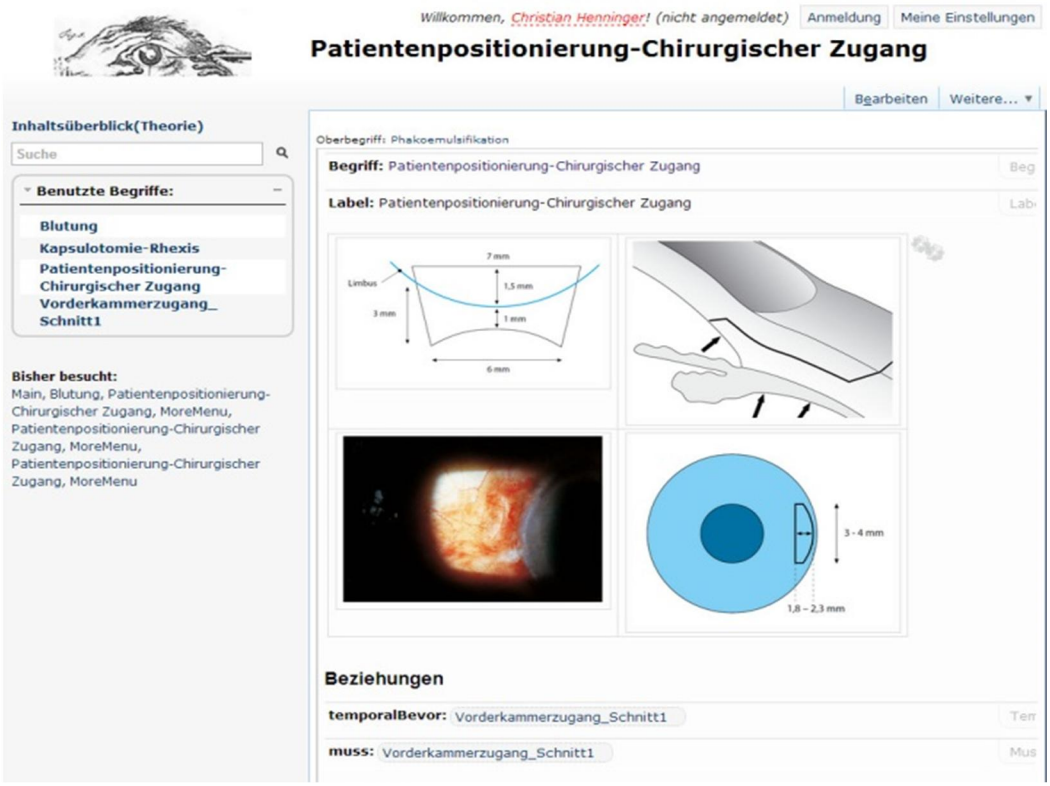


Figure 1: Screenshot of the Semantic Wiki KnowWE

Expansion through new Connection Types

In order to build a semantic net and to provide the users more expressive types of knowledge, more additional relation types were added:

"subconcept": A refinement of the given concept ,used to arrange the concepts in a hierarchical order.

"has to": Connection between complications, which may occur during the operation and their necessary treatments.

"can": A relation used to identify possible reactions to the given state of the patient.

"cave": This relation is used to connect concepts that should be urgently considered.

"before": A relegation used to represent a time period between two concepts.

The resulting ontology is represented in RDF(S) [RDFS, 2004].

At the moment, the ontology consists of 381 concepts. Those are connected by 331 "subconcept", 60 "has to", 49 "*associated with*", 44 "can", 25 "before" and 26 "cave" relations.
In cooperation with the physicians the requirements for the visualizations were identified to guide the development. In conclusion the following use cases were defined:

Use cases

1. Obtain an overview of the knowledge base by reducing complexity by using visualization methods.
2. Obtain an overview of the processes and dependencies between procedure steps of ophthalmic surgery.
3. Browsing through the entire knowledge base to identify interesting spots.
4. Retrieve detailed information on special relations between concepts and procedure steps on demand.
5. Help the user to find quickly the category of a concept.

3 Visualization approaches

In this section we describe a number of different visualization approaches and we discuss their applicability with respect to the described use cases.
All visualizations try to implement the well-known visualization mantra by Shneiderman: Overview First, Zoom and Filter, Then Details-on-Demand [Shneiderman, 1996].

Hierarchical Forest Visualization

The Hierarchical-Forest-Visualization (see Figure 2) is an approach to represent a pool of relationships in the knowledge base according to the first and second use case. It is based on the classical representation of a graph with hierarchical levels, used in file managers for instance. The graph view makes it possible to form a hierarchical structure of a concept and a selected relation. The user can see an overview of all connected concepts by a specified relation. The overview does not exclude a focused view on sub-concepts. It is also possible to select a certain sub-concept from the overview and open the related sub-concepts. The higher level of a concept shows the broader outline of related concepts. The key aspect of this visualization is the combination of two different relations in order to show a more specific structure in the overview. Both relations have a different layout direction in the visualization. The horizontal direction represents an order likewise a time oriented process. The vertical direction represents a sectioning of the concept by an arbitrary relation.

The implementation was realized with the JavaScript library jsPlumb. It provides a way to visually connect elements on a web page [jsPlumb, 2013].

The following visualizations in the next sections were realized with d3js. This JavaScript library was developed in order to manipulate documents based on data [d3js, 2013].

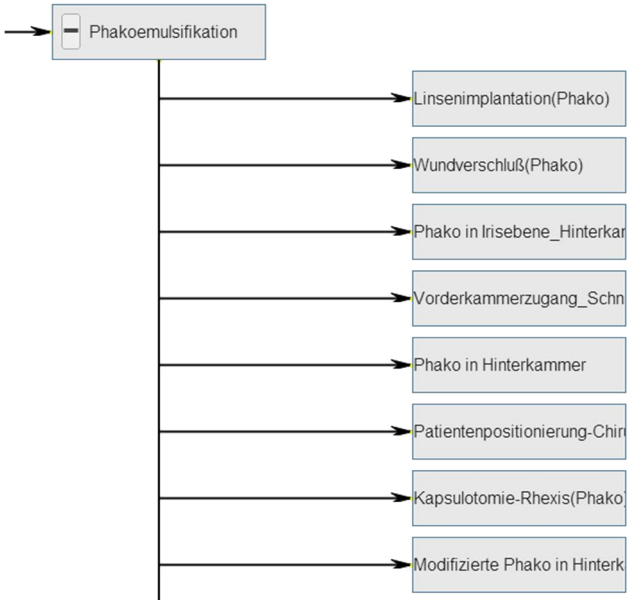


Figure 3: Hierarchical Forest Visualization

Collapsible Tree

In contrast to the Hierarchical-Forest Visualization the idea of the Collapsible Tree Visualization, shown in Figure 3, is to visualize the hierarchy depth of the knowledge base by concentrating on a single type of relation. In this case the use cases three and four were implemented. By selecting a concept it can be expanded to show the following sub-concepts.

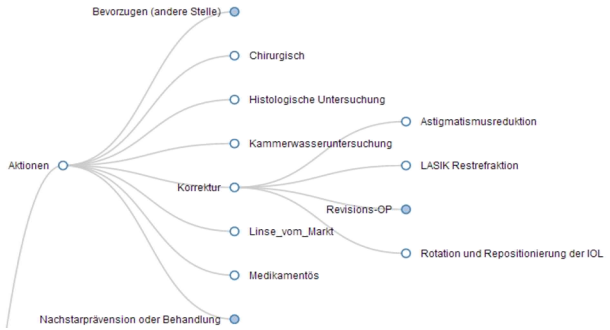


Figure 4: Collapsible Tree Visualization

The user can browse to a topic of interest. Reciprocally the user can also collapse branches of the graph. These techniques allow a suitable display on a single screen without a restraint to scroll. The reason for choosing a classic tree structure to visualize ontologies is that ontologies are graphs and can be often presented as trees. Using this technique is quite common in many research fields and often helps to organize large hierarchical information and bring it to a general overview [Song *et al*., 2010].

Circle Pack Visualization

The Circle Pack Visualization implements the use cases three and five and shows an entire overview of the whole knowledge base.

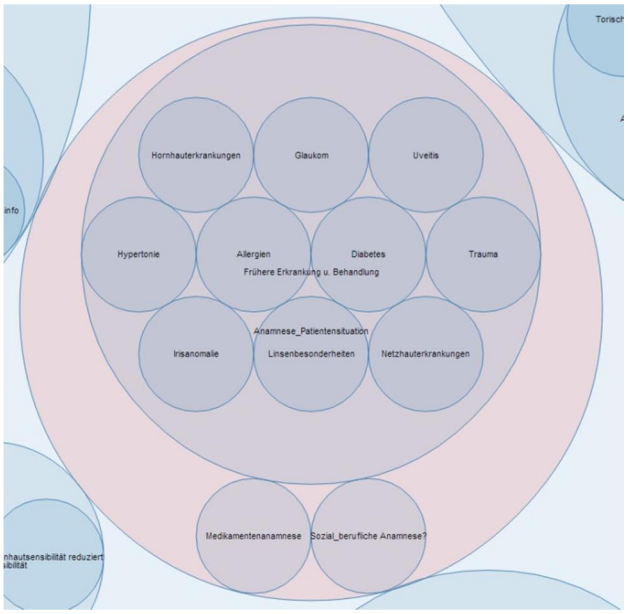


Figure 2: Circle Pack Visualization

As seen before the hierarchical view of concepts is well represented by a tree structure. However, the view becomes confusing very quickly by presenting the entire content of a large knowledge base. The tree diagram becomes too large when too many nodes and branches must be placed on a single page. Addressing those disadvantages, the Circle Pack Visualization provides a useful alternative by representing hierarchical relations through containment. It is possible to see an overview of the overall structure and the position of a certain concept. Concepts are displayed as circles. Child-concepts are located inside their parents. For a better orientation the selected concept is highlighted in the overview (see Figure 4). To increase the readability and to avoid cluttering, only the labels of bigger circles are displayed. The user can zoom by clicking on circles to display the labels of the included circles. The main reason for selecting this kind of visualization is the big advantage that large amounts of hierarchically structured data can be visualized with a clear representation of structural relationships [Wang *et al*., 2006]. Another advantage is the use of size to display the amount of contained sub-concepts.

Wheel Visualization

The Wheel Visualization in Figure 5 shows an overview of the entire knowledge base. This deals with the use case number two and four. The advantage of this ordering is a maximum use of space: many concepts are presented on a single page. The main concept is placed in the middle. The hierarchy depth is presented by circles around the selected concept. Child-nodes use sub-divisions of the space of their parent. The size of Siblings depends on the amount of children they contain themselves. By selecting

a concept it becomes the new center of the wheel. Only the children of the new concept build the new wheel. By using this zoom function, all labels can be read easily and the subsection can be explored.

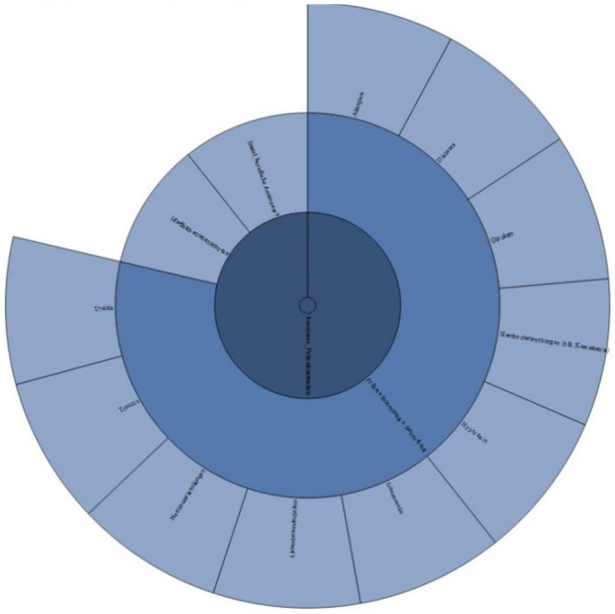


Figure 6: Wheel Visualization

4 Case Study

In order to evaluate the results with a domain specialist, a physician working in an ophthalmic clinic reviewed the developed system. We provided new versions in small iterations and changed details based on the users' feedback. The surgeon takes also part in the development of the knowledge base.

Generally the user was satisfied to obtain a visual representation of the ontology. The visualization was further used in the process of knowledge acquisition.

All visualizations suffered from cluttered labels due to long concept names in this domain. To avoid that, the labels of the included concepts will be abbreviated to get a shorter description name. In order to obtain a clear arrangement on the display further abbreviation techniques described in [Stum et al, 1991] will be applied later in the development process.
The hierarchical forest view was the preferred option by the user. He stated that the hierarchical view of knowledge fit the medicals needs best. The extra division of the tree into process steps simplified the comprehension noticeable. It was also the only visualization which was understood without explanations. This opinion is also backed by several surveys, for example in [Rivadeneira and Bederson, 2003].
We recognized that as more data was added to that view, it became more and more complicated to get an overview. It required scrolling and became too complex in more detailed tasks.
The expandable tree is very similar to the user's mental concept. He understands the concept that the hierarchical relation is displayed.
The situation when the knowledge base becomes bigger is also difficult: Not all concepts fit on the screen. On the plus side the expandability enables the user to show only the concepts of interest.
The Circle Pack Visualization needed some explanation: The user stated that he implies values to the different sizes of the circles. He also didn't realize that the leaf nodes were highlighted in a different color.
In our opinion the Circle Pack Visualization is a great choice to get a quick overview about the amount of concepts in different paths. It also shows the level in which the concept is located. The view needs further work to make it easier to understand and it needs to be evaluated whether it adds advantages to the user.

The Wheel Visualization confused the physician at the first glance. He doubted the suitability of this kind of visualization. Therefore we need to focus on simplifying this approach or finding more suitable alternatives.
In summary the visualizations Hierarchical Forrest and Collapsible Tree are best for the physician to obtain a quick overview of all concepts.
The Circle Pack Visualization can assist the Knowledge Engineering process by highlighting over- and under-populated topics.
The Wheel Visualization suits best when the expert tries to obtain detailed information on specific concepts.

Visualization	Use Cases	Feedback
Hierarchical-Forest	1, 2	+
Circle Pack	3, 5	-
Collapsible Tree	3, 4	O
Wheel	2, 4	O

5 Conclusion

Related work

A general overview of visualization of ontologies is given in [Fluit *et al*., 2003].This paper gives an overview of the current state-of-the-art tools that help visualizing ontologies and evaluates those to find out their weaknesses. Based on those results it recommends requirements for a tool for best user experience.

In [Menge, 2007] a visualization add-on was developed and added to the predecessor system. The main focus is to visualize rules and their derivation. The implemented concepts consisted of pie-charts, Cluster Maps and tree charts. Menge suggested the evaluation of cone respectively disc-charts to visualize larger amounts of data. In conclusion, she proposes to visualize the concept of an entire knowledgebase, which is done in our project.

In [Cvjetković et al, 1991] a development process for a web based ontology view is described. The ontologies displayed are limited to ordinary hierarchical trees which are only displayed as trees. The technologies used and the architecture is very similar to those used here.

Summary

We presented visualization methods to be used during the development and use of a decision support and tutoring system. The different approaches are useful in a number of use cases.
In general, the visualization added value for the user and should be considered, when a knowledge based system is

created. Especially the "knowledge formalization continuum" [Baumeister et al, 2011], which enables users to contribute knowledge to an existing system, can profit from the visualization. To get the best option in our case we need to further adjust the systems to the users' needs.

Future work

To verify the results and to evaluate the benefit of the visualizations, data and experience from the daily use of the system is required. By applying ontologies from different domains which contain bigger amounts of data the plugin can be tested whether the visualization is also useful in states where the knowledge base has increased. Also the performance can be measured and optimized. This will help to achieve an acceleration of response times for a faster way of displaying of the results. It may also reveal if it is necessary to develop specified representations.

Acknowledgments

The research project is a cooperation of the Karlsruhe Institute of Technology, Germany (KIT) and the denkbares GmbH. It is funded as a ZIM-KOOP by the German Federal Ministry of Economics and Technology (BMWI)[2].
The authors also thank the project executing organization AiF[3] in Berlin, which is responsible for the allocation of the budget to the research center and the commercial company.

References

[Baumeister et al, 2011] J. Baumeister, J. Reutelshoefer and Frank Puppe. Engineering intelligent systems on the knowledge formalization continuum. In *International Journal of Applied Mathematics and Computer Science (AMCS) 21 (2011),* pages 27–39, Versita Warsaw, Poland, March 2011.

[Cvjetković et al, 1991] Vladimir Cvjetković, Marija Đokić, Branko Arsić. Ontology Visualization - Graphical web user interface for ontologies. In *Advanced Research in Scientific Areas 2012,* pages 1999 – 2004, Virtual Conference, Slovakia, December 2012.

[d3js, 2013] d3js.org. Data-Driven Documents. homepage: http://www.d3js.org. 2013

[Dieng-Kuntz et al, 2006] Rose Dieng-Kuntz, David Minier, Marek Růžička, Frédéric Corby, Olivier Corby, Laurent Alamarguy, Building and using a medical ontology for knowledge management and cooperative work in a health care network. In *Computers in Biology and Medicine*, pages 871–892, Volume 36, Issues 7–8, July–August 2006.

[Fluit *et al*., 2003] Christiaan Fluit, Marta Sabou and Frank van Harmelen . Supporting User Tasks through Visualisation of Light-weight Ontologies. In *Handbook on Ontologies in Information Systems*, pages 417–432, Springer-Verlag, Berlin Heidelberg, 2003.

[jsPlumb, 2013] jsPlumb. http://www.jsplumbtoolkit.com/jquery/demo.html . 2013

[KnowWE, 2013] d3web - The Open-Source Diagnostic Platform. homepage: http://www.d3web.de. 2013

[Menge, 2007] Martina Menge. Kategorisierung aktueller Visualisierungstechniken und Implementierung für Wissensbasierte Systeme. Diploma thesis, Bayerische Julius-Maximilians-Universität Würzburg. Würzburg, 2007.

[RDFS, 2004] RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/. 2004.

[Rivadeneira and Bederson, 2003] Rivadeneira, W., Bederson, B. B. A Study of Search Result Clustering Interfaces: Comparing Textual and Zoomable Interfaces, University of Maryland HCIL Technical Report HCIL-2003-36, October 2003.

[Shneiderman, 1996] Ben Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages,* pages 336 – 343, 1996.

[Song *et al*., 2010] Song, Hyunjoo and Kim, Bohyoung and Lee, Bongshin and Seo, Jinwook. A comparative evaluation on tree visualization methods for hierarchical structures with large fan-outs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 223--232, ACM, New York, NY, USA, 2010.

[Stum et al, 1991] G. Stum, P. Demasco, and K. F. McCoy. Automatic Abbreviation Generation. In *RESNA 14th Annual Conference (Kansas City, MO),* pages 97 – 99, Washington, D.C., RESNA PRESS 1991.

[Wang *et al*., 2006] Wang, Weixin and Wang, Hui and Dai, Guozhong and Wang, Hongan. Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 517-520, ACM, New York, NY, USA, 2006.

[2] http://www.zim-bmwi.de
[3] http://www.aif.de

OpenMathMap: Accessing Math via Interactive Maps

Jan Wilken Dörrie, Michael Kohlhase
Computer Science, Jacobs University Bremen
<firstinitial>.<lastname>@jacobs-university.de

Abstract

World Math literature is growing at an alarming rate (3.3M journal articles today increasing by 120k a year). While much of that can be retrieved online, we lack technologies to navigate and understand the space of math literature. The OpenMathMap project wants to develop and deploy novel interfaces that empower interested parties to find their way. We conjecture that such maps can act as cognitively adequate access mechanisms to many large-coverage Mathematical Knowledge Managements systems.

The first concrete interface is an interactive map generated from publication data. We have developed a prototype map generation service based on MSC classifications and deployed the maps resulting from ZBMath data in OpenStreetMap. It is accessible at http://map.mathweb.org/.

1 Introduction

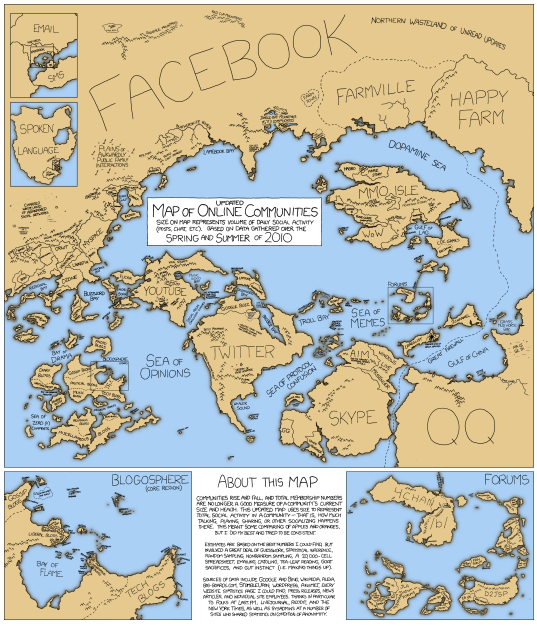


Figure 1: Map of Online Communities, XKCD 2010 http://xkcd.com/802/

In the information age fueled by the Internet, the problem of information and knowledge foraging changed from retrieving documents to finding out about them. In particular, navigating the space of available documents efficiently becomes an important subtask.

Even in science, the times where single individuals could have an overview over all of science are long past. Even in the Renaissance polymaths like Leonardo da Vinci were considered a rare exception. The scientific community has developed various tools to work around this problem: encyclopedias, survey articles, classification systems, and review services. But with the proliferation of scientific publication – 50 million articles in 2010 [Jinha [2010]] with a doubling time of 8-15 years these tools start collapsing under the sheer mass of information. Internet-age tools like search engines, bibsonomies, and citation databases solve (part of) the information retrieval and navigation problems by providing word-based search and browsing along citations. Note that these tools are "myopic" in the sense that they only give very local view of the immediate surroundings of a word or document.

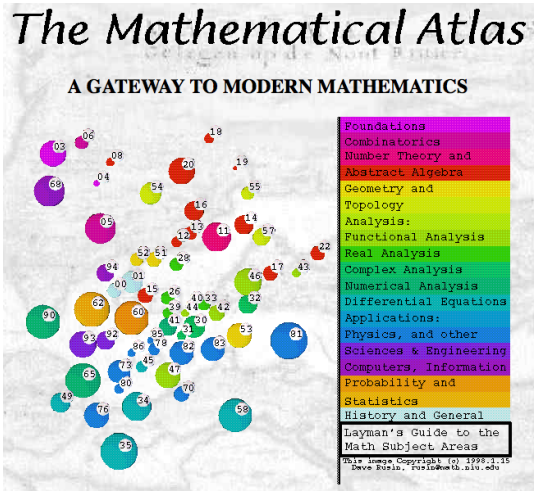


Figure 2: Dave Rusin's Math Atlas

Classification systems like the Math Subject Classification (MSC, see [American Mathematical Society]), take a more global stance, but they lack user interfaces that give information foragers an intuitive sense of direction and locality that is so helpful to humans in navigation tasks. In the MathSearch project we are currently rethinking access to mathematical knowledge and resources. As a first experiment, we are building a global, map-based navigation service for mathematics.

The main idea is that humans are very skilled in spatial navigation and in particular have learned to use map repre-

sentation to navigate spaces and locate targets. Concretely, we want to create a map of mathematics like the one in Figure 1 used to visualize usage patterns of online communities. We want to base the map on ideas from Dave Rusin's Math Atlas [Rusin] (created 1998, last updated 2001, see also Figure 2), which uses topics from the Math Subject Classification for map regions and calculates the positioning and relative sizes from topic interconnections and the numbers of publications.

Acknowledgements Work on the concepts presented here has been partially supported by the Leibniz association under grant SAW-2012-FIZ_KA-2. The authors are indebted to Wolfram Sperber for the publication data for Zentralblatt Math and Patrick Ion for initial discussions and to Lars Linsen for supervision on data visualization matters.

2 Creating a Map from MSC Data

In the creation of the map we made use of the 2010 Mathematics Subject Classification [American Mathematical Society] jointly developed by the American Mathematical Society and Zentralblatt Math. The results are 63 top level classes, 528 second level classes and 5607 third level classes summing up to 6198 classes in total. Zentralblatt MATH provided us with the metadata for 3.3 million articles in mathematics.

Map Geometry The first step in map creation is to compute the geometry from the publication data. In the current incarnation, the geometry should adequately represent the relative sizes and proximities of the MSC classes, where we define the similarity of two classes as $s(i,j) = |\text{MSC}_i \cap \text{MSC}_j| / |\text{MSC}_i \cup \text{MSC}_j|$.

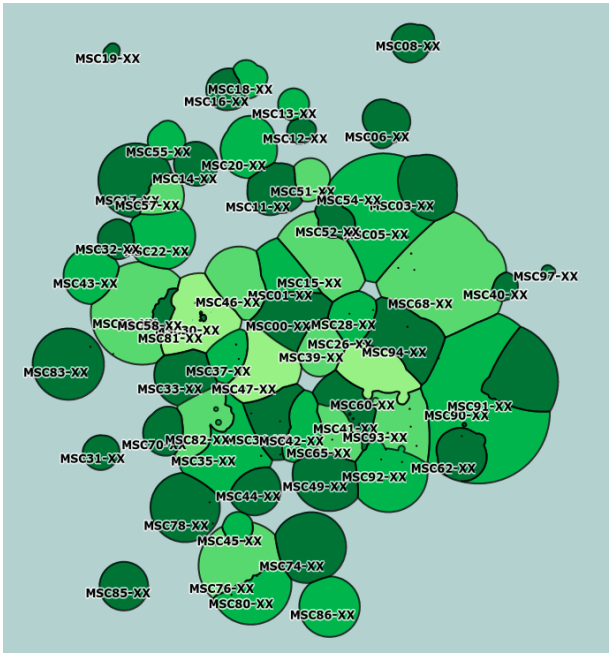


Figure 3: Geometry of the Math Subject Classifications

For the initial version of the map geometry (see Figure 3), we calculate the similarity between every pair of top-level MSCs and obtain a similarity matrix of size 63×63. We applied multidimensional scaling (MDS) to obtain two-dimensional coordinates for each MSC. Computations were executed via Matlab's `mdscale` method, which takes a $n \times n$ (dis)similarity matrix D and the target dimension p as argument and returns a $n \times p$ - sized configuration vector Y.

To visualize the size of a given MSC class in terms of "map area", we have to assign any given point in 2D space to a MSC class. We use a radial basis function whose origin is given by MDS and obtain the map geometry in figure 3.

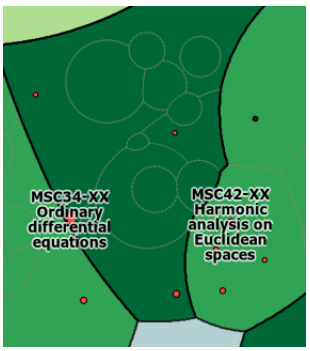


Figure 4: Adding Settlements

As the MDS computation becomes intractable for larger similarity matrices we opt for a hierarchical approach to determining finer-grained map geometries (taking second-level and leaf MSC classes into account). Here we apply the same procedure as above, but add "boundary classes" from the neighboring MSCs.

Next we populate map geometry with "cities", "towns", and "villages": we simply view every classified paper as an "inhabitant", and compute the "center of gravity" of (the MDS coordinates of) its MSC codes. As the number MSC combinations is finite, this will yields a finite number of settlements, which can be visualized by size; see the red dots in Figure 4 on the left.

Mapmaking & Deployment The next step is to convert the geometry data from the last section into a map that has the features we are used to. Note that there is no encoding of the height in Figure 3, this leaves room for visualizing additional information. We are currently experimenting with encoding the "activity level" of an area with this: We can compute the "elevation of an area" by counting the (relative) number of publications in that area e.g. in the last year. This makes research hotspots peaks that can serve as additional landmarks in the map.

Interactive Services & Mashups Having our map deployed on OpenStreetMap (OSM) already gives us some base-level interactivity: zooming, and shifting. Additional location-based interactions can be implemented by adding custom JavaScript to the pages served by OSM subject to availability of date. One immediate example is the generation of custom queries for publication databases like Zentralblatt Math [ZBMath]. Another service might be to localize mathematicians by their publication record and give them "home address" according to their primary research topic (based on the center of gravity of their publications). Similarly, research trajectories of mathematicians could be plotted on the map by computing yearly centers of gravity. Finally, we could use the math maps as a target for mashups of external services. For instance, the search results of a mathematical search engine could be shown by localizing them on the OpenMathMap service.

3 Conclusion & Future Work

We have presented a novel access method to mathematical knowledge and resources that makes use of the highly

evolved cognitive skills of spatial representations in humans. We have implemented a first prototype (http://map.mathweb.org/) that deploys maps computed from mathematical publication data in a standard map server and instruments it with information services. This prototype is just a first step we want to use in experimentation in human-oriented access methods to mathematics. We could imagine that connections between mathematical areas could be implemented as roads, highways or air/sea connections (possibly depending on their salience), important theorems could be entered/visualized as landmarks, and finally, we could imagine to go from interactive map servers to much more immersive environments (from Minecraft to second life).

Finally, we acknowledge that the motivation for the OpenMathMap project was a cognitive question, which we have answered with a technical system.

Even though first feedback from mathematicians ranged from puzzled to enthusiastic (with an emphasis on the latter), we will have to systematically evaluate whether OpenMathMap-like systems and services can help with mathematician's day-to-day navigation problems and access tasks, or if OpenMathMap is essentially a useless, but fun gadget.

References

American Mathematical Society. Mathematics Subject Classification MSC2010. Web page at http://msc2010.org. Seen November 2011.

Arif Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

Dave Rusin. The mathematical atlas, a gateway to modern mathematics. Web page at http://www.math-atlas.org/. Seen November 2009.

Zentralblatt MATH. Web page at http://www.zentralblatt-math.org/zbmath. Seen November 2012.