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). 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, & Yavuz, 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](image)

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](image)

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 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-
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

Fig. 3. Case Representation
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).
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 comparing 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>
<thead>
<tr>
<th>MAE</th>
<th>RMSE</th>
<th>RMSE-MAE</th>
<th>SMAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBR</td>
<td>15,761</td>
<td>20,423</td>
<td>4,661</td>
</tr>
<tr>
<td>Regression</td>
<td>35,018</td>
<td>42,394</td>
<td>7,375</td>
</tr>
</tbody>
</table>

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 progress 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


