Recommendation in Higher Education Using Data Mining Techniques

César Vialardi[1], Javier Bravo[2], Leila Shafti[2], Álvaro Ortigosa[2]
cvialar@correo.ulima.edu.pe, {Javier.Bravo, Leila.Shafti, Alvaro.Ortigosa}@uam.es
[1]Universidad de Lima
[2]Universidad Autónoma de Madrid

Abstract. One of the main problems faced by university students is to take the right decision in relation to their academic itinerary based on available information (for example courses, schedules, sections, classrooms and professors). In this context, this work proposes the use of a recommendation system based on data mining techniques to help students to take decisions on their academic itineraries. More specifically, it provides support for the student to better choose how many and which courses to enrol on, having as basis the experience of previous students with similar academic achievements. For this purpose, we have analyzed real data corresponding to seven years of student enrolment at the School of System Engineering at Universidad de Lima. Based on this analysis, a recommendation system was developed.

1.-Introduction

A university curriculum is generally flexible. The study program to obtain a university degree is conformed by several courses, which are distributed in academic terms. Prior to the beginning of each term, the student should enrol on one or more courses of which some are compulsory and other optional, corresponding to the period according to his/her progress; the succession of courses enrolled in each term made by a student during his/her career is called the student's academic itinerary. An academic itinerary is successful when the student, after realising his/her successive enrolments, obtains good results in each enrolled courses, allowing thus to finish him with his/her career in the exact time and with good results.

In this work the case of the School of System Engineering at Universidad de Lima was analyzed. In this institution, enrolment is done through a Web system. Even though students with better academic performance have priority on choosing groups, there are enough vacancies for all the students. In this sense, students can enrol on some or in all the courses available for his/her study plan or curriculum (a course is available for a student when he/she satisfied all the requirements for enrolment and is able to take the course).

The enrolment of a student in a course only depends on his/her decision. Previously, the student can require advice from a professor with experience, in order to know, based on his/her academic record, how many and which of the available courses he/she should enrol on. Nevertheless, students rarely require these advices from professors; most of the time the enrolment is based only on the student experience and on the information available.

However, many students do not have enough experience for taking enrolment decisions, as they do not know to associate time, effort and intellectual abilities required to successfully culminate each course. Many times the choosing criteria are closely related to the time required to finish the studies, mainly due to students' maturity. Certainly, the university offers the required quantitative information (available courses, sections, classrooms and professors), but qualitative information (implicit information regarding the experience of previous students) is lost.

Collaborative recommendation systems are agents that suggest options [4] for the user to choose among them. They are based on the idea that individuals with approximately the same profile generally select and/or prefer the same things. The systems are highly accepted and offer good outcomes for a large number of applications.

In the education environment, a recommendation system is an intelligent agent that suggests different alternatives to students, having as starting point previous actions from other students with approximately the same characteristics, such as academic performance and other personal information. It is known that before taking a course, the student have to enrol on the course; the most notorious of this process is not enrolment itself, but the previous decision that has to be taken, mainly related to how many and which course are going to be taken. In this work, we show a collaborative recommendation system based on data mining techniques [7] applied to the educational environment. The aim of this work is to offer students key elements to take better decisions in the enrolment process, using as basis the academic performance of other students with similar profiles, in order to obtain good results in each courses pertaining to its academic itinerary.

For this work we had used data of enrolments since 2002. The data is composed of demographic information of each student, enrolment in courses, grades obtained, number of courses taken at each academic term, average grade and cumulative grade per academic term. After filtering and cleaning the data, we applied the learning algorithm C 4.5[13], obtaining rules that are used for the system to suggest the student if his/her enrolment in certain course has good probabilities of success or not [17]. With this information, students will have a supporting tool that will help them taking the best decisions previous to their enrolment.

Evaluations made on the performance of the rules used by the recommender system show that they are expected to predict correctly student results in approximately 80% of the cases.

The rest of this paper is organized as follows: section 2 gives an overview of related works applying data mining in education environments. In section 3 we describe the recommendation systems and their relation with data mining techniques. In section 4 we had described data processing. In section 5 we explain the sequence of experiments required for this domain. Section 6 shows the analysis of results. Finally, section 7 outlines the conclusions and future work.

2.-Related Work

Data mining techniques are useful when huge amounts of data have to be classified and analyzed [9]. Nowadays, it is a very common situation in many scenarios, such as web information exploitation [16]. In the last years, a number of works have focused on the use of data mining techniques in the context of educational environment [14]. The most widespread techniques are: classification algorithms [3] and association rules [2]. Although the interest for using data mining in this context is growing, little work has been done regarding the use of these techniques in education.

The use of data mining is more common in educational environments based on e-Learning, for instance Educational Adaptive Hypermedia (EAH) courses. These techniques are used to discover the patterns used by students in web courses, thus helping professors and students to optimize the use of such systems. In this sense [5] these techniques support the improvement of EAH courses, applying decision tree analysis, finding the most relevant branches of the tree afterwards. These branches are presented to the professor in order to improve the course design.

Many authors have also researched the application of data mining techniques to Recommender Systems. In [1], several examples where data mining techniques are used to learn a user model (based on previous ratings) and classify unseen items are explained. Recommendation systems link users with items [15], associating the content of the recommended item or opinion of other individuals with the actions or opinions of the original users of the system. Recommendation techniques are classified in three different categories [12]: Rule-based Filtering System, Content-filtering System and Collaborative Filtering System.

Many and diverse algorithms can be applied to recommendation systems [1]. The Rule-based Filtering Systems are based on classic filtering techniques, which are information search and retrieve. Differently, collaborative filtering systems use *classification* [3], *clustering* [11], *association* [2] and *sequential patterns* [10] to discover new and interesting models that can help to suggest recommendations based on different user profiles. These systems used for educational environments has been used based on decision making have had little research or development as means to help students in taking decisions. Our system includes courses taken by the students, their grades, the registered courses in the semester, and the grade point average before registering in the course. Data mining techniques are used, in our case, as a basis for the system, since it provides precise recommendations to the students in relation to their academic performance. A similar idea is used in [6], in this case they will introduce *OrieB*, a *CRS* working in the Academic Orientation domain, to support advisors helping students of secondary school to make decisions about their academic future. *OrieB* utilizes the students' grades as input data in order to suggest their academic possibilities by providing qualitative recommendations based on the fuzzy linguistic approach.

3.-Recommendation using Data Mining

The objective of the system is to predict how convenient it is for a particular student to take a specific course, using as basis results obtained by other students with similar profile who had taken that course. To achieve this, data was organized in a table; each row represents data from a student and a course. In this way, if a certain student had taken C courses, in the table there will be C records with data about this student.

The result of the layout of the data is an *m x n* table (m records and n attributes) students-attribute; the columns have data of when the student took the course: number of courses taken simultaneously, name of course, grade obtained and accumulated grade point average (GPA) of the student until the previous semester. The class to be considered in the application of the supervised learning algorithm is the grade. This has been discretized following current norms of the institution: failure (from 0 to 10.99) and success (from 11.00 to 20).

Likewise, as the number of courses per curriculum is limited, there are not scalability and dispersion problems inherent to this type of representation in traditional collaborative filtering systems [4].

Figure 1 shows the architecture of the Recommendation System in the context of the enrolment system. Firstly, the recommendation system uses the data of the historical database of students and results obtained by them, with the goal of obtaining the rules. These rules are generated by the Pattern Discovery module. In this module, the C4.5 sub-module uses the training data (Pre-processing and Filtering prepares the training data) as an input for generating the decision tree. This tree is used by the Production Rules to generate the rules. Finally, the system provides recommendations based on these rules. An example of a set of production rules is exhibited on the left side of figure 1.

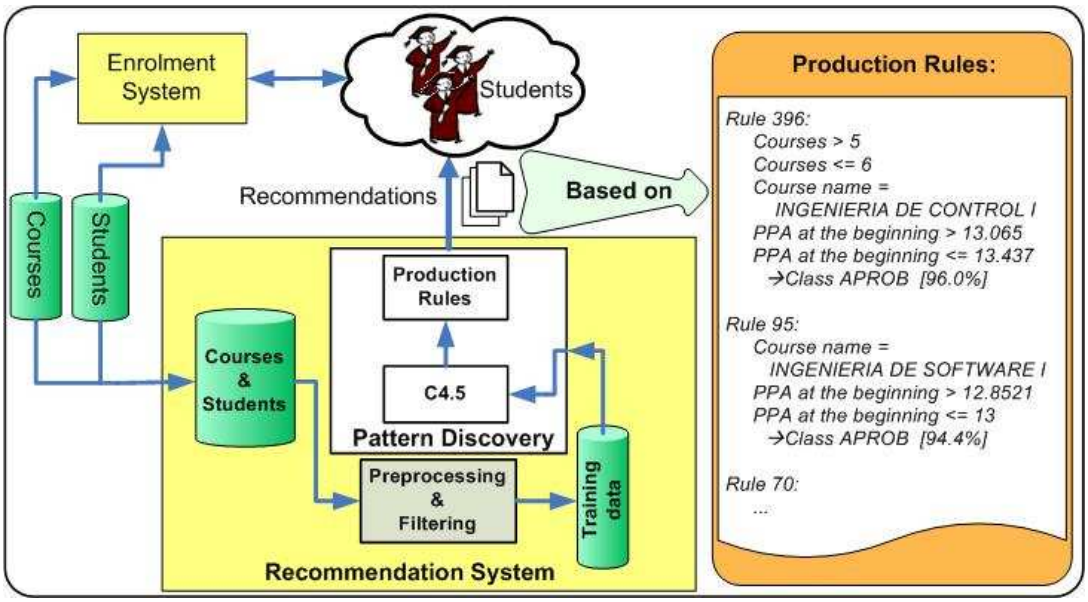

Fig. 1 Recommendation system architecture

4.- Data pre-processing

In the entire data mining process, it is of great relevance the data cleaning process in order to eliminate irrelevant items. The discovery of patterns will be only useful if the data represented in files offers a real representation of the academic performance and the actions and/or decisions taken by the student in the past [8].

Initially, the University provided us a database of 100274 records corresponding to 3230 students. After filtering process of the data 58871 records were left. The data supplied is only from students at the School of Systems Engineering, enrolled through the years 2002 to 2008.

The main objective of our research is to discover patterns that will be used to suggest positive or negative recommendations to a student previously to his/her enrolment at given course, taking as basis grades from other students with similar academic yields. In this sense, knowing the role of each attribute and the implicit relations among them, we have decided to consider that automatic learning will be performed with attributes Courses enrolled in, Name of course, Accumulative GPA starting the term, Grade

5.- Pattern extraction and evaluation

The objective is to develop a system able to predict the failure or success of a student in a course using a classifier obtained from the analysis of a set of historical data related to the academic yield of other students, who took the same course in the past.

We tried out several techniques and configurations, seeking classifiers models to optimize predictions about students' outcomes. Of these tries, two distinctive factors of the work were taken into consideration:

- The method used to learn the classifier should consider that on real situations, conditions could change from year to year and the classifier could reflect "old" patterns. For example, recommendations made on the first term of academic year 2009 will be done using historical data until 2008.

- As with any learnt classifier, it is expected to have a percentage of error in the predictions. It will be worst to recommend a student to enrol in a course that he/she will not pass, than recommend not to enrol in a course that he/she would pass.

These criteria lead the configuration of trials performed to determine patterns for a correct prediction of academic performance. Four assays were performed; following is a description of them. As it was mentioned, these assays were performed in 58871 records after filtering and cleaning the data, corresponding to 2867 students who carried out their enrolment between the years 2002 and 2008.

The first trial was the application of the algorithm C4.5[13] with the training set. This set corresponds to instances between 2002 and 2007, that is to say, 50488 instances were

analyzed. It is worth to mentioning that instances representing enrolments during 2008 (8383 instances) were taken as unseen instances. For this reason, they are used to test the accuracy of the model. The model obtained in this trial has the number of false positives significantly greater than the number of false negatives. This fact is crucial for this research since the knowledge of experts in this domain indicates that this recommender system would be useful if the false positives are lower than false negatives. In other words, the number of false positives is more important than the false negatives. Although the accuracy of the resulting model was high (more than 80%), a second trial was needed.

The second assay consisted of using the same algorithm with the same previous conditions, but using re-sampling on the training data set. On this way, the class distribution of the instances was biased towards a uniform distribution. The goal was to increases the chances of negative prediction, even if it would imply a worse performance of the classifier. Luckily the result of this assay showed that false positives were decremented and the accuracy was only 5% less than the accuracy of the previous trial.

The validation of the model was made in the third assay by executing C4.5 algorithm with a supplied experiment set (unseen instances). It is worth to noting that using enrolment data from year 2008 to test the classifier model learnt with data between 2002 and 2007 replicates how the model will be used in real situations: old data is used to predict outcomes of new students. This trial demonstrated that the accuracy of the model is enough to consider it adequate to predict the success or failure in a course by a student.

Finally, the algorithm C4.5 was applied to the entire set of data, from 2002 to 2008, using re-sampling. As a result, patterns that will be effective for the recommendation system were obtained.

6.- Analysis of Results

The analysis of the results of the assays, as well as the main statistics of the prediction model, is explained in this section.

6.1. - First assay

Using the first classifier, 41086 instances were correctly classified (81.38%), and 9402 were incorrectly classified (18.62%). The confusion matrix showed that false positives (students that failed but were classified as passing) were 7217, representing 76.76% of wrongly classified and 14.29% of total of classified students. False negatives (students that passed but were classified as failing) were 2185, representing 23.24% of wrongly classified and 4.32% of the total of classified students.

As the most important in a recommendation system is the effectiveness it achieves with users and giving the particular domain, the experts in the domain consider that in this first assay the value of representing false positives is too high in relation with the false negatives.

6.2.- Second assay

The second model classified correctly 62470 instances (75.80%). This percentage means that accuracy in this trial was lower than in the first one, but it was still adequate. The instances incorrectly classified were 19944 representing 24.20%. The confusion matrix displays that the false positives were 7262 representing 36.41% of incorrectly classified and 8.81% of total classified students. False negatives were 12682, representing 63.59% of incorrectly classified and 15.39% of total classified students.

6.3.- Third assay

The third model classified correctly 6193 instances, representing 73.9%. In this trial accuracy of the model was lower than in the two previous ones. Instances correctly classified were 2190 representing 26.1%; from the confusion matrix, it can be deduced that false positives were 435 representing 19.8% of incorrectly classified and 5.2% of total classified students. False negatives were 1755 representing 80.2% of wrongly classified and 20.4% of total classified students.

6.4.- Fourth assay:

This last trial was performed with the minimum number of items for branches in the C4.5 method set to 20. Table 6 presents a set of interesting production rules.

Table 1. Rules fourth phase results

Rule 198:	Name of course = INTEROPERABILIDAD Y ARQ. DEL SOFTWARE Accumulative GPA starting the term > 10.0256 Accumulative GPA starting the term <= 10.7428 -> class APROB [96.2%]
Rule 396	Courses enrolled in > 5 Courses enrolled in <= 6 Name of course = INGENIERIA DE CONTROL I Accumulative GPA starting the term > 13.0652 Accumulative GPA starting the term <= 13.4371 -> class APROB [96.0%]
Rule 221	Courses enrolled in > 5 Courses enrolled in <= 7 Name of course = DINAMICA DE SISTEMAS Accumulative GPA starting the term > 11.173 Accumulative GPA starting the term <= 11.7333 -> class APROB [95.3%]
Rule 95:	Name of course = INGENIERIA DE SOFTWARE I Accumulative GPA starting the term > 12.8521 Accumulative GPA starting the term <= 13 -> class APROB [94.4%]

7. - Conclusions and future work

In this section some of the main conclusions and contributions of the work are summarized, and some possible future development lines are commented.
As it has been emphasized, the most important point of this research is the acquisition of knowledge from students' academic performance. The main purpose is to provide support

for new students in order they can choose better academic itineraries. Recommendation systems are familiar to this process; this is the technique used in the recommendation engine that stresses accuracy and effectiveness. In this sense, the research has focused in testing, using real data, the application of techniques and tools in data mining to support students when enrolling on a new term. To achieve this goal, we had analyzed real data from students that had taken the same courses in the past, using techniques such as decision trees to discovery trends, patterns and rules that will be used as support for decision taking in the itinerary of a certain university career.

It was found that using data mining, it is possible to develop a model representing the behaviour of students in their way through different academic itineraries. This facilitates a proper vision of the behaviour and performance of the group of students at certain university career and, at the same time, allows feeding the system to offer recommendations for students to increase their effectiveness and relevance at decision taking in relation with the courses to be enrolling on.

We presented four assays by using the same technique with four different configurations, in order to show the advantages and disadvantages of the technique used as well as to detect classifiers that would perform better in real settings. The first trial was a classic 10 fold cross validation over 50488 instances corresponding to enrolments from 2002 to 2007.

However, the same technique was also applied with other setting, penalizing more the error produced by false positives giving more weight to the data corresponding to instances whose class was *failure*. Assay results show that global accuracy was 75.80%. Although global accuracy is lower at this second trial than in the first one, the objective was accomplished, as the number of false positives decreased in relation to false negatives, assuring its effectiveness.
Due to the success of the second assay, a third essay was carried out. The main objective was that the system learns with the record of enrolments made from 2002 to 2007 (training set), considering the set of records representing enrolment in 2008 as the experiment set (unseen set). In this way, we simulated the fact of implementing and testing the recommendation system for students enrolled in 2008.

Finally, a fourth assay was developed, by applying the learning algorithm to all the instances, obtaining 77.3% of global accuracy. The global accuracy of this last trial was greater than the second and the third ones.

It is meaningful to emphasise that assay presented in this work, besides using real data, shows that used tools are very powerful techniques, but also they have weak points when looking for patterns in a domain of knowledge with participation of human factor and has direct relevant consequences in the data.

In addition, pattern detection offers two types of information. On one hand, the student can infer that system's recommendation is related with his/her global academic performance or to certain courses. Therefore, the student could freely decide, taking into consideration more subjective factors, to enrol or not. For the university the system has

relevant information related to students' academic results in one or several courses, information that would not be available applying descriptive statistical techniques. It is worth to mentioning that analyzed information could be used as input in an eventual curriculum modification.

In this way, the main objective of this work is to support students through recommendations, in the complex process of deciding how many and which courses enrol on, taking into consideration academic performance and other similar characteristics. But a second important benefit is that it can offer useful information to meaningfully improve academic performance of students, using as basis a good study plan for the academic period.

Even though the obtained results had been satisfactory, it is necessary to mention that for future works it will be necessary to test the system with data from other faculties in order to know consistency and convergence. We want to establish effectiveness thresholds in the use of these techniques to obtain more and better outcomes in the application of data mining techniques for recommendation systems in this domain of application.

Other important aspect is to obtain the relation between the different courses. Currently, we are working to obtain (from data used in this research) patterns that relate courses with students' academic performance. This type of information will eventually represent an additional element to improve the recommendations offered by the system.

Acknowledgements

This work has been funded by Spanish Ministry of Science and Education through the HADA project TIN2007-64718. César Vialardi is also funded by Fundación Carolina. We would like to thank Eduardo Pérez and Jorge Chue who contributed in this work with their helpful comments.

References

[1] Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. *ACM Transactions on information Systems,* 2005, 23(1), p. 103-145.

[2] Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. *ACM SIGMOD Conference on Management of Data*, 1993, p. 207-216.

[3] Arabie, P., Hubert, J., and De Soete, G. Clustering and Classification. (Eds.) *World Scientific Publishers Company*, 1996. London.

[4] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Item – Based Collaborative Filtering Recommendation Algorithms, *Proceedings of the 10th international conference on World Wide Web*, 2001, p. 285-295.

[5] Bravo, J., Vialardi, C., and Ortigosa, A. A Problem-Oriented Method for Supporting AEH Authors through Data Mining. *Proc. of International Workshop on Applying Data Mining in E-learning (ADML'07) held at the Second European Conference on Technology Enhanced Learning (EC-TEL 2007)*, 2007, p. 53-62.

[6] Castellano, E. and Martínez, L. ORIEB, A CRS for Academic Orientation Using Qualitative Assessments. *Proceedings of the IADIS International Conference E-Learning*, 2008, p 38-42.

[7] Chen, M., Han. J., and Yu, P. Data mining: an overview from Databases Perpective. *IEEE Transaction on Knowledge and Data Engineering*, 1996, p. 833-866.

[8] Cooley, R., Mobasher, B., and Srivasta, J. Data Preparation For Mining Word Wide Web *Browsing Patterns. Knowledge and Information Systems*, 1999, 1(1), p. 5-32.

[9] Fayyad, U., Piatetsky-Shapiri, G., and Smyth, P. From Data mining to knowledge Discovery in Databases. *AAAI*, 1997, p. 37-54.

[10] Han, J., Pei, J., and Yan, X. Sequential Pattern Mining by Pattern-Growth: Principles and Extensions. *Series in studies in Fuzziness and soft computing*, 2005, 180, p.183-220.

[11] Jain, A.K., Murty, M.N., and Flynn, P.J. Data Clustering. *A Review. ACM Computing Surveys*. 1999, 31(3), p. 264-323.

[12] Mobasher, B. Data Mining for Personalization. The Adaptive Web: Methods and Strategies of Web Personalization, Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.). *Springer-Verlag*, 2007. Berlin Heidelberg, p. 1-46.

[13] Quinlan, J.R. C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*, California, USA, 1993.

[14] Romero, C. and Ventura, S. Educational Data Mining: a Survey from 1995 to 2005. *Expert Systems with Applications*, 2007, 33(1), p.135-146.

[15] Schafer, J.B. The application of data-mining to recommender systems. J. Wang (Eds.), *Encyclopedia of data warehousing and mining*, 2005 Hershey, PA. Idea Group, p. 44-48.

[16] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*, 2000, 1(2), p. 12-23.

[17] Superby, J.F., Vandamme, J.P., and Meskens, N. Determination of Factors Influencing the Achievement of the First-year University Students using Data Mining Methods. *Workshop on Educational Data Mining*, 2006, p. 37-44.