December 2006

Knowledge Discovery in Academic Registrar Data Bases using Source Mining: Data and Text

Ma. Teresa Rios
Tecnológico de Monterrey

Francisco J. Cantu
Tecnológico de Monterrey

Knowledge Discovery in Academic Registrar Data Bases using Source Mining: Data and Text

Ma. Teresa Rios-Quezada
Tecnológico de Monterrey, Academic Affairs
riostere@hotmail.com

Francisco J. Cantu-Ortiz
Tecnológico de Monterrey, Campus Monterrey
fcantu@itesm.mx

ABSTRACT

In this paper we describe a knowledge-based system for extracting knowledge from academic and registrar databases using *source mining*, where the sources are data or text. Other sources not included in this research are image, sound or gestures. Patterns of student behaviour were obtained by examining data from student attributes such as city of birth, scholarship needs, field of knowledge and major, student gender and other attributes, for various undergraduate academic programs offered by the campus of Tecnológico de Monterrey university system across the country. These patterns proved useful in predicting student enrolment and designing advertising campaigns. We use text mining techniques to match and compare course description from universities with which we have student ex-change programs for course revalidation. Equivalence of 64 courses were obtained by using text mining techniques for matching course descriptions for universities like Michigan State, Carnegie Mellon and New Mexico State. These equivalences helped the International Programs Office in developing course revalidation. Data mining techniques employed include C4.5 decision tree learning and feed-forward neural networks as implemented in the *SIPINA* intelligent environment (Sipina Research). Text mining techniques utilized are based on statistical and syntactic-semantic analysis and include Clasi-Tex (Clasitex), IBM Intelligent Miner for Text (IntelligentMiner), Text Roller (TextRoller), and Free Text Technologies Master Text

Keywords

Text mining, data mining, source mining, knowledge discovery and registrar databases

INTRODUCTION

Student enrollment is an important task for universities, colleges and schools. There are offices at universities devoted to attracting the best students and for this purpose they design campaigns and financial aid programs. In addition, agencies such as the US News and World report produce annual reports of rankings of the best colleges and schools in the USA. These rankings are revised by students and their parents to make the best decision in going to college. In this paper we describe an action research case study called *source mining* where the source is either a database or a repository of documents. Other sources not included in this study are images, sounds or gestures. The goal of the data mining side is to learn about student behavior in selecting a major and in scholarship assignation to attract the students with the highest potential. For the text mining side, since our university runs exchange programs with more than 300 universities from North America and Europe, we need to decide in an informed manner when a course taken at an associate university is equivalent to a course within the students' curriculum. For this comparison we developed a study to find course similarities using text mining techniques (Rios-Quezada, 2003). This paper is organized as follows: first, we present the background and related work then we describe the data mining case study and analyze the results, next we explain the text mining case study and discuss the results, finally, we summarized the utility of theses studies and present the conclusions.

BACKGROUND

Knowledge discovery in data bases (KDD) is multi-disciplinary and active field of research in machine learning, statistical inference, natural language understanding, database systems and pattern recognition. Applications of these techniques are widespread in broadcast companies, banks, financial, manufacturing, and service firms (Berry and Linoff, 1997). Those companies use KDD technologies for conducting business intelligence, market analysis, customer relationship management (CRM) and user profiling among other applications Berry and Linoff, 1999). Colleges and universities are a kind of institution that provide education services to a base of customers including potential students, enrolled students and their parents. Learning about the behavior of those customers is useful for attracting students, managing the scholarship program and retaining them at the university. Also, exchange programs among universities are a feature that is offered to students in order to provide international experiences. Establishing equivalences among course taken in different places is a problem for the registrar offices, especially when the language spoke is not English. Text mining and text comparison technique are a useful aid in establishing course equivalences for course revalidation (Faloustos et al., 2000; Nasukawa and Nagano, 2001).

LEARNING STUDENT BEHAVIOR BY MINING REGISTRAR DATABASES

First, we present a case in knowledge discovery in data bases for learning student behavior by applying data mining techniques to registrar databases. We present the problem definition, the data mining methodology that was utilized and then we analyze the results.

Problem definition

Tecnológico de Monterrey is a university system of 28 campuses in different cities of the country. It offers undergraduate and graduate degrees in areas of engineering, information technologies, business management, humanities and social sciences and medicine. (ITESM, 2006). This study focuses on the engineering undergraduate programs of the various campuses.

The main research questions of this study is the following: *Are there patterns in the engineering student population of the various campuses that can help in attracting students from geographical areas, designing scholarship programs as well as international ex-change programs?. Is the gender of the student population a factor to be considered in designing campaigns for student enrollment?*

Research methodology

The research methodology employed to answer these questions is based on the data mining life cycle: (1) Identification of the business problem, (2) Preparation of data, (3) Model selection, (4) knowledge extraction, and (5) Analysis of results (Nikhil and Lakhmi, 2005; Robles et al, 2005). We have already identified and defined the problem. In the following sections we describe the processes for the data preparation, model selection, knowledge extraction and analysis of results.

Preparation of data

The Registrar office provided data bases of 14 undergraduate engineering majors from 26 campuses for years 1998 to 2001. Special attention was paid to the *systems and industrial engineering major* which is one of the most populated programs and is offered by all of the campuses. There are dozens of attributes for these programs in the registrar data bases. A data mining table was prepared in which the columns are the attributes of interest which are grouped in the following categories: Student enrollment per campus and major, students with a type of scholarship, student gender per campus and major, students participating in international exchange programs, freshmen and non-freshmen students per campus and major, out-of state and term. Some of these attributes are discrete and categorical and some are numerical. This is important in choosing the modeling technique for knowledge extraction. The data mining table was prepared using statistical analysis techniques. For instance, figure 1 shows the distribution of students with a scholarship for each of the campuses between 1998 and 2000; figure 2 displays the distribution of students participating in international exchange programs per campus between 1998 and 2001; figure 3 presents the distribution of out-of-state students per campus between 1998 and 2001.

The data mining table contains 17 thousand registers with one register row per student. The prediction variables are engineering major, gender, campus, in-state, scholarship, year of study, and participation in international exchange programs.

Model selection

An analysis of the attributes as well as of the prediction variables was done in order to select the data mining modeling techniques. Decision trees and neural networks were considered appropriate techniques for data modeling. A plan of experiments was designed and executed for cleaning the data and tuning the parameters of the algorithms. Various data mining tools were analyzed. After many trials, the SIPINA RESEARCH data mining tool now called TANAGRA, was chosen for the knowledge extraction process. SIPINA implements decision trees and neural networks (SIPINA, 2006).

Figure 1. Student distribution of scholarships per campus from 1998 to 2000

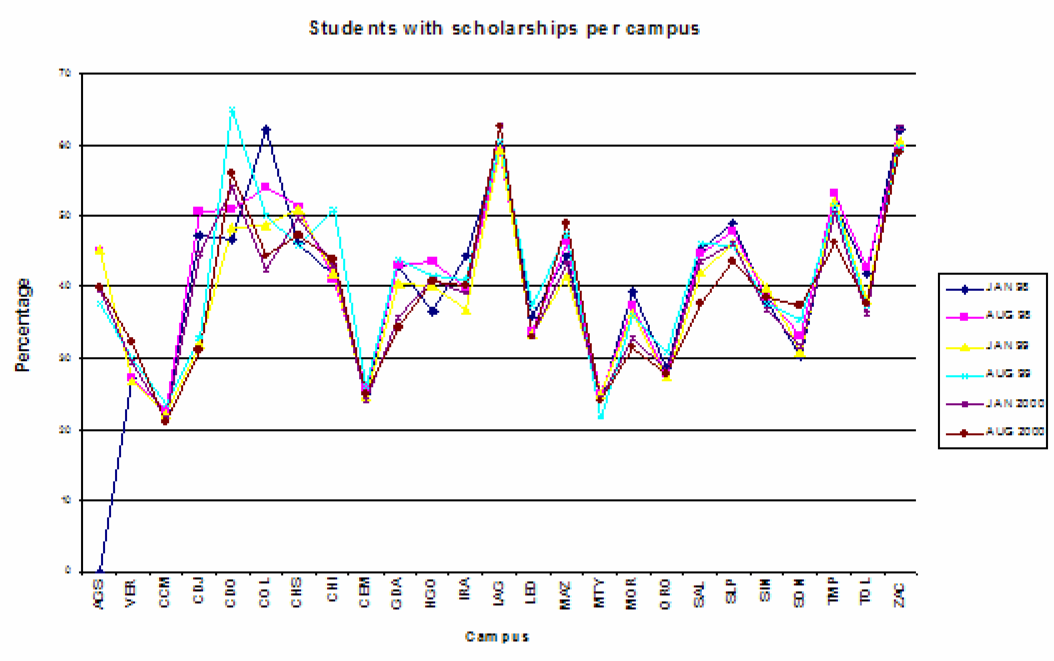


Figure 2. Student distribution of international exchange programs from 1998 to 2001

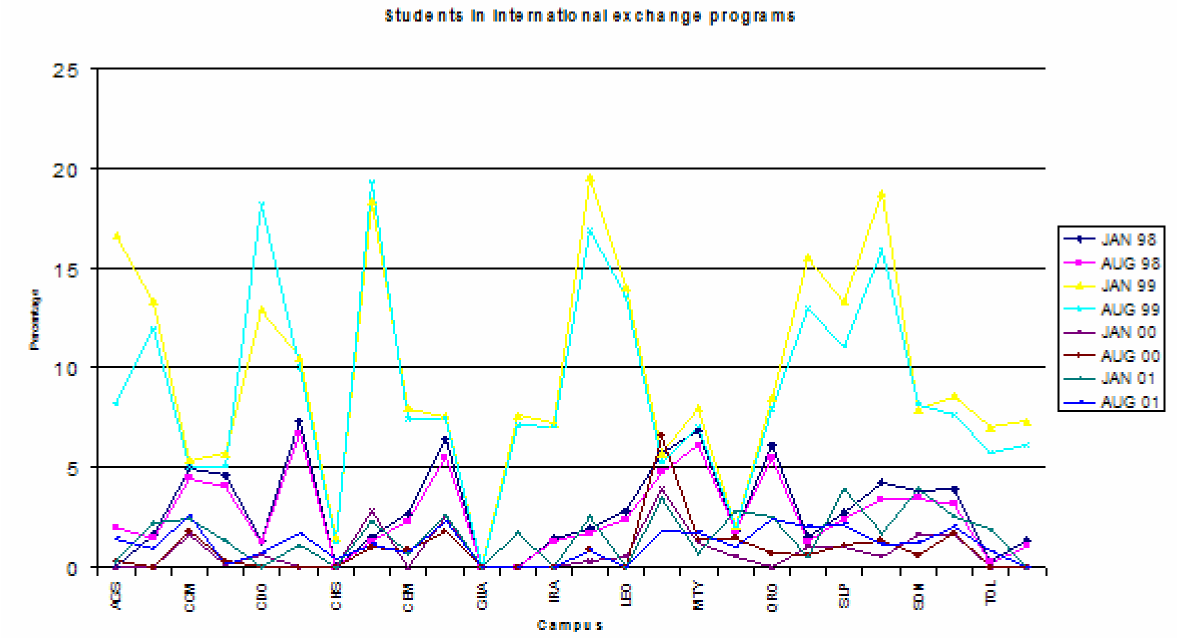
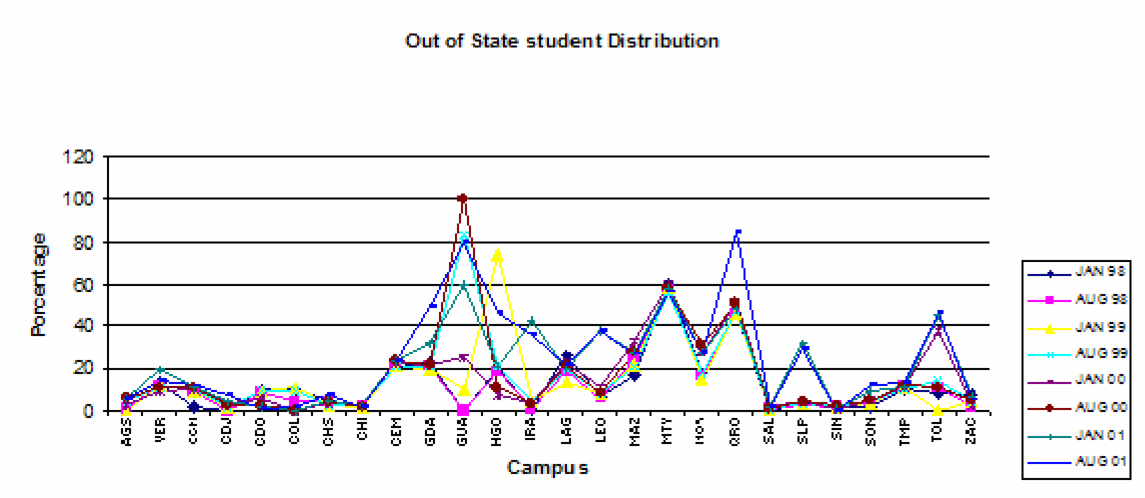


Figure 3. Out-of-state student distribution



Knowledge extraction

A set of experiments was planned and executed using decision trees and the C4.5 entropy-based algorithm. Levels of confidence were set at different percentages. The data mining table was split in 60% for training data, 30% for test data and 10% for validation. Since term is a time variable it was modeled as a time series. The output was a set of decision rules in the format if (conditions) then (conclusions). Figure 4 shows a set of rules produced by SIPINA for some of the experiments. The format of the rules has been edited to improve readability.

Figure 4. Type of rules generated by the decision tree

If the Campus population is greater than 5,000 students then 55% of the students are out-of-state

If the Campus population is less than 4,000 students then at least 95% of the student population is local (in-state)

If the Campus population is less than 4,000 students then at least 40% of the student population owns a scholarship

If the student population is local then at least 40% of the students own a scholarship

Another set of experiments was planned and executed using neural networks. The prediction variables were the ones found by the decision tree experiments as the ones with the most predictive power. The neural network is of type feed-forward with back-propagation and hidden layers. Experiments were done in which the number of hidden layers varied between 1 and 10 and the number of variables per hidden layer was also between 2 and 10. Figure 5 displays the parameters of one of the experiments with neural networks. The weights of the neural net were the default values of generated by SIPINA.

Figure 5. Neural network parameters

Method: Multilayer Perceptron

Maximum number of iterations: 5000

Maximum error: 0.05

Number of hidden layers: between 1 and 10

Number of nodes per hidden layer: between 5 and 10

Attributes: Term, Campus, student population in civil engineering, student population in mechanical and electrical engineering

Analysis of Results

The main results of the data mining analysis are new knowledge about the student behavior that is useful in designing campaigns for student attraction and scholarship assignation. The main results can be summarized as follows:

- Scholarship programs should be oriented for male or female engineering students in disciplines like chemistry, civil, mechanical or electrical engineering
- Scholarship programs are a strategy of small campus for attracting good local students
- Out-of-state students are mainly males with a type of scholarship in hard disciplines
- Female students prefer majoring in architecture and industrial design. Scholarships and local campuses attract higher numbers of these students
- Systems and Industrial Engineering students save the elective courses of their curriculum for international exchange in the areas of humanities, agricultural and business engineering

This is just a sample of the type of knowledge that was found in analysis the abundant results that came out from the data mining experiments.

TEXT MINING OF COURSE DESCRIPTION FOR REVALIDATION IN STUDENT EXCHANGE

We now present a case in text mining of course description for course revalidating in student exchange. We present the problem definition, the text mining methodology that was utilized and then we analyze the results.

Problem definition

Tecnológico de Monterrey keeps international exchange agreements with more that 300 universities worldwide of which more than 100 are active every year, especially in the United States, Canada, Europe and the Far East. There is a set of elective courses for he student to revalidate as a minor in a predefined discipline. International exchange program is a way of revalidating those courses. This study focuses on the exchange programs that Tecnológico de Monterrey holds with three USA universities: Michigan State University (MSU), Carnegie Mellon University (CMU) and New Mexico State University (NMSU). The majors chosen are the undergraduate programs in business management, international business, marketing and systems-industrial engineering. These are the four majors with the highest number of students in exchange programs.

The problem can be stated as follows: *Given a set of courses that a Tec's students takes abroad, how do we know that those courses are equivalent to the courses the student must have taken to complete a predefined minor?*

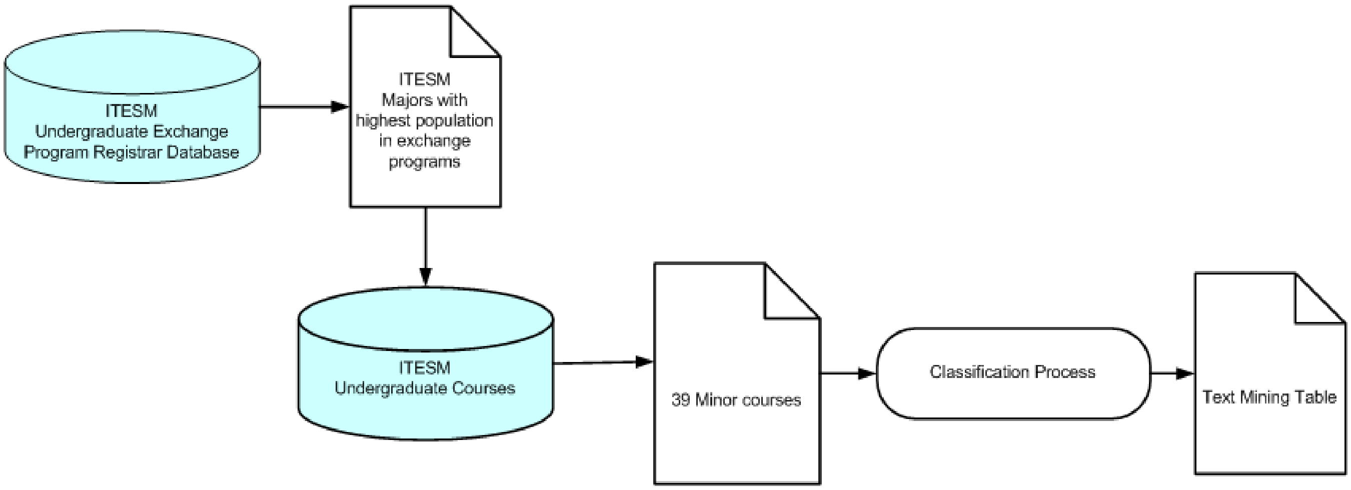
Research methodology

The research methodology employed to answer this question is based on concepts of texts mining and natural language processing and includes the following steps: (1) identification of the business problem, (2) preparation of data, (3) model selection, (4) knowledge processing and (5) analysis of results (Yeates et al, 1998; Guzman A, 1998). We have already identified and defined the problem. In the following sections we describe the processes for data preparation, model selection, knowledge processing and analysis of results.

Data preparation

We prepared a data base of 39 minor courses from the four undergraduate majors chosen for this study, namely business management, international business, marketing and systems-industrial engineering. These courses were classified by academic department. We obtained another data base of minor courses of the abroad universities from their Web pages. Figure 6 outlines the main data preparation steps previous to text mining.

Figure 6. Text mining



Model selection

We did a search for models, algorithms and tools for text processing, semantic analysis and text classification (Matheus C. et al, 1996; Piatetsky-Shapiro and Matheus, 1994). For text processing and semantic analysis we used the models of IBM's *Intelligent Miner for Text*. For text mining and text classification we used the tools *Text Roller*, and *Master Text*. We also studied statistical text processing tools such as *ClasiTex* (Guzman A. 1998).

Knowledge Processing

Once we had the text mining table with the course description and the models and tools for text mining the next step is the application of these tools to the text mining table for knowledge processing and extraction. This process is summarized in the following steps:

(1) For each course description in the text mining table, find the main theme of that course. The output of this step is stored in a course themes table.

(2) For each pair of course themes, one from Tecnológico de Monterrey and one from abroad, find equivalences by comparing the themes

(3) Repeat step 2 for all pair of local and abroad courses

The result of step 1 is a table of course themes produced by the tool Text Roller. Table 1 displays the main themes for a set of minor courses of the industrial engineering undergraduate major. The last column lists the minor's set of courses. The results of steps 2 and 3 are a table of course equivalences obtained by using the tool Master Text in a semi-automatic way. This means that user validation is required to guide the equivalence checking. Table 2 displays an example of course equivalences for the courses between Tecnológico de Monterrey and NMSU.

Analysis of Results

The main result of the text mining analysis is comparison of minor courses between a local and a foreign university for course revalidation. The main results can be summarized as follows:

- Three USA universities with which Tecnológico de Monterrey has exchange agreement were included in the study: CMU, NMSU and MSU
- Four undergraduate majors with the most number of exchange students we included in the study: business management, international business, marketing and systems-industrial engineering.
- A total of 39 minor courses from various departments taken by exchange students were included in the study

Major	Course	Minor course	Minor courses
Industrial engineering	Administración de la producción I	The role of inventory systems. Inventory role-models according to demand.	Total Quality Management
Industrial engineering	Administración de la producción II	Strategy and management of goods and/or services. Resource planning of goods and/or services. Operations analysis of goods and/or services. Manufacturing management.	Total Quality Management
Industrial engineering	Administración de proyectos	Project Administration	None
Industrial engineering	Diseño de sistemas	Principles and characteristics of systems thinking. Mechanics of systems thinking. Mathematical modeling. Systems thinking modeling. System archetypes. Sensibility analysis	Systems
Industrial engineering	Factibilidad de proyectos	Project feasibility	None
Industrial engineering	Investigación de operaciones III	Support systems for product planning. Support systems for production scheduling. Support systems for production control.	Total Quality Management
Industrial engineering	Laboratorio de producción	Measuring operations. Process analysis and design.	Total Quality Management
Industrial engineering	Laboratorio de sistemas integrados de manufactura	Manufacturing planning and control through AMNET.	Total Quality Management
Industrial engineering	Metodología de sistemas	Diagnosis of soft systems. Planning and control in human systems	None
Industrial engineering	Modelación estructural de sistemas	Introduction to decision-making. The modeling process.	None
Industrial engineering	Planeación de plantas industriales	Strategic facility planning. Product management.	Total Quality Management
Industrial engineering	Sistemas de calidad	Total quality management for improving performance.	Total Quality Management
Industrial engineering	Sistemas de planeación	Planning	None
Industrial engineering	Sistemas integrados de manufactura	Manufacturing systems. Group technology. Scheduling of manufacturing operations. Manufacturing control. Numeric control.	Total Quality Management
Industrial engineering	Evaluación de proyectos	Engineering economy.	None

Table 1. Minor courses for Industrial Engineering Major

- Existing text mining tools facilitated the data preparation and knowledge extraction processes.
- Course equivalences were found for the three universities and the four majors.
- The knowledge extracted, that is, course equivalences proved useful for international programs and registry and academic affair offices.

Major	Local course	Main theme	Carnegie Mellon course
International business	Análisis de la competitividad internacional	Development of strategies of international commercialization from the point of view of the exporting company	70365 International Trade and International Law 70342 Managing Across Cultures
International business	Estrategias de comercialización internacional		
International business	Proyecto de comercialización Internacional		
International business	Seminario integrador de comercio internacional		
International business and Business management	Logística empresarial	Productive/operational manufacturing Processes and services in a corporation.	70371 Production I
Business management	Negociaciones internacionales	International negociations	70365 International Trade and International Law 70342 Managing Across Cultures
Business management	Seminario de administración estratégica	Strategic Management	70440 Business Leadership & Strategy
International business and Business management	Administración electrónica de negocios (e-business)	Information systems, e-commerce	70451 Management Information Systems
Industrial engineering	Factibilidad de proyectos	Project feasibility	46531 Managerial Economics
Industrial engineering	Laboratorio de producción	Measuring operations. Process analysis and design.	39405 Engineering Design: The Creation of Products and Process
Industrial engineering	Laboratorio de sistemas integrados de manufactura	Manufacturing planning and control through AMNET.	70371 Production I
Industrial engineering	Metodología de sistemas	Diagnosis of soft systems. Planning and control in human systems	70311 Organizational Behavior
Industrial engineering	Planeación de plantas industriales	Strategic facility planning. Product management.	39405 Engineering Design: The Creation of Products and Process
Industrial engineering	Proyectos de ingeniería	Project engineering	46531 Managerial Economics
Industrial engineering	Sistemas de información	Information system development	70451 Management Information Systems 88350 Computational Modeling of Organizational Technology and Society
IIS, LIN, LAE	Evaluación de proyectos	Engineering economy.	46531 Managerial Economics

Table 2. Course equivalences for Tecnológico and Carnegie Mellon

RELATED WORK

There is abundant research and applications in data and text mining (Nikhil and Lakhmi, 2005). Some works have been reported in the use of data mining for university cases like learning about student (Sanchez et al., 2005), targeting best students (Yiming Ba, et al 2000), monitoring students behavior in order to get a grade predictive model (Mierle and Keir, 2005), improving teaching techniques (Spacco, Jaime, 2005) or support decision making (Tarapanoff, Kira, 2001, Janson Luke, 2004). However, the data mining case reported in this work focuses on student profiling and behavior discovery for promotion campaign and scholarship management. To the best of our knowledge this is one of the first works on the use of text mining for the comparison of course descriptions for course accreditation in a university environment (Rios, 2003).

CONCLUSION

We have presented research results of what we call *source mining*, whereby the source of the mining process may be either data, text, image, sound or other type of source. In this paper we have focused on data and text mining describing two research projects in registrar and academic data bases in a university environment. The knowledge found with both techniques proved useful in knowing about student behavior for campaign design and in finding equivalences in courses taken by students in abroad universities and show the advantages of using knowledge extraction technology for competitiveness and business intelligence decisions.

ACKNOWLEDGEMENTS

The acknowledge the support of Tecnológico de Monterrey and the Vice Rectory for Academic Affairs, The International Programs Office, the Registrar Office and their Officers Hilda Cruz and Barbara Valle. We also thank the Graduate Program on Information Technology, the Center for Artificial Intelligence and the Data Mining for Business Intelligence Research Group.

REFERENCES

1. Berry, Michael J.A., Linoff Gordon, *Data Mining Techniques for Marketing, Sales and Customer Support*, John Wiley & Sons, EEUU,1997
2. Berry, Michael J.A., Linoff Gordon, Mastering data Mining: The Art and Science of Customer Relationship Management, John Wiley & Sons, EEUU, 1999.
3. Carnegie Mellon University, Pittsburgh Pennsilvania, USA. Schedule of Classes, Fall 2002, https://acis.as.cmu.edu/gale2/open/Schedule/SOCServlet?Formname=ByDept
4. Clasitex. Guzmán Adolfo, *Clasitex++: Una herramienta para el análisis inteligente de textos*, http://www.cic.ipn.mx/~aguzman/Public/clasitex.html
5. Faloustos, C., Gibson G., Mitchell, T, Moore A., Thrun, S., *Data Minig at CALD-CMU:Tools, Experiences and Research Directions*, Center for Automated Learning and Discovery (CALD), Carnegie Mellon University, EEUU, Summer 2000
6. Free Text Technologies ,http://www.insight.com.ru/
7. Guzmán Adolfo, *Finding the main themes in a Spanish document*, in Expert Systems with Applications 14 (1998) pp. 138-149, Elsevier, 1998
8. Intelligent Miner for Text, http://wwww-4.ibm.com/software/dataiminer/fortext/
9. ITESM: Tecnológico de Monterrey http://www.itesm.mx, 2006
10. Janson Luke, Ong Wai Kit, Vijanth S. Asirvadam, and Suresh K. Krishnan, *Data Mining Approach on Nationwide School Exam System*, Proceedings of the 2004 International Conference on Information and Communication Technologies (ICT 2004), November 18-19, published by Assumption University, Thailand , 2004.
11. Kroeze Jan H., Matthee Machdel C., Bothma Theo J. D., *Differentiating data- and text-mining terminology* In Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology SAICSIT , 93-101, September 03

12. Matheus Christopher J., Piatetsky_Shapiro, Gregory, McNeill Dwight, *Selecting and Reporting What Is Interesting*, en Advances in Knowledge Discovery and Data Mining, pp. 495-515, MIT Press, EEUU, 1996.
13. Michigan State University, Archived Description of Courses, 2001-2002, http://www.reg.msu.edu/Read/DescCourses/2001-02.asp
14. Mierle Keir, Laven Kevin, Roweis Sam, Wilson Greg, *Mining Software Repositories (MSR): Mining student CVS repositories for performance indicators* , Proceedings of the 2005 international workshop on Mining software repositories MSR '05, Volume 30 Issue 4 , ACM SIGSOFT Software Engineering Notes , May 2005
15. Nasukawa T. and T. Nagano, T., *Text analysis and knowledge mining system,* IBM Systems Journal, volume 40, number 4, 2001
16. New Mexico state University, Course Descriptions, http://www.nmsu.edu/aggieland/majors/courses.html
17. Piatetsky-Shapiro, G. Matheus, C.J., *The Interestigness of Deviations*, in Proceedings of the 1994 Knowledge Discovery in databases Workshop, ed. U. Fayyad and R. Uthurusamy, Tech. Report WS-94-03, Menlo Park, Calif, AAAI Press, 1994
18. Nikhil R. Pal, Lakhmi Jain (eds). Advanced Techniques in Knowledge Discovery and Data Mining, Springer, 2005
19. Rios-Quezada, M.T., Hallazgo de conocimiento en bases de datos escolares utilizando herramientas de minería de texto y datos, Tecnológico de Monterrey, Master Thesis, 2003
20. Robles, A., Cantu, F.J., Morales, R. A Bayesian Reasoning Framework for On-line Business Information Systems. Proceedings of the Eleventh Americas Conference on Information Systems, Omaha, NE, USA August 11th-14th, 2005
21. Sanchez A., Gutierrez M., and Meneses C. Using Data Mining to Support University Decision Process: A Case in a Chilean University. Proceedings of the Eleventh Americas Conference on Information Systems, Omaha, NE, USA August 11th-14th, 2005
22. Sipina Research, http://eric.univ-lyon2.fr/~ricco/sipina.html, 2006
23. Spacco Jaime, Strecker Jaymie, Hovemeyer David, Pugh William, Mining Software Repositories (MSR): Software repository mining with Marmoset: an automated programming project snapshot and testing system, , Proceedings of the 2005 international workshop on Mining software repositories MSR '05, Volume 30 Issue 4 , ACM SIGSOFT Software Engineering Notes , Publisher: ACM Press, May 2005
24. Stuart A. Yeates, David Bainbridge, and Ian Witten, *Using Compression to Identify Acronyms in Text*. In James A Storer and Martin Cohn editors, Proceedings of the Data Compression Conference (DCC), Snowbird, UTAH, 28-30 March 2000
25. Tarapanoff, Kira, Quoniam Luc , de Araújo Júnior Rogério Henrique, Alvares Lillian, *Intelligence obtained by applying data mining to a database of French theses on the subject of Brazil*, Information Research, Vol. 7 No. 1, October 2001.
26. Yiming Ma, Bing Liu, Ching Kian Wong, Philip S. Yu, Shuik Ming Lee , *Targeting the right students using data mining* Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Publisher: ACM Pres, August 2000.