Issues in Informing Science and Information Technology Volume 6, 2009

Finding Diamonds in Data: Reflections on Teaching Data Mining from the Coal Face

Shana R. Ponelis
School of Information Studies, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

ponelis@uwm.edu

Abstract

Making sense of the exponentially expanding sources of structured electronic data collected by organizations is increasingly difficult. Data mining is the extraction of implicit, previously unknown, and potentially useful information from large volumes of such data to support decision-making in organizations and has led to an increase in demand for students who have an understanding of data mining techniques and can apply them to organizations' data. Thus data mining is an increasingly important component of the Information Systems curriculum in order to meet this skills demand. This paper describes the development of a curriculum for an elective data mining course in an Information Systems graduate program based on the only available model curriculum from the ACM SIGKDD over a two year period and concludes with student feedback and lecturer reflection. This paper will be useful to educators responsible for developing curricula and teaching data mining to IS graduate students; in addition, it serves as instructor feedback to the authors of the ACM SIGKDD model curriculum.

Keywords: data mining, education, curriculum, curriculum development, graduate, postgraduate.

Introduction

Data that has relevance for managerial decisions are accumulating at an incredible rate due to a host of technological advances: electronic data capture has become inexpensive and ubiquitous as a by-product of innovations such as the internet, e-commerce, electronic banking, point-of-sale devices, bar-code readers, and intelligent machines. Such data is often stored in data warehouses and data marts specifically intended for management decision support with supporting theory and techniques to manage such large volumes of data. Due to this explosion of data, many decision makers have realized the importance of techniques, such as data mining, that produce timely, relevant information to enable informed decision-making. As a result, data mining is an increasingly important component of the Information Systems curriculum in order to meet this skills demand.

This aim of this paper is to describe how to teach students to find diamonds in data: the development of a curriculum for an elective data mining course in an Information Systems graduate program based on model curricula over a two year period. The paper is structured in the following manner: first, data mining is defined. Second, the model curricula that pertain to Information Systems in general and data mining in particular are examined. Thereafter, the curriculum development of the elective graduate data mining course at the University of Pretoria is discussed in detail. The paper concludes with the evaluation from the students and the reflection from the lecturers.

Data Mining

Data mining is a rapidly growing field that is concerned with developing techniques to assist managers to intelligently exploit these large volumes of data. According to Chakrabarti et al. (2006, p. 2) the central endeavor in data mining is to "extract knowledge from data" with this knowledge being captured "in a human-understandable structure." This extraction of knowledge from the data is done through the exploration and analysis by automatic means of large quantities of data to discover actionable patterns and rules which is what must be captured in a form that humans can understand. The University of North Texas defines data mining as follows in its course glossary (http://www.coba.unt.edu/itds/courses/dsci4520/DMglossary.htm#anchor314309) to be "an information extraction activity whose goal is to discover hidden facts contained in databases." Furthermore, data mining finds "patterns and subtle relationships in data and infers rules that allow the prediction of future results [own emphasis]."

Data mining, therefore, will automatically detect previously unknown patterns and rules in large volumes of data that can be extrapolated to future events in a format that can be understood by humans in order to support decision-making. It offers a way to use massive quantities of data that businesses generate with the goal of improving marketing, sales, customer support through better understanding of customers. A number of successful applications have been reported in areas such as credit rating, fraud detection, database marketing, customer relationship management, and stock market investments.

Given the central tenets of data mining its evolution from the disciplines of statistics and artificial intelligence makes sense: a combination of machine learning, statistical analysis, modeling techniques and databases is used in the process of discovering unknown patterns and relationships. Having defined data mining together with its application, the next section looks at model curricula for data mining courses at tertiary education level.

Model Curricula

The IS2002 Model Curriculum and Guidelines for Undergraduate Degree Programs in Information Systems (Gorgone, Davis, Valaich, et al., 2002) was jointly developed by a number of organisations, namely, the Association for Computing Machinery (ACM), the Association for Information Systems (AIS) and the Association of Information Technology Professionals (AITP). Furthermore, the curriculum is endorsed by the ACM SIG on Management Information Systems (SIGMIS), the International Academy for Information Management (IAIM), INFORMS Information Systems Society (INFORMS-IS), the AITP SIG on Education (EDSIG), the International Association for Computer Information Systems (IACIS), the Society for Information Management (SIM), the Decision Sciences Institute (DSI), and the IEEE Computer Society.

Data mining, however, is mentioned only once in the widely endorsed IS2002 model curriculum and only as one of a number of topics in the course specification of *IS 2002.8 Physical Design and Implementation with DBMS* [database management system] (Gorgone, Davis, Valaich, et al., 2002, p. 30) After completing such a course students should demonstrate their mastery of the design process acquired in earlier courses by designing and constructing a physical system using database software to implement the logical design.

Since the undergraduate model curriculum doesn't offer sufficient guidance for an entire full course on data mining, the graduate model curriculum might give an indication as to what students need to be prepared for. The Model Curriculum and Guidelines for Graduate Degree Programs in Information Systems or MSIS 2000 (Gorgone & Gray, 1999) again jointly developed by the Association for Computing Machinery (ACM) and Association for Information Systems (AIS) is a model for a Master's Degree in Information Systems within the context of the USA and Canada. Again, data mining is mentioned as part of the description of the Data Management course (MSIS2000.1) but no specific topics address data mining *per se*; rather the emphasis is on management of an organisation's data resources i.e. databases.

The omission of topics to exploit data contained in the database in both these curricula is interesting to note. It raises the question: should data mining be included as a focus area, i.e. a separate elective course in the senior undergraduate IS curriculum or should the curricula be updated to include data mining or at least data exploitation as a topic? There are a number of institutions that already offer data warehousing and data mining either as a combined course e.g., the Master of Business Informatics program at Dhurakij Pundit University in Thailand, or as separate course, e.g., in the undergraduate Bachelor of Business Administration and the Master of Science in Decision Technologies program at the University of North Texas, and the Master of Science in Information Quality program at Donaghey College of Engineering and Information Technology. Answering this question is not within the scope of this paper but is worthwhile examining nevertheless.

The only model curriculum that covers data mining in more detail is that of the ACM Special Interest Group on Knowledge Discovery in Databases (SIGKDD). SIGKDD convened a Curriculum Committee to "design a sample curriculum for data mining that gives recommendations for educating the next generation of students in data mining" (Chakrabarti et al., 2006, p. 1) aimed at senior level undergraduate students which translates to the South African Honours degree level . The committee consisted of higher education professors, researchers and practitioners from industry and representatives from government agencies – all who have extensive involvement in data mining in their respective roles and can thus inform the curriculum from different perspectives.

The point of departure of this curriculum is that "the teaching of data mining should concentrate on long-lasting scientific principles and concepts of the field" (Chakrabarti et al., 2006, p. 2) implying that a solid foundation is needed rather than merely focusing on recent research and the latest techniques. As a result the curriculum consists of two parts, namely (1) Foundations and (2) Advanced Topics. The first part (1) contains basic material that the committee believes should be covered in *any* introductory course on data mining whereas the second part (2) is "a comprehensive collection of material that can be sampled to complete an introductory course or selections of which can form the basis for an advanced course in data mining" (Chakrabarti et al., 2006, pp. 1-2).

The foundations (or Course I) entails the following units (which are also referred to as modules) (Chakrabarti et al., 2006):

1. Introduction: basic concepts of data mining, including motivation, definition, the relationships of data mining with database systems, statistics, machine learning, different kinds of data repositories on which data mining can be performed, and the current trends and developments of data mining (the material can probably be introduced by case studies.)
2. Data pre-processing:
   - Why pre-process the data?
   - Basic data cleaning techniques
   - Data integration and transformation
   - Data reduction methods

3. Data warehousing, the dimensional model, and OLAP for data mining
4. Association, correlation, and frequent pattern analysis techniques
5. Classification techniques such as decision trees
6. Cluster and outlier analysis techniques
7. Techniques for mining time-series and sequence data, for example, regression analysis
8. Text mining and web mining techniques
9. Visual data mining techniques
10. Data mining: industry reports, privacy, social impacts, and data mining system products

The advanced units (or Course II) include (Chakrabarti et al., 2006):

1. Advanced data pre-processing, namely, advanced data reduction methods
2. Data warehousing, OLAP, and data generalization
3. Advanced association, correlation, and frequent pattern analysis
4. Advanced classification
5. Advanced cluster analysis
6. Advanced time-series and sequential data mining
7. Techniques for mining stream data
8. Mining spatial, spatio-temporal, and multimedia data
9. Mining biological data
10. Text mining, emphasizing the new issues which arise
11. Hypertext and web mining
12. Data mining languages, standards, and system architectures
13. Data mining applications, for example, in financial data analysis, retail industry, the telecommunication industry, intrusion detection, and scientific and statistical applications (note: Some of these themes, if concrete and good materials are available, should go into the Foundations part as case studies.)
14. Data mining and society
15. Trends in data mining

Each of the above units are unpacked in the model curriculum (Chakrabarti et al., 2006, pp. 3-9). The curriculum indicates that a "standard 12-week one semester introductory course on data mining (offered to either senior undergraduate or first-year graduate students) could cover all the units in Foundations and a selected set of units from the Advanced Topics" (Chakrabarti et al., 2006, p. 3) but recognises that the interdisciplinary nature of data mining can give rise to a multiplicity of educational goals and, therefore, that courses in different disciplines will have different emphases (Chakrabarti et al., 2006, p. 9).

Since data mining is not only a theoretical discipline but ultimately requires that human understandable and actionable results be generated skill and experience in the practical application thereof is essential. According to the model curriculum, "laboratories and exercises give students an opportunity to carry out experiments that illustrate topics in a realistic setting and at the same time learn the specifics of the software used." In addition, students can also be assigned to work on projects that are too large to be completed during a single lecture. Several potential categories of projects are mentioned although innovative ideas and suggestions are also encouraged (Chakrabarti et al., 2006, p. 10):

1. Apply data mining techniques using data mining software and statistics analysis software tools.

2. Implementation, refinement, and comparison of performance of several different data mining methods.

3. Proposing, implementing and testing data mining algorithms and functions.
4. Use some sample data sets to implement and test data mining functions, such as KDD CUP data sets, UC-Irvine Machine Learning/KDD Repository, DBLP database, and other selected Web data sets.

Given that data mining draws so heavily from the disciplines of databases, statistics, and modeling techniques, it makes sense that the model curriculum assumes that students have basic background knowledge of the following (Chakrabarti et al., 2006, pp. 2-3):

1. Conceptual database design, data models, query languages such as SQL, and transactions.
2. Statistics, specifically, basic probability, distributions, hypothesis testing, expected values, ANOVA, and estimation of a distribution parameter.
3. Linear Algebra, including vectors and matrices, vector spaces, matrix inversion, and solving linear equations.
4. Familiarity with basic data structures and general maturity of students to understand algorithms written in pseudo-code.

Apart from the diverse prior knowledge requirements, the overlap between statistics, databases and various business disciplines, furthermore represents the challenge of developing and teaching a data mining course as it can "be taught in many different ways." (Mrdalj, 2007, p. 134). Furthermore, Chakrabarti, S. et al (2006:9) state that because a data mining course can be taught in different fields with different emphases it cannot be expected that "the material will be covered in full spectrum with similar emphasis." The plan is to add modules to the model curriculum over time based on "feedback of instructors who have taught materials in specific fields" (Chakrabarti et al, 2006, p. 9).

Having examined the available model curricula to support the development of a data mining-specific curriculum, the next section examines the curriculum developed at the University of Pretoria within the Information Systems programme.

Data Mining at the University of Pretoria

As previously noted, the degrees conferred in the South African higher education system differ somewhat from that in the United States. After successful completion of three years of study a bachelor degree is conferred, and the fourth year, leading to the conferment of what is called an Honours degree, is optional. Thus the Honours degree is seen as an additional, finishing year of the undergraduate programme rather than what is traditionally thought of as graduate, or in the South African context, postgraduate study:

> "Our undergraduate programme leading to the degree of BCom (Informatics) extends over three years. The honours course should be seen as a finishing fourth year of the undergraduate program, containing material that is aimed at enhancing the technical background of students and broadening their horizons in respect of the information technology and information systems field" (University of Pretoria, 2008a).

The data mining course is an elective course in this optional fourth year of study. In addition, there is an underlying intention of the course is to narrow the gap between academic theory and industry, there should be an emphasis on practical aspects, data mining tool demonstrations and selection, as well as the application thereof. Thus the course should provide students with working knowledge of data mining concepts, techniques, and tools with emphasis on business applications i.e. answering business questions in order to support decision-making.

In order to enrol for an Honours degree, a prospective student must be in possession of a Bachelors of Commerce degree in Informatics, Information Systems or Business Information Systems. In addition a student needs "an adequate knowledge of management, financial and economic sciences as well as statistics as determined by the head of the department concerned in consultation with the Dean" (University of Pretoria, 2008b, p. 111). Given that the preceding undergarduate degree is in business rather than science, the assumption of mathematical knowledge as specified by the ACM SIGKDD curriculum does not hold and therefore this particular curriculum needs to take this into account.

Course Objectives

The course aims to cover the most important data mining techniques as indicated in the model curriculum as well as providing background knowledge on how to conduct a data mining project. Based on the definitions and purpose of data mining described earlier as well the particular context and intention of the degree program in which the data mining course is being offered the desired learning outcomes upon completion of this course are for a student to be able to:

1. Understand data mining and its positioning within the business intelligence framework.
2. Identify suitable techniques to be applied for data mining projects.
3. Describe the use and available techniques commercial data mining products available in South Africa.
4. Perform data mining on real-world data.

Learning outcomes 1 and 2 address the required theoretical basis with learning outcomes 3 and 4 addressing the practical application thereof. Given the desired learning outcomes four streams were identified according to which the course was structured; each stream's number below corresponds to the number of the desired learning outcome listed above:

1. Process,
2. Techniques,
3. Product demonstrations, and
4. Projects.

These four streams are echoed in the catalog description for the elective data mining course entitled *Knowledge Acquisition and Sharing* (course code INF791) is:

> "In this information age a lot of data is captured every day and recorded in databases, but the wealth of this data is kept locked in the databases because relatively little mining is performed on this data. This course introduces you to data mining in terms of:
>
> - The data mining process - how do you mine data?
> - The data mining techniques - an overview of the data mining techniques that can be used
> - Practical data mining experience - a practical project mining real industry data to find unknown patterns
> - Product overviews - product demonstrations by data mining vendors" (University of Pretoria, 2008a).

Table 1 maps the foundation units of the ACM SIGKDD model curriculum onto these four learning outcome streams. It is clear from the table that the streams identified based on the desired learning outcomes would be able to cover all the units deemed necessary by the model curriculum. Note that unit 3 is marked as not applicable (N/A) because a separate graduate course is presented on data warehousing (Advanced Database Systems, course code INF785) covering advanced database design and management, database architectures and languages, data warehousing and data marts and current trends. As a result no data warehousing content is included in the data mining course and successful completion of the data warehousing course is highly recommended.

Table 1: Mapping of the foundation units of the ACM SIGKDD model curriculum onto the four learning outcome streams

Curriculum Part	Unit	Theory		Practice	
		Process	**Techniques**	**Products**	**Projects**
Foundations (Course I)	1. Introduction	✓			
	2. Data pre-processing	✓			
	3. Data warehousing, the dimensional model, and OLAP for data mining	N/A			
	4. Association, correlation, and frequent pattern analysis techniques		✓		
	5. Classification techniques such as decision trees		✓		
	6. Cluster and outlier analysis techniques		✓		
	7. Techniques for mining time-series and sequence data, for example, regression analysis		✓		
	8. Text mining and web mining techniques		✓		
	9. Visual data mining techniques		✓		
	10. Data mining: industry reports, privacy, social impacts, and data mining system products			✓	
Advanced Topics (Course II)	13. Data mining applications, for example, in financial data analysis, retail industry, the telecommunication industry, intrusion detection, and scientific and statistical applications (Note: Some of these themes, if concrete and good materials are available, should go into the Foundations part as case studies.)	✓			✓

Course Schedule

The Honors programme is presented in a bi-weekly modular format over 16 weeks, which means that classes are take place 8 times for 2 hours for a total contact time of 16 hours. In the first year, the course was structured such that time was allocated during each contact session to each stream. Although conceptually simple this confused students somewhat.

In the next year, the format was changed: in the first 2 weeks a basic introduction to data mining is given after which the selected techniques are discussed in detail for 3 weeks. Finally, in the remaining 3 weeks are spent on tool selection and product demonstrations by vendors. The projects are discussed in class but students are expected to work on completion outside of class. The actual course schedules for both years are listed in Appendix A.

Textbooks

The model curriculum makes no recommendations on suitable textbooks. There is no shortage of data mining related textbooks: a list of more than 50 introductory textbooks on data mining and knowledge discovery and professional and business-oriented books are available on the website of KDnuggets.com (http://www.kdnuggets.com/publications/books-professional.html). However, the interdisciplinary nature of data mining is evident in the nature of the books: either the content is of a very mathematical, technical nature, or the content focusses on business applications without examining the various techniques' algorithms in some depth. Given that this particular course requires both practical business application and some detail on the techniques' algorithms it makes it rather difficult to select an appropriate textbook.

In the first iteration of the data mining course a single textbook was adopted: *Data Mining: Concepts and Techniques* (Han & Kamber, 2006) with a single recommended text: *Data Mining Solutions: Methods and Tools for Solving Real-World Problems* (Westphal & Blaxton, n.d.). The primary reason for adapting this textbook is its popularity as an introductory text. In the second year no text was required since the previously adopted textbook by Han and Kamber (2006) was evaluated as too complex and difficult to understand as well as being too expensive (due to the relative weakness of the exchange rate of the South African Rand to the US Dollar textbooks are proportionately much more expensive in South Africa); students were encouraged to read the recommended texts by Han and Kamber (2006), Westphal and Blaxton (n.d.) as well as additional texts entitled *Mastering Data Mining: The Art and Science of Customer Relationship Management* by Berry and Linoff (2000) and *Introduction to Business Data Mining* by Olson and Shi (2007). The text by Olson and Shi (2007) was used as the central text with additional material from the other recommended texts. The book is structured as follows:

- Part I: introduction to the material;
- Part II: description and demonstration of basic data mining algorithms;
- Part III: business applications of data mining; and
- Part IV: overview of the developing areas in this field, including web mining, text mining, and the ethical aspects of data mining.

This time around the students were even less satisfied with the study material as it was considered to basic and the lack of a single required text confused students. The lecturers believe that although this book to a large extent meets the need of the course, not using it as the required text was the cause of the problem.

Teaching Methods

Various teaching methods were used: lectures, case discussions to demonstrate business applications, product demonstrations by vendors, and team project(s). In class lectures cover material from the text(s) and other sources using PowerPoint slides, which are available online afterwards as lecture notes. Students are, however, encouraged to attend class regularly and take notes as not all material can be found in either the text(s) or additional sources and these will also not provide enough insight to be able to have a comprehensive understanding of the course content.

In the first year in-class quizzes were included to encourage attendance and ensure students paid attention to the lectures. In the second year this was replaced by in-class student presentations where students formed teams of between 4 and 8 students and where given specific techniques to investigate and present in class with commentary by the lecturers. Data mining as a discipline originated to solve practical business problems, therefore, case studies are used to demonstrate a variety of business applications and how it can be add business value. In addition, students are more likely to remember these narratives, including Grose (2006), Schoeman, Matthee and Van Der Merwe (2006) and Loveman (2003).

Product Demonstrations

Product demonstrations by vendors serve create an awareness of data mining software available in the South African marketplace where most students will eventually be employed. In addition, in the second iteration these demonstrations were linked to a team project on how to select an appropriate data mining tool (see below for more detail).

Projects

Business-oriented team projects are intended to help students to identify with real-world problems and data mining applications. Teams were assigned team project(s) that are to be completed outside of contact sessions. The objective underlying the team project(s) is to enable students to assimilate classroom material in a real-life situation. The main purpose is to apply newly acquired data mining expertise by experience first-hand how data mining techniques can be used" as a proof-of-concept to support decision making in an organisation. In terms of the ACM SIGKDD model curriculum, the projects require students to apply data mining techniques using data mining software and statistics analysis software tools by proposing, implementing and testing data mining algorithms using sample data sets to identify and implement selected data mining functions. More detailed descriptions of the projects are listed in Appendix B.

Student Feedback

Upon completion of the course, students are given an opportunity to give feedback on the course on a standard departmental questionnaire covering course content, lecturers and assessment. These questionnaires are used as a tool to identify specific areas that need to be adjusted in order to improve the course offering, keeping in mind that assessment, reflection and improvement is ongoing. Such areas include study material, level of difficulty, work load, etc. Table 2 lists the enrolments and response rates for the feedback forms for both years.

Table 2: Response rates for student feedback forms

	YEAR I	YEAR II	VARIANCE
Enrolled students (N)	82	83	1
Completed student feedback forms (n)	43	34	-9
Response rate	52.4%	40.9%	-11.5%

Since a standard student feedback form is used for all courses in the Honours degree programme, the ability to draw conclusions specific to the data mining course from it is limited. That being said, students' evaluation of the module content over the two years remained fairly constant and within acceptable levels for the department (3.55 and 3.59 respectively on a scale of 1 to 5, 5 being excellent). Although the evaluation of the organization of the course as well as the study material declined, there was a marked increase in the evaluation of the provided material for completing the course and of the usefulness of the course for students' existing or soon-to-be started careers. Overall the evaluation covered the entire spectrum: from too difficult to too easy, from confusing to well-organized as can be seen in Table 3.

Table 3: Student feedback on the data mining course

COURSE CONTENT	YEAR I	YEAR II	VARI-ANCE
Organization of module (clarity of syllabus, out-comes, time table, requirements, etc.)	3.63	3.47	-5%
Study material (prescribed books, notes, slides, etc.)	3.51	3.24	-9%
Level of difficulty (compared to other honors courses)	3.52	3.47	-2%
Work load (each module carries 20 credits implying 200 notional hours)	3.67	3.59	-2%
Usefulness of material for completing course	3.42	3.68	7%
Usefulness of course for career	3.56	4.09	13%
Average	**3.55**	**3.59**	**1%**

Lecturer Reflection

As discussed earlier, data mining presents unique challenges when it comes to teaching (Chakrabarti et al, 2006, p. 9; Mrdalj, 2007, p. 134). The experience of the lecturers have indeed been that it is quite challenging to develop a course. There is little guidance in the IS undergraduate (IS2002) or graduate (MSIS) model curricula for a course focused on data mining. The ACM SIGKDD model curriculum definitely provides broad guidelines but it remains a difficult task as it is necessary to take a specific context into consideration; particularly for a business-focused degree where not all students do not have in-depth preparation in statistics or knowledge of mathematical algoritms, it is difficult to find a balance between being too superficial and too technical is quite tough..

Another issue is that there is no single suitable textbook for an IS course although there are many books on the subject of data mining. Since the students in the course expressed an overall preference for having a textbook compiling a course-specific reader may be the most suitable solution.

Quizzes are a good way of encouraging class attendance but do not promote in-depth understanding as it is similar to that of a reading-level quiz; an alternative to encouraging class attendance needs to be found that ensures a more in-depth engagement and understanding of the material. On the other hand, the inclusion of software demonstrations by vendors and real-world projects enhanced the students' understanding of how to apply data mining and thus the usefulness of the course for students' careers. Student presentations would have worked better if fewer teams presented on a particular technique allowing more time for discussion and commentary by the lecturers although it does also encourage participation.

Conclusion

Teaching graduate IS students to find diamonds in data is not a straightforward exercise. This paper presented the development of a curriculum based on the ACM SIGKDD model curriculum for a graduate IS course as a result of the omission of detailed guidelines in the widely accepted IS model curricula. This begs the question that consideration be given to including more such details to guide educators and ensure basic standards, especially given the unique challenges of the field. Given that it data mining can "be taught in many different ways" it provides further motivation to argue that data mining should indeed be included more extensively in model curricula for IS as it is conceivable that these different ways can result in widely varying standards within the IS community.

References

Berry, M. J. A., & Linoff, G. (2000). *Mastering data mining: The art and science of customer relationship management*. New York: Wiley.

Chakrabarti, S. et al. (2006). *Data mining curriculum: A proposal (Version 0.91)*. Intensive Working Group of ACM SIGKDD Curriculum Committee. Retrieved March 25, 2008, from http://www.sigkdd.org/curriculum/CURMay06.pdf

Gorgone, J. T., & Gray, P. (Eds.). (1999). *MSIS 2000: Model curriculum and guidelines for graduate degree programs in information systems*.

Gorgone, J. T., Davis, G. B., Valacich, J. S., Topi, H., Feinstein, D. L., & Longenecker, Jr., H. E. (Eds.). (2002). *IS 2002: Model curriculum and guidelines for undergraduate degree programs in information systems*.

Grose, T. K. (2006, October 9). Rapid response. *Time*, 25.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Loveman, G. (2003, May). Diamonds in the data mine. *Harvard Business Review*, 109-113.

Mrdalj, S. (2007). Teaching an applied business intelligence course. *Issues in Information Systems, VIII*(1), 134-138.

Olson, D., & Shi, Y. (2007). *Introduction to business data mining*. McGraw-Hill.

Schoeman, J. H., Matthee, M. C., & Van der Merwe, P. (2006). The viability of business data mining in the sports environment: Cricket match analysis as application. *South African Journal for Research in Sport, Physical Education and Recreation*, *28*(1), 109-125.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4):13-22.

University of Pretoria. (2008a). Honours Program > Department of Informatics > University of Pretoria. Retrieved March 31, 2008, from http://web.up.ac.za/default.asp?ipkCategoryID=2032&sub=1&parentid=2018&subid=2023&ipklookid=7

University of Pretoria. (2008b). *Yearbook: Faculty of engineering, built environment and information technology* (Section II).

Westphal, C., & Blaxton, T. (n.d.). *Data mining solutions: Methods and tools for solving real-world problems*.

Appendix A: Course Schedules

The course schedules for each of the two years that is under consideration in this paper are listed in Tables 4 and 5 respectively.

Table 4: Course schedule for Year I

Week	Technique	Process	Product Demonstrations
1	Ch 1 - Introduction		
2	Ch 5 Concept Description	Problem Definition	Visual Analysis
3	Ch 6 Association Rules	Ch 2 Data Sources	GIS
4	Ch 7 Classification	Ch 3 Data Pre-processing	Knowledge Factory
5	Ch 8 Clustering	Ch 4 Data Mining Primitives	SAS Analysis
6	Ch 9 Mining Complex Types	Ch 4 DMQL	SAS Neural Networks
7	Ch 9 Mining Complex Types	Verification and Presentation	Text Mining
8	Ch 10 - Conclusion		

Table 5: Course schedule for Year II

Week	Stream	Topics
1	Process	• Approach, structure and schedule • Introduction and applications of data mining • Class discussion
2	Process	• DM methodology: CRISP-DM • Brief on presentations • Team selection
3	Technique	• Presentations: Correlation analysis (Market-basket Analysis) and Link analysis
4	Technique	• Presentations: ANN and Regression
5	Technique	• Presentations: Decision Trees and Cluster detection • Discussion on data mining project
6	Product demonstration Project	• Tool selection and vendor selection assignment • Work on data mining project
7	Product demonstration	• Product demonstration by vendors
8	N/A	• Conclusion and exam information • Submission of vendor selection assignment • Submission of data mining project

Appendix B: Detailed Description of Projects

Projects in Year I

1. Retail proof of concept

Data is sampled from the actual database of an importer of a branded fashion accessory that is then distributed to department and speciality stores. Transactions are either orders or consignment-based. Students are required to clean the data and then perform data generalisation on customers and products and perform any class comparison using the sales data. The required deliverable is a presentation as if presenting to the actual client, thus it should contain a short, not too technical description of process and the results as well as the recommendations based on the results. The business value of data mining should be emphasized.

2. Lottery results

Students are required to download data from South Africa's National Lottery website (http://www.nationallottery.co.za/lotto/results.aspx, this URL is no longer active since the license was revoked by the government and reissued to another company) in MS Excel format, import the data into a database and sample data from this database themselves. They are the required to use the apriori algorithm to determine the most common combinations. Again, students are required to present their findings giving a short description of process, their results and an interpretation thereof (although this project delivered some interesting results, no students (or lecturers) subsequently won the lottery.)

Projects in Year II

1. Undergraduate enrolment data

Students are provided with anonymized enrolment data containing data on student demographics, grades, an so on. Student teams are informed that the University of Pretoria would like them to perform a proof-of-concept, i.e., show them what data mining can do for them. They need need to decide on appropriate data mining technique(s) based on the data and also to select an open source tools with which to apply their selected technique(s) on the data. The deliverable takes the form of a presentation to present to the university's management i.e., they must present actionable but unknown patterns.

2. Tool selection assignment (linked to vendor demonstrations)

Students are instructed to read the following articles:

- Gartner. 2007. *Magic Quadrant for Customer Data Mining, 2Q7.*
- METAgroup. 2004. *Data Mining Tools: METAspectrum Evaluation.*
- Data Mining (Analytic) Tools (May 2007) [online]. URL: http://www.kdnuggets.com/polls/2007/data_mining_software_tools.htm

After the class discussion on tool selection frameworks each team must construct a tool selection framework based on a particular case study (each team receives a different case study) that is then used to evaluate the vendors' demonstrations in class.

Biography

Shana Ponelis is currently a lecturer with the School of Information Studies at the *University of Wisconsin-Milwaukee*. Prior to this appointment she was a senior lecturer with the Department of Informatics at the *University of Pretoria*, a consultant with *Atos KPMG Consulting* and *Andersen* in the Business Consulting practice specializing in data warehousing and business intelligence, and a business analyst at a large financial institution in South Africa focusing on business intelligence. She has published a number of articles in at several international conferences and is a member of a number of professional bodies, including the Association for Information Systems (AIS) and Association for Computing Machinery (ACM).