Data mining in the e-learning domain

Margo Hanna

The author

Margo Hanna is Education Liaison Officer for e-Learning, Knowsley Council and University of Liverpool, Wigan, UK.

Keywords

Higher education, Classification, Data encapsulation, Databases

Abstract

Higher education (HE) is becoming a big business, with huge investments in IT technology supporting online learning. With the awareness of the knowledge economy has come a growing consciousness that HE constitutes a large industry or economic sector in its own right. In a marketing fashion, we understand that some customers present much greater profit potential than others. But, how will we find those high-potential customers in a database that contains hundreds of data items for each of millions of customers? Data mining software can help find the "high-profit" gems buried in mountains of information. However, merely identifying the best prospects is not enough to improve customer value. One must somehow fit the data mining results into the execution of the content management system that enhances the profitability of customer relationships. However, data mining is not yet engaged into e-learning systems. This paper describes how we can profit from the integration of data mining and the e-learning technology.

Electronic access

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1065-0741.htm

Campus-Wide Information Systems
Volume 21 · Number 1 · 2004 · pp. 29-34
 · ISSN 1065-0741
DOI 10.1108/10650740410512301

The future of e-learning

The world is going fast towards online learning, so we can see a lot of open universities that provide courses online through the Internet. At the same time, we can observe lots of advertisements for online professional certificates, so one can study and take the exam and get certified (without the previous hassle of attending classes, travelling), whenever and wherever he/she wants. Managing and tracking students, courses, degrees, grades, certificates, universities and learning providers requires a massive content management system to manage, track, and control the whole system with all possible implications and complications. Every organization would look at e-learning information from different angles according to the mission, vision and objectives of the organization and whether it looks for profit or has other national objectives.

Mining e-learning

Data mining applied to the Web has the potential to be quite beneficial. Web mining is mining of data related to the Web. This may be the data actually present in Web pages or data related to the Web activity. Web data can be classified into:

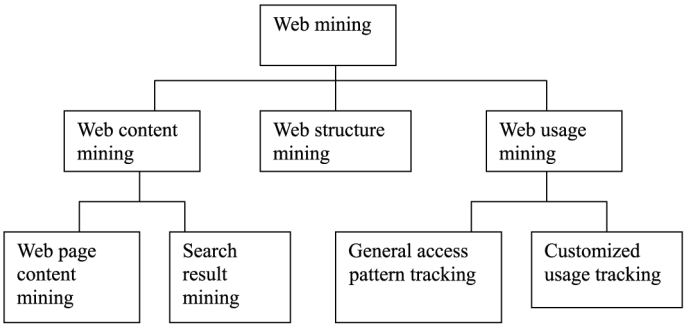
- content of Web pages;
- intrapage structure includes the HTML or XML code for the page;
- interpage structure is the linkage structure between Web pages;
- usage data that describe how Web pages are accessed by visitors; and
- user profiles include demographic and registration information obtained about users (Dunham, 2003, p. 192) (see Figure 1).

Web content mining

Web content mining can be thought of as extending the work performed by the basic search engines. There are many different techniques that can be used to search the Internet. Most search engines are keyword-based. Web content mining goes

Campus-Wide Information Systems
Volume 21 · Number 1 · 2004 · 29-34

Figure 1 Web mining taxonomy



beyond this basic IR technology. It can improve on traditional search engines through such techniques as concept hierarchies and synonyms, user profiles, and analyzing the links between pages. Data mining techniques can be used to help search engines provide the efficiency, effectiveness, and scalability needed.

Agent-based approaches have software systems (agents) that perform the content mining. In the simplest case, search engines belong to this class, as do intelligent search engines, information filtering, and personalized Web agents. Intelligent search agents go beyond the simple search engines and use other techniques besides keyword searching to accomplish a search. For example, they may use user profiles or knowledge concerning specified domains. Personalized Web agents use information about user preferences to direct their search. The database approaches view the Web data as belonging to a database. There have been approaches that view the Web as a multilevel database, and there have been many query languages that target the Web.

Many Web content mining activities have centered on techniques to summarize the information found. In the simplest case, inverted file indices are created on keywords. Simple search engines retrieve relevant documents usually using a keyword-based search retrieval technique (Dunham, 2003, pp. 195-8).

Personalization

With personalization, Web access or the contents of a Web page are modified to better fit the desires of the user. This may involve creating Web pages that are unique per user or using the desires of a user to determine what Web documents to retrieve. In our e-learning system, information of courses could be retrieved in a personalized fashion per learner according to his/her profile.

With personalization, advertisements to be sent to a potential customer are chosen based on specific knowledge concerning that customer. Unlike targeting, personalization may be performed on the target Web page. The goal here is to entice a current customer to purchase something he/she may not have thought about purchasing. Perhaps the simplest example of personalization is the use of a visitor's name when he/she visits a page. Personalization is almost the opposite of targeting. With targeting, businesses display advertisements at other sites visited by their users (Dunham, 2003, p. 199).

Education is a big business

Today, technology leaders in universities need to develop broad-based support for enterprise-wide e-education services that integrate as many online transactions as possible through core standards-based applications. Leaders also want choice and flexibility, and the ability to customize applications to meet local needs and user preferences – and to do it all as efficiently as possible with the least amount of organizational pain and fiscal outlay in the process of installing and learning new systems.

Higher education is a big business. With the awareness of the knowledge economy has come a growing consciousness that higher education constitutes a large industry or economic sector in its own right. Higher education in the USA alone is now a $240 billion business, and the total amount spent annually on all post-secondary education is around $300 billion and growing. Higher education now forms a large business ecosystem with many vendors, publishers, and service providers dedicated to serving its needs. When you add up core higher education spending and its multiple spillover effects, the total economic impact of higher education in the USA is over $1.2 trillion (Irvine, 2002, pp. 5-7).

Campus-Wide Information Systems
Volume 21 · Number 1 · 2004 · 29-34

Why data mining?

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, research projects, and market analysis, to engineering design and science exploration.

Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities: data collection and database creation, data management (storage, retrieval, transaction processing), and data analysis and understanding (data warehousing and data mining).

Data can now be stored in many different types of databases. One database architecture that has recently emerged is the data warehouse. A repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision making. Data warehouse technology includes data cleansing, data integration, and OLAP, that is analysis techniques with functionalities such as summerization, consolidation, and aggregation, as well as the ability to view information from different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data change over time.

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. The best-growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. Consequently, important decisions are often made based not on the information rich data stored in databases but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. In addition, consider current expert system technologies, which typically rely on users or domain experts to input knowledge manually into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time consuming and costly. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research.

Data mining functionalities for e-learning domain

In the e-learning domain, we are interested in managing mainly two groups of users: the learners as well as the learning providers, whether private training companies, governmental organizations and local authorities providing training for their employees or universities who aim to publish their courses and make them accessible online via the Internet. As for learners, databases should store all personal details including name, age, gender, address, postcode, and educational-relevant details such as qualifications. Moreover having information like work experience, career objectives, income range, previous courses taken and courses of interest would be of great value to be able to predict future behaviour of different classes of employed professional people. Also other information such as personal interests and hobbies would be very valuable for data mining tool in order to discover hidden patterns by building intelligent models based on the huge amount of data.

At the same time, databases that keep records of the online learning providers' information do exist. Databases store the following information per learning provider:

- name of the company;
- type of the company;
- size of the company;
- courses provided;
- number of courses;
- areas/subjects of courses;
- targeted learners (audience);
- number of hours (duration) per course;

Campus-Wide Information Systems
Volume 21 · Number 1 · 2004 · 29-34

- cost;
- prerequisites of each course;
- course objectives;
- contents; and
- course path.

It would be more complicated in the case of universities, as we have to track courses and students distributed among departments, by course tutor, lecturer, course prerequisites, course pathway, course contents, course assignments and course evaluation. The multidimensional database is essential in order to be able to view all the information from different angles and fulfil different managerial levels. In this case, system administration becomes extremely complicated.

Applying data mining would enable us to help the learners who are interested in certain areas by suggesting relevant or complimentary courses of which they might not be aware in an efficient way, providing them with a personalized registration Web page. As for the learning provider, they will have the chance to view data of learners and courses from different angles in order to have the full picture, enabling them to make the most profitable decision via targeting the class of users of interest to them and investing more in courses that are highly required by their targeted classes of learners. For example, a charity would base the decision of targeting learners differently from business companies who look for the profit. Also this would be different from local authorities who aim at achieving certain national objectives. It means that applying data mining with e-learning would be of great value enriching the management with valuable information and knowledge that would lead to efficient decision making. Moreover, data mining models build models that help predict future behaviours, which would enhance the decision-making process of efficiency.

As for universities, it will draw the attention to areas in which management should invest more, in terms of degrees, professional certificates, courses, and modules. It would highlight the demanding areas from the market point of view; it might help in suggesting certain complementary courses, it might provide advice on certain courses that might be valuable for each class of learners leading to more development and enhancing the competitive edge of the learning providers in their learning strategies.

Concept/class description: characterization and discrimination

Data can be associated with classes or concepts. For example, concepts of customers include bigSpenders and budgetSpenders. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept description (Dunham, 2003, p. 200). In our e-learning domain, it is valuable to classify learners by courses this way: interested-in-Course1, interested in-Course2, etc. We can also have five-star-department, three-star-department, top-ten-courses, top-ten-universities and whatever sort of intelligent reporting and queries.

Data characterization is a summarization of the general characteristics of features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query. For example, to study the characteristics of the top ten courses which achieved 20 per cent higher hits by online learners this year, the data related to such courses can be collected by executing an SQL query. The data mining system should be able to produce a description summarizing the characteristics of online learners who spend, for example, more than £1,000 a year.

The output of data characterization can be presented in various forms such as pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables (Dunham, 2003, p. 200).

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. For example, we might be interested in comparing the general features of the courses whose sales increased by 20 per cent this year with those whose sales decreased by 30 per cent during the same period. Discrimination descriptions should include comparative measures that help distinguish between the target and contrasting classes.

Data mining system should be able to compare two or more groups of learners or of learning providers or of universities. The resulting description could be a general comparative profile of each group.

Association analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. It is widely used for market analysis and transaction data analysis (Dunham, 2003, p. 201). Given the e-learning relational database, a data mining system may find association rules such as: learners 16 to 20 years old register for multi-media courses, while learners 20 to 29 years old with an income of £20k to £25k register for professional Web development courses.

Classification and prediction

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes and concepts, to be able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.

The derived model can be represented in the form of classification rules, decision tree, mathematical formula, neural networks, etc. (Dunham, 2003, p. 202).

We might be interested in classifying learners to bigSpenders, budgetSpenders, or classify courses to expensiveCourses, cheapCourses, or maybe attractiveCourses and normalCourses, etc.

Cluster analysis

Clustering analyses data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison with one another, but are very dissimilar to objects in other clusters. Each cluster can be viewed as a class of objects, from which rules can be derived (Dunham, 2003, p. 203).

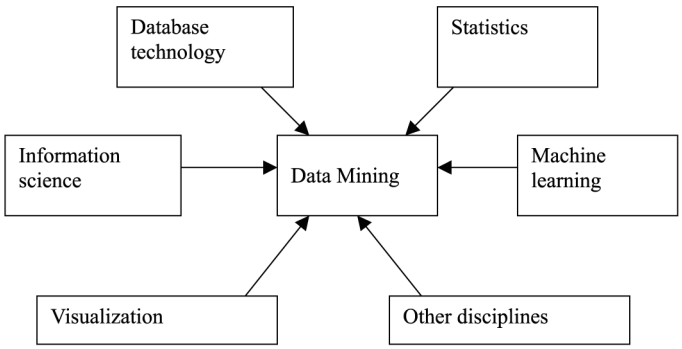
Data mining systems versus statistical analysis

Data mining is an interdisciplinary field, the confluence of disciplines including database systems, statistics, machine learning, visualization and information science. Depending on the data mining approach used, techniques may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high performance computing. Depending on the kinds of data to be mined or on the given data mining applications, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology. Also, there is one crucial point about data mining: once one has built the model and loaded the data. he/she can view this data from different angles depending on the number of dimensions entered to build the model. Therefore, the model is built once and then the data mining tool visualizes results, depending on your selection of the dimension(s), while in statistical packages one needs to figure out the sort of report and work out the whole thing every time for every dimension or group of dimensions needed to discover or predict any relevant behaviour (see Figure 2).

Conclusion and future work

Mining online learning events is becoming a promising area for research and development, particularly when the business in education is growing impressively.

Figure 2 Data mining system components



Data mining applications is a major component of the whole content management system that enables different managerial levels to track, understand, and manage the wealth of information stored in multiple data sources.

There are plans to use a statistical package versus data mining tool on an e-learning database to show the impressive accurate real time results of the data mining tool over the statistical package. And in this concern, we should consider the technology (data mining tool) in exactly the same way as the content, as the quality of data is a major concern which requires special handling before uploading this data and getting started with building the model.

References

Dunham, M.H. (2003), *Data Mining: Introductory and Advanced Topics*, Prentice-Hall, Upper Saddle River, NJ.

Irvine, M. (2002), "The emerging e-education landscape, a blackboard strategic white paper", Blackboard, Washington, DC, available at: www.blackboard.com

Further reading

Debevoise, N.T. (1999), *The Data Warehouse Method*, Prentice-Hall, Upper Saddle River, NJ.

Groth, R. (2000), *Data Mining: Building Competitive Advantage*, Prentice-Hall, Upper Saddle River, NJ.

Han, J. and Kamber, M. (2001), *Data Mining: Concepts & Techniques*, Morgan Kaufmann, San Francisco, CA.

Minaei-Bidgoli, B. and Punch, W.F. III (2003), "Using genetic algorithms for data mining optimization in an educational Web-based system", *GECCO 2003*, pp. 2252-63, available at: www.lon-capa.org