Web Mining as a Tool for Understanding Online Learning

Jiye Ai
University of Missouri-Columbia
Columbia, MO USA
jadb3@mizzou.edu

James Laffey
University of Missouri-Columbia
Columbia, MO USA
LaffeyJ@missouri.edu

Abstract

After an introduction to Web mining and e-learning and a brief review of Web mining applications in business and education, this paper presents an experiment with pattern classification for student performance prediction in a WebCT learning environment. The results illustrate that recognition for a certain class (with good grades) on a large data set can be obtained by a classifier built from a small size data set. The paper concludes that Web mining can be an approach to build knowledge about E-learning and has potential to help improve learning performance.

Keywords: e-learning, Web mining, Course Management Systems (CMS), Data mining, WebCT

Introduction

The World Wide Web (WWW) is a vast resource of multiple types of information in varied formats. Researchers are beginning to investigate human behavior in this distributed Web data warehouse and are trying to build models for understanding human behavior in virtual environments. Data mining, often called Web mining when applied to the Internet, is a process of extracting hidden predictive information and discovering meaningful patterns, profiles, and trends from large databases. Etzioni (1996, p. 1) defined Web mining as: "... the use of data mining techniques to automatically discover and extract information from Web documents and services." Web mining is an iterative process of discovering knowledge and is proving to be a valuable strategy for understanding consumer and business activity on the Web.

Online learning (E-learning) systems accessible through the Internet are intranets that represent self-contained versions of the data warehouses and human behavior found more broadly across the Internet. E-learning systems have great potential to improve education through extending educational opportunities for those who can not use time and place-bound traditional courses and through offering new interactive learning services and functions that enhance the traditional classroom. E-Learning systems offer students Web-based texts, images, and multimedia, and also offer instructors and students ways to communicate with each other asynchronously and synchronously. E-learning systems provide multiple ways of learning (self-paced, collaborative, tutorial) within a common application as well as providing the potential for rich media and complex interactions. The software applications used most frequently to implement E-learning in higher education are called Course Management Systems (CMS). Examples of CMS include Blackboard and WebCT. E-learning systems are rich in content and invite

complex forms of activity and interactivity. Understanding these forms of behavior and building models of patterns of behavior represent challenges for educational researchers.

Potential of Web Mining

Web usage mining (hereafter called Web mining) approaches can be applied to CMS-based E-learning and have promise for helping explain system usage. Web mining can be used to explore and investigate patterns of activity. Articulating and identifying patterns of use in E-learning systems may provide better understanding of how students undertake Web-based learning and guidance for better organization of online learning activities. There are many potential benefits of using Web mining for exploring learning behavior and patterns in E-learning using CMS. However, there is scant literature on the use of Web mining with CMS. This article is a case report of how one Web mining method, classification could be applied against a CMS data set. This case report shows how patterns emerge from the application of Web mining approaches. The key purpose of this report is to illustrate the potential of Web mining and to identify issues in its application in currently available CMS. As a way of setting the context for the case report the following examples show how Web mining could potentially benefit E-Learning.

(1) Understand learner behavior

University administrators and instructors may be able to improve the implementation of E-learning systems by understanding the dynamic behavior of students in the Web systems.

2) Determine E-learning system effectiveness:

Patterns of behavior may be associated with system performance and enable more customized system configuration. Administrators and instructors may be able to discover high and low use areas of the E-learning system and adjust resources to optimize the technical performance of the system.

(3) Measure the success of instructional efforts:

In E-learning systems, students use email, the Web forum, feedback forms, etc. to express their concerns and ask questions. These data are completely recorded in the E-learning system. Web mining can provide quantitative feedback to instructors about the outcomes of their activity.

These 3 examples show how Web mining could provide new insights about student activity in CMS and to address information needs as well as suggest customization of approaches for implementing E-Learning by administrators and instructors. The examples illustrate how, although we are in the early stages of understanding the use of Web mining in E-Learning, broad and rich benefits may accrue from better understanding patterns of behavior in CMS through the use of techniques such as Web mining.

Literature Survey

Web mining has been most often used for developing business and marketing intelligence. For example, Web mining is frequently used by online retailers to leverage their online customer data in order to predict customer behavior. The business benefits that Web mining brings to digital service providers include personalization, collaborative filtering, enhanced customer support, product and service strategy definition, particle marketing (marketing or customizing a product for one customer) and fraud detection. In sum, the objectives and benefits are the ability to become more customer centric by delivering the best and most appropriate service to individual customers at the most appropriate moment.

There are several efforts to develop Web mining algorithms and systems for e-business. Chakrabarti (1998) made pioneering efforts in Web structure mining that is an examination of the use of hyperlinks and document structure. However, these Web structure approaches only took into account hyperlink

information and paid little attention to the Web content. Cooley, Mobasher, and Srivastava (1997) illustrated that Web usage mining is an excellent approach for achieving the goal of making dynamic recommendations to a Web user based on his/her profile of usage. This behavioral data has been useful for applications like cross-sales and up-sales in e-commerce. Buchner and Mulvenna (1998) presented a knowledge discovery process to identify marketing intelligence from Web data. Based on three classification labels with "non customer", "visitor once" and "visitor regular", the company could provide a special offer to attract potential online shoppers. The company also used association rules and sequential patterns to discover customer navigation behavior so that online shoppers who followed certain paths could be rewarded for their loyalty to the Web site. Padmanabhan (1998) used Web server logs to generate beliefs about the patterns of accessing Web pages for a given Web site. Padmanabhan identified 15 beliefs about the data in three groups: "(1) Usage of coupons, e.g. "young shoppers with high income tend not to use coupons". (2) Purchase of diet vs. regular drinks, e.g., "shoppers in households with children tend to purchase regular beverages more than diet". (3) Day of shopping, e.g. "professionals tend to shop more on weekends than on weekdays".

Relative to Web mining activity in business, there has not been much Web or data mining application in education. However, some work is available to inform our efforts. One study (Luan, 2002) focused on student enrollments in community colleges, and reports a case study in which data mining was used to monitor and predict community college students' transfer to four-year institutions. The model developed in this case study represents a profile of students who have transferred so as to predict which students currently enrolled in a community college are likely to transfer. These predictions allow the college to personalize and time their interactions and interventions with these students, who may need certain assistance and support. In this case study, Luan selected a set of features to investigate:

- Demographics: age, gender, ethnicity, high school, zip codes, planned employment hours, education status at initial enrollment
- Financial aid
- Transfer status (doubled as the reference variable)
- Vocational, basic skill, science, and liberal arts courses taken
- Total units earned and grade points by course type

Luan showed how data mining could be applied to the college data on an annual basis so that the model could be used to repeatedly monitor student transfer status. Across a range of data mining analyses, the accuracy rate for predicting students who had transferred was at least 77.5%, and the rate for predicting students who would not transfer was at least 70.0%. (The number of students in the dataset is 32 thousand.) The potential of Web or data mining for efficiently drawing insights from educational records and supporting and informing practices in education seems quite substantial, but is still virtually unexplored. In the following table, the first 2 columns show how Luan matched questions that are frequently asked in the business world with likely counterparts for education. The third column was added by the authors and shows possible extensions of these questions that may be appropriate for looking at E-Learning courses.

Table 1. Comparison of data mining questions in business, education and e-learning.

Business	Education	E-learning
Who are my most profitable customers?	Who are the students taking most credit hours?	Who are the students with highest frequency of logging-in?
Who are my repeat Website visitors?	Who are the ones likely to return for more classes?	Who are the students most engaged on discussion boards?
Who are my loyal customers?	Who are the persisters at our university, college?	What pages do the students access most?
What clients are likely to defect to my rivals?	What type of courses can we offer to attract more students?	What kinds of students are likely to get a high score online?

Note: modified from Luan (2002) Table2.3. Comparison of Data Mining Questions in Education and the Corporate World, p28

Research Goals

Given the limited use of Web mining in education and the potential benefits to online education of making it more student centric that might emerge from effective utilization of Web mining, we set two research goals for this project: (1) to see if patterns of behavior could be used to predict achievement in online learning in a set of CMS data, and (2) to develop a better understanding about the process of applying Web mining to E-learning systems, as well as the constraints of using datasets from existing current versions of CMS. While Web mining is a recognized approach for building knowledge and value in business and commercial information systems, its application in education is not well understood. In this sense, this research is primarily exploratory and while the objective is to build new insights about learning activity, an equally important objective is to examine the fit of Web mining approaches to CMS. What are the challenges in extracting data from CMS and applying Web mining approaches? What strategies for data formatting and data analysis are needed for building meaningful insights? What changes are needed in CMS or Web mining solutions to improve the yield of Web mining in E-leaning systems? A key outcome of this research will be suggestions for understanding how best to utilize Web mining in E-learning. The Web mining approaches applied in this study will focus on understanding learner behavior. For example, we will examine student profiles, frequency of access to learning resources, the clustering of students with similar patterns and the cross relationship of student behaviors.

Process and Approaches

Web mining is a multi-stage process that requires understanding how data are stored, formatted and accessed within a data set, and work stages of selection, preprocessing, transformation and mining. These processes are described below using WebCT as an illustrative context.

Selection

These data are obtained from the course management system “WebCT”. The principle forms of data in WebCT are shown below:

(1) The *User Profile* holds the demographic data of students including user ID, gender, academic level, etc. These data can be accessed via the student management tool in WebCT.

(2) *Usage* data represent access to Web pages. These data items include IP address, page reference, time of access, etc. The access log file is the major source of usage data for Web mining in WebCT.

(3) *Structure* describes the hierarchy of Web pages, primarily linking of the content. The data for representing structure can be collected from different data source locations: server side, client side, proxy server and database. In WebCT, the data are obtained from the server side.

Preprocessing

This step primarily includes data cleaning. There are several steps in the data cleaning process. First, all the entries of the images (graphs) have to be removed from the file because they are not included in the pattern discovery. Second, the entries with HTTP status code such as 404 which means, "resource not found on the server" are deleted. Third, the requests from the Web proxy are removed. The reason for excluding requests from the Web proxy is that they are mechanisms that take the place of the server answering the client's request, and no insight into user behavior can be identified from them.

Transformation

The data are transformed into formats that can be used by the different mining applications. Here are the most common steps in the transformation process: user identification, session identification, traversal path completion, and learning activity mapping. Integration with other data such as a backend database can also be considered.

Mining

There are various data mining techniques such as statistics, classification, association rules, sequential patterns, and clustering which can apply to the Web domain. Classification is the form of data mining used in this study and is a technique that uses a set of pre-classified examples to develop a model that can classify the population of records. There are many algorithms for classification such as decision tree, neural network classification, etc. The classification algorithm starts with a training set of pre-defined example transactions. The classifier training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination. The algorithm encodes these parameters into a model called a classifier. After an effective classifier is developed, it will be used in a predictive mode to classify new records into these same pre-defined classes. For instance, one classifier that is capable of identifying student performance could be used to help in the decision of whether to provide a specific recommendation to an individual student. In this study, the decision tree software C4.5 (Quinlan, 1993) is used, which is shown in detail in the case report. C4.5 is an algorithm introduced by Quinlan for inducing Decision Trees from data.

Data

In this research, the population is the undergraduate students in a large enrollment course of a research university in the Midwest. A binary decision tree was built to classify the access log file from a course on WebCT. The course was a blended course (face to face and online). The total number of the students in this course was 748.

The course data were examined to identify which and to what extent student behaviors and attributes could predict grades. One form of predictor used in the study was seasonality which is a term used to represent periodic variations over time. An example of seasonality is that sales vary throughout a year and peak during the Christmas season. One of the key contributions of this study is the inclusion of seasonal effect as an educational attribute and comparisons among different sample sizes of training data for determining a good classifier.

The log file contains over 90,000 entries. An example of an entry in the log file in WebCT is shown here:

> ???.???.???.?? - ****** [17/Jan/2005:18:10:26 -0600] "GET /SCRIPT/stat_1200_lr2/scripts/student/serve_home?_homepage+START HTTP/1.1" 200 3051 "-" "Mozilla/4.0 (compatible; MSIE 5.0; Mac_PowerPC)"
>
> (Note: student ID is masked by *****, and IP address is masked by ???*)*

The student's grade is used as the criterion variable for evaluating students' performance in the course, and students are divided into three classes based on their overall grades: Good, Medium, and Poor. Table 2 shows those classes.

Table 2. Classes

ID	Name	Description	Value	Distribution
1	Good	Grade A and B	0	480/748
2	Medium	Grade C and D	1	216/748
3	Poor	Grade F, W, NA	2	52/748

Results and Discussion

One challenge in Web usage mining by classification is that it is difficult to identify the better attributes before building the classifier. This is confounded in online learning because the WebCT file system does not hold data in ways that are easily associated with important educational constructs. For example, there is no representation of the URL for some function pages such as discussion page, etc. and the URL hierarchy is simple, which means the depths of links on a page may not be well represented. To address these issues the researcher must construct plausible variables from the WebCT logs. For example, the variable Access Period was constructed to represent the distinction between student's access to the system between midnight and 8:00 AM and other access between 8:00 AM and midnight. Similarly, "Test date" represents whether an entry of the log file was recorded on the date when there was a test, and "Lecturing date" represents whether an entry of the log file was recorded on the date when there was a lecture. After examining the available information on the WebCT site, attributes were constructed for this study: test date, lecturing date, college, academic level and the others shown in Table 3.

Table 3. Attributes

ID	Name	Category	Description	Value
1	Test date	Time	Access to system on a test date	0 for yes, 1 for no
2	Access period		Regular time or after midnight	0: 0am-7:59:59am 1: 8am-11:59:59pm
3	Lecturing date		Access to system on a lecturing date	0 for yes, 1 for no
4	Hit	Page & file	Number of page hits per session	discrete
5	PDF		Ratio of PDF file access per session	continuous
6	Home		Ratio of access to course home page per session	continuous
7	Marks		Ratio of access to course grade book per session	continuous
8	Tree		Ratio of access to serve_tree per session	continuous
9	URL depth		The longest URL distance from the first page per session	continuous
10	College	Student info.	Student's college	JOURN, EDUC, A&S, HP, NURS, HES, BUS,AF&NR, NATR, ENGR,NA, SOWK.
11	Academic level		Student's academic level	FR, SR, JR, SO, NA

Most attributes are chosen based on an estimation of what Web usage might be important and related to learning activity and a balance of finding what data for representing those attributes is actually available. The attributes "Access period", "Test date" and "Lecturing date" are included in the analysis to determine if there is seasonality in online learning behaviors and test our assumption that students access course Web sites in a seasonal fashion. The following figures display statistics showing some seasonal effects. Figure 1 shows that in the time period of a month the number of hits varied and reached higher levels on lecturing dates and the test date. In Figure 2, the number of hits is shown to maintain a high level from hours 9 to 24 over a day.

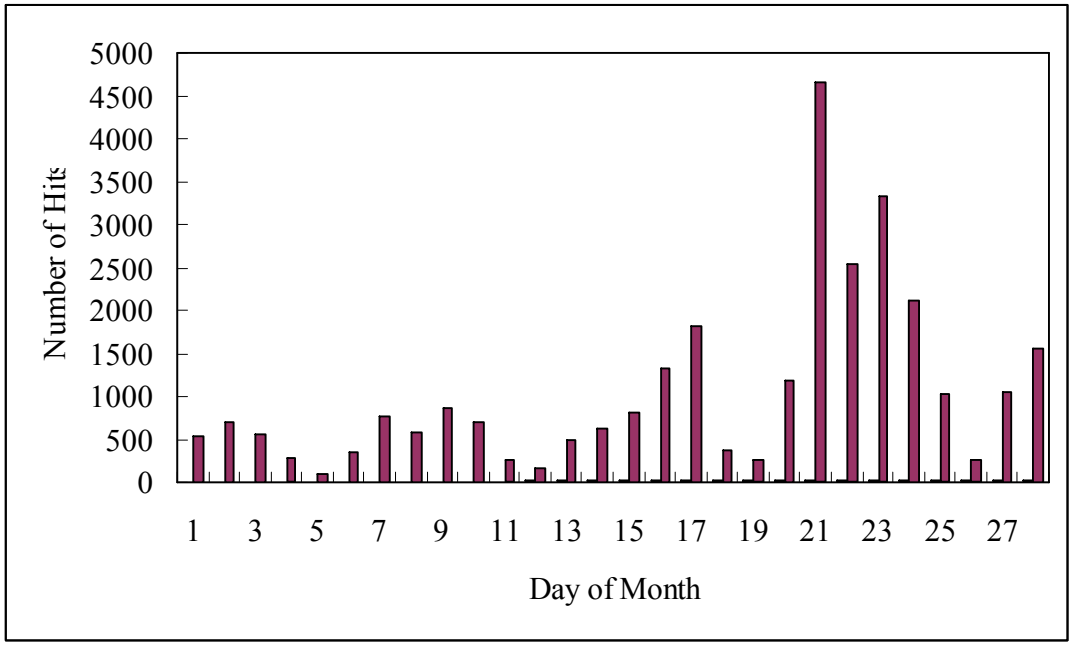

Figure 1. Number of hits in February 2005.
Note: Testing date was 2 and Lecturing dates were 2,7,9,14,16,21,23, and 28.

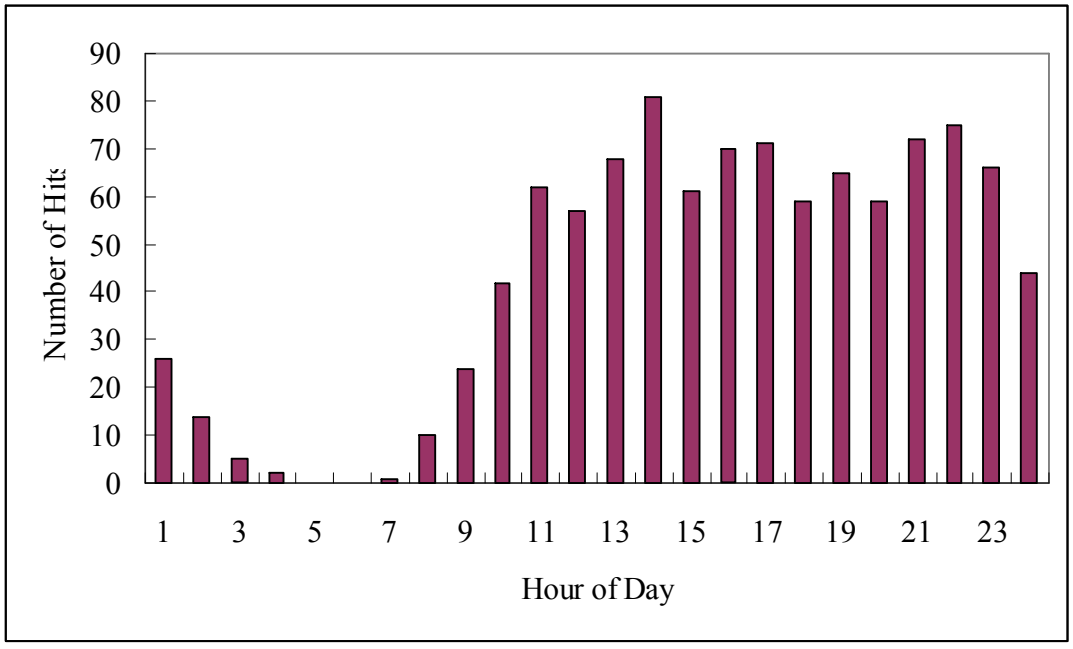

Figure 2. Average number of hits over a day in February 2005.

After data pre-processing, 17,317-sesssion log items were available. Table 4 displays a sample of this log data for the attributes listed in Table 3.

Table 4. A sample of the log data after data processing

Attribute 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	Class
0, 1, 0, 27, 0.038461538461538, 1.0815199894884E-07, 0, 2.3741690408438E-06, 3, A&S, SO	1
1, 0, 1, 2, 0, 0.79166666779325, 0.03125, 0.083333358064261, 1, JOURN, FR	0
1, 1, 0, 7, 0, 0.14901960784516, 0.024206349206349, 7.2948091928E-07, 1, JOURN, FR	2

We selected one-week, two-week, three-week and one-month log data respectively as the training data. The purpose of these selections is to identify whether a small size data set can build a relatively good classifier that has precision and recall for analyzing student's accessing behaviors in an e-learning system. Precision is the ratio of the number of the correctly classified class to the total number of the classified class (including correct and not correct). Recall is the ratio of the number of the correct classified class to the total number of that class. Table 5 reports the final results and shows that one week of data is fairly close to as accurate for predicting class 1 students as are the larger data sets.

Table 5. Recognition accuracy on 17,317-sesssion data

	Precision %		Recall %	
Classifier	Class 1	Class 2	Class 1	Class 2
One week	72.2	44.8	91.9	16.8
Two weeks	72.7	42.6	90.3	25.6
Three weeks	73.3	51.1	93.3	19.5
One month	73.0	57.2	95.9	16.7
All	75.7	76.2	97.1	27.2

To the extent that instructors can use student behavior to predict student performance, they would then be able to adopt more student-centric strategies that address the needs of individual students. Using the method shown here a couple of week's log files can be used to forecast the final grade of the student for the whole semester at 70% accuracy. Based on that, hypothetically a new or customized recommendation could be made to students forecast to do poorly, or modified instructional strategies might be considered by an instructor if the number of students forecast to do poorly is high.

Conclusions

In this research, we explained the use of Web mining approaches in CMS and identified some illustrative learning patterns that can be found by using Web-mining approaches. Although some interesting patterns

were found, the exploratory state of Web mining tools in education suggests replication and confirmation from other forms of research to build a context for understanding and drawing implications from the data. The primary findings of this research are to suggest that Web mining can be an approach that educational researchers can use, and when combined with other forms of data collection has potential for adding to the way we build knowledge about E-learning. A second contribution of the current study is to draw implications for how to improve the process of Web mining e-learning data sets.

The current research has shown that Web mining has promise for identifying patterns within the large datasets of CMS that may be valuable for teaching and learning. One implication of these patterns is that some form of personalization may be possible and may lead to improved learning and teaching processes. For example, student achievement might be improved if a software agent could monitor patterns of student activity and match those patterns with patterns associated with high performing students and then trigger mechanisms such as making suggestions to students and instructors for changing behaviors. These results and our ability to use Web mining in E-learning are quite preliminary, and there is a need for further exploration and possible adaptation of the forms and usage of Web mining to best suit education.

References

Buchner, A. and Mulvenna, M.D. (1998). Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. *SIGMOD Record,* 27(4):54-61

Chakrabarti, S., Dom, B., and Indyk P. (1998). Enhanced hypertext categorization using hyperlinks. In *Proc. of the ACM SIGMOD Conference on Management of Data*, Seattle, Washington: ACM Press**,** pp. 307—318

Cooley, R. Mobasher, B. Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. *In Proc. of IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA, pp558~567

Etzioni, O. (1996).The World-Wide Web: Quagmire or Goldmine? *Communications of the ACM*, vol. 39, no. 11, pp65-68.

Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 2002, 113 (Spring), pp17-36.

Padmanabhan, B. and Tuzhilin, A. (1998). A Belief-Driven Method for Discovering Unexpected Patterns. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). AAAI, Newport Beach, California, USA, pp94--100.

Quinlan, J. (1993). C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA

Manuscript received 26 Feb 2006; revision received 25 May 2007.