IJICT International Journal of Information & Communication Technology

ITRC

Volume 3- Number 1- January 2011 (67-72)

Technical Note

Improving the Performance of E-Learning Systems Using Extracted Web Server Log File Suggestions

Sadegh Sulaimany
IT &Computer Engineering Department
University of Kurdistan
Sanandaj, Iran
S.Sulaimany@uok.ac.ir

Behrooz Maghsoudi
Computer Engineering Department
Islamic Azad University
Zanjan Branch
Zanjan, Iran
Behroozmagsodi@Gmail.com

Ali Amiri
Computer Engineering Department
Zanjan University
Zanjan, Iran
A_Amiri@iust.ac.ir

Received: October 7, 2010- Accepted: December 27, 2010

Abstract—**There are many opportunities to improve E-Learning web-based applications with regard to continue challenges e.g. their lack of interaction between teachers and students and structural evaluation of the presented learning activity. In this paper, E-Learning system web server log data and information about its control panel are provided; then the gathered data were preprocessed. After extracting association rules form these data and selecting results using more confidence coefficients, they were used as a virtual consultant for improving the E-Learning system. MySQL was used as the database for storing and extracting patterns. Extracting association rules with more confidence coefficients was done using Weka Software. The selected E-Learning web application was Moodle and the virtual university case study was Iran University of Science and Technology. The obtained results showed the needs which cannot be granted by a human consultant easily and cannot be inferred from current application log directly.**

Keywords- E-Learning systems, Web server log data, Data mining

I. INTRODUCTION

Today, the World Wide Web is one of the most important media for collecting, sharing and distributing information. Specifically, in higher education, web-based technologies facilitate the distribution of findings. Web-based education provides various fortunes and possibilities for educational courses such as [1, 2]:

- Supporting almost every classroom activity
- Facilitating interaction and feedback
- Supporting various types of learning; Associative learning, Discussion-based learning, Student-based learning and Source-based Learning
- Making time and location flexible
- Overcoming limitation on the number of classroom participants
- Sharing and reusing resources

Special systems that are designed for managing and tracking learners, course design, evaluation, etc. are called Learning Management Systems or LMS. In spite of their advantages, LMSs suffer from specific

aspects of performance decrease. Thus, it is necessary to make them productive and achieve the maximum benefit of their potential as well as having a good understanding of their defects in order to increase their performance efficiency.

One of the most important defects of LMSs with regard to professors and teachers is the lack of close student-educator relationship. Educators cannot track and evaluate all user activities and course structure and its effects on the learning process. Users may not also be able to state their reason of unwillingness to work with the system during interaction, which leads to system inefficiency [2, 3].

It is clear that LMS, as a web application, needs a special web server platform to run on, and to be accessible via the Internet. IIS (Internet Information Service), from Microsoft, and open-source Apache, are the most common web servers today [4]. Log files of such web servers provide a collection of raw data about user behavior while browsing the web, e.g. resource referring pattern and its usage. Some E-learning applications like Moodle also provide comprehensive logs form user usage of the system and his or her behavior. These data can be used for several purposes such as data mining, which is the idea of this paper. Using data mining over web server Log files has many advantages as follows:

- Professors can track and evaluate students' E-learning process and find most frequent browsing patterns of students with respect to the presented courses, which enables professors to find their course structure performance.
- Desired activities and resources help personalize E-learning environment for students.
- Webmaster (Website administrator) is informed of the performance parameters such as pages with errors, web server normal traffic and required CPU usage for the web server, etc. and can make users more satisfied by applying desirable changes.
- Stockholders are more informed of user behavior; hence, they organize human and academic resources and try to improve E-learning services.

The data provided form web servers or E-learning software logs are raw and stockholders and students do not know much about extractable information and the relationship among different parts of them. This paper attempts to extract associative relationships from raw log files of E-learning web servers using data mining techniques. Examples from the extracted rules that can be used for advising students are as follows:

IF (Browser = IE6 AND Entrance Time = Night AND Gender = Female) THEN Grade =15

IF (Occupation = Staff AND Entrance Time = Night AND Gender = Male) THEN Grade = 9

IF (Grade = 12–15 AND Gender = Male) THEN Website Options = Weak

Clearly, not every extracted rule is useful. The extracted rules should be filtered and prioritized based on their value and usability.

II. Literature Review

Many works have been done so far to improve web-based E-learning systems. Clustering methods have focused on categorizing learners based on common attributes [5]. Learners are clustered according to their learning styles and results are used for improving educational level. [6] grouped questions and exams based on students' answering styles. Also, [7] categorized learners based on their web site entrance time. Paper results show that noticeable achievements have been achieved via clustering.

Prediction is another common method that can forecast the value of a specific field using educational data. Student reaction was predicted against learning policies in [8] while, in [9], students' final grades were foreseen and [10] predicted success ratio of E-learners in LMSs.

Associative rules try to extract rules from relations among the data. This method was used in [11] in order to make related recommendations for students. [12] also used associative rules for troubleshooting and giving recommendations for students in learning contexts. Another related work in [13] tried to investigate the relations among learner behaviors using associative rules extracted by data mining.

III. Method

Most resources have a 3-step process for mining web server log files [2,14,15]:

1. Preprocessing
2. Pattern mining
3. Pattern analysis

Raw data available from the present investigated E-learning application log file is usually diverse and incomplete; and it is difficult to use them directly for useful pattern acquisition. Not all of them are necessarily useful for improving LMS functionality. These logs are stored in MySql tables including information such as time, date, client IP address, server name, server port number, user authentication, exact name and address of accessed resources on the web server, request status like success, failure or redirect, bytes sent, bytes received, HTTP protocol version, client browser type and version, etc [16].

Such data need preprocessing and filtering to be used in the data mining process. Web server log data cleaning takes around 80% to 95% of time and effort of the preprocessing phase (Fig. 1) [1].

The method presented here for extracting rules is based on the following state chart diagram (Fig. 2). In this paper, data storage is done using MySql and preprocessing is done by Weka. Weka is also used for finding associative rules and investigating confidence values.

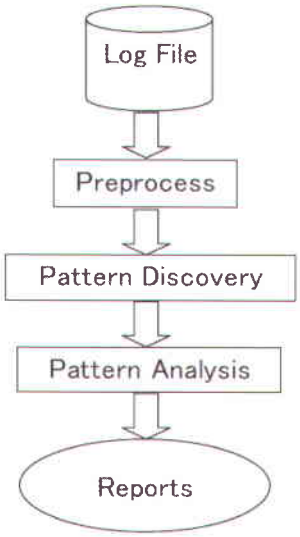

Fig. 1 Web server log mining general system structure (the idea of the pictrue was derived from [14])

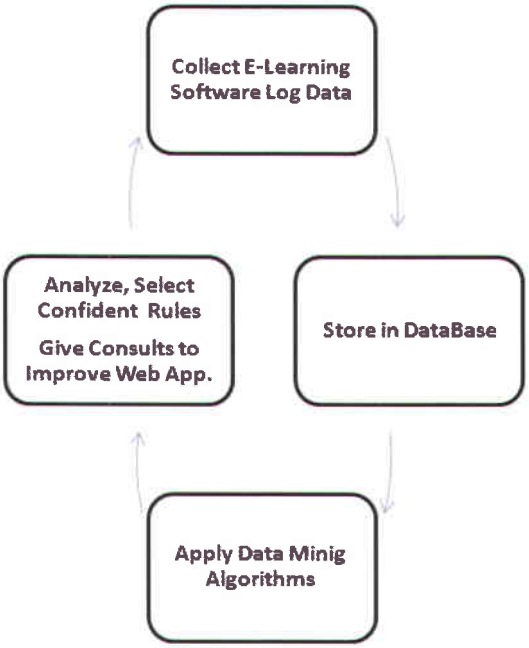

Fig. 2 Diagram of method discussed in this paper

IV. Data Prepartion

A. Data Collection

The present method has generality in working with any web server log files; but, since the current E-learning software was appropriate for case study and its options for logging were extensive, its log file was used which was created directly using the E-learning software itself. Moodle, as the selected E-learning software for this study, is an open-source learning course management system to help educators create effective online learning communities. Moodle keeps detailed logs of all activities done by students including keeping track of what materials students have accessed. Moodle logs every click that students make for navigational purposes and has a modest built-in log viewing system (see Fig. 3). Log files can be filtered based on course, participant, date and activity.

Two steps should be passed in order to collect required data form Moodle database:

1. Selecting suitable fields
2. Filling an abstract table

At the first step, it is necessary for a data miner to select meaningful tables from among the large number of Moodle database tables and separate the required properties form each table. The most important tables used for this paper are mdl_chat s, mdl_forum, mdl_forum_posts dl_fossions and mdl_user.

The second step should include collecting and integrating several types of student information from different tables. This leads to data mining facilitation and cleaning, especially for the following steps. It was done for the fields in Table 2 and these fields were used for finding associative rules. Finally, they were integrated into a single table by means of SQL commands.

Although information was available from 800 students in 50 courses corresponding to different Moodle courses in Iran University of Science and Technology (IUST), only 4 courses and 200 students were selected because of their higher number of Moodle activities.

V. Data Cleaning

There are three main tasks for data cleaning: deletion of noise data, data digitalization and field type converting. Noise data are the ones with big variance. They are far from society mean and they are not suitable for data mining. So, they are removed from the system. For example, if there are 15 to 20 units for 95% of student registrations and a small number of them currently have 5 to 8 units, then these students should be removed.

Digitalization is another important task. Data are organized based on some specified rules. Categories with few numbers of data entries should be deleted because they have explicit differences with other normal data. In this paper, presence time at web application is divided to 3 zones of low, medium and high.

The last step is done in order to identify and correct data errors and inconsistency and increase data quality for performing data mining operations. Data errors may be a result of syntax errors, wrong data or null values. Duplicate and repeated entries should also be removed to provide a better data set. Fig. 4 shows Weka output on data dispersal. Horizontal axis represents different parts of website such as courses, forums, blogs, etc. Each color represents a special action in its area. The data far from the center of focus was considered noise and should be deleted. For example, black arrow in Fig. 4 is pointing to the message part of website and red color displays delete action. Thus, it can be inferred that a few number of users have used the delete action on messages; hence, this small number of users do not affect the data mining process and should be deleted. The data pointed with blue arrow in Fig. 2 also show a special one which is far from its center and is not suitable for data mining; thus it is better to be ignored.

id	time	userid	ip	course	module	cmid	action	url
1	1190371124	2	127.0.0.1	1	user	0	update	view.php?id=2&course=1
2	1190371258	2	127.0.0.1	1	course	0	view	view.php?id=1
3	1190371419	2	127.0.0.1	1	user	0	view	view.php?id=2&course=1
4	1190371477	2	127.0.0.1	1	user	0	update	view.php?id=2&course=1
5	1190371477	2	127.0.0.1	1	user	0	view	view.php?id=2&course=1
6	1190371483	2	127.0.0.1	1	user	0	logout	view.php?id=2&course=1
7	1190371503	2	127.0.0.1	1	user	0	login	view.php?id=2&course=1
8	1190371504	2	127.0.0.1	1	course	0	view	view.php?id=1
9	1190371793	2	127.0.0.1	1	course	0	view	view.php?id=1
10	1190371817	2	127.0.0.1	1	user	0	logout	view.php?id=2&course=1
11	1190371832	2	127.0.0.1	1	user	0	login	view.php?id=2&course=1
12	1190371833	2	127.0.0.1	1	course	0	view	view.php?id=1
13	1190373026	2	127.0.0.1	1	course	0	view	view.php?id=1
14	1190373090	2	127.0.0.1	1	course	0	view	view.php?id=1
15	1190373343	2	127.0.0.1	1	course	0	view	view.php?id=1
16	1190289170	2	127.0.0.1	1	course	0	new	view.php?id=2

Fig. 3 Sample Moodle log file structure

Table 1 Important Moodle Tables

Name	Description
mdl_user	Information on all users.
mdl_user_students	Information on all students.
mdl_log	Logs every user's action.
mdl_assignement ssii	Information about each assignment.
gnment_subm	Information on assignments submitted.
mdl_chat s	Information on all chatrooms.
dl_chat_userm	Which users are in which rooms
dl_choice	Information on all choices.
mdl_glossary	Information on all glossaries.
mdl_survey	Information on all surveys.
mdl_wiki	Information on all wikies.
mdl_forum	Information on all forums.
mdl_forum_posts dl_fossions	Stores all posts to the forums.
rum_discu	Stores all forums discussions.
mdl_message ms	Stores all current messages.
dl_ message_readmdl_ quiz	Stores all read messages. Information about all quizzes.

Table 2 Fields selected for this paper

Field Name	Value
ID	Student Identification Number
Log Time	Time duration of being active at web site
Post	Number of topics posted on website
Search	Number of searches in website
Archive	Times course archives are visited
Virtual	Times the user participates in Virtual Courses
Source	Times course resources are visited
Forum	Times the forum is viewed.
Assignment	Times homeworks are referred to.
Time in n	Number of nighttime visits.
Time in m	Number of daytime visits.
Grade	Final grade at related course.

VI. Extracting Appropriate Patterns

Virtual learning environments make several explicit and implicit problems for learners. Extracted patterns can be helpful in solving many student problems with E-Learning web-based applications. As explained later, some results of this paper are not discoverable easily from E-learning web application interface and normal usage by means of human factor. Also some other result rules is more important than their system stockholders my think or predict.

After collecting data from different resources and cleaning and making them consistent, it is necessary to extract associative, implicit and hidden rules. Finding association rules means discovering relations between attributes and their values from a database as follows [17]:

IF <Some conditions are true> THEN <Predict value for other related attributes>

After the "If" statement, "Antecedent" is called, and after the "Then" statement, "Consequent" is called. Metrics used for finding associative rules normally have the following features:

1. Support count: Adjusted records qualify all the condition rules.
2. Comprehensibility: Number of existing attributes for a rule.

Such algorithms have two phases:

1. Finding frequent item sets
2. Making minimum confidence

Finding minimum support and minimum confidence parameters is also another difficulty of these algorithms.

After performing Apriori algorithm on existing data, 200 rules with different confidence coefficients were extracted. Then 35 rule were selected out of the

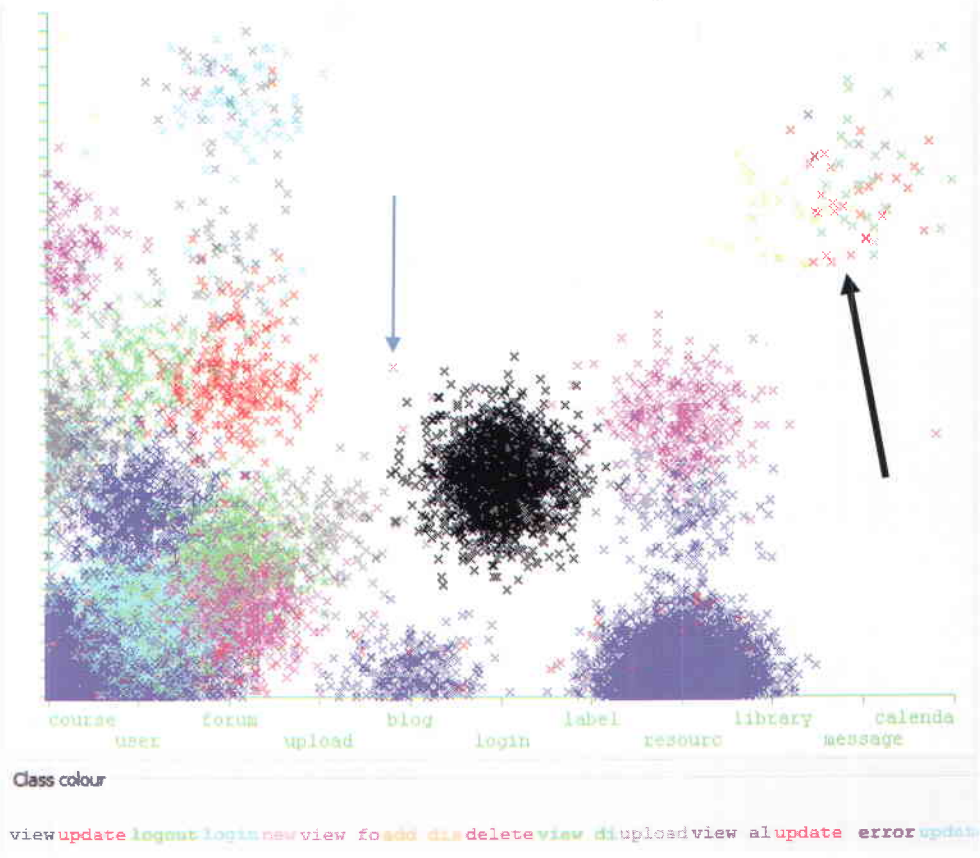

Fig. 4 Data dispersal in Weka for noise elimination

rules whose their confidence coefficients was greater than 60.

Some of the selected rules that were extracted by Apriori algorithm using Weka are shown below:

If (time= h and Source=m and Forum =h) then grade= c Accuracy: 0.70

This rule shows that if being active at the website takes long and student visits many forums, and if his or her course of study resources is at a medium level, then her or his grade will be between 10 and 15 out of 20.

If (virtual=m and source=l and search= m) then grade =b Accuracy: 0.80

According to this extracted rule, if students' presence at a virtual class is at a medium level and its resources of study are rare, and the search level is high, then the final grade will be between 10 and 15.

If (time in n=h and virtual=m) then Forum=h Accuracy: 0.60

This rule defines the high access level to forums for students with a long-time activity at night and average class attendance.

If (assignment=m and post=h and timein=h) time in m=l

According to this rule, if students' are active during the night and the number of posts they send is high, while their homework review is average, then their daytime reference rate is low.

Here, only some examples of extracted rules are briefly presented while several extra examples can be shown with complete results. Clearly, inferences of rules also depend on the responsive point of view.

VII. Conclusion and Future Works

A new method was presented for virtually consulting electronic learners based on data mining algorithms and raw log data available from an E-learning web-based application web server. Some hidden and implicit conclusions were made available that helped learners and stockholders to use the system better or change it and make it more useful. Guidelines are mostly a combination of E-learning control panel options and user behaviors. Another feature of the findings of this paper is their capability of being used in analyzing the feasibility of E-learning web-based applications options. All cleaning, data preparation and rule extraction steps were done using Weka. Future works can include more accurate predictions and information by prioritizing selected attributes for data mining and selecting and using more important ones.

VIII. Acknowledgment

Many thanks to Mr. Farzan Badakhshan, head of IEEE student branch at University of Kurdistan who revised the earlier version of the paper.

References

[1] K. R. Suneetha and R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File,"

International Journal of Computer Science and Network Security, vol. 9, p. 6, 2009.

[2] O. R. Za"iane, "Web Usage Mining for a Better Web-Based Learning Environment," 2002. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.799&rep=rep1&type=pdf

[3] M. E. Zorrilla, et al., Web Usage Mining Project for Improving Web-based Learning Sites, *Lecture Notes in Computer Science, 2005, Volume 3643*, 2006.

[4] Wikipedia Website, Web server, http://en.wikipedia.org/wiki/Web_server

[5] T. Tang and G. McCalla, "Utilizing Artificial Learners to Help Overcome the Cold-Start Problem in a Pedagogically-Oriented Paper Recommendation System," in Proceedings of the International Conference on Adaptive Hypermedia, 2004, pp. 245-254.

[6] J. Spacco, T. Winters, and T. Payne, "Inferring use cases from unit testing," in Proceedings of the AAAI Workshop on Educational Data Mining, New York, USA, 2006, pp. 1-7.

[7] F. H. Wang and H. M. Shao, "Effective personalized recommendation based on timeframed navigation clustering and association," Expert Systems with Applications, vol. 27, no. 3, 2004, pp. 365–377.

[8] G. Chen, C. Liu, K. Ou, and B. Liu, "Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology," Journal of Educational Computing Research, vol. 23, no. 3, 2000, pp. 305–332.

[9] Behrouz Minaei-Bidgoli and Bill Punch, "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System," Genetic and Evolutionary Computation, vol. 2, 2003, pp. 2252–2263.

[10] W. Hamalainen and M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems," in Proceedings of the 8th international conference in intelligent tutoring systems, Taiwan, 2006, pp. 525– 534.

[11] O. Zaïane, "Building a recommender agent for e-learning systems," in Proceedings of the International Conference on Computers in Education, 2002, pp. 55–59.

[12] G. J. Hwang, C. L. Hsiao, and C. R. Tseng, "A computer-assisted approach to diagnosing student learning problems in science," Journal of Information Science and Engineering, vol. 19, 2003, pp. 229–248.

[13] P. Yu, C. Own, and L. Lin, "On learning behavior analysis of web based interactive environment," in Proceedings of the Workshop on Implementing Curricular Change in Engineering Education, Oslo, Norway, 2001, pp. 1-10.

[14] Z. Yang, *et al.*, "An Effective System for Mining Web Log," 2006. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.4578

[15] L. Guo, *et al.*, "Use Web Usage Mining to Assist Online E-Learning Assessment," in *IEEE International Conference on Advanced Learning Technologies (ICALT'04)*, 2004.

[16] M. H. A. Wahab, *et al.*, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm," *World Academy of Science, Engineering and Technology*, vol. 48, 2008.

[17] Han. J. and Kamber. M., *Data Mining concepts and techniques* , Morgan Kaufmann publishers, 2006.

Sadegh Sulaimany got his M.Sc. degree in Computer Science from the Amirkabir University of Technology in 2006. Recently he is a lecturer and researcher at Information Technology and Computer Engineering Department of University of Kurdistan, Iran. His current research interests are E-Learning, Computing Education and some topics in Web Applications.

Behrooz Maghsoudi is a M.Sc. student at the Computer Engineering Group of Engineering Department of Islamic Azad University of Zanjan from 2010. He has published several papers about data mining. His research interest is educational data mining. He is also lecturer at Azad University Sanandaj branch.

Ali Amiri received the Ph.D. degree in Artificial Intelligence in 2010 from Iran University of Science and Technology (IUST), Tehran, Iran. Since 2011, he has worked in the Computer Engineering Group of Zanjan University as an Assistant Professor. His research interests include data mining, video analysis and pattern recognition.