DISCOVERING USAGE PATTERNS FROM WEB SERVER LOGS

Fadzilah Siraj, Nooraini Yusoff and Mohd Helmy Abd. Wahab

Department of Computer Science
Faculty of Information Technology
Universiti Utara Malaysia
06010 Sintok
Kedah, Malaysia

Email: fad173@uum.edu.my, nooraini@uum.edu.my, mhzcyber@yahoo.com

Abstract: *As the amount of information available on the World Wide Web (WWW) increases rapidly, the number of sites that hold particular information also increases. In order to have some insights o the site usage, system administrator needs tools that can aid in his usage site's analysis. To achieve this goal, the use of web mining too is necessary to discover the usage pattern of a particular site. For the purpose of this study, server logs from the educational portal were retrieved, pre-processed and analyzed. Information collected by the Web servers are kept in the server logs and used as the main source of data for analyzing users' navigation patterns. Once the server logs have been preprocessed and sessions have been obtained, there are several kinds of access pattern mining that can be performed, depending on the needs of the analyst. In this study, data mining technique known as Generalized Association Rule was used in order to get some insights into website usage pattern. The findings from this study provide an overview of the usage pattern of particular educational portal. The study also demonstrates how Generalized Association Rule can be used in site usage analysis. Such a technique enables the discovery of hidden information within the web server logs using data mining technique.*

INTRODUCTION

As the amount of information available on the World Wide Web (WWW) increases rapidly, the number of sites that hold particular information also increases. To date, more than 1,000,000,000 pages are indexed by search engines, and finding the desired information on WWW is not an easy task (Pat *et al.*, 2002). Hence with the explosive growth in size and usage of the WWW,

discovering and analyzing useful information from the WWW becomes a practical necessity. For example, it is important for Web administrators to track and analyze the navigation patterns of Web site visitors. The scale of the web data exceeds any conventional databases, and therefore there is a need to analyze the data available on the web. In addition, there are also needs from the users of the web and business built around the web to benefit more from the web data. For example many users still complain from the poor performance of the websites and the difficulty to obtain their goal in the current websites because of the poor site structure or mismatches between site design and user needs (Pramudiono, 2004).

Research in web mining is at cross road of several research communities such as database, information retrieval, and within artificial intelligence (AI), especially in sub areas of machine learning, natural language processing (Kosala and Blockeel, 2000) and in business and e-commerce domain areas (Mobasher *et al.*, 1996). Webmining can be broadly defined as the discovery and analysis of useful information from the WWW (Mobasher *et al.*, 1996). It is currently one of popular techniques for analyzing data from WWW (Paramudiono, 2004) that uses data mining techniques to automatically discover and extract information from Web documents. Web mining can be broadly defoned as the discovery and analysis of useful information from the WWW (Mobasher, 1996). It is an integrated technology of various research fields including computational linguistics, statistivs, informativs, artificial intelligence (AI) and knowledge discovery (Fayyad *et al.*, 1996; Lee and Liu, 2001), Srivastava *et al.* (2002) classified Web Mining into three categories: Web content mining, Web Structure Mining and Web Usage Mining. These categoried of Web Mining are illustrated in Fig. 1.

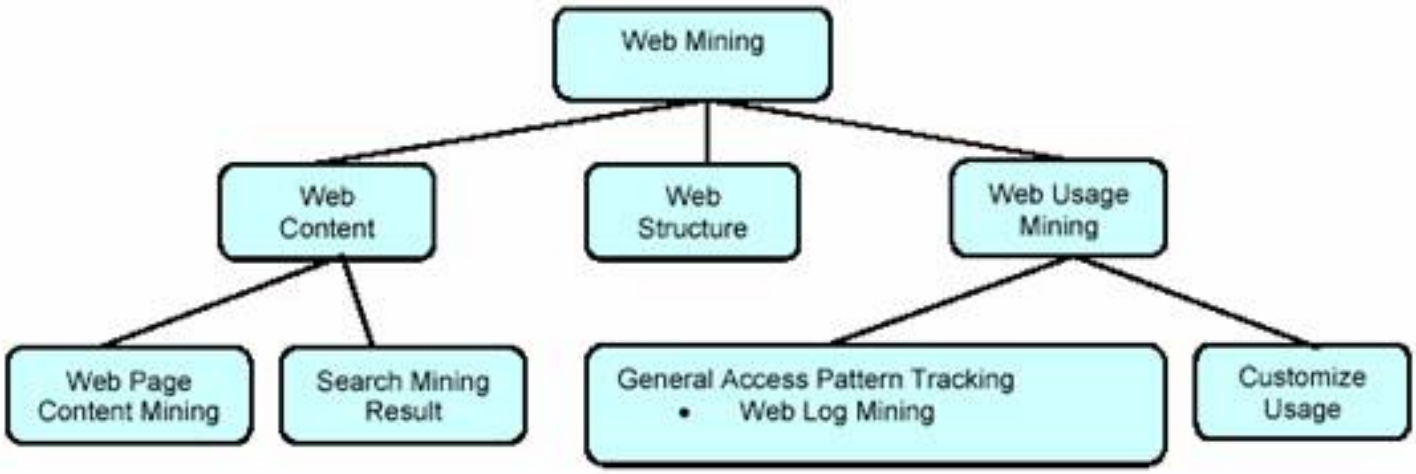


Figure 1: Taxonomy of Web Mining

Web Content Mining involves mining web data contents (Madria *et al.*, 1999) ranging from the HTML based document and XML-based documents found in the web servers to the mining of data and knowledge from the data source. It consists of two domain areas: Web Page Content Mining and Search Mining Result.

Web Structure Mining (Chakrabarti *et al.*, 1999) used to discover the model underlying the link structures of the web. Based on the structural data used, Web Structure Mining cn be further divided into two types, namely Hyperlinks and Document Structure. This Web Mining technique reveals more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the documents (*Pal et al.*, 2002).

While content mining and structure mining utilize the real or primary data on the web, Web Usage Mining focuses on discovery of meaningful patterns from data generated from clients-server transactions on one or more web servers in order to study the navigation behaviour and access patterns of website visitors (Mobasher *et al.*, 1999). Web usage data includes data from server access logs, proxy server logs, browser logs, user profiles, registration files, users' sessions or transaction, users' queries, bookmark folders, mouse clips and scrolls, and any data generated by the interaction of users and the web (Pal *et al.*, 2002). Cooley *et al.*, (1997) suggested that web usage mining includes data related to the usage of the pages of

a website such as IP address, page references and the date and time access. As mentioned before, the mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally the data will be classified into the usage data that resides in the Web clients, proxy server and servers (Srivastava *et al*., 2000).

This study focuses on web usage mining, which analyzes the history of users' behaviour in the form of access patterns recorded in web access logs of web server. Organizations often generate and collect large volume of data in their daily operations. Most of this information is usually greeted automatically by web servers and collected in server access logs. Other sources of user information include referrer logs which contains information about the referring pages for each page referred, and user registration or survey data gathered via tools such as CGI scripts.

To have some insights on the site usage, system administrator needs tools that can aid in his usage site's analysis. To achieve this goal, the use of web usage mining tool is necessary to discover the usage pattern of a particular site.

WEB USAGE MINING

Web usage mining is a research field that focuses on the development of techniques and tools to study users' web navigation behavior. Understanding the visitors' navigation preferences is an essential step in the study of the quality of an electronic commerce site. In fact, understanding the most likely access patterns of users allows the service provider to customise and adapt the site's interface for the individual user (Perkowitz and Etzioni, 1997), and to improve the site's static structure within the underlying hypertext system, (Rosenfeld and Morville, 1998).

When web users interact with a site, data recording their behavior is stored in web server logs. These log files may contain invaluable information characterizing the users' experience in the site. In addition, since in a medium size site log files amount to several megabytes a day, there is a necessity of techniques and tools to

help take advantage of their content.

The web usage mining process could be classified into two commonly used approaches (Borges and Levene, 1999). The first approach maps usage data of the Web Server into relational tables before an adapted data mining techniques is performed. The second approach uses the log data directly by utilizing special pre-processing techniques. Mining behavioural patterns from web log data needs data cleansing, user identification, session identification and path completion.

In order to have some insights on the site usage, system administrator needs tools that can aid in his usage site's analysis. To achieve this goal, the use of web mining tool is necessary to discover the usage pattern of a particular site. For the purpose of this study, server logs from the educational portal were retrieved, pre-processed and analyzed. Information collected by the Web servers are kept in the server logs and used as the main source of data for analyzing users' navigation patterns.

Log Files and Data Preparation

Web log file analysis is used by IT administrators to ensure adequate bandwith and server capacity on their organizations website (Wilson, 1999). Log file data can offer valuable insight into web site usage. It reflects actual usage in natural working condition, compared too the artificial setting of a usability lab. It represents the activity of many users, over potentially long period of time, compared to a limited number of users for an hour or two each.

The flow of the system used in this study is illustrated in Fig. 2

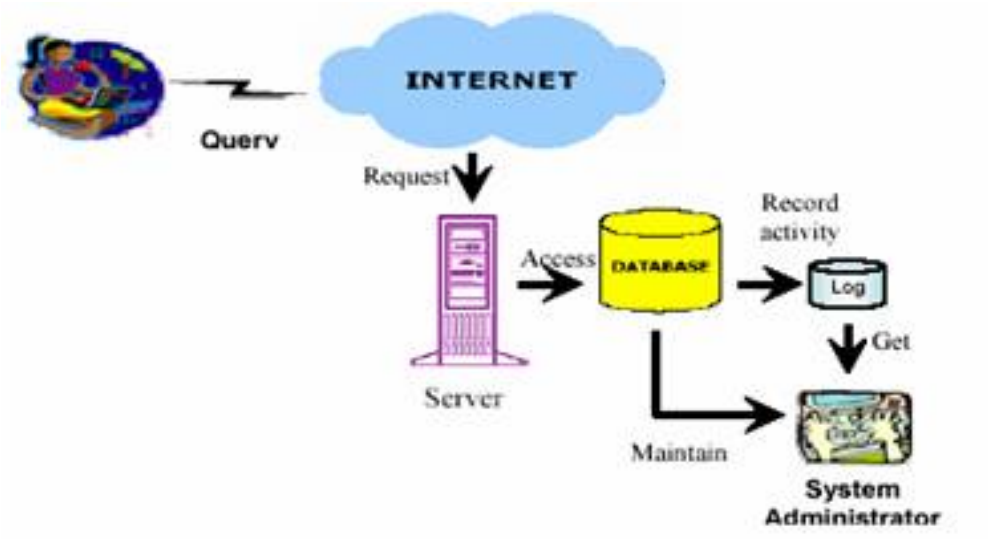


Figure 2: Flow how the log file is created

Adopting from Xue *et. al* (2002), the Generalized Association Rule is depicted in Figure 3.

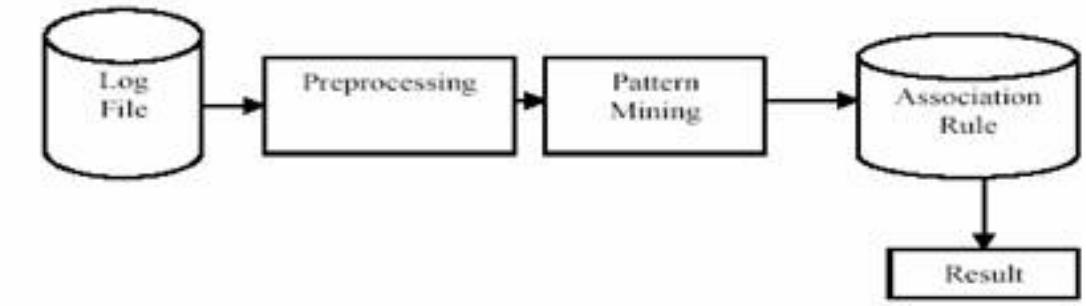


Figure 3: Generalized Association Rules

For the analysis purposes, the log file dated 24th November 2003 was retrieved from the portal server www.tutor.com.my. One of the main problems encountered when dealing with the log files is the amount of data needs to be preprocessed (Drott, 1998). Sample of a single entry log file is displayed in Fig. 4. There are 19 attributes identified from the log file, in this study, the attributes such as cookies, hostname, server IP will be removed.

```
2003-11-23 16:00:13 210.186.180.199 - CSLNTSVR20 202.190.126.85 80 GET
/tutor/include/style03.css - 304 141 469 16 HTTP/1.1 www.tutor.com.my
Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+98;+Win+9x+4.90)
ASPSESSIONIDCSTSBQDC=NBKBCPIBBJHCMMFIKMLNNKFD;+browser=done;+ASPSESS
IONIDAQRRCQCC=LBDGBPIBDFCOKHMLHEHNKFBN http://www.tutor.com.my/
```

Figure 4: Single entry of raw log file

From the technical point of view, Web usage mining is the application of data mining techniques to usage logs of large data repositories. The purpose of it is to produce result that can be used to improve and optimize the content of a site (Drott, 1998). However, before applying data mining algorithm, data preprocessing must be performed to convert the raw data into data abstraction necessary for the further processing (see Table 1).

Table 1: Preprocessed Log File

Trans	ClientIP	Datetime	Method	ServerIP	Port	URI Stem
0	202.185.122.151	11/23/2003 4:00:01 PM	GET	202.190.126.85	80	/index.asp
1	202.185.122.151	11/23/2003 4:00:08 PM	GET	202.190.126.85	80	/index.asp
2	210.186.180.199	11/23/2003 4:00:10 PM	GET	202.190.126.85	80	/index.asp
3	210.186.180.199	11/23/2003 4:00:13 PM	GET	202.190.126.85	80	/tutor/include/style03.css
4	210.186.180.199	11/23/2003 4:00:13 PM	GET	202.190.126.85	80	/tutor/include/detectBrowser_cookie.js

Web Usage Mining Techniques

Various techniques have been used in web usage mining such as adaptive neural network to visualize the Website usage patterns (Perotti, 2003) and path analysis to determine most frequently visited paths in a Website (Mobasher, 1996). In addition, association rules were applied to database of transactions in order to find the support and confidence for each transaction that consists of a set of item in the database (Mobasher, 1996; Agrawal and Srikant, 1994), rule classification to

classify new data items that are added to the database (Mehta *et al.*, 1996; Cheeseman and Stutz, 1996; Han *et al.*, 1993; Weiss and Kullikowski, 1991) and sequential patterns were used to predict user visit patterns (Mannila *et al.*, 1995; Srikant and Agrawal, 1997). Other techniques such as clustering analysis allows one to group together clients or data items that have similar characteristics to facilitate the development and execution of future marketing strategies in both online and offline (Kaufman and Rousseew, 1990; Fisher, 1995; Ng and Han, 1994).

One of data mining techniques that is commonly used in web mining is association rules. Since its introduction in 1993 (Agrawal *et al.*, 1994), the task of association rule mining has received a great deal of attention. In brief, an association rule is an expression X => Y, where X and Y are sets of items. The meaning of such rules is quite intuitive: Given a database D of transactions where each transaction T º D is a set of items, X => Y expresses that whenever a transaction T contains X than T probably contains Y also. The probability or rule confidence is defined as the percentage of transactions containing Y in addition to X with regard to the overall number of transactions containing X.

Drott (1998) offers ways of manipulating the log files in designing the site to reduce the file size and answer specific questions such as those involved in tracking links, searches, paths, and initial contact by the user. This research also points out the limitations of the data exist within the log files, provide basic suggestions on how to view the data while Stout (1997) provides suggestion on how to design website tailored to specific user audiences. Wong *et al.* (2001) proposed a new methodology based on case based reasoning approach to discover user access patterns by mining fuzzy association rules from the historical web log data.

Mobasher *et al.* (1996) has developed a prototype known as proposed a framework for web mining and developed a prototype known as WEBMINER. They identified that WEBMINER does not have the facility to perform cluster analysis, association rules and sequential pattern discovery. In conjunction with Mobasher *et al.* (1996) an effective techniques for capturing user profiles based on

association rules discovery and usage based clustering was proposed by Mobasher *et al.* (1999). The results from the two methods were integrated with current status of on-going activity to perform real time personalization while Toolan and Kushmerick (2002) uses web usage mining to deliver Personalized Site Maps that are specialized to the interest of each individual visitor. Xue *et al.* (2001) have used re-ranking method and generalized association rules to extract access patterns of site's pattern usage. In addition, generalized association rule has been used by Nanopoulos *et al*. (2000) on web perfetching. Yang (2002) has conducted a comparative study on different kinds of sequential association rules for web document prediction and shows that the existing approaches can be classified into two important dimensions, namely the type of antecedents of rules and on the other hand the criterion for selecting prediction rules. Lin *et. al* (2001) investigate the use of association rule mining as an underlying technology for collaborative recommender systems.

In the Web domain, the pages, which are most often referenced, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions (Cooley
et. al., 1999). Cooley (2000) pointed that in the term of the Web usage mining, the association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. The support is the percentage of the transactions that contain a given pattern.

For the purpose of mining the results of Tutor.com, a Generalized Association Rule is adopted from Xue *et. al* (2002). The flow of this rule is as shown in Fig. 3.

RESULTS

The web usage mining in this study can be divided into two separate stages. The first stage is the preprocessing and data preparation, including data cleaning, filtering and transaction identification. The second is the mining stages in which usage pattern are discovered by a generalized association rule mining. Based on

the client IP, the distribution of countries accessing Tutor.com is illustrated in fig. 5. The pie chart indicates that United States and Malaysia are the two top countries that access Tutor.com within a day. One interesting finding from the graph is that the American visited Tutor.com more often than Malaysian.

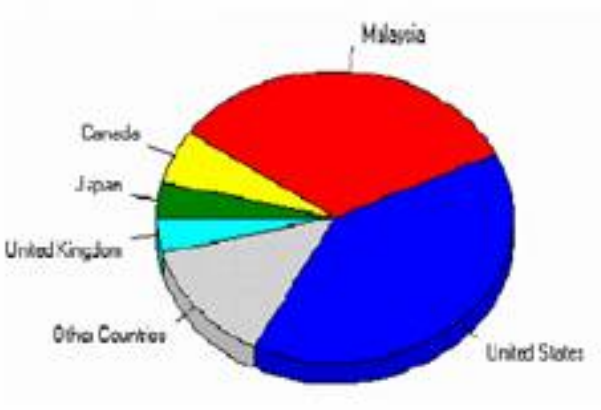


Figure 5: Top country access to Tutor.com

Based on Universal Resource Indicator (URI) stem, the users only accessed the host http://www.tutor.com.my without navigating other options provided by Tutor.com (see Fig. 6)

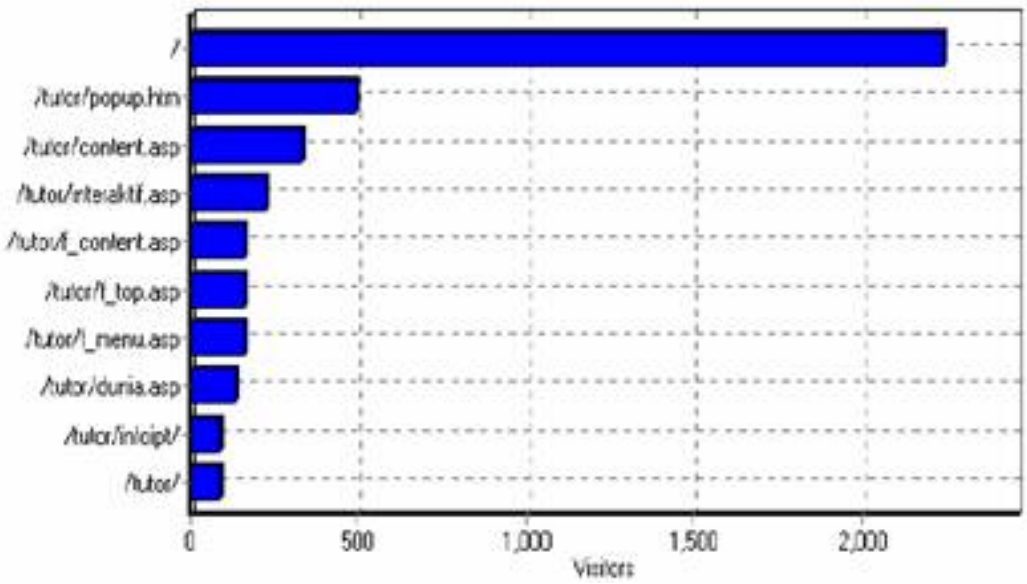


Figure 6: Top visitors

The first level of Tutor.com consists of eleven element option as shown in Fig. 7. The bar charts shown in Fig. 7 reveal that Estidotmy has the highest support and

confidence values. This implies that the most popular option is the news option. The second popular is the game option. The result illustrated in Fig. 7 also indicates that Tuisyen is one of the least popular options. This might be due to the fact that user's access to Tuisyen option based on users' queries. The query session was stored in a separate database. Therefore this session cannot be accessed through the server logs. Another option that is emphasized in this study is question bank. The analysis of this option is further explained in Fig. 8.

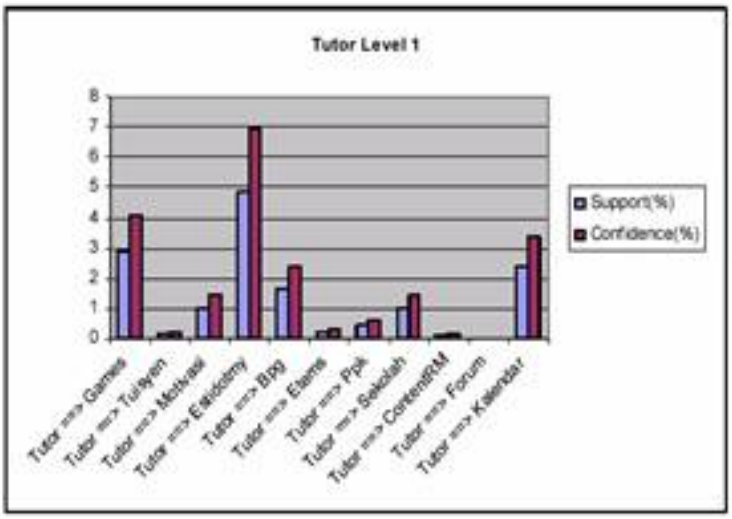


Figure 7: Support and Confidence for level 1 for Portal Tutor.com

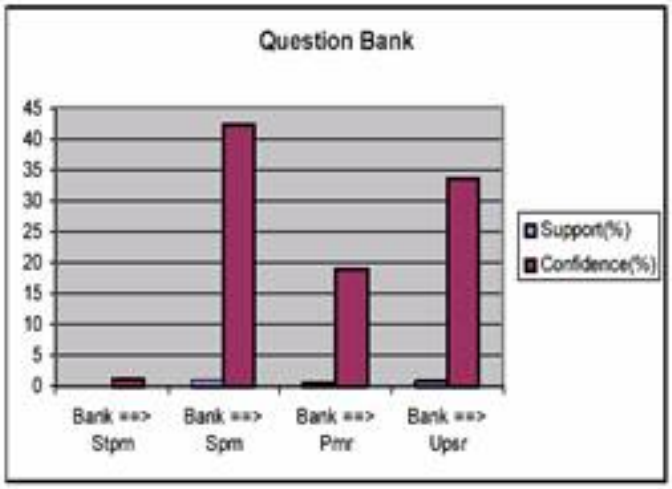


Figure 8 : Support and Confidence for Question Bank

It is interesting to note that SPM question bank scores the highest confidence

compared to other examination question banks. The least confidence is shown by STPM question bank. The graph in Fig. 8 also indicates that UPSR question bank is also popular option. Since the server logs were retrieved on 24 November 2003, this implies that both SPM and UPSR question bank are still popular even after the UPSR and SPM examinations.

To determine which SPM subject are more popular to users, further association rules mining was performed on SPM access (refer to Fig. 9). The results indicate that Mathematics question bank has been accessed more frequent compared with other subjects. The next popular subject is Additional Mathematics, followed by Chemistry. The least popular subject is History. Hence, the findings in Fig. 9 show that science question bank at Tutor.com has been visited more frequently than arts subjects.

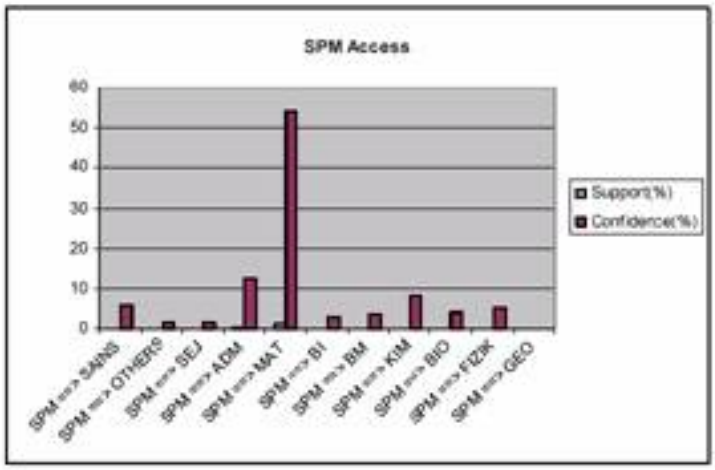


Figure 9 : Support and Confidence for SPM

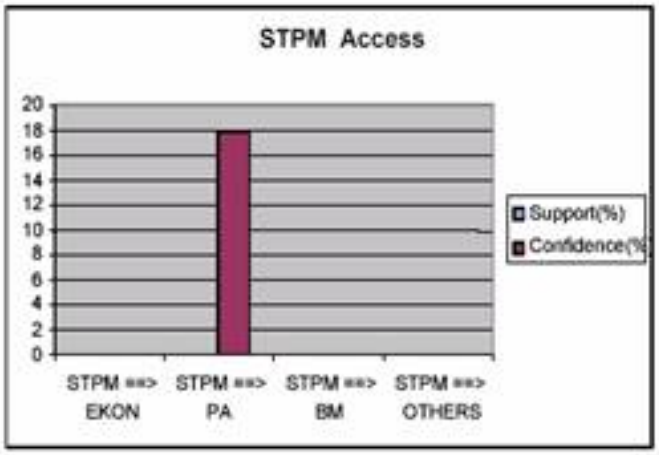
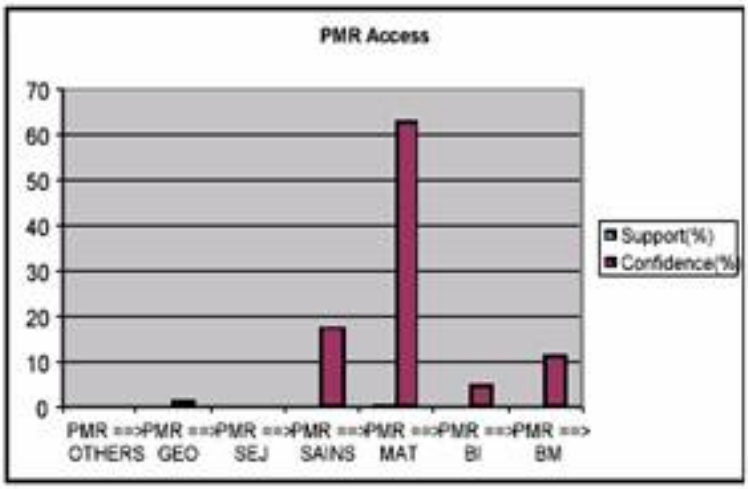


Figure 10 : Support and Confidence for STPM

Result exhibited in Fig. 8 indicate that STPM question bank was the least popular. Further analysis on STPM's access reveals that only one subject, that is General Paper (PA) question bank was referred by Tutor.com's users.

The PMR access results shown in Fig. 11 also indicate that Mathematics is the most popular subject among PMR's examination subjects. The next popular subject at PMR level is a science subject. Comparing Fig. 9 and Fig. 11, the results exhibit that Mathematics and Science subjects are more popular than arts.

Figure 12: Support and Confidence for UPSR



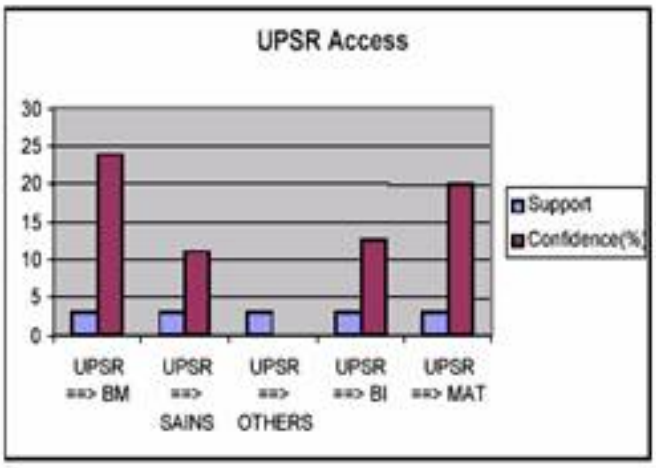


Figure 11: Support and Confidence for PMR

For UPSR's access, Bahasa Melayu subject is more popular followed by Mathematics and Science. One possible explanation for Bahasa Melayu preferences may be due the fact that there are two Bahasa Melayu's papers in UPSR examination.

CONCLUSION

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent year (Kerkhofs *et al.*, 2001). Commercial companies as well as academic researchers have developed an extensive array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behaviour on a particular web site. Performing this kind of investigation on the web site can provide information that can be used to better accommodate the user's needs.

Web usage mining has been applied to several applications such as business and finance (Lee and Liu, 2001), E-commerce (Srivastava *et al.*, 2000), information

retrieval (Pal *ei al.*, 2002). In this study, generalized association rules have been applied to web server log from Tutor.com. An important finding from this study is that Mathematics subject generally popular from SPM, PMR and UPSR levels. On the contrary, arts subjects are not popular to Tutor.com users. The system administrator may consider to evaluate the content and the link for such subjects, so that the real problem can be identified.

Future work on this study includes more sophisticated techniques for data preprocessing and the identification of access sessions, in order to alleviate common problems of Web Usage Mining. Other algorithms for pattern discovery will also be included in the system, in order to provide alternative methods such as apriori algorithm and adaptive clustering technique can be explored for further enhanced analysis.

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. of the 20th VLDB Conference*. pp 487-499.

Borges, J. and Levene, M. (1999). Data Mining of User Navigation Patterns. *Proceedings of the WEBKDD '99 Workshop on Web Usage Analysis and User Profiling*. pp.31 – 36.

Chakrabarti, S., Dom, B., Gibson, D., Klienberg, J., Kumar, S., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Mining the Link Structure of The World Wide Web. *IEEE Computer*. Vol. 32. No. 8. pp. 60 – 67.

Cheeseman, P. and Stutz, J. (1996). Baysian classification (autoclass): Theory and results. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R.

Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press. pp. 153-180.

Cooley, R., Mobasher, B., and Srivastava, J. (1997).Web Mining: Information and Pattern Discovery on the World Wide Web. *Technical Report TR 97-027.*

Cooley, R., Tan, P.N., and Srivastava, J. (1999). WebSIFT: The Web Site Information Filter System. *Proceedings of the Web Usage Analysis and User Profiling Workshop(WebKDD '99).*

Cooley, R. (2000). Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. *PhD thesis*. Dept. of Computer Science, University of Minnesota

Drott, M. C. (1998). Using Web Server Logs to Improve Site Design. *Association for Computing Machinery (ACM) Proceeding of the Sixteenth Annual International Conference on Computer Documentation. pp. 43 – 50.*

Fayyad, U. M., Piatetski-Shapiro, G., Smith, P. (1996). From Data Mining to Knowledge Discovery: An Overview. *Advances in Knowledge Discovery and Data Mining*, pp. 1 – 34.

Fisher, D. (1995). Optimization and simplification of hierarchical clusterings. *Proc. of the*
First Int'l Conference on Knowledge Discovery and Data Mining. pp. 118-123.

Han, J., Cai, Y., and Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Eng.* Vol. 5. pp. 29-40.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to*

Cluster Analysis. John Wiley & Sons.

Kerkhofs, J., Vanhoof, K., and Pannemas, D. (2001). Web Usage Mining on Proxy Server: A Case Study. *Technical Report. Limburg University Centre.*

Kosala, R. and Blockeel, H. (2000). Web Mining Research: A Survey. *ACM SIGKDD*. Vol. 2. Issue 1, pp. 1 – 14.

Lee, R. S. T and Liu, J. N. K. (2001). iJADE eMINER: A Web-Based Mining Agent Based on Intelligent Java Agent Development Environment on Internet Shopping. *PAKDD 2001*. LNAI 2035. pp 28 – 40.

Lin, W. Alvarez, S. A., and Ruiz, C. (2001). Efficient Adaptive – Support Association Rule Mining for Recommender Systems. *Kluwer Academic Publisher.*

Mannila, H., Toivonen, H., and Verkamo, A. I. (1995). Discovering frequent episodes in
Sequences. *Proc. Of the First Int'l Conference on Knowledge Discovery and Data*
Mining. pp. 210 – 215.

Mehta, M., Agrawal, R., and Rissanen, J. (1996). SLIQ: A fast scalable classifier for data
mining. *Proc. of the Fifth Int'l Conference on Extending Database Technology*.

Mobasher, B., Jain, E., Han, E., and Srivastava, J. (1996). Web mining: Pattern discovery
from world wide web transactions. *Technical Report TR 96-050.*

Mobasher, B., Cooley, R., and Srivastava, J. (1999). Creating adaptive web sites through usage-based clustering of URLs. *In Proceeding of the 1999 IEEE Knowledge and Data Engineering Exchage Workshop (KDEX'99)*

(Nov.).

Madria, S., Bhowmick, S. S., Ng, W. K., and Lim, E. P. (1999). Research Issue in Web Data Mining. *Data Warehousing and Knowledge Discovery.*

Nanopoulos, A. Katsaros,D., Manalopoulos, Y. (2000). Effective Prediction for Web User Access: A Data Mining Approach. Data Engineering Lab. Department of Informatics, Aristotle University, Greece.

Ng, R. and Han, J. (1994). Efficient and effective clustering method for spatial data mining. *Proc. of the 20th VLDB Conference*. pp. 144-155.

Pal, S. K., Talwar, V., and Mitra, P. (2002). Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions. *IEEE.*

Perotti, V. (2003). Techniques for visualizing website usage patterns with an adaptive neural network. *The ACM Digital Library*. pp 35 – 40.

Perkowitz, M. and Etzioni, O. (1998). Adaptive sites: Automatically Synthesizing Web Pages. *Proceedings of the fifteenth National Conference on Artificial Intelligence*. pp. 727 – 732.

Pramudiono, I. (2004). Parallel Platform for Large Scale Web Usage Mining. *Phd Thesis.* Department of Computer Science, University of Tokyo.

Rosenfeld, L. and Morville, P. (1998). Information Architecture for the World Wide Web. O'Reilly, Cambridge.

Srikant, R., Vu, Q., and Agrawal, R. (1997). Mining Association Rules with Item Constraints. *American Association of Artificial Intelligence (AAAI).*

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. N. (2000). Web Usage Mining: Discovery and Application of Usage Patterns from Web Data. *ACM SIGKDD Explorations.* Vol. 1. Issue 2. pp. 12.

Stout, R. (1997). *Web Site Stats: tracking hits and analyzing traffic*. Osborne McGraw-
Hill: Berkeley.

Toolan, F., and Kuhmerick, N. (2002). Mining Web Logs for Personalized Site Maps. *First International Wokshop on Mining for Enhanced Web Search.*

Weiss, S. M. and Kulikowski, C. A. (1991). Computer Systems that Learn: Classification
and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann.

Wilson, T. (1999). Web Traffic Analysis Turns Management Data to Business Data. *TechWeb*.
http://www.internetk.com/story/INW19990402S0006 Date Accessed: 25 February 2004

Wong, C., Shiu, S., and Pal, S. (2001). Mining Fuzzy Association Rules for Web Access
Case Adaptation. *Proceeding of the Workshop Program at The Fourth International Conference on Case Based Reasoning 2001*.

Xue, G. R., Zeng, H. J., Chen, Z., Ma, W. Y., and Lu, C. J. (2002). Log Mining to Improve the performance of Site Search. Third Int. Conf. of WISEw ’02.

Yang, Q. (2002). Building Association Rule-Based Sequential Classifiers for Web Document Prediction.