11

Discovering Web Server Logs Patterns Using Generalized Association Rules Algorithm

Mohd Helmy Abd Wahab[1], Mohd Norzali Haji Mohd[2]
and Mohamad Farhan Mohamad Mohsin[3]
[1,2]*Universiti Tun Hussein Onn Malaysia*
[3]*Universiti Utara Malaysia*
Malaysia

1. Introduction

With the explosive growth of data available on the World Wide Web (WWW), discovery and analysis of useful information from the World Wide Web becomes a practical necessity. Data Mining is primarily concerned with the discovery of knowledge and aims to provide answers to questions that people do not know how to ask. It is not an automatic process but one that exhaustively explores very large volumes of data to determine otherwise hidden relationships. The process extracts high quality information that can be used to draw conclusions based on relationships or patterns within the data.

Using the techniques used in Data Mining, Web Mining applies the techniques to the Internet by analyzing server logs and other personalized data collected from customers to provide meaningful information and knowledge. Web access pattern, which is the sequence of accesses pursued by users frequently, is a kind of interesting and useful knowledge in practice (Pei, 2000). Today web browsers provide easy access to myriad sources of text and multimedia data. With approximately 4.3 billion documents online and 20 million new web pages published each day (Tanasa and Trousse, 2004), more than 1 000 000 000 pages are indexed by search engines, and finding the desired information is not an easy task (Pal *et al.*, 2002). Web Mining is now a popular term of techniques to analyze the data from World Wide Web (Pramudiono, 2004). A widely accepted definition of the web mining is the application of data mining techniques to web data. With regard to the type of web data, web mining can be classified into three types: Web Content Mining, Web Structure Mining and Web Usage Mining.

As an important extension of data mining, Web mining is an integrated technology of various research fields including computational linguistics, statistics, informatics, 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 (see Figure 1).

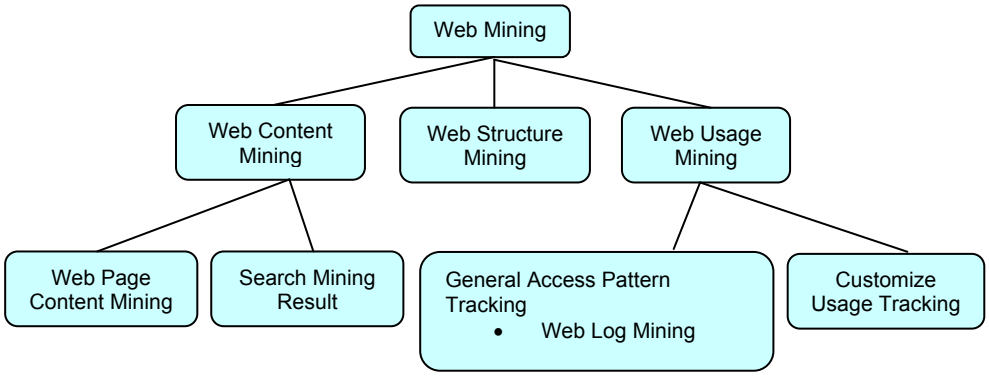

Fig. 1. Taxonomy of Web Mining

Content Mining involves mining web data contents (Madria, 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. Content Mining consists of two domain areas: Web Page Content Mining and Search Mining Result. Content data corresponds to the collection of facts from a Web page that was designed to convey information to the users (Srivastava *et al.*, 2000). It may consist of text, images, audio, video, or structured records such as lists and tables.

Web Structure Mining (Chakrabarti *et al.,* 1999) used to discover the model underlying the link structures of the Web. This type of mining can be further divided into two types based on the structural data used.

- *Hyperlinks*: A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A Hyperlink that connects to a different part of the same Web page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis of which provides an up-to-date survey (Desikan *et al.*, 2002).
- *Document Structure*: In addition, the content within a Web page can also be organized in a tree structure format, based on the various HTML and XML tags within the page. Mining efforts have focused on automatically extracting document object model (DOM) structures out of documents (Wang and Lui, 1998; Moh *et al.,* 2000).

The amount of information available on World Wide Web (WWW) and databases has increased and is still rapidly increasing (Mohamadian, 2001). For a particular website, normally hundred thousand of users will be accessing a particular site. The administrator of a system has an access to the server log. However, the pattern of site usage cannot be analyzed without the use of a tool. Therefore, Data Mining method would ease the System Administrator to mine the usage patterns of a particular site. The work applied on Education Portal organized by Utusan Malaysia as one of the most popular web portal that offer a variety of services such as question bank, quiz, notes, games and etc.

The findings from this study provide an overview of the usage pattern of Tutor.com Portal. The study also demonstrates the use of Generalized Association Rules in Web Usage

Mining. The outcome of this study can be used by the Tutor.com's System administrator as a guideline in enhancing the use of Tutor.com Web Portals.

2. Association Rule

Today, there are several efficient algorithms that cope with the popular and computationally expensive task of association rule mining. One of data mining techniques that are 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 $X \rightarrow Y$, 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 $T \epsilon D$ is a set of items, $X \rightarrow Y$ 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. That is, the rule confidence can be understood as the conditional probability $p(Y \subseteq T \mid X \subseteq T)$.

Different methods for building association-rule based prediction models using the web logs have been proposed by (Yang *et al.*, 2000) and the proposed model had empirically tested on real web logs. However, there has been no systematic study on the relative merits of these methods. While Borgelt and Kruse (2002) describe an implementation of the well-known apriori algorithm for the induction of association rules that is based on the concept of a prefix tree. A traditional model of association rule mining (Tao *et al.*, 2003) is adapted to handle weighted association rule mining problems where each item is allowed to have a weight. This research goal is to steer the web mining focus to those significant relationships involving items with significant weights rather than being flooded in the combinatorial explosion of insignificant relationships.

In practices, users are often interested in subset of association rules. For example, they may only want rules that contains a specific item or rules that contains children of specific item in hierarchy. While such constraint can be applied as a post-processing step, integrating them into the mining algorithm can dramatically reduce the execution time. Several fast algorithms for mining association rules have been developed by Srikant *et al.* (1997) to discover this problem.

Web mining using Association rules is one of the most commonly used techniques so far. Since this study is exploratory in nature, the association rules technique as employed. To this end, a generalized association rule is selected due to its better performance when compared with the standard association rule (Xue *et al.* (2001; Dunham, 2002).

3. Related Work

3.1 The Challenge

The challenge to discover valuable knowledge from chaotic WWW has become a challenging task to research community. Explosive growth in size and usage of the World Wide Web has made it necessary for Web site 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. 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). However, data mining techniques are not easily applicable to Web data due to problems both related with the technology underlying the Web and the lack of standards in the design and implementation of Web pages.

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). Web mining can be broadly defined as the discovery and analysis of useful information from the WWW (Mobasher *et al.*, 1996). In general, web mining can be classified into web structure mining, web content mining and web usage mining. This study focuses on web usage mining, which analyzes the history of user behavior 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 reference, and user registration or survey data gathered via tools such as CGI scripts.

3.2 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.

Currently, several commercial log analysis tools are available (Stout, 1997). However, these tools have limited analysis capabilities producing only results such as summary statistics and frequency counts of page visits. Statistical analysis is the most common method to extract knowledge about visitors to a Web site (Srivastava, 2000). By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of navigational path while Nakayama *et al*. (2000) also try to discover the gap between the website designer's expectations and visitor behavior. Nakayama *et al.* (2000) uses the inter-page conceptual relevance to estimate the former, and the inter-page access co-occurrence to estimate the latter. The approach focused on website design improvement by using multiple regressions to predict hyperlink traversal from page layout features.

In the meantime the research community has been studying techniques to take full advantage of the information available in log files and the more relevant of such techniques are presented. Various techniques have been used in web usage mining such as adaptive

neural network to visualize the Website usage patterns (Perotti, 2003) and these visual representations support the identification of clusters of Web page that are frequently visited together by users.

4. Steps in Manipulating Log Files

This section presents the step involved in the server log process. The section also discusses the preprocessing of the server logs and pattern mining. In addition, the algorithm of generalized association rules is also presented. However, the lack of standardizing the data preprocessing, we adopting methodology for intersite Web Usage Mining as proposed by Tanasa and Trousse (2004).

4.1 Raw Log Files

The log files are text files that can range in size from 1KB to 100MB, depending on the traffic at a given a website (Novak and Hoffman, 1996). In determining the amount of traffic a site receives during a specified period of time, it is important to understand what exactly; the log files are counting and tracking. In particular, there is a critical distinction between a hit and access, wherein: -

- A hit is any file from web site that a user downloads. A Hit can be text document, image, movie, or a sound file. If a user downloads a web pages that has 6 images on it, then that user "hit" the web site seven times (6 images + 1 text page).
- An access, or sometimes called a page hit, is an entire page download by a user regardless of the number of images, sounds, or movies. If a user downloads a web page that has 6 images on it, then that user just accessed one page of the web site.

In this study, raw log file were collected from Portal Pendidikan Utusan Malaysia or known as Tutor.com. This portal focuses on education and provides more information related to education purposes such as Tutorials, Question Banks, Teaching Guidelines, and etc. For the analysis purposes, data that consists of 82 683 records was retrieved from the server and needed to be preprocessed.

The raw log files consists of 19 attributes such as Date, Time, Client IP, AuthUser, ServerName, ServerIP, ServerPort, Request Method, URI-Stem, URI-Query, Protocol Status, Time Taken, Bytes Sent, Bytes Received, Protocol Version, Host, User Agent, Cookies, Referer. One of the main problems encountered when dealing with the log files is the amount of data needs to be preprocessed (Drott, 1998). A sample of a single entry log file is displayed in Figure 4.2

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;+ASPSESSIONIDAQRRCQCC=LBDGBPIBDFCOK HMLHEHNKFBN http://www.tutor.com.my/

Figure 4.1: Single entry of raw log file

4.2 Preprocessing

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). In this phase, the starting point and critical point for successful log mining is data preprocessing. The required tasks are data cleaning, user identification and session identification. One of the most popular tools for data preprocessing is Log Data Processor (LODAP) which designed to preprocessing the log data (Castellano, Fanelli, and Torsello, 2007). However, we prefer by writing our own algorithm to perform data cleaning which specifically based on the requirement for generalized association rules. The algorithm can be illustrated below (Mohd Helmy Abd Wahab et. al., 2008)

```
1   Const ForReading = 1
2   Const ForWriting = 2
3
4   Sub ReadLog( Physical-Path, ModeFile-1, TypeOfLogFile, ModeFile
5           2,StrTypeOfLogFormat)
6
7   RecordCounter = 0
8   Set  LogReader = Server.CreateObject("IISLog")
9                   LogReader.OpenLogFile LogFilePath, ModeFile-1,
10                    TypeOfLogFile, ModeFile-2, StrTypeOfLogFormat
11
12  LogReader.ReadLogRecord
13
14  While NOT LogReader.EndOfLogRecord
15
16  Retrieve Log Attributes
17  ……
18  ……
19  RecordCounter = RecordCounter + 1
20  LogReader.ReadLogRecord
21  Loop
22
23  LogReader.CloseLogFile
24  End Sub
```

An entry of Web server log contains the time stamp of a traversal from a source to a target page, the IP address of the originating host, the type of request (GET and POST) and other data. Many entries that are considered uninteresting for mining were removed from the data files. The filtering is an application dependent. While in most cases accesses to embedded content such as image and scripts are filtered out. 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.).

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

Table 1. Preprocessed Log File

Table 1 exhibits the sample of preprocessed log file. All attributes after preprocessing cannot be shown in the table above due to the space restriction. After preprocessing completed, the pattern mining was performed to mine the access pattern.

4.3 Generalized Association Rules

In this phase, Generalized Association Rules is used to mine the data in order to obtain the support and confidence for each rule. Generalized association rule is one of the commonly used web usage mining technique. Concept hierarchy is used to illustrate the relationship between options provided by Tutor.com. Concept hierarchy shows the set of relationship between different items, generalized association rules allow rules at different levels (Dunham, 2002). For the exploratory purpose, this study uses the generalized association rules in mining patterns from server logs of Tutor.com due to the hierarchical structure of the options provided. From the log file, a structure of the web site is shown in Figure 2.

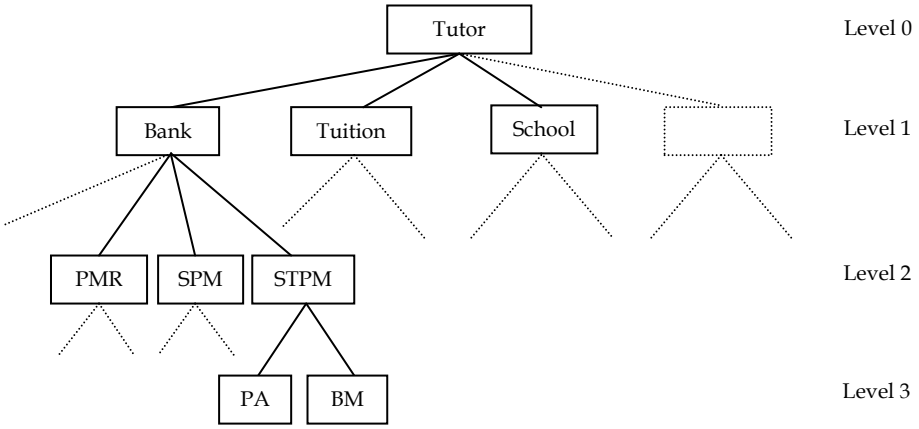

Fig. 2. Hierarchy structure for Portal Pendidikan Utusan Malaysia

Figure 2 depicts the hierarchy structure for Portal Pendidikan Utusan Malaysia. The structure was captured from the Server logs. This study only expands the structure until

level 3, specifically focusing on Question Banks and Tuition which have been selected for applying Generalized Association Rules. Since the Tuition modules are updated daily, it is not possible to apply the generalized association rules because these modules are based on URI query and access to data is from separate database. Due to this problem, only Question Banks were selected for mining the patterns using Generalized Association Rules. Small simulator program was written using Active Server Pages in order to mine the rules and to calculate the support and confidence.

Generalized association rules were applied to mine the useful patterns using support and confidence counting. From the server logs, hierarchy of the websites is determined. To perform this task, generalized association rules is applied until level 3. Comparing with the standard association rules, generalized association rules allow rules at different levels (Dunham, 2002). Generalized association rules were also used to tackle the data diversity problems (Xue *et al*. 2002).

To perform generalized association rules, server logs must be cleaned or filtered. Many entries are considered uninteresting for mining were removed. In the entries such as images *(e.g. *.gif, *.jpg)* and scripts *(*.css, *.js)* are removed.

An algorithm for cleaning the entries of server logs is presented below:

Read *record in database*
For *each record in database*
Read *fields (URI-Stem)*
If *fields = {*.gif, *.jpg, *.css}* ***then***
Remove records
Else
Save records
End if
Next *record*

Fig. 3. Algorithm for data cleaning

To generate the rules, the simulator program is executed and the page will prompt the server logs (see Figure 4)

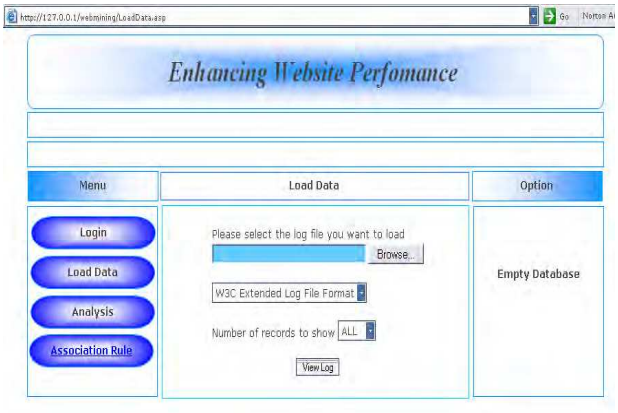

Fig. 4. Starter Prompt for Log File Simulator

Figure 4 illustrates the interface for input the log file, in this simulator; log file will be prompted before the analysis begun. This simulator performs the generalized association rules only. After loading the files generalized association rules can be executed. Rule produced from this simulator is shown in Figure 5.

Fig. 5. Rule produced by the simulator

Figure 5 illustrates the rule produced by the simulator. In this study, the total transaction before cleaning is 82683 records and after cleaning total transaction becomes 38524.

4.3.1 Counting Occurrence Algorithm

Before support and confidence for each rules is determined, number of occurrence for each rules must be calculated. An algorithm for counting the number of occurrence is shown below:

Read *record in database*
For each *record in database*
  If Filter $Item_{Level1} \cap Item_{Level2} <> 0$ ***then***
    $Counter = Counter + 1$
  End if
Next *record*

Fig. 6. Counting Occurrences Algorithm

Figure 6 illustrates an algorithm for calculating the number of occurrence based on the rules for each record and level. The implementation of this algorithm is as follows (see Figure 7).

From Level 1:
Identify child from parent
*For example: **Tutor → QuestionBank***
***Tutor** (parent node)*
***QuestionBank** (child node in level 1)*
*Combine parent and node using **AND** operator then fit in Filter Function*

Fig. 7. Implementation of counting occurrences

4.3.2 Algorithm for Support

Support measures how often the rules occur in database. To determine the support for each rules produced, several arguments have been identified in calculating the support such as Total Transaction in database and number of occurrences for each rules. The formula for support is shown in Figure 8.

Input :-
Total Transaction in DB
No. of occurrences each item {x,y}

$$Support = \frac{Number\ of\ occurrences\ \{x,y\}}{Total\ Transaction\ in\ DB}$$

Fig. 8. Support calculation formula

Based on the Figure 8 below, an example of calculation of support from the mined rules is shown below (see Figure 9).

Total Transaction in Database = 38524 (After cleaning process)
Total occurrence for Tutor → Estidotmy = 310
*Support (Tutor → Estidotmy) = (310 / 38524) * 100*
Support (Tutor → Estidotmy) = 0.80

Fig. 9. Implementation of Support Counting

Based on the Figure 9, the support for rules Tutor → Estidotmy has a value 0.80 percent. This means the probability the Tutor → Estidotmy occur in database is 0.80%.

4.3.3 Algorithm for Confidence

Based on the examples, support measures how often the rules occur in database while confidence measures strength of the rules. Typically, large confidence values and a smaller support are used (Dunham, 2002). Formula for calculating the confidence value is shown in Figure 10.

Input :-

$$Confidence = \frac{Total\ occurrence\ for\ item\ X\ and\ Y}{Total\ occurrence\ for\ item\ X}$$

Total occurrence for item X
Total occurrence for item X and Y

Fig. 10. Confidence calculation formula

Based on Figure 10, the example of calculation of confidence value from rules produced by simulator (see Figure 11).

Calculating confidence for rules Tutor → Estidotmy
Number of occurrence for Tutor = 21179
Number of occurrence for Tutor → Estidotmy = 310
*Confidence (Tutor → Estidotmy) = (310/ 21179) * 100*
Confidence (Tutor → Estidotmy) = 1.46

Fig. 11. Implementation of Confidence Counting

Since the confidence from the above calculation shows that Confidence > Support, it means the rules for Tutor → Estidotmy are interesting rules.
When applying generalized association rules, any unused transaction from the database such as images access *(e.g. .gif, .jpg*), scripts *(e.g. .css, .js)* were removed.

5. Results and Analysis

After generating the rules from the system, the result will be illustrated using graphical representation in order to perform analysis. In addition, the support and confidences of different levels of server portal accessed will be illustrated using bar charts. The system administrator could make a decision from the result illustrated in order to improve or enhance the content, link, site navigation and facilities.

5.1 Support and Confidence for Structure Level 1

The first level of Portal Pendidikan Utusan or Tutor.com consists of eleven elements option as shown in Figure 12. However, the table for support and confidence for Level 1 one can be referred in Table 2.

Rules	Level	Occurrences	Support	Confidence
Tutor --> Kalendar	1	393	1.02	1.86
Tutor --> Forum	1	0	0.00	0.00
Tutor --> ContentRM	1	79	0.21	0.37
Tutor --> Sekolah	1	173	0.45	0.82
Tutor --> Ppk	1	139	0.36	0.66

Tutor -->Etems	1	132	0.34	0.62
Tutor --> Bank	1	313	0.81	1.48
Tutor --> Estidotmy	1	310	0.80	1.46
Tutor --> Motivasi	1	120	0.31	0.57
Tutor --> Tuisyen	1	124	0.32	0.59
Tutor --> Games	1	1118	2.90	5.28

Table 2. Rule Structure level 1

The bar charts shown in Figure 12 reveal that Game option has the highest support, which score 2.90%, and confidence, which score 5.28 %. This implies that the most popular option is the entertainment option. The second popular is the Calendar option, which scores 1.02% for support and 1.86% for confidence. The result illustrated in Figure 12 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 has been emphasized in this study is the question bank. The analysis of this option is further explained in Figure 13

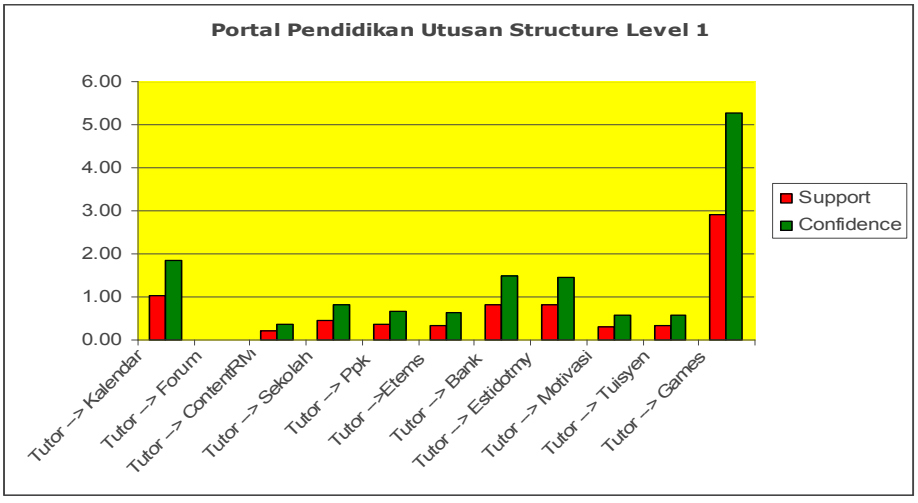

Fig. 12. Support and Confidence for level 1 for Portal Tutor.com

5.2 Support and Confidence for Structure Level 2

Fig 13 shows the graph in the Question Banks. It is interesting to note that SPM question bank scores the highest confidence, which score 48.34 % compared to other examination in question banks. The least confidence is shown by STPM question bank, which score 1.31%. The graph in Figure 13 also indicates that UPSR question bank is also a popular option. Since the server logs were processed, this implies that both SPM and UPSR question bank are still popular even after the UPSR and SPM examinations. Table contains the support and confidence can be seen on Table 3.

Rules	Level	Occurrences	Support	Confidence
Bank --> Upsr	2	537	1.39	33.54
Bank --> Pmr	2	194	0.50	12.12
Bank --> Spm	2	774	2.01	48.34
Bank --> Stpm	2	21	0.05	1.31

Table 3. Rule Structure level 2

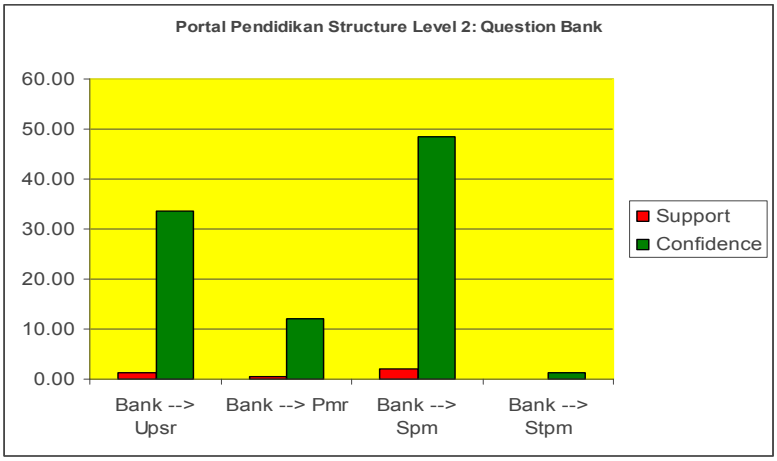

Fig. 13. Support and Confidence for Question Banks

5.3 Support and Confidence for Structure Level 3

Structure of level 3 contains the support and confidence can be seen on Table 4.

Rules	Level	Occurrences	Support	Confidence
Upsr --> BM	3	178	0.46	32.13
Upsr --> BI	3	94	0.24	16.97
Upsr --> Sains	3	82	0.21	14.80
Upsr --> Mat	3	159	0.41	28.70

Rules	Level	Occurrences	Support	Confidence
Pmr --> BM	3	58	0.15	25.66
Pmr --> BI	3	25	0.06	11.06
Pmr --> Mat	3	83	0.22	36.73
Pmr --> Sains	3	45	0.12	19.91
Pmr --> Sej	3	0	0.00	0.00
Pmr --> Geo	3	1	0.00	0.44
Pmr --> Others	3	14	0.04	6.19

Rules	Level	Occurrences	Support	Confidence
Spm --> BM	3	62	0.16	7.69
Spm --> BI	3	46	0.12	5.71
Spm --> Mat	3	134	0.35	16.63
Spm --> ADM	3	157	0.41	19.48
Spm --> Sains	3	64	0.17	7.94
Spm --> Sej	3	26	0.07	3.23
Spm --> Geo	3	0	0.00	0.00
Spm --> Fizik	3	84	0.22	10.42
Spm --> Bio	3	64	0.17	7.94
Spm --> Kim	3	138	0.36	17.12
Spm --> Others	3	24	0.06	2.98
Stpm --> BM	3	3	0.01	1.95
Stpm --> PA	3	27	0.07	17.53
Stpm --> Ekon	3	0	0.00	0.00
Stpm --> Others	3	0	0.00	0.00

Table 4. Rule Structure level 3

For UPSR's access as illustrated in Figure 14, Bahasa Melayu (32.13%) subject is more popular followed by Mathematics (28.70%) and Science (14.80%). One possible explanation for Bahasa Melayu preferences may be due the fact that there are two Bahasa Melayu's papers in UPSR examination.

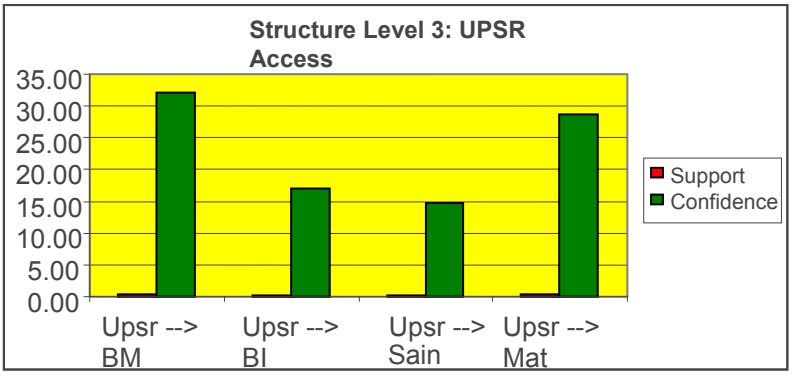

Fig. 14. Support and Confidence for UPSR

The PMR access results shown in Figure 15 also indicate that Mathematics (36.73%) is the most popular subject among PMR's examination subjects. The next popular subject at PMR level is a Bahasa Melayu (25.66%) subject. Comparing Figure 14 and Figure 15, the results exhibit that Mathematics and Bahasa Melayu subjects are more popular than other subjects.

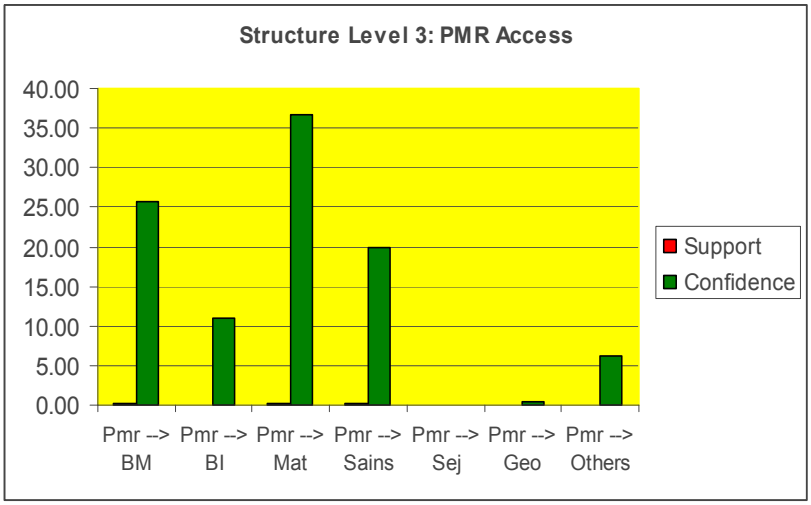

Fig. 15. Support and Confidence for PMR

To identify which SPM subject are more popular to users, further association rules mining was performed on SPM access (see Figure 16). The results indicate that Additional Mathematics question bank has been accessed more frequent which score 19.48% compared with other subjects. The next popular subject is Chemistry, followed by Mathematic. The least popular subject is History. Hence, the findings in Figure 16 show that science stream subject question bank at Tutor.com has been visited more frequently than arts stream subjects.

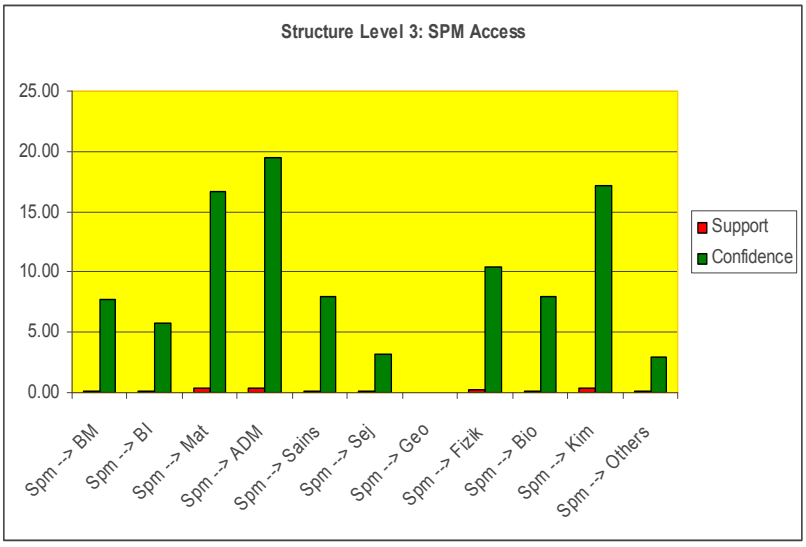

Fig. 16. Support and Confidence for SPM

Results exhibited in Figure 17 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 to by Tutor.com's users and the least subject referred is Bahasa Melayu subject.

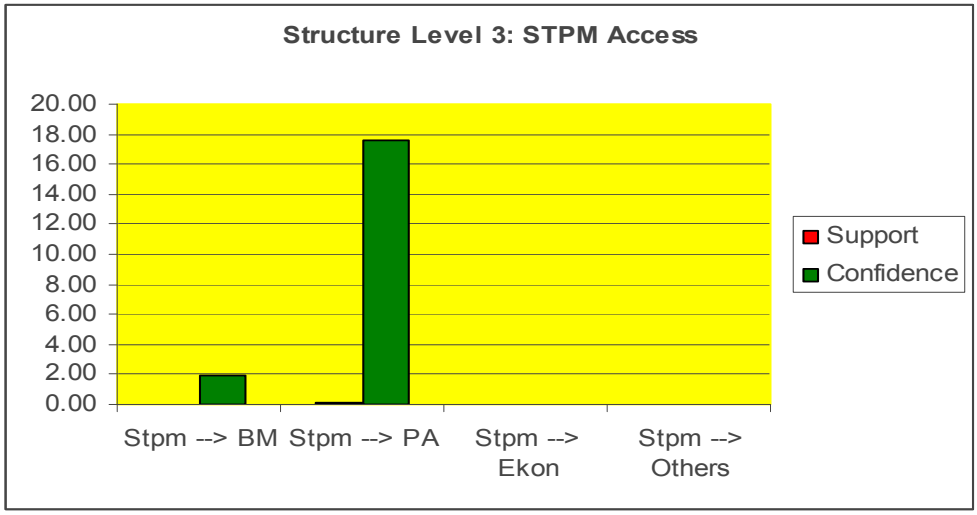

Fig. 17. Support and Confidence for STPM

6. 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, 2000), information retrieval (Pal, 2002) and Academic and Industry (Srivastava *et al.*, 2000). In this study, generalized association rules have been applied to web server log from Tutor.com.

It is important to mention that the most efforts have relied on relatively simple techniques which can be inadequate for real user profile data since noise in the data has to be firstly tackled. Thus, there is a need for robust methods that integrates different intelligent techniques that are free of any assumptions about the noise contamination rate.

7. References

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

Borgelt, C. and Kruse, R. (2002). Induction of Association Rules: Apriori Implementation. *15th Conference on Computational Statistics (CompStat 2002)*

Castellano, G., Fanelli, A. M., and Torsello, M. A. (2007). Log Data Preparation for Mining Web Usage Patterns. In Proceeding of IADIS International Conference on Applied Computing 2007. Pp. 371 – 378.

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.

Desikan, P., Srivastava, J., Kumar, V., Tan, P. N. (2002). Hyperlink Analysis – Techniques & Applications. *Army High Performance Computing Center Technical Report.*

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.*

Dunham, M. H. (2002). *Data Mining: Introductory and Advanced Analysis*. New Jersey: Prenhall.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). Advances in Knowledge Discovery and Data Mining. *AAAI Press/ The MIT Press*.

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 (iJADE) on Internet Shopping. *PAKDD 2001*. LNAI 2035. pp. 28 – 41.

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

Mobasher, B., Jain, E., Han, E., and Srivastava, J. (1996). Web mining: Pattern discovery from World Wide Web Transactions. *Technical Report TR 96-050.*

Moh, C-H., Lim, E-P., Ng, W. K. (2000). DTD-Miner: A Tool for Mining DTD from XML Documents. *WECWIS 2000.* pp. 144 – 151.

Mohammadian, M. (2001). Intelligent Data Mining and Information Retrieval from World Wide Web for E-Business Applications.
http://www.ssgrr.it/en/ssgrr2002w/papers/230.pdf

Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, and Mohamad Farhan Mohamad Mohsin. (2008). Data Preprocessing for Web Server Logs for Generalized Associations Rule Algorithm. In Proceeding of World Academy of Science, Engineering and Technology. Vol. 36. pp. 970 – 978.

Nakayama, T., Kato, H., and Yamane, Y. (2000). Discovering the Gaps Between Web Site Designers' Expectations and Users' Behaviour. *Proc. Of the Ninth Int'l World Wide Web Conference.*

Novak and Hoffman. (1996). *New Metrics for New Media: Toward the Development of Web Measurement Standards*
http://www2000.ogsm.vanderbilt.edu/novak/web.standards/webstand.html
[Date Accessed: 28 February 2009].

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

Pei, J., Han, J., Asl, B. M., and Zhu, H. (2000). Mining Access Patterns Efficiently from Web Logs.

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

Perotti, V. (2003). Techniques for Visualizing Website Usage Patterns With an Adaptive Neural Network. *The ACM Digital Library*. Pp 35 – 40.

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., Desikan, P., and Kumar, V. (2002). Web Mining: Accomplishments and Future Directions.

Srivastava, J., Cooley, R., Tan, P. –N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*. Vol. 1. No. 2. pp. 12 – 33.

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

Tao, F., Murtagh, F., and Farid, M. (2003). Weighted Association Rule Mining using Weighted Support and Significant Framework. *SIGKDD 2003.*

Tanasa, D. and Trousse, B. (2004). Advanced Data Preprocessing for Intersites Web Usage Mining. IEEE Intelligent System, pp. 59 – 65.

Wang and Liu, H. (1998). Discovering Typical Structures of Documents: A Roadmap Approach. *Proceeding of the ACM SIGIR Symposium on Information Retrieval.*

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. *Journal of Data Mining and Knowledge Discovery.*

New Advanced Technologies

Edited by Aleksandar Lazinica

ISBN 978-953-307-067-4
Hard cover, 350 pages
Publisher InTech
Published online 01, March, 2010
Published in print edition March, 2010

This book collects original and innovative research studies concerning advanced technologies in a very wide range of applications. The book is compiled of 22 chapters written by researchers from different areas and different parts of the world. The book will therefore have an international readership of a wide spectrum.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd and Mohamad Farhan Mohamad Mohsin (2010). Discovering Web Server Logs Patterns Using Generalized Association Rules Algorithm, New Advanced Technologies, Aleksandar Lazinica (Ed.), ISBN: 978-953-307-067-4, InTech, Available from: http://www.intechopen.com/books/new-advanced-technologies/discovering-web-server-logs-patterns-using-generalized-association-rules-algorithm

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821