Proceedings of the International Conference on Information Technology and Multimedia at UNITEN (ICIMU' 05), Malaysia
22nd – 24th November 2005

Web Usage Mining in Online Community for Evaluating Staff Performance

Nureize bt Arbaiy, Azizul Azhar b Ramli, Zurinah bt Suradi and Noor Haslina bt Yusoff
Fakulti Teknologi Maklumat dan Multimedia
and
Fakulti Pengurusan Teknologi
Kolej Universiti Teknologi Tun Hussein Onn,
86400 Parit Raja, Batu Pahat, Johor
{nureize, azizulr, zurinah, noor}@kuittho.edu.my
Tel: 07-4538000
Fax: 07-4532199

ABSTRACT

Evaluating performances are perhaps the most obvious and most frequently cited issues in dealing with human capital in organizations. Recently, one of the ways to measure performance is through individual's contribution on the Net. Activities captured by the web site, such as Online Community Portal, are useful information for measuring performance. In this case, data collected in the log file of a website server provide valuable information to web administrator and web designer. However, since the data accumulated were in large quantity, analyzing them create problem. Nevertheless, web mining can be utilized to minimize the problem, as it has been identified as one of the processes to analyze these types of data. The output from this process is identification of usage patterns of the user's access. In this study, Association Rule Mining (ARM), a technique to mine data in log file is used. This technique allocates rules that satisfy user defined constraints on minimum support and confidence with respect to a given dataset. The result of the study shows that users in the faculty prefer to participate in the teaching and learning option as compared to other options. This result would simplify and assist appraiser and superior in the faculty to evaluate work performances of faculty members objectively and clearly. In addition, users of the website can actively participate and inform/share their work performances within the faculty.

Keywords: Web Usage Mining, Performance Evaluation, Data Mining, Online Community

1. INTRODUCTION

The performance of an organization depends on the performance of its members within the organization. However, organizations have to rely on supervisors/superiors to sort out how well individuals/subordinates under their supervisions perform. The aim is that superior can disentangle the effects of job changes, collective action, luck and their own likes and dislikes to make an accurate judgment of how well their subordinates performed over a period of time. But, this goal is rarely realized. The appraisers (or the supervisors) bring their own biases and information-processing problems to the task of performance appraisal, thus the appraisal are often flawed. Instead of creating sense out of a very complex situation, they add in confusion and complexity in measuring performance. This paper proposes a possibility to evaluate individual performance objectively and effectively when much of the information and knowledge related to performance are captured within organization's knowledge based in an Online Community Portal using Web Usage Mining approach.

Online Community is a platform that enables the capturing of enormous content of information in the World Wide Web. Members in the community share and contribute knowledge and information that are of common interest and needs. At the same time, the superior within the community may use the venue to evaluate his/her subordinates' performances in the online community. However, a major drawback from this massive dataset is its inability to manually inspect and analyze the raw data. The resulting growths in on-line information combined with the almost unstructured web data necessitate the development of powerful computationally efficient web mining tools [1]. Thus, it is an obvious candidate for data mining research.

The principle of data mining applied to the web has the potential to be beneficial to an organization. Web Mining is the mining of data related to World Wide Web. It has been used in three distinct ways; Web Content Mining, Web Structure Mining and Web Usage Mining. In this study, Web Usage Mining using basic Association Rules algorithm – Apriori Algorithm, is used to identify and measure staff performance in the Faculty of Information Technology and Multimedia (FTMM), KUITTHO based on the Online Community Portal (E-FTMM) log files that were captured in the portal server.

The E-FTMM portal was configured using Microsoft Windows SharePoint that is a collaborative portal solution. The SharePoint Portal Server extends the capabilities of Microsoft Windows SharePoint Services by providing organization and management tools for SharePoint sites. The E-FTMM portal provides a platform for the faculty members to share knowledge, contribute and communicate with each other. At the same time, the portal endows functions and activities that are consistent and inline to the staff evaluation guidelines.

To date, identifying and measuring performance of staff in the university, generally and the faculty, specifically have been very subjective with no proper and clear benchmark for the evaluation. These lead to job frustration and job dysfunctional amongst staff. A transparent performance evaluation measure may bring potential benefits for the organization and to the staff. Thus, in a learning organization, the performance evaluation should be designed to enable the organization and individuals within the organization to pursue excellence in performing job responsibilities and tasks. Management should be able to promote quality performance and ongoing professional learning and growth. By supporting and monitoring staff contributions using the online community portal, the faculty, specifically and the university, generally can utilized the portal for improvement and growth. At the same time, the system administrator of the online community portal may evaluate and propose a better design of web structure in order to enhance the performance of the web usage itself within the system.

2. RELATED WORK

Performance measurement is a tool that uses statistical evidence to determine progress toward specific defined organizational objectives. [2] states that performance measures are the indicators of the work performed and the results achieved in an activity, process, or organizational unit. According to [3], the purpose of performance measurement is to improve performance. It would seem to make sense to pay particular attention to the areas in which individuals/subordinates are doing inadequately. For that reason, it can be treated as an ongoing process and an iterative process that progresses and has no end. An organization's commitment to performance measurement is a tacit agreement to continually build, change, and improve in order to ensure the survival and growth of the organization

Data Mining and the Web were developed as independent technology areas in the middle of 1990s. The widespread use of the Web nowadays witnesses data overloading and in order to make full use of data in the Web, Web Mining is useful. Essentially, Web Mining is mining databases on the web or mining the usage patterns so that relevant information can be retrieved and provided to the user. In web usage mining, usage patterns of various users and trends of usage are tracked, and predictions are made about users need [4]. Eventually this Web Usage Mining and analysis will help the website administrators to manage the site properly.

[5] and [6] identify taxonomy for Web mining into two categories which are getting patterns form Web data and getting Web logs. This taxonomy was expanded to include three areas that are known as Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining involves mining web data contents [7] ranging from the HTML based document and XML-based documents accumulated in the web servers. The goal of Web Content Mining is mainly to assist or to improve information finding or filtering the information. On the contrary, Web Structure Mining attempts to identify the model that based on the hyperlinks topology and document structure on the Web. Meanwhile, the Web Structure Mining aims to generate structural summary about web sites and web pages which in turn, identify the structure information of the Web. Whereas, Web Usage Mining focuses on the discovery of user access patterns from web server logs. Web server logs generate log files which records web server activity. [8] suggests 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.

One of data mining techniques that are commonly used in web mining is Association Rules. 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 $T \in D$ 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. And, according to [9] the task of Association Rule mining has received a great deal of attention.

As we enter the age of Web technology, the amount of information available on the World Wide Web (WWW) has rapidly increased [10]. In general, for a particular website more than hundred thousand of users will access it. This server transactions activity will be recorded in a server log files and log files produced by the web servers are in the form of text files. This file consists of HTTP transactions that processed by hash coding URLs, IPs, client environment and cookies. The Web logs provide information to ensure adequate bandwidth and server capacity on their organizations website. Log file data can

offers valuable insight of web site usage among users. It reflects actual usage in natural working condition, compared to the artificial setting of a usability lab and log data file represents the activity of many users, over potentially long period of time.

3. METHODOLOGY

The high level process of Web Usage Mining proposed by [11] is adapted in this study. It consists of data pre-processing, pattern discovery (Association Rules) and pattern analysis (see figure 1). The results are then discussed in two sections. Firstly, the pattern discovery that explains general descriptions of the access pattern and users using descriptive statistic. Secondly, the Association Rules that generates the rules with the percentages of support and confidence for the different portal path.

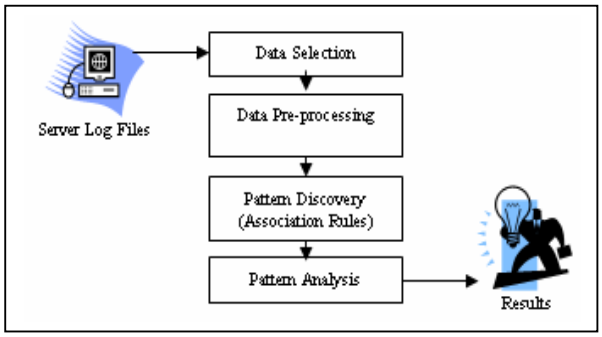

FIGURE 1 - WEB USAGE MINING METHODOLOGY
(Adapted by Srivastava, 2000)

Server Log Files (Data Selection)

Fig. 2 shows how a system administrator can gather information from the server logs. Basically, when a user sends queries to the server, the requested information is retrieved from the database. At the same time, the user session including the URL, Client's IP address, accessing date and time, query stem will be recorded in the server logs. Thus, this server logs will be pre-processed and mined in order to get some insight into the usage of a server site as well as the users' behavior.

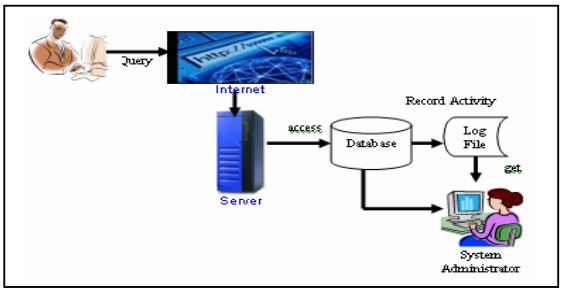

FIGURE 2 – LOG FILE FORMATION

The server log files dated from 1st August 2005 to 26th August 2005 were retrieved from Online Community Portal (E-FTMM) server. A sample of a single entry log file contains the following information:

D902F7E5-71CD-4621-838B-6D0BDB9ACFD6 14:45:18 http://ftmm/research shared documents/ickm'05/ickm'05(azizulazhar).ppt mdzaki Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322) kÒ˜□□ "□□ W ´ È □ J ”

3.1 Pre-processing

Pre-processing is a task of converting the usage, content, and structure information from the data sources into data abstractions necessary for pattern discovery. Before data mining algorithm can be applied, data pre-processing must be performed to convert the raw data into data abstraction necessary for the further processing such as Pattern Analysis. This process involves data extraction and data cleaning tasks.

Data cleaning is part of the data pre-processing tasks. Several important steps are put into action in this process; that is

i. Splitting the logs because the format for Microsoft SharePoint combines the Internet Protocol (IP) address with the Unified Resource Locator(URL)

ii. Removing the Agent Logs from each transaction lines. Agent Logs supply data on the browser, browser version, and operating system of the accessing user [12].
e.g.: Mozilla/4.0

iii. Removing all the transaction contains of pictures extension such as .bmp, .gif, .jpg

iv. Reformating the logs according to Internet Information Service (IIS) format which contains IP address, userid, time, method/URL/Protocol/Status/Size/Referrer/ Agent. This step is taken in order to use the Sawmill 7 tool for descriptive statistics result and Association Rules pattern purposes.

3.2 Pattern Mining - Association Rule

In the Web domain, the pages, which are most often referred, can be put in one single server session by applying the association rule generation. Association Rule mining techniques are used to discover unordered correlation between items found in a database of transactions [13]. [14] noted that in 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.

ARunner 1.0 is used to produce the set of related Association Rules. It is a prototype of *Apriori* algorithm tools that support GUI for *Apriori* application. Fig. 3 shows the interface of *ARunner* 1.0. There are three target types provided: item set, rules and hyperedge. The value for minimal and maximum number of item, percentages of support and confidence can be customized before the *Apriori* process proceeds.

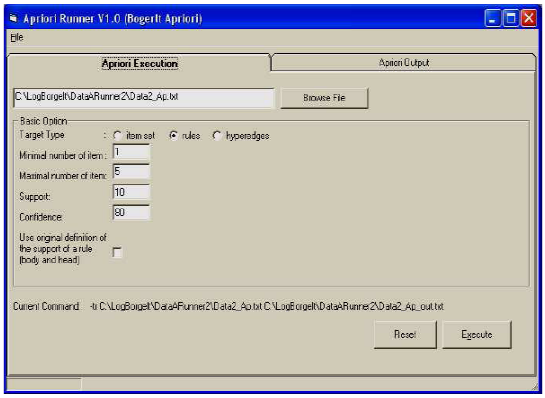

FIGURE 3 – A-RUNNER 1.0

4. RESULTS AND FINDINGS

The results of this study are divided into two sections. The first section discusses the Pattern Mining that is concerns about the general descriptions of the access pattern and users behaviours of E-FTMM portal. For this purposes, Sawmill 7.0.7 is used during the pattern mining process. This tool is used to generate the statistical output. In the second section, ARunner 1.0 is discussed as it is used to produce the Association Rules.

4.1 General Pattern Analysis Results

The Online Community Portal has several options. Each option has a meaningful function. This study concentrates on the Community of Practice option which is sub divided into 7 working areas categories. These categories are general guidelines for staff performance evaluation in the university. The seven categories are teaching and learning, staff development, research and development, writing and publication, supervision, consultation and community service.

Table 1 tabulates the frequency of hit pages from the seven(7) working areas captured from the E-FTMM Portal. While figure 4 graphically illustrates the result. The total number of pages hit by staff is 458 from August 1st to August 26th, 2005.

TABLE 1: PAGES HIT BY WORKING AREAS

Working Areas	Path Option	Frequencies
Teaching and Learning	ftmm/pnpdefault.aspxftmm	179
Research and Development	ftmm/researchdefault.aspxftmm	70
Supervision	ftmm/seliadefault.aspxftmm	58
Writing and Publication	ftmm/tulisdefault.aspxftmm	46
Staff Development	ftmm/hrddefault.aspxftmm	44
Community Service	ftmm/khidmatdefault.aspxftmm	42
Consultation	ftmm/rundingdefault.aspxftmm	19
Total		458

Fig. 4 shows that staff frequently visited pages on teaching and learning. There are 179 hits recorded as compared to the other options. This reflects that these pages are considered by the staff as the main area in performing their activities in the faculty.

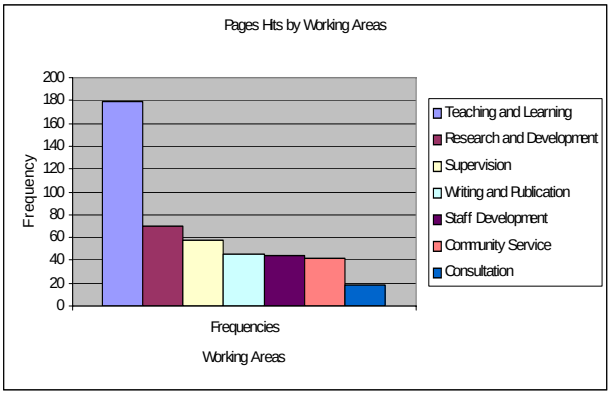

FIGURE 4: PAGES HIT BY 7 WORKING AREAS

In contrast, the consultation option shows the lowest rate of pages visited by the staff. The page has only 19 hits (see Table 1). This shows that the consultation task is less favourable than the other tasks in this 'young' faculty. The other 5 working areas show a fair participation amongst staff selected in this study.

Table 2 shows the distribution of pages hit by the staff in the 7 working areas according to User IDs. From the table, the participation of each staff in the faculty can be monitored. Subsequently, this information can be used to improved the areas with lower participation and strengthen the other areas of activities.

TABLE 2: DISTRIBUTION OF PAGES HIT BY 7 WORKING AREAS BY USERID

UserID	Teaching & Learning	Research & Development	Supervision	Writing & Publication	Staff Development	Community Service	Consultation	
Aaaa	13	1	2	6	1	2	0	**25**
Bbbb	0	3	0	2	3	0	0	**8**
Cccc	64	16	5	5	5	6	3	**104**
Dddd	19	0	3	0	0	0	0	**22**
Ffff	0	0	0	0	0	3	0	**3**
Eeee	10	6	4	2	4	3	4	**33**
Gggg	11	5	13	8	7	5	4	**53**
Hhhh	1	3	1	4	1	1	1	**12**
Jjjj	6	4	0	3	0	5	0	**18**
Kkkk	37	15	17	7	19	9	6	**110**
Llll	4	0	0	0	0	0	0	**4**
Mmmm	0	0	0	1	0	0	0	**1**
Nnnn	0	0	0	2	0	0	0	**2**
Oooo	2	0	1	1	1	1	1	**7**
Pppp	1	17	0	4	0	1	0	**23**
Qqqq	0	0	0	1	1	6	0	**8**
Rrrr	4	0	9	0	2	0	0	**15**
Ssss	7	0	3	0	0	0	0	**10**
	179	**70**	**58**	**46**	**44**	**42**	**19**	**458**

4.2 Association Rules

Association Rules generate the rules based on percentages of support and confidence in the different levels in the E-FTMM portal, by applying the Association Rules technique. The outcomes from this research can be used by Web administrator to plan necessary improvement and enhancement of the Web services. In addition, the improvement of Web sites, including their contents, structure, presentation and delivery to the Community portal would enhance the performance of the web.

Association Rules is used to extract rules in the form of $X \Rightarrow Y$ (if X then Y) quantified with a confidence (proportion of occurrences that verifies Y among occurrences that verifies X) and a support (proportion of occurrences that verifies X and Y among all occurrences). Consequently, a problem encountered where the log data files captured from Microsoft Share Point Server contains insufficient log's attributes. From the observation, the log data files did not record the IP Address for clients (who are FTMM's staff). This maybe an effect from the implementation of Dynamic Host Configuration Protocol (DHCP) technology or the limitation on the IP setting of the server.

Apart from that, to produce Association Rules, unique IP addresses must initially be recorded and identified in order for the staff access pattern behavior in the Community of Practice (E-FTMM) can be monitored.

In this case, Association Rules can not be produced due to the pitfall of unrecorded client IP addresses in the E-FTMM log file. In practice, the ARunner tool is used to produce the set of related Association Rules. The attribute for log files such as IP addresses and pages should be organized in rows according to client's IP addresses before implementatation using ARunner. This tool, then processes the log files to produce Association Rules as a result.

In this case, the Association Rules result can not be processed due to the limitation or insufficient attributes in the E-FTMM log files data. The log data files attributes should be recorded in the IIS format. Thus, in the future study, the researchers shall attempt to examine the IP addresses settings on the Microsoft Share Point Server. Nonetheless, the data mining technique used in the study offers an avenue that would assist the performance evaluation to be carried out objectively in an IT environment.

5. CONCLUSION

The findings from this study provide an overview of the usage pattern of the E-FTMM portal. The result of this study is based on a 26-days-time range on a selected staff who participated in the Online community in the faculty's portal. Based on the results, we envisage that with proper support and monitoring from the management, performance evaluation can be carried out objectively using the data mining techniques applied to the log files in the server.

6. REFERENCES

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

[2] Brinker, B.J., 1997. *Performance Measurement*

[3] Stevens, D.F., 1996. *Principles of Effective Performance Measurement*, Ernest Orlando Lawrence Berkeley National Laboratory Berkeley, CA 94720

[4] Thuraisingham, B., 2003. Web Data Mining and Applications in Business Intelligence and Counter-Terrorism, CRC Press, London

[5] Cooley, R., Taxonomy for Web Mining, private communication, BedFord, MA, August 1998

[6] Shrivastava, J., Web Mining Proceedings of the Next-Generation Data Mining Workshop, Baltimore, 2002

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

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

[9] Agrawal, R. and Srikant, R. 1994. Fast Algorithms For Mining Association Rules. Proc. of the 20th VLDB Conference. pp 487--499.

[10] Mohammadian, M. 2001. Intelligent Data Mining and Information Retrieval from World Wide Web for E-Business Applications. URL http://www.ssgrr.it/en/ssgrr2002w/papers/230.pdf, 2004

[11] Shrivastava, J., Web Mining Proceedings of the Next-Generation Data Mining Workshop, Baltimore, 2002

[12] Bertot, J.C., McClure, C.R., Moen, W.E., and Rubin, J. 1997. Web Usage Statistics: Measurement Issues and Analytical Techniques. Government Information Quarterly. 14(4). Pp 373-395

[13] Cooley, R., ,Mobasher, B., and Srivastava, J. 1999. Data preparation for mining world wide Web browsing patterns. *Knowledge and Information Systems*. Vol 1. No.1.

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