Processing Web Logs in order to Mine Web Usage Patterns

Craig P. Oosthuizen, Janet Wesson & Charmain Cilliers
Department of Computer Science and Information Systems
University of Port Elizabeth, PO Box 1600, Port Elizabeth, 6000
Tel: (041) 504 2323, Fax: (041) 504 2831
Email: {Craig.Oosthuizen, Janet.Wesson, Charmain.Cilliers}@upe.ac.za

Abstract - Web usage mining is the application of data mining techniques to web clickstream data in order to extract usage patterns. These patterns can then be analysed to determine whether a web site is being used as it was intended. When visiting a web site, the only information left behind by users is the trace through the pages they accessed. In order to determine which pages of the web site were accessed and how various web pages were reached, requires examining the raw data recorded in the log files created by the web server. This paper discusses the first phase of web usage mining in order to identify the web usage patterns of the Computer Science & Information Systems web site at the University of Port Elizabeth.

Index Terms – data mining, data warehousing, log file analysis, web usage mining.

I. Introduction

As more organisations make use of the Internet and the World Wide Web to convey information, the conventional approaches to web site evaluation need to be revised [3]. The Department of Computer Science & Information Systems (CS&IS) at the University of Port Elizabeth is an example of an organization which uses a large Intranet to convey information to students. In order for it to be used as it was intended, it must be structured correctly.

The objective of this paper is to discuss the development of a system to analyse the web usage of the CS&IS web site in order to identify firstly, any usage patterns that may exist and secondly, any potential problems with the information architecture. From this analysis, recommendations will be made with regard to the web site design. Thereafter a visualisation tool will be developed to effectively visualise the web usage patterns which were identified.

Before the web site can be improved, its current usage needs to be evaluated. Usability evaluation is the process of collecting data about the usability of a design by a specified group of users for a specific activity within a specified environment [5]. Using traditional methods, in order to evaluate the usage of the web site, a group of users would have to be selected. Certain activities that they are expected to perform would have to be identified and their actions recorded while performing these tasks. It is impractical and costly to carry this out each time the quality of the web site is evaluated. It would be ideal to evaluate the web site based on the data that is automatically recorded by the web server. All activities that take place on a web site are recorded in a file called the web server log. By using data mining techniques, it is possible to extract information from these files which reflect the web site's quality.

II. Web Usage Mining

A web usage pattern or navigation pattern is the sequence of web pages or scripts accessed by a user during a user session [5]. The representation of these pages or scripts originates from the web server log. A user session is the clickstream of web pages for a single user across the entire web [7]. However, only the portion of each user session relating to the CS&IS web site will be used for analysis, since access information is not publicly available from the vast majority of web servers. The set of pageviews in a user session for a particular web site is referred to as a server session (also commonly known as a visit). Web usage mining can be broken down into three main phases, namely preprocessing, pattern discovery and pattern analysis [4], of which the first is discussed below.

A. Preprocessing

The preprocessing phase involves preparing the data for analysis. Part of the preprocessing phase also includes cleaning the server log to eliminate all of the irrelevant items [2]. This can be done by checking the suffix of the URL name and deleting the entries which are of no assistance to the analysis, such as JPG, GIF etc. The preprocessing phase includes the identification of the relevant data in the web site logs and the storage of this data.

B. Data Identification

A structure diagram of the CS&IS web site was drawn up to show the relationship of the web pages to each other. An extract of this diagram is shown in Figure 1. The home page is shown on the left-hand side. Moving towards the right-hand side is the sequence of web pages the user can follow in order to reach a desired page. The leaf node indicated in blue represents a link to another web page within the site. Each access is recorded in the log file.

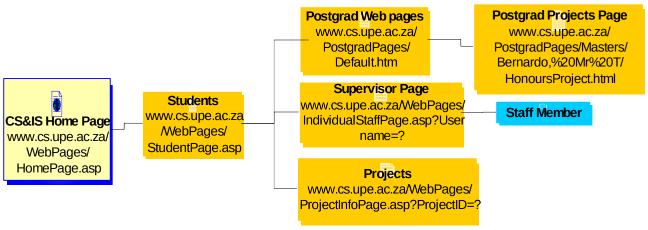

Figure 1. CS&IS Web Site Structure Diagram Extract

Each entry in the log file consists of a sequence of fields relating to a single HTTP transaction with the various fields separated by a space. If a field is unused in a particular entry, it is marked with a dash "-". Table 1 describes the data that will be necessary for the web usage mining.

Attribute	Description
Date	The date on which the activity occurred.
Time	The time the activity occurred.
IP Address	The IP address of the client that accessed the server.
Username	The name of the user who accessed the server
URL	The resource accessed: for example, an HTML page, a CGI program, or a script.
User Agent	The browser used on the client.

Table 1. Description of web log file attributes.

C. *Data Storage*

The results of preprocessing the web server logs will be stored in a data warehouse to facilitate easy retrieval and analysis. Figure 2 illustrates the structure of the data warehouse that will be used to facilitate the mining of web usage data (patterns).

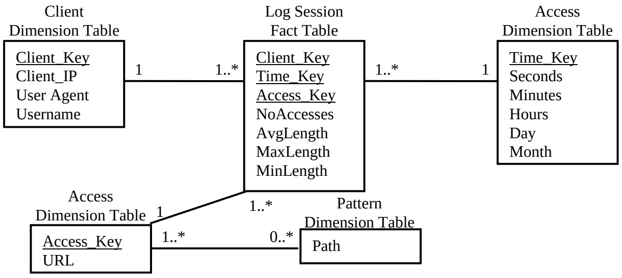

Figure 2. Data Warehouse Snowflake Schema

III. Future Work

Once the data warehouse has been created and populated, various statistical and data mining techniques will be used in order to identify any web usage patterns that exist [1]. An existing application that may be able to assist with this pattern discovery phase is 123LogAnalyzer [8]. These patterns will then be analysed, interpreted and used to determine how well the web site is being used. A graphical representation of these patterns will also be created. An example of how this can be done is shown in Figure 3. The usage pattern is indicated by the dotted line and it can be seen that the shortest route to the target page was not followed. This could be interpreted as a problem as the information on the web site may have been misleading to the user.

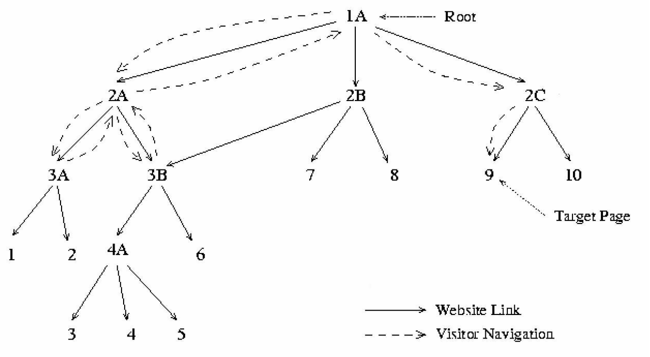

Figure 3. Visualisation Example [6]

IV. Conclusion

Web usage mining can be used to determine web usage patterns on a web site. A data warehouse is needed to store the relevant data in the log files created by the web server. This data represents the page accesses that take place on the CS&IS web site. This data warehouse will assist with the discovery of web usage patterns and determining problems with the information architecture of the web site.

Acknowledgment

We would like to thank the Telkom Centre of Excellence Programme and the Department of Computer Science and Information Systems at the University of Port Elizabeth for making this research possible.

References

[1] BECKER, K. and VANZIN, M. (2003): Discovering interesting Usage Patterns in Web-based Learning Environments. *Utility, Usability and Complexity of e-Information Systems*:57-73.8-9 December 2003.

[2] COOLEY, R. (2003): The use of web structure and content to identify subjectively interesting web usage patterns. *ACM Transactions on Internet Technology(TOIT)* **3**(2):93-116.May 2003.

[3] COOLEY, R., MOBASHER, B. and SRIVASTAVA, J. (1999): Web Mining: Information and Pattern Discovery on the World Wide Web. http://www-users.cs.umn.edu/~mobasher/webminer/survey/survey.html

[4] EIRINAKI, M. and VAZIARGIANNIS, M. (2003): Web mining for web personalization. *Acm Transactions on Internet Technology(TOIT)* **3**(1):1-27.February 2003.

[5] SPILIOPOULOU, M. (2000): Web usage mining for Web site evaluation. *Communications of the ACM* **43**(8):127-134.August.

[6] SRIKANT, R. and YANG, Y. (2001): Mining Web Logs to Improve Website Organization.May 2001.

[7] SRIVASTAVA, J., COOLEY, R., DESHPANDE, M. and TAN, P.-N. (2000): Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD Explorations Newsletter* **Vol I**(2):12-23.January.

[8] ZY COMPUTING, I. (2003): 123 Log Analyzer. San Jose, USA. http://www.123loganalyzer.com

Craig P. Oosthuizen received his B.Sc Hons degree in 2003 from the University of Port Elizabeth. He is presently doing his M.Sc in Computer Science at UPE.