Perplexities in Discovering Navigation Patterns from Server Log

Navin Kumar Tyagi [1], A. K. Solanki [2], Manoj Kumar Sharma[3]

[1]*Bhagwant University Ajmer (Rajasthan), India*
[2]*Department of Comp. Sc. & Engg. BIET, Jhansi (UP), India*
[3]*Department of Comp. Sc. & Engg. MIT, Bulandshahr (UP), India*

[1]nt_1974@rediffmail.com
[3]manojcs2005@rediffmail.com
[2]directormiet09@gmail.com

***Abstract*— Web navigation patterns discovered from usage data can be used to build prediction model to recommend interesting web pages to the users. A user session may have one or more transactions. Identification of transactions or user behaviors from session data is difficult because web pages cannot be classified strictly as navigation or content pages. In order to identify transactions from log data, data preprocessing activities are required. In this paper we review some problems along with their solutions that arise during the discovery of navigation and content pages from server log.**

***Keywords*— Web navigation patterns, Transactions, Data preprocessing, Web pages, Usage data.**

I. Introduction

The need for web navigation aids is like a paradox in this age of information technology and users can access information anytime, anywhere by making the retrieval, production and distribution of information with the help of this technology. On the basis of web data, web mining can be categorized as web structure mining, web content mining & web usage mining. Web usage mining is the application of data mining techniques to large web data repositories [1]. Web data is collected in web server when user accesses the web and represented in standard formats. Web usage data in common log formats[2]consists attributes like IP address, access date and time, request method (GET or POST), URL of page accessed, transfer protocol, success return code etc. In order to discover access patterns, preprocessing is necessary as raw data coming from the web server is incomplete and only few fields are available for pattern discovery. Main objective of this paper is to understand the difficulties and their possible solutions (heuristics) in preprocessing of usage data. On preprocessed data different data mining techniques [3] like statistical analysis, association rules, sequential patterns and clustering can be applied to discover usage access patterns.

This paper is organized as follows. In section II we discuss some related work in this domain. In section III we discuss how web navigation patterns can be used to build prediction based systems. In section IV we reviewed some difficulties involved in preprocessing and conclusion is given in section V.

II. Related Work

In this section we present the main related works in this particular area. In the recent years, much attention has been given on web usage mining by the researchers. However, data preprocessing has received far less attention than it deserves. Methods for user identification, session zing, page view identification, path completion and episode identification are presented in [4].In some other related work [5], the authors compared time based and referrer based heuristics for visits reconstruction. In [6], Marquardt et al. presented the application of web usage mining in the e-learning area which targets on the preprocessing phase. They redefined the notion of visit from the e-learning point of view. In their approach, a learning session or visit can span over several days if this period corresponds to a given learning period. Algorithms for different data preprocessing activities like data cleaning and data reduction are presented in [7].

III. Web Navigation Prediction

Web navigation prediction takes the form of recommended hyperlinks for users to access the information of their interest. Depending on the system, predictions can either be site specific, or unrestricted to the hypermedia. In site specific prediction, although coverage is much narrower, but the system designer has much greater control of possible system inputs, and there is a better chance of constructing an accurate prediction model, while the converse situation holds in the later case.

Personalization can applicable to any web browsing activity, not just e-commerce. Web personalization can be defined as any action that tailors the web experience to a particular user, or set of users. The experience can be something as casual as browsing a website or as significant as trading stocks or purchasing a car. The actions may range from simply making the presentation more pleasing to anticipate the needs of a user and providing customized information.

Most personalization systems for the web are categorized into three major categories: manual decision rule systems, collaborative filtering systems, and content based filtering agents. The new generation of web personalization tools attempts to incorporate techniques for pattern discovery from web usage data. For example, some collaborative filtering systems like net perceptions are experimenting with obtaining implicit user ratings from usage data. Web usage mining systems run any number of data mining algorithms on usage or click stream data gathered from one or more websites to discover user profiles. The increasing focus on web usage data is due to several factors. The input is not a subjective description of the users by the users themselves, and therefore is not prone to biases. The profiles are dynamically obtained from user patterns, and thus the system performance does not degrade over time as the profiles age. Moreover, using content similarity alone as a way to obtain aggregate profiles may result in missing important semantic relationships among web objects. Thus, web usage mining can reduce the need for obtaining subjective user ratings or registration-based personal preferences.

A. *Website Topology*

Topology of the website is determined mainly for two reasons: one it helps in one of the heuristics used to determine pasts user navigation paths by starting a new path if a page access is found which is not reachable by hyperlink from any of the currently visited pages and second the site topology may be useful at the time of prediction by favouring those recommendable pages which are far from the user's current navigation path. Determination of the topology of the website and contents of the HTML pages involved discovering the web pages and their contents, and the site's link structure. The process consist two steps: (1) design and use of a web crawler to discover, parse and code the HTML pages of the website and (2) construction of a graph structure from the discovered HTML pages to represent the site topology. The HTML pages found by the crawler can be used to construct the website topology, which is coded as a graph that have vertices as web pages (nodes), and edges as directed hyperlinks. For each node, the directed hyperlinks are simply the outgoing links which are found on the pages. The graph is constructed using an adjacency-lists representation. The representation comprised an array of linked lists, with each array index representing a vertex, and the entries in each linked list, the nodes that the current vertex linked to. Figure 1 illustrates the representation used.

A → B
B → B, C, I
C → D, F
D → E
E → F, G
F → G, H
G → F
H →
I → B, H, F

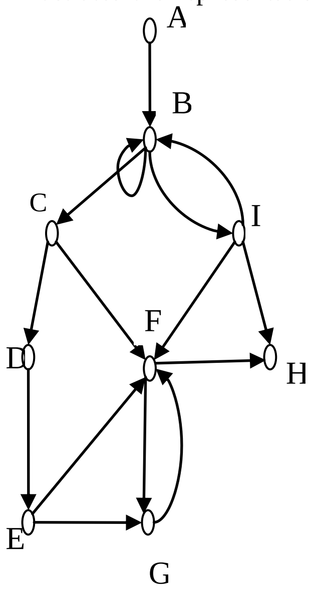

Fig. 1 Graph representation of website

IV. Analysis of problems to determine navigation patterns

There are some difficulties involved in the discovery of navigation and content pages from web server user access logs. One difficulty is the identification of users of the website and second is to determine all the web pages (navigation paths) accessed by each user, each time they browse the website. One more difficulty is to determine the various navigation behaviors (transactions) exhibited by the user for a given session. These difficulties, and heuristics used to address them, are presented below as the user identification problem, navigation path identification problem, and transactions identification problem

A. *User Identification Problem*

It is difficult to identify a user on a website because most of the web servers do not require user authentication. Even websites that require user authentication, there is no certainty that the information provided by users is accurate and reliable because users are very particular about privacy. In order to protect themselves users may register multiple times or provide misleading information. One method to identify users is by using the IP addresses of their machines as a first step. Problems in using IP addresses is that if the client machine is shared by more than one user, which makes the task of identifying different users difficult. On the other hand identifying users whose machines are located behind a proxy server or firewall, which results in a single IP, address (IP address of the proxy server or firewall) reaching the web server is difficult.

B. *Navigation Path Identification Problem*

It is difficult to determine all of the pages that a user visited during a session due to the problems related to caching between user browser and the web server. One problem is related to the working of the HTTP protocol [8]. A single HTML page requested from a web server may contain other URLs, like image, sound, video files, and applets, embedded in it. In order to retrieve these additional files, the HTTP protocol makes a separate connection to the web server to get each of them, and a separate entry is made into the server log files for each of these requests. The problem this introduces to navigation path identification takes several forms. The number of additional pages retrieved may differ as it depends on the number of embedded URLs which the web page designer placed in the requested page. This has an impact on path identification as it results in varying path lengths for the same number of pages requested and the order in which the embedded pages are retrieved is not the same, even if the only difference between two web pages is the order in which the embedded documents are placed in them. This would have an impact on path identification if the sequence of pages accessed is an important consideration. Some of the embedded pages could be pages that the user may sometimes explicitly request, and the web server logs would have no way of distinguishing a page that is explicitly requested from one that just happens to be embedded in a requested page. Another problem that hinders accurate tracking of a user's navigation behaviour is the failure of some user requests to be recorded in web server logs. This problem results from the presence of caches at proxy servers and at the user's browser. Caches are used to improve client-server interaction by making available, cached pages, instead of requiring a new connection to the web server for every page request. In figure 2, three visitors are accessing a website. Two of them connected to web server through a proxy server, and the third connected directly to the web server. Proxy server and the machines used by Visitor A and Visitor B have caches, while Visitor C's machine has no local cache. If Visitor A requests Page W, followed by Page X, and then uses his browser's Back button to return to Page W, only page accesses for Page W and Page X will be recorded in the log file because the second access to Page W is provided by the local browser cache. If visitor C access the same sequence of pages, it will result in three page accesses recorded in the server logs, two for Page W and one for Page X because every page request from Visitor C will reach the web server (without local cache). Another problem caused by browser caching can be observed if we consider a scenario illustrated in Figure 3. If Visitor A access pages W and X, followed by the browser's back button, also visits page Z, then if referrer logs are not with the web server, it is not possible to determine from the server logs if he reached Page Z via the hyperlink on Page W, or that on Page X.

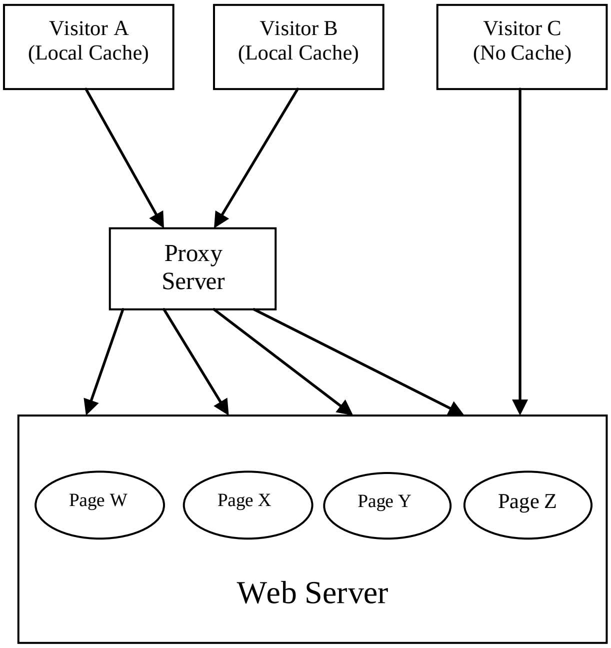

Fig. 2 caching effect

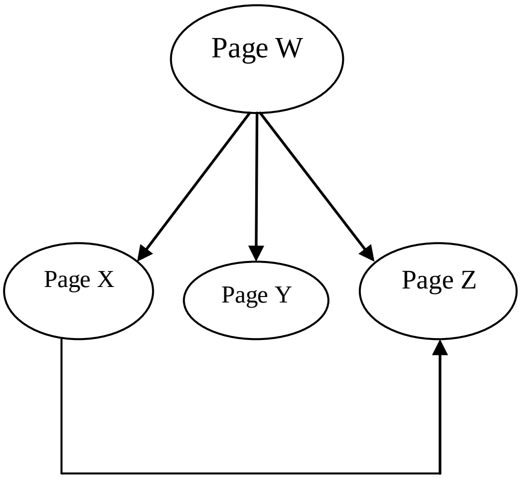

Fig. 3 Problem in linked page accesses

Caching at the proxy server can create another problem which is illustrated by examining what happens if Visitor A and Visitor B requests for the same page, the first from Visitor A and the second from Visitor B. In this case only one entry will register in the logs because the first request from Visitor A places the page in the proxy server cache and the request of Visitor B will be satisfied from the proxy server cache. One more difficulty in determining navigation paths when one session ends, and another begins. In order to overcome this problem session boundaries can be defined by using a timeout mechanism to determine when a current session must end [9, 10]. Thirty minutes is commonly used to define the maximum session window for a user. Hence, for the same IP address, the first web page access beyond the 30 minute window is assumed to the start of a new session. Some of the heuristics for identifying unique users and user sessions from web server logs are as follows.

• For the same IP address, if a request comes from a different browser or operating system, then it can be assumed that there is a different user with the same IP address.

• For a given IP address if a page is requested which is not directly reachable by hyperlink from one of the pages already visited by the user, it may be assumed that the request is coming from a different user . This assumption may fail if the users access the page by directly typing the URL. But it is not a typical browsing strategy and this type of errors is expected to be few which should not severely affect the user or session identification process. A least recently used (LRU) policy can be used if the multiple users with same IP address made the request for a page. Knowledge of whether a page is reachable from another page can be obtained from the topology of the website, which shows how pages are linked to each other.

C. *Transactions Identification Problem*

A user session may have one or more transactions. Identification of transactions or user behaviors from session data is difficult because web pages cannot be classified strictly as navigation or content pages. Some pages which serve as navigation pages for some navigation behaviors may serve as content pages for other behaviors, and vice versa. The concept of maximal forward reference can be used as a useful indicator of content pages from a stream of navigation and content pages. Forward and backward references are tracked to determine maximal forward reference (content) pages, each of which denotes the end of a transaction. Figure 4 shows a graph representation of a small portion of a website consisting nine web pages a – i as nodes, and edges connecting nodes which are linked to each other. If a user navigates this portion of the website by visiting pages in the order a-b-c-d-c-f-g-f-i-b, then the maximal forward references are d, g and i. The next forward reference following a maximal forward reference indicates the start of a new navigation path, while the maximal forward reference page at the end of a navigation path is assumed to denote a page of interest for that navigation behaviour.

.

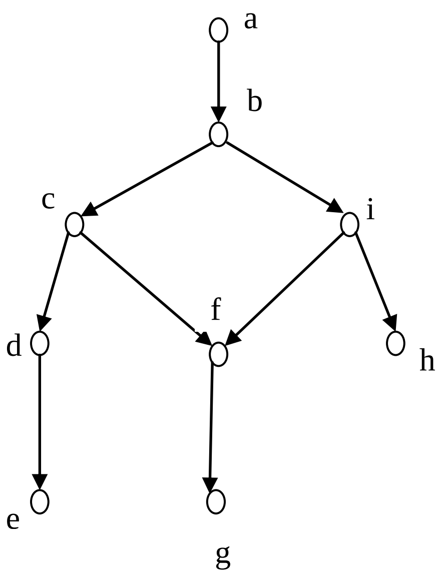

Fig. 4 Maximal forward reference

V. CONCLUSION

Web navigation patterns discovered from usage data can be used to build prediction model to recommend interesting web pages to the users. Web navigation prediction takes the form of recommended hyperlinks for users to follow. In this paper we have reviewed some difficulties as the user identification problem, navigation path identification problem, and transaction identification problem involved in the discovery of navigation and content pages from web server logs. We have also presented heuristics to address these problems.

REFERENCES

[1] Robert Cooly, Bamshad Mobasher, Jaideep Srivastava .Web mining: Information and Pattern Discovery on the World Wide Web. Proceedings of international conference on tools with artificial intelligence, Newport Beach, IEEE, pp.558-567, 1997.

[2] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava. Data Preparation for Mining World Wide Web browsing Pattern Knowledge and Information Systems, vol.1, no.1, pp.5-32, 1999.

[3] http://www.w3.org/Daemon/user/config/logging.html # common - log – file -format.

[4] Karuna P. Joshi, Anupam Joshi and Yelena Yesha. On using a Ware-house to Analyze Web Logs. Distributed and Parallel Databases, 13(2), pp.161-180, 2003.

[5] B. Berendt, B. Mobasher, M. Nakagawa and M. Spiliopoulou. The Impact of Site Structure and User Environment and Session Reconstruction in Web usage analysis. Proceedings of the forth web KDD 2002 workshop at the ACM – SIGKDD Conference on Knowledge Discovery in Databases (KDD 2002), Edmonton, Alberta, Canada, 2002.

[6] C. Marquardt, K. Becker, and D. Ruiz. A Preprocessing Tool for Web usage mining in the Distance Education Domain

Proceedings of the International Database Engineering and Application Symposium (IDEAS' 04), pp. 78-87, 2004.

[7] Navin Kumar Tyagi, A.K. Solanki and Sanjay Tyagi. An Algorithmic Approach to Data Preprocessing in Web Usage Mining. International Journal of Information Technology and Knowledge Management, pp.279-283, 2010.

[8] R.Fielding, J. Gettys, J.Mogul, and H. Frystk. Hypertext transfer protocol -- HTTP/1.1, 1999.

[9] M.S.Chen, J.S. Park, & P.S.Yu. Data mining for path traversal patterns in a Web environment. Proceedings of the 16th International Conference on Distributed Computing Systems, pp. 385-392, 1996.

[10] S.Schechter, M. Krishnan, & M.Smith. Using path profiles to predict HTTP requests. Proceedings of 7th International World Wide Web Conference. Brisbane, Australia, 1998.