A Study of Web Navigation Pattern Using Clustering Algorithm in Web Log Files

Mrs.V.Sujatha, Dr.Punithavalli

ABSTRACT -Web user navigation pattern is a heavily researched area in the field of web usage mining with wide range of applications. Web usage mining is the process of applying data mining techniques to the discovery of usage pattern from data extracted from web log files. Discovering hidden information from Web log data is called Web usage mining. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behavior of the users. This can be used for advertising purposes, for creating dynamic user profiles etc. In this paper four types of clustering approaches are investigated in web log files to improve the quality of clustering for user navigation pattern in web usage mining systems, for predicting user's intuition in the large web sites.

Index Term - Classification, Clustering, Web mining, Weblog data, and Web usage mining.

1. INTRODUCTION

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. New approaches should be used which better fit the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area. Web mining The term web mining is coined by Etzioni in 1996, to signify the use of data mining techniques to automatically discover web documents and services, uncover general pattern on the web and to observe user behavior (viewing, book marking and browsing history).Web mining is the process of finding out what users are looking for on the internet .Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web usage mining is classified into three and are web content mining, web structure mining, web usage mining.

Mrs.V.Sujatha, Asst.Prof,
CMS COLLEGE OF SCIENCE & COMMERCE
COIMBATORE,INDIA,sujatha.padmakumar@rediffmail.com
Dr.Punithavali, SNS College of arts & science, Coimbatore, India

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the web. As mentioned before the mined data in this category are the secondary data on the web as the result of interaction. These data could range very widely but generally it is classified into usage data that resides in the web client, proxy server and servers.

The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services ecommerce, to personalize the Web portals or to improve the Web structure and Web server performance. The first stage is preprocessing, next stage is pattern discovery and the last stage is pattern analysis.

2. WEB USAGE MINING ARCHITECTURE

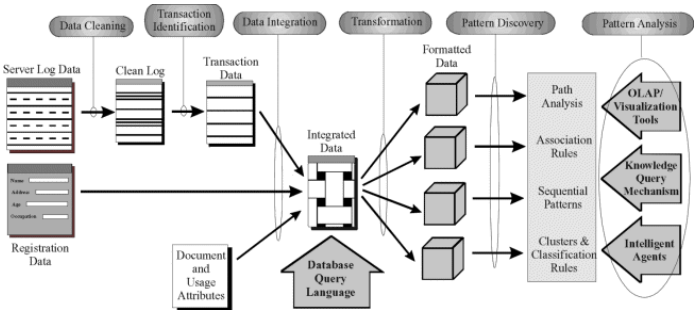

Fig 1: Web Structure Mining

2.1. Preprocessing

Pre-processing "consists of converting the usage, content, and structure information contained in the various

available data sources into the data abstractions necessary for pattern discovery". This step can break into at least four sub steps: Data Cleaning, User Identification, Session Identification and Formatting. Unneeded data will be deleted from raw data in web log files in the data cleaning step. At least two log file formats exists: Common Log File format (CLF) and Extended Log File format ([16] for more details). Our university log file consists of these fields: Date, Time, client IP address, Method, URI stem, Protocol status, Bytes sent, Protocol version, Host, User Agent and Referrer.

Fig 2: An example of raw log file

2.2 Pattern Discovery

* Statistical Analysis such as frequency analysis, mean, median, etc.

* Clustering of users help to discover groups of users with similar navigation patterns (provide personalized Web content).

* Classification is the technique to map a data item into one of several predefined classes.

* Association Rules discover correlations among pages accessed together by a client.

* Sequential Patterns extract frequently occurring inter-session patterns such that the presence of a set of items s followed by another item in time order.

* Dependency Modeling determines if there are any significant dependencies among the variables in the Web.

2.3 Pattern Analysis

Pattern Analysis is the final stage of WUM (Web Usage Mining), which involves the validation and interpretation of the mined pattern.

Validation: to eliminate the irrelevant rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process.

Interpretation: the output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations.

3. RELATED WORK

Identifying Web browsing strategies is a crucial step in Website design and evaluation, and requires approaches that provide information on both the extent of any particular type of user behavior and the motivations for such behavior [9].Pattern discovery from web data is the key component of web mining and it converge algorithms and techniques from several research areas. Baraglia and Palmerini (2002) proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. Liu and Keselj (2007) proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users' future requests and Mobasher (2003) presents a Web Personalizer system which provides dynamic recommendations, as a list of hypertext links, to users. Jespersen et al. (2002) [10] proposed a hybrid approach for analyzing the visitor click sequences. Jalali et al. (2008a [7] and 2008b [8]) proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. Dixit and Gadge (2010) [5] presented another user navigation pattern mining system based on the graph partitioning. An undirected graph based on connectivity between Referrer and URI pages was presented along with a preprocessing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph. Ant-based clustering due to its flexibility

and self-organization has been applied in a variety of areas from problems arising in e-commerce to circuit design, and text-mining to web-mining, etc (Jianbin et al., 2000. The various works proposed in this area with particular emphasize on web usage mining, clustering and classification was provided in this section. In this present work, research work is one another attempt made to propose a hybrid system that uses clustering and classification methods to discover the user's navigation pattern and analyze them from the server's web log file.

4. OVERVIEW OF CLUSTERING ALGORITHM.

4.1 Ant-based Clustering

Deneubourg et al. in [22] proposed ant-based clustering and sorting. In the case of ant-based clustering and sorting, two related types of natural ant behaviors are modeled. When clustering, ants gather items to form heaps. And when sorting, ants discriminate between different kinds of items and spatially arrange them according to their properties [23]. Lumer and Faieta in [24] proposed ant-based data clustering algorithm, which resembles the ant behavior described in [22].

4.2 Fuzzy Clustering Algorithm

Fuzzy clustering algorithm is one of the approaches to derive user categories by capturing the similar user interests from web usage data available in log files. In particular CARD+ a fuzzy relational clustering algorithm that works on data quantifying similarity between user interest, with main two activities, and are the first method is to create the relation matrix containing the dissimilarity values among all pairs of users and the next approach is to categorize the user by grouping the similar user.

4.3 Graph

Graph partitioning theoretic approach is presented by Perkowitz and Etzioni [6], who have developed a system that helps in making Web sites adaptive, i.e., automatically improving their Organization and presentation by mining usage logs. The core element of this system is a new Clustering method, called cluster mining, which is implemented in the Page Gather algorithm. Page Gather receives user sessions as input, represented as sets of pages that have been visited. Using these data, the algorithm creates a graph, as signing pages to nodes. An edge is added between two nodes if the corresponding pages co-occur in more than a certain number of sessions. Clusters are defined either in terms of cliques, or connected components. Clusters defined as cliques prove to be more coherent, while connected component clusters are larger, but faster to compute and easier to find. A new index page is created from each cluster with hyperlinks to all the pages in the cluster. The main advantage of Page Gather is that it creates overlapping clusters. Furthermore, in contrast to the other clustering methods, the clusters generated by this method group together characteristic features of the users directly. Thus, each cluster is a behavioral pattern, associating pages in a Web site. However, being a graph based algorithm, it is rather computationally expensive, especially in the case where cliques are computed.

4.4 Page Cluster

In Page clustering algorithm page ratings are calculated, then the web pages with similar ratings still do not necessarily have similar contents or navigational functions. By taking into consideration the incoming links and the transition probabilities on them, to cluster Web pages having similar incoming links and ratings together to integrate with search results and give them more semantic meanings. We define incoming link similarity of two Web pages as the accumulated difference of transition probabilities on their incoming links. By setting a threshold, Web pages are clustered together based on both incoming links and ratings. The clustering algorithm reflects the observation that Web pages, that have links in a similar set of pages and receive a similar number of hits from these pages, tend to have similar contents or navigational functions. Each cluster of pages can be given a description based on concept learning.

	A	B	C	D	E
1	117.254.157.152	1			
2	117.254.157.152	2			
3	117.254.157.152	2			
37	203.223.188.114	2			
38	203.223.188.114	2			
39	203.223.188.114	2			
71	59.92.102.117	3			
72	59.92.102.117	3			
73	75.15.85.21	3			
142	85.25.124.4	3			
143	122.178.146.123	3			
144	122.178.146.123	3			
145	118.94.8.197	3			
146	117.254.157.152	4			
147	117.204.97.156	5			
148	117.204.97.156	6			
149	117.204.97.156	6			

Fig 3: clusters group

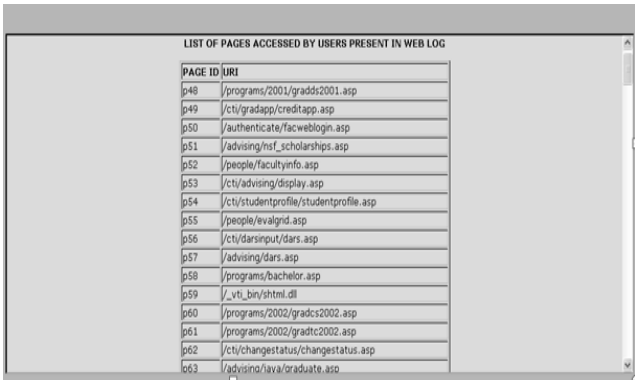

Fig : 4 List of page visited.

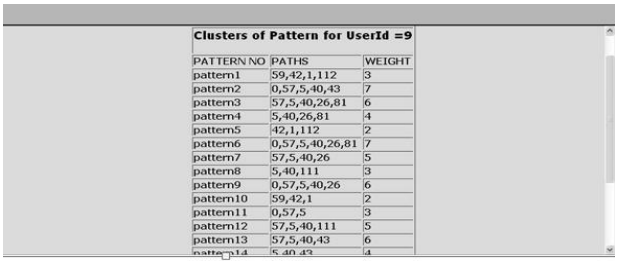

Fig. 5. List of Page id and corresponding pages/url

5. CONCLUSIONS

This paper deals with the problem of discovering hidden information from large amount of Web log data collected by web servers. The contribution of the paper is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behavior.

6. REFERENCES

[1] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on Web usage mining," Communications of the ACM, vol. 43, pp. 142-151, 2000.

[2] F. Masseglia, P. Poncelet, and R. Cicchetti, An Efficient Algorithm for Web Usage Mining, *Networking and Information Systems Journal (NIS)*, 2(5-6), pp. 571-603, 1999.

[3] R. Cooley, Web Usage Mining: Discovery and Application of Interesting patterns from *Web* Data, Ph. D. Thesis, University of Minnesota, Department of Computer Science, 2000.

[4] P. Pirolli, J. Pitkow, and R. Rao, Silk From a Sow's Ear: Extracting Usable Structures from the Web, *Proceeding on Human Factors in Computing* Systems *(CHI'96)*, ACM Press, pp. 118-125, 1996.

[5] M. Spiliopoulou, and L.C. Faulstich, WUM: A Web Utilization Miner,p*roceeding of EDBT Workshop on the Web and Data Bases (WebDB'98)*, Springer Verlag, pp. 109-115, 1999.

[6] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 1(2), pp. 12-23, 2000.

[7] F. Masseglia , P. Poncelet, M. Teisseire, A. Marascu, Web usage mining: extracting *unexpected* periods from web logs, *Data Min Knowl Disc*, 16, pp.39–65, 2008.

[8] M. *Spiliopoulou* , L.C. Faulstich , K. Winkler , A data miner analyzing the navigational behavior of web users, *Proceeding of the workshop on machine learning in user modeling of the ACAI'99 international Conference Creta, Greece, 1*99.

[9] F. Bonchi , F. Giannotti , C. *Gozzi* , G. Manco, M. Nanni , D. Pedreschi, C. Renso , S. Ruggieri, Web log data warehousing and mining for intelligent web caching , *Data Knowl Eng*, 39(2), pp. 165–189, 2001.

[10] B. Hay , G. Wets, K. Vanhoof , Mining navigation patterns using a Sequence alignment method , *Knowl Inf Syst*, 6(2), pp.150–163, 2004.

[11] Zhu, J., Hong, J., *Hughes*, J.G. 2002 Using Markov chains for link Prediction in adaptive web sites. *Proceeding of soft-ware: first International conference on computing in an imperfect world*, Belfast, UK, pp. 60–73, 2002.

[12] M. *Nakagawa*, B. Mobasher, Impact of site characteristics on recommendation models based on association rules and sequential patterns. *Proceeding of the IJCAI'03* workshop *on intelligent techniques for web* personalization, Mexico, 2003.

[13] R. Srikant, R. Agrawal, Mining sequential patterns: generalizations and performance improvements. *Proceeding of the 5th international conference on extending* database *technology (EDBT'96)*, pp. 3–17, France, 1996.

[14] A. Mueller, Fast sequential and parallel algorithms for association rules mining: a *comparison*, Technical report CS-TR-3515, Department of Computer Science, University of Maryland-College Park, 1995.

[15] A. Abraham, V. Ramos, Web Usage Mining Using Artificial

Ant
Colony Clustering and Genetic Programming, *Congress on*
Evolutionary *Computation (CEC)*, IEEE 2003.
[16] W3C extended log file format. Available at
http://www.w3.org/TR/WD-logfile
[17] WCA. Web *characterization* terminology & definitions. Available at
http://www.w3.org/1999/05/WCA-terms/ .
[18] M. Eirinaki, M.Vazirgiannis, Web Mining for Web Personalization,
Athens University of Economics and Business, 2003.
[19] J. Huysmans, B. Baesens , J. Vanthienen , Web Usage Mining: A
Practical Study, Katholieke Universities Leuven, Dept. of Applied Economic Sciences, 2003.
[20] RFC 1413. Identification Protocol. Available at
http://www.rfceditor.org/rfc/rfc1413.txt .
[21] L. Catledge, J. Pitkow, Characterizing browsing behaviors on the
World Wide Web, Computer *Networks and ISDN* Systems, 27(6),1999.
[22] J. Deneubourg -L., S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, L. Chrétien, The dynamics of collective sorting: robot-like ants and ant-like robots. Proceeding of the first international conference on simulation of adaptive behavior, pp. 356–365, MIT Press, 1991.
[23] J. Handl, B. Meyer, Ant-based and Swarm-based clustering, Swarm Intelligence, 1, pp. 95–113, 2007.
[24] E. Lumer, B. Faieta, Diversity and adaptation in populations of
clustering ants. Proceeding of the third international conference on
Simulation of adaptive behaviour, pp. 501–508, MIT Press, 1994.