Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.32 – 37

REVIEW ARTICLE

COMPREHENSIVE FRAMEWORK FOR PATTERN ANALYSIS THROUGH WEB LOGS USING WEB MINING: A REVIEW

Bhaiyalal Birla[1], Sachin Patel[2], Hemlata Sunhare[3]
[1]PCST, Indore, India
[2]PCST, Indore, India
[3]PCST, Indore, India

[1]itbirla15@gmail.com; [2]er.sachinpcst@gmail.com; [3]sunhare.hemlata@gmail.com

Abstract— Web server log repositories are great source of knowledge, which keeps the record of web usage patterns of different web users. The Web usage pattern analysis is the process of identifying browsing patterns by analyzing the user's navigational behavior. World Wide Web is a large amount of data provider and a very large source of information. Users are increasing day by day for accessing web sites. For effective and efficient handling, web mining coupled with suggestion techniques provides personalized contents at the disposal of users. Web Mining is an area of Data Mining dealing with the extraction of interesting knowledge from the Web. Here we are presenting a personalization process based on Web usage mining. This paper reviews the process of discovering useful patterns from the web server log file. In this a host of Web usage mining activities required for this process, including the pre-processing and integration of data from multiple sources, and common pattern discovery techniques that are applied to the integrated usage data.

Key Terms: - Log Repositories; Web usage pattern; Web server log; Web Mining; Web Personalization, Web-Usage Mining; Data Mining; Personalization and Pattern Discovery

I. INTRODUCTION AND BACKGROUND

Web mining is an appliance of data mining techniques to large web log data repositories [2].This term was coined by Etzioni in 1996[16]. The tremendous growth in the number and the complexity of information resources and services on the Web has made Web personalization an indispensable tool for both Web-based organizations and for the end users. The ability of a site to engage visitors at a deeper level, and to successfully guide them to useful and pertinent information, is now viewed as one of the key factors in the site's ultimate success. Web personalization can be described as any action that makes the Web experience of a user customized to the user's taste or preferences. Principal elements of Web personalization include modelling of Web objects (such as pages or products) and subjects (such as users or customers), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization.

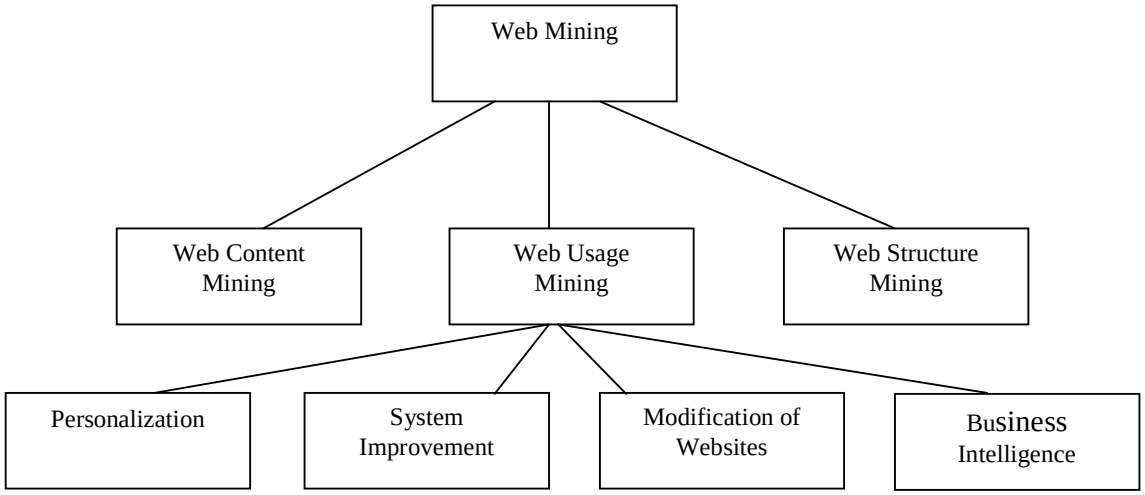

Figure 1: Web Mining Categories

There are several well-known drawbacks to content-based or rule-based filtering techniques for personalization. The type of input is often a subjective description of the users by the users themselves, and thus is prone to biases. The profiles are often static, obtained through user registration, and thus the system performance degrades over time as the profiles age. Furthermore, using content similarity alone may result in missing important "pragmatic" relationships among Web objects based on how they are accessed by users. Collaborative filtering [Herlocker et al., 1999; Konstan et al., 1997; Shardanand and Maes, 1995] has tried to address some of these issues, and, in fact, has become the predominant commercial approach in most successful e-commerce systems. These techniques generally involve matching the ratings of a current user for objects (e.g., movies or products) with those of similar users (nearest neighbours) in order to produce recommendations for objects not yet rated by the user. The primary technique used to accomplish this task is the *k*- Nearest-Neighbour (*k*NN) classification approach which compares a target user's record with the historical records of other users in order to find the top *k* users who have similar tastes or interests.

II. DATA PREPARATION AND MODELLING

The data preparation process is often the most time consuming and computationally intensive step in the knowledge discovery process. Web usage mining is no exception: in fact, the data preparation process in Web usage mining, often requires the use of especial algorithms and heuristics not commonly employed in other domains. This process is critical to the successful extraction of useful patterns from the data. In this section we discuss some of the issues and concepts related to data modelling and preparation in Web usage mining.

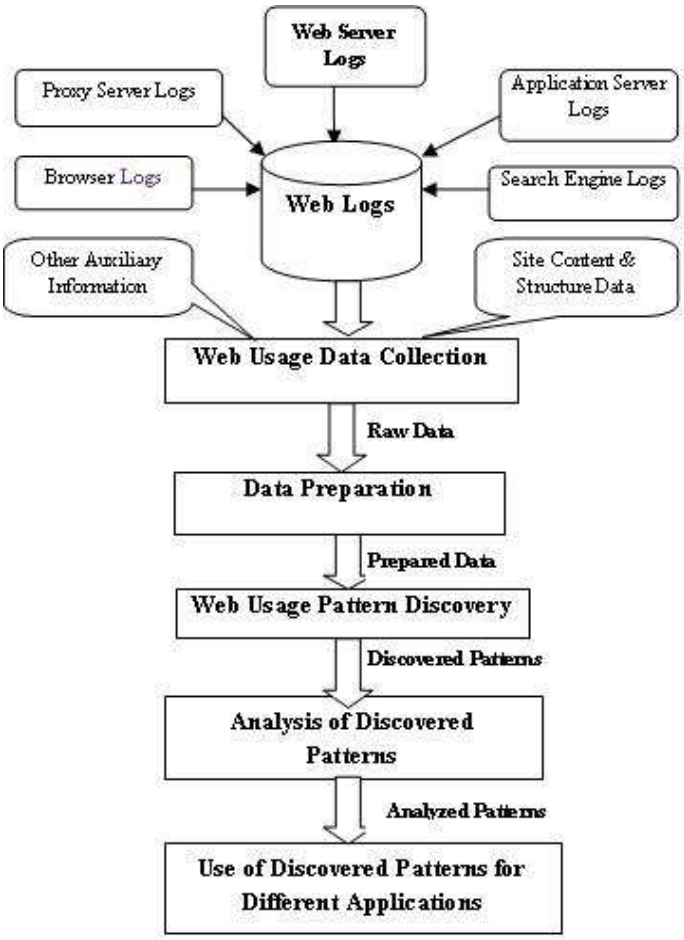

Figure 2: Web Log Mining Process

III. SOURCES AND TYPES OF DATA

The primary data sources used in Web usage mining are the server log files, which include Web server access logs and application server logs. Additional data sources that are also essential for both data preparation and pattern discovery include the site files and Meta data, operational databases, application templates, and domain knowledge. Generally speaking, the data obtained through these sources can be categorized into four groups.

IV. DATA MINING TASKS FOR WEB USAGE DATA

We now focus on specific data mining and pattern discovery tasks that are often employed when dealing with Web usage data. Our goal is not to give detailed descriptions of all applicable data mining techniques, but to provide some relevant background information and to illustrate how some of these techniques can be applied to Web usage data. In the next section, we present several approaches to leverage the discovered patterns for predictive Web usage mining applications such as personalization.

V. CLUSTERING APPROACHES

In general, there are two types of clustering that can be performed on usage transaction data: clustering the transactions (or users), themselves, or clustering page views. Each of these approaches is useful in different applications, and in particular, both approaches can be used for Web personalization. There has been a significant amount of work on the applications of clustering in Web usage mining, e-marketing, personalization, and collaborative filtering.

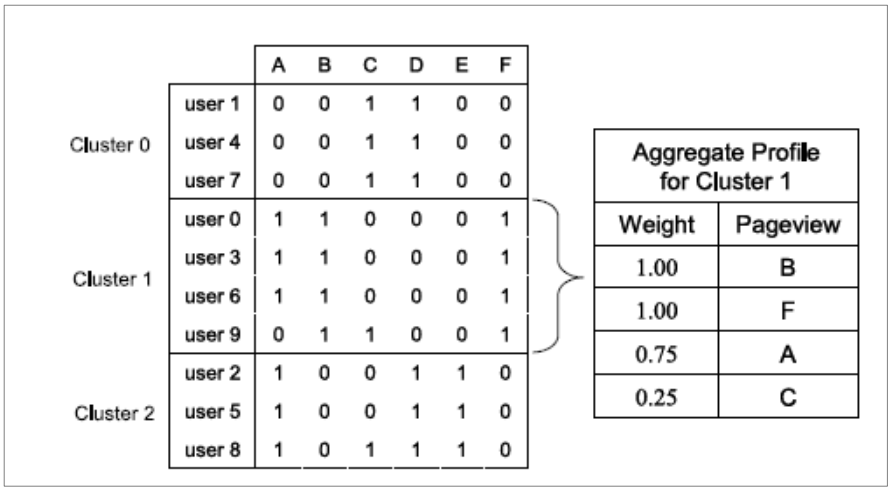

Figure 3:- deriving aggregate usage profiles from transaction clusters

For example, an algorithm called PageGather has been used to discover significant groups of pages based on user access patterns [Perkowitz and Etzioni, 1998]. This algorithm uses, as its basis, clustering of pages based the Clique (complete link) clustering technique. The resulting clusters are used to automatically synthesize alternative static index pages for a site, each reflecting possible interests of one user segment. Clustering of user rating records has also been used as a prior step to collaborative filtering in order to remedy the scalability problems of the *k*-nearest-neighbour algorithm [O'Conner and Herlocker, 1999]. Both transaction clustering and page view clustering have been used as an integrated part of a Web personalization framework based on Web usage mining.

VI. USING THE DISCOVERED PATTERNS FOR PERSONALIZATION

As noted in the Introduction section, the goal of the recommendation engine is to match the active user session with the aggregate profiles discovered through Web usage mining, and to recommend a set of objects to the user. We refer to the set of recommended object (represented by pageviews) as the recommendation set. In this section we explore the recommendation procedures to perform the matching between the discovered aggregate profiles and an active user's session. Specifically, we present several effective recommendation algorithms based on clustering (which can be seen as an extension of standard *k*NN-based collaborative filtering), association rule mining (AR), and sequential pattern (SP) or contiguous sequential pattern (CSP) discovery. In the cases of AR, SP, and CSP, we consider efficient and scalable data structures for storing frequent item-set and sequential patterns, as well as a recommendation generation algorithms that use these data structures to directly produce real-time recommendations (without the apriori generation of rule).

VII. WEB LOG FORMATS

World Wide Web consortium is an organization to provide standard format for web server log files, but there exist some other proprietary formats also. For example IIS provides six different log file formats which are used to track and analyze information about IIS-based sites and services such as:

- W3C Extended Log File Format
- W3C Centralized Logging
- NCSA Common Log File Format
- IIS Log File Format
- ODBC Logging
- Centralized Binary Logging

In addition to the six available formats, custom log file format can also be configured. A log file in the W3C extended format contains a sequence of lines containing ASCII characters. Every line in a log file may include either a command or an entry. Commands or directives are begins with # character and contains the information about logging process, software, version etc. But entries are sequence of fields corresponding to a single HTTP

transaction. Different fields in entries are separated by white space. If a particular field is not recorded in a log file entry then it is represented by dash "-" marks. The following directives are defined in the W3C Extended format [8]; the following are examples of different log file formats recorded by systems:

1. NCSA Common log file format:

172.21.13.45-REDMOND\fred[08/Apr/1997:17:39:04-0800]"GET/scripts/iisadmin/ism.dll?http/serv HTTP/1.0"200 3401

Description of headers- Remote host address, Remote log name (This value is always a hyphen), User name, Date, time, and Greenwich mean time (GMT) offset, Request and protocol version, Service status code (A value of 200 indicates that the request was fulfilled successfully), Bytes sent

2. W3C Extended log format produced by the Microsoft Internet Information Server (IIS):

#Software: Microsoft Internet Information Server 4.0#Version: 1.0 #Date: 2011-07-05 22:48:39 #Fields: date time, c-ip, cs-username, s-ip, cs-method, csuri-stem, cs-uri-query, sc-status, sc-bytes, cs-bytes, timetaken cs-version, cs-User-Agent, cs-Cookie, cs-Referrer.2011-07-05 22:48:39 206.175.82.5 -208.201.133.173 GET/global/images/topborder.gif - 200540 324 157 HTTP/1.0Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+95)USERID=CustomerA;+IMPID=01234http://yourturn.rollingstone.com/webx?98@@webx1.html

Description of headers: c- Client, s –Server, r –Remote, cs -Client to Server, sc -Server to Client, sr -Server to Remote Server(this prefix is used by proxies), rs-Remote Server to Server(this prefix is used by proxies),x-
(Application specific identifier)

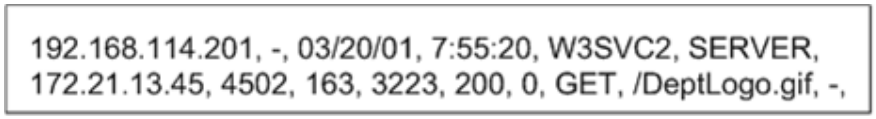

Figure 4: Microsoft IIS Log File Format

This is the identifying information that the client browser reports about itself. More recent entries are appended to the end of the file. This information can be stored in a single file, or distributed into different logs files, such as an access log, error log, or referrer log. Web usage mining research focuses on finding patterns of navigational behavior from users visiting website. The extracted knowledge of user's navigational behavior from web log file, which is recorded in any of the above format, may be used to answer different queries like efficiency of web site in delivering information, users view point about web site structure, prediction of users next visit, fulfillment of needs of different users, user satisfaction, web content personalization and many more such type of information to facilitate web administrators in taking decision. Flourishing websites can be tailored to meet user preferences both in the appearance of information and in significance of the content that best fits the user requirement.

VIII. CONCLUSION

The results of analysis may be used to for following purposes to increase popularity of web site amongst its visitors, to increase e usefulness of web pages for medium of revenue generation, for diagnostic purposes such as for detection of system errors, tarnished and wrecked links. Other web server logs may be used for similar kind of studies to increase the effectiveness of web portals or to better understanding of user behavior.In this chapter we have attempted to present a comprehensive view of the personalization process based on Web usage mining. We have discussed a host of Web usage mining activities necessary for this process, including the pre-processing and integration of data from multiple sources, and pattern discovery techniques that are applied to the integrated usage data. We have also presented a number of specific recommendation algorithms for combining the discovered knowledge with the current status of a user's activity in a Web site to provide personalized content to a user. The approaches we have detailed show how pattern discovery techniques such as clustering, association rule mining, and sequential pattern discovery, performed on Web usage data, can be leveraged effectively as an integrated part of a Web personalization system.

REFERENCES

[1] R. Agarwal, C. Aggarwal, and V. Prasad. A Tree Projection Algorithm for Generation of Frequent Itemsets. In Proceedings of the High Performance Data Mining Workshop, Puerto Rico, April 1999.

[2] C. C. Aggarwal, J. L. Wolf, and P. S. Yu. A New Method for Similarity Indexing for Market Data. In Proceedings of the 1999 ACM SIGMOD Conference, Philadelphia, PA, June 1999.

[3] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile, Sept 1994.

[4] R. Agrawal and R. Srikant. Mining Sequential Patterns. In Proceedings of the International Conference on Data Engineering (ICDE'95), Taipei, Taiwan, March 1995.

[5] A. Banerjee and J. Ghosh. Clickstream Clustering Using Weighted Longest Common Subsequences. In Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, Chicago, Illinois, April 2001.
B. Berendt, A. Hotho, and G. Stumme. Towards Semantic Web Mining. In Proceedings of the First International Semantic Web Conference (ISWC02), Sardinia, Italy, June 2002.

[6] B. Berendt, B. Mobasher, M. Nakagawa, and M. Spiliopoulou. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. In Proceedings of the 4th WebKDD 2002 Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2000), Edmonton, Alberta, Canada, July 2002b.

[7] B. Berendt and M. Spiliopoulou. Analysing Navigation Behaviour in Web Sites Integrating Multiple Information Systems. VLDB Journal, Special Issue on Databases and the Web, 9(1):56–75, 2000.

[8] J. Borges and M. Levene. Data Mining of User Navigation Patterns. In B. Masand and M. Spiliopoulou, editors, Web Usage Analysis and User Profiling: Proceedings of the WEBKDD'99 Workshop, LNAI 1836, pages 92–111. Springer-Verlag, 1999.

[9] A. Buchner and M. D. Mulvenna. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. SIGMOD Record, 4(27):54–61, 1999.

[10] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining Content-based and Collaborative Filters in an Online Newspaper. In Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, California, August 1999.

[11] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns fromWeb Data. Ph. d. dissertation, Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, 2000.

[12] R. Cooley, B. Mobasher, and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems, 1(1):5–32, 1999.

[13] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to Construct Knowledge Bases from theWorldWideWeb. Artificial Intelligence, 118(1-2):69–113, 2000.

[14] H. Dai and B. Mobasher. Using Ontologies to Discover Domain-LevelWeb Usage Profiles. In Proceedings of the 2nd SemanticWeb Mining Workshop at ECML/PKDD 2002, Helsinki, Finland, August 2002.

[15] M. Deshpande and G. Karypis. Selective Markov Models for Predicting Web-Page Accesses. In Proceedings of the First International SIAM Conference on Data Mining, Chicago, April 2001.

[16] W. B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, NJ, 1992.

