A Complete framework of Web Usage Mining

Shilpa G V [1]
[1]Asst.Professor
Vemana IT, VTU
shilpakishore@yahoo.com

Abstract: **With the increasing popularity of the world wide Web for its huge repository of web data like we pages and links. Due to the growth of web tremendous data are added daily to web log as approximately one million pages. Several users can access the web pages and links freely according to their interest from the web. Thus the web log files are growing at a faster rate and becoming huge in size. Thus web usage mining applies mining techniques on log data to extract user behaviors which is used in various applications like e-commerce, personalization, creating attractive web sites etc., In this paper we give the taxonomy of Web Mining , various Data Sources, Stages of Web Usage Mining and Applications.**

Keywords: Web Mining, Web Usage Mining, Personalization, Web logs.

I. INTRODUCTION

World Wide Web is a huge repository of web data which pertains to web pages and links. Hence it gives large amount of information that is freely offered for users to access. The users face the following problems [9] while interacting with the Web:

a. Finding relevant information: When user search for specific information on the web search tools have the following problems like (i) Low precision due to the presence of irrelevant search results. (ii) Low recall due to inability to index all the information on the web.
b. Creating new knowledge out of the information available on the Web
c. Personalization of the information: This problem arises when Internet Users differ in the contents and presentations they are interested in while interacting with the web.
d. Learning about consumers or individual users: This is a sub problem pertaining to personalization where data has to be personalized pertaining to intended consumers' interest or to individual user.

The above problems and can be resolved by using Web mining techniques.

Web mining is applying Data mining techniques to web data present in the web log server in order to extract knowledge It is categorized into Web 1.Content Mining 2. Structure Mining 3.Usage Mining. The taxonomy of Web mining is as shown in Fig 1:

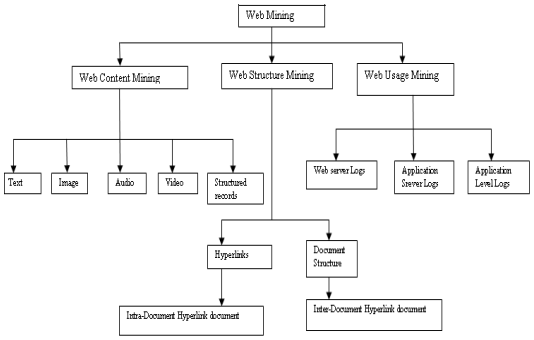

Fig 1. Taxonomy of Web Mining

- *Web Content Mining:* It is the method of extracting text, images, audio, video etc. from the web page.

- *Web Structure Mining:* It is the process of discovering structure information. This is used to improve [19] the structure of the web pages. E.g. Links pointing to documents indicates popularity of the document and Links coming out of the document indicates richness of variety of topics covered on the documents.

- *Web Usage Mining:* It is also known as web log mining on large web log repositories to discover interesting usage patterns and website usage statistics that can be used for various website design tasks. From web logs. Web logs record the web data access information of the Users. The web log data are growing at a faster rate because of the tremendous usage of web. The data in web have to be organized and handled efficiently. So the data mining techniques were implemented on web data leading to Web Data Mining.

II. DATA SOURCES

The main sources [2] of data for web usage mining are Web Servers, Proxy Servers and Web Clients.

A. Web Servers:

These are the server side sources of collecting data in their log files. The information in log files will be usually represented in standard format e.g.: Common Log Format, LogML Extended Log Format. This log contains name, IP address of the remote host, request date and time, the client request exactly as it came from the client etc., Sometimes databases are used for collecting large information. Major issue here is the identification of users page requests during navigation through the web site. The most common approaches is to use cookies. If cookies are not available various other heuristics are employed.

B. Proxy Server:

It acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

C. Web Clients:
Usage data is collected in the client side by using Java Applet, JavaScript or even modified browsers. The problems of users session identification and caching (like the use of back button) can be avoided in the client side.

III. WEB USAGE MINING STAGES

A. Data Collection [1]:

During this stage, usage data from various Data sources [2] are gathered. Data sources can be Server-side, Client-side, proxy servers or an organizations database.

(i) *Server level collection* collects client requests and stores in server as web log. The Server follows the common log format as " ipaddress username password date/timestamp url version status-code bytes-sent"

(ii) *Cookies* are unique ID generated by web server for individual users and hence it automatically tracks the site visitors. Next time when the user sends the request his ID will be also sent to the server.

(iii) *Explicit User Input* data is collected through registration forms. Hence not reliable since there are chances of incorrect data or users neglect those sites.

(iv) *Client Side Collection* Browsers are modified to record the browsing behaviors. Remote agents are used to collect user browsing information. It is advantageous than server side since it overcomes both the caching and session identification problems.

(v) *Proxy level collection* is the data collected from intermediate server between browsers and web servers. Access log from proxy servers are of same format as web server log and it records the web page request and response for the server.

B. Data Preprocessing [15]:

Information in the web is heterogeneous and unstructured. So in preprocessing phase it transforms raw click stream data into a set of user profiles. Hence preprocessing discover user behavior patterns.

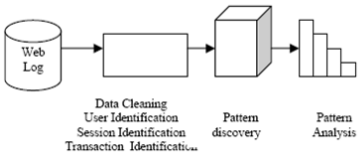

Fig 2: Stages of Web Usage Mining

- *Data Cleaning* is a process of removing irrelevant items such as jpeg, gif files or sound files which will be automatically downloaded and stored in the log file. Eg [2] requests for graphical page content, requests of any other file which may be included in web page or even navigation sessions by web robots and web spiders. Eliminating graphical requests and files are easy. But navigation patterns of robots and web spiders must be explicitly identified. This is done referring to the user agent or checking the robot text file. Some heuristics methods are also used to separate robots sessions from actual user sessions. Data cleaning enhances quality which helps in analysis. Data cleaning is performed on Web Log file [7] which will be in text format as shown in Table 1.

- *User Identification* is identification of users from Log files. Various methods [15] are followed (i) By assigning different user id to different IP address. (ii) In Proxy Servers many users will share same IP address. In that case consider referrer and user agent information. If the IP address of a user is same as previous entry and user agent is different than the user is assumed as a new user. If IP address and user agent are same then consider referrer URL and site topology.
- *Session Identification* As long as user is connected to website it is called session. In this user may have single or multiple sessions during a period. User can have multiple click streams during this session, which will be identified and portioned into logical clusters. This process is called Sessionization or Reconstruction. Here the issue is to identify when session is finished as HTTP protocol is stateless. By using three heuristics [2] methods identification of sessions termination can be done. Two were based on time between users' page requests and third was based on information about referrer. The most commonly used method is timeout threshold.

Table 1. Log file

IP address	Rfc:931	Authuser	Date & Time of request	Request	Status	Bytes	Referer	User_agent
128.101.35.92	-	-	[09/Mar/2002:00:03:18 -0600]	"GET /~harum/ HTTP/1.0"	200	3014	http://www.cs.umn.edu/	Mozilla/4.7 [en] (X11; I; SunOS 5.8 sun4u)

IP address: IP address of the remote host
Date:date and time of the request
Rfc:931:the remote login name of the user
Status: the HTTP response code returned to the client
Authuser:the username as which the user
Bytes: The number of bytes transferred
has authenticated himself
Referer:The url the client was on before requesting your url
User_agent: The software the client claims to be using

- *Path Completion [6]* Client or proxy side caching can often result in missing access references to those pages that have been cached. Example if a user goes back to a page A during the same session, the second access will not be recorded in server logs but will be cached in the client side. So missing pages are added by checking the log record from recent history.

C. Pattern Discovery[5] A variety of data mining techniques like association rules, classification, clustering and sequential patterns are performed for pattern discovery on the user transactions. Association rules [2] is most widely used technique in web usage mining. The rule X→Y states the transactions which contain items in X are likely to contain also the items in Y. when applied to web usage mining the result has the form "A.html, B.html → C.html" which states that if user visits A.html and B.html, it is very likely that in the same session he has visited C.html. Sequential patterns are used to identify sequential navigation patterns that appear in users sessions frequently from large amount of sequential data. E.g. 70% of users who first visited A.html and then B.html have also accessed page C.html in the same session. There are 2 algorithms to extract sequential patterns 1.Based on association rule mining E.g. used AprioriAll and GSP are two extensions of the Apriori algorithm for association rule mining. 2. Tree structures and Markov chains E.g. WAP-tree are used to represent navigation patterns. Classification is mapping of data into one several predefined classes. It can be done by using techniques such as Decision Tree based methods, Naïve Bayes and Bayesian Belief Networks etc., Association rule techniques are applied to databases of transactions where each transaction consists of a set of items. Using Apriori algorithm the frequently accessed item sets from the transaction databases by the user are discovered. Clustering is the technique which groups users browsing similar patterns. Such knowledge is useful in E-commerce applications. Concept based clustering [2] estimates group in conjunction with the time spent on web pages. Sequence alignment measures similarities among item sets. Genetic algorithms improve results of clustering through user feedback. Fuzzy Artificial Immune system and clustering techniques improve users profiles obtained through clustering. Multi model clustering is a technique which builds clusters by using multiple information data features. Association rule hypergraph partitioning in clusters to identify interesting group of user's behaviors.

D. Pattern Analysis is the process of filtering out uninterested rules or patterns from the set found in pattern discovery phase. In this stage various tools are provided to transform information into Knowledge. The tools are Knowledge Query Mechanism such as SQL Example WEBMINER is the most common method of pattern analysis, OLAP/Visualization tools and Intelligent Agents/ Expert Systems.

IV. SOFTWARE

There are many commercial tools [2] to perform analysis on log data collected from web servers. Accrue provides packages to web analytics. Accrue G2 that allows advanced information extraction and integration from different sources like CRM data, web server logs etc. Accure Insight 5 is used as web analytics for e-business. Pilot hit list acquired by Accure offers an efficient web analytics software for medium size companies. Luminous from web server, proxy server and client side. Net Tracker performs e-business analysis and allow integrations with CRM solutions. IBM provides Surfaid Analytics that perform OLAP operation. Web Hound is the analysis tool by SAS extracts information from web logs and performs click-stream analysis.

V. WEB USAGE MINING APPLICATIONS

- Personalized experience in B2C e-commerce – Amazon.com
  Usage of clustering, association analysis, temporal sequence analysis [4] etc., will identify users past behavior which will deepen and broaden customer relationships, to build customer loyalty, automate proactively market products to customers, to track customers response to marketing efforts.

- Web Search-Google
  Google is one of the famous search engine [4] which does content analysis and Hyperlink analysis to determine relevant pages to a query which improves quality and quickness of the search facility. Google Toolbar is another service provided by Google that makes search simple by providing additional features

such as highlighting the query words on the returned web pages.

- Understanding auction behavior – eBay
  eBay uses Web mining techniques [4] to analyze bidding behavior to determine if a bid is fraudulent.

- Personalized Portal for the Web – MyYahoo
  Web site designed to have the look-and-feel [4] and content personalized to the needs of an individual end user.

VI. FUTURE ENHANCEMENT

Web Usage Mining for personalizing the web using semantic web which integrates semantics with the unstructured data on web so that intelligent techniques can be applied to get more efficient and improved results. So personalize the system that requires knowledge in a machine interpretable form that results to retrieve more relevant data to the goal set by the user. Hence primary challenge for the next generation of personalization systems will regard the integration of semantic knowledge from domain ontologies into various stages of Web Usage Mining

VII. CONCLUSION

Web Usage mining has emerged as the essential tool for realizing web personalization as per user interest and business-optimal Web services. This article provides a framework in Web Usage Mining, focusing on different stages in Web Usage Mining taxonomy, Data Sources, Web Usage Mining stages and Applications. System improvement can be done by understanding the web traffic behavior by mining log data. The quality of a web site server can be evaluated by user accesses to the website. Web personalization will customize the information or services provided by a web site to an individual.

REFERENCES

[1] A. V.Chitraa and Dr. Antony Selvdoss Davamani "A Survey on Preprocessing Methods for Web Usage Data", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.
[2] Federico Michele Facca, Pier Luca Lanzi "Mining interesting knowledge from weblogs : a survey" Data & Knowledge Engineering 53 (2005) 225-241.
[3] Web Usage Mining Process and Techniques Mrs. Bharati Kharade(Techniques).
[4] Web Mining – Concepts, Applications and Research Directions by Jaideep Srivastava, Prasanna Desikan, Vipin Kumar , Department of Computer Science, USA.
[5] Web Mining: Accomplishments & Future Directions by Jaideep srivastava University of Minnesota USA.
[6] "Web Data Mining : Exploring Hyperlinks, Contents and Usage Data", Springer Chapter written by Bamshad Mobasher.
[7] Data Preprocessing in Web Usage Mining by Vijayashri Losarwar, Dr. Madhuri Joshi International Conference on Artificial Intelligence and Embedded Systems July 15-16, 2012 Singapore.
[8] Semantic Web Personalization : A Survey Ayesha Ameen, Khaleel Ur Rahman Khan, B. Padmaja Rani IISTE Vol 2, No 6, 2012.
[9] Web Mining Research: A Survey by Raymond Kosala and Hendrick Blockeel in SIGKDD Explorations volume 2.