Web Log Based Analysis of User's Browsing Behavior

Ashwini Ladekar[1], Dhanashree Raikar[2],Pooja Pawar[3]

B.E Student, Department of Computer, JSPM's BSIOTR , Wagholi,Pune, India[1]

B.E Student, Department of Computer, JSPM's BSIOTR , Wagholi,Pune, India[2]

B.E Student, Department of Computer, JSPM's BSIOTR , Wagholi, Pune,India[3]

Abstract: The paper discusses about user browsing behavior and interest, and web mining technology, web log data which is the source of information from the system and the apriori algorithm which has been used. Web graph is created by incorporation of user browsing behavior. The system does not collect the user's feedback, hence the privacy is not affected. The paper also confers about how the estimation of user behavior is carried out based on the study of web logs.

Keywords— Web log, data mining, page interest estimation, apriori algorithm

I. INTRODUCTION

With ample amounts of information present on the World Wide Web (WWW), issues relating to acquiring useful data from the Web has mounted the attention among researchers in the field of knowledge discovering and mining of data. In today's agonistically business environment, Web services have become an implicit need for the organizations for discovering patterns. Knowledge acquired from the data which is collected helps in developing strategies for business. To create faithful customers and gain militant advantage organizations are implementing value-added services. By providing personalized products and services, the companies are creating long-term relationships with users. By focusing on each individuals need this type of personalization can be achieved. Web mining helps to retrieve such knowledge for personalization and improved Web services. Web mining pertains to Knowledge Discovery in Data (KDD) from the web. That means, it is the process of application of data mining technique to retrieve useful information from immense amount of data available from web. Web mining and data mining objective being same both try searching for variable and useful information from web log and databases.

In retrieval of data from the Web, the personalized search carried out by a particular user forms an important research for personalized search engine. Commonly, there are two ways to collect user interest. The first approach is to take a feedback from the user in terms of his interest level. But all the user's are not interested to give the feedback so this approach is a bit inconvenient so we use the second approach of user interest based on his browsing behavior. Without users knowledge, the degree of his interest is evaluated. The second approach has become one of the important approach for collecting the interest of the user. This paper deals with web usage mining and analyzing the user's browsing behavior.

II. Web mining

When we are surfing on the internet we come across data which is not completely useful as per our needs. Mining is the technique of excavation for discovering knowledge. The process of extraction of knowledge from data is known as data mining. While the process of extracting information and data and patterns from web is known as web mining. Three methods as shown in fig1 are relevant for web mining-

- Web content mining
- Web structure mining
- Web usage mining

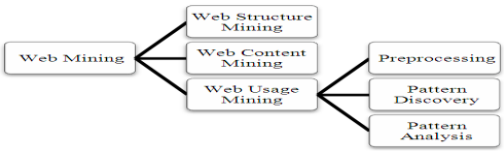

Fig1. Classification of Web Mining

ISSN: 2278 – 1323

III. Web usage mining:

Web Usage Mining can be used to make search significant by determining frequent access behavior for users, needed links can be identified to improve the overall performance for future accesses. Web Usage mining has been defined as the purpose of data mining techniques to discover usage patterns from Web data in order to recognize andserve the needs of Web based applications. Web usage mining consists of three parts, namely preprocessing, pattern discovery, and pattern analysis. Web Usage Mining may be applied to data stored in logs files. A log file contains logical data related to the user requests to a website. Web usage mining may be used to developa website structure or giving recommendations to visitors [9]. The aim in web usage mining is to discover and retrieve logical and attractive patterns from a large dataset. In web mining, Web data contains different kinds of logical information, including web structure data, web log data, and user profiles data. Web mining is the appliance of data mining techniques to retrieve knowledge in web data, where structure or usage data is used in the mining process. Web usage mining has several application areas such as web pre-fetching, link prediction, site reorganization and web personalization. Most important phases of web usage mining is finding useful patterns from web log data by using pattern discovery techniques such as Apriori, FP-Growth algorithm[10]. (PPR1)

A. Web usage mining process:

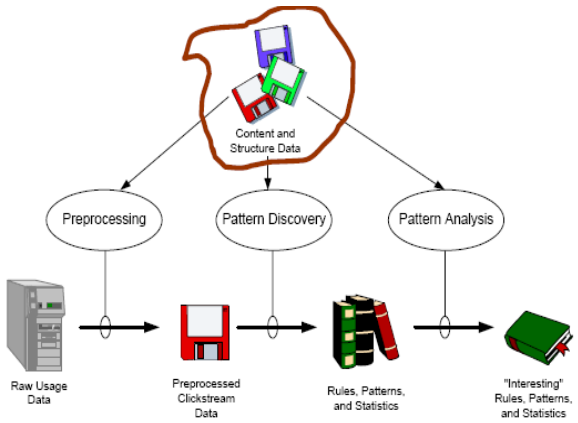

Fig2. Web Usage Mining Process

B. Hyperlink

Hyperlink is the interaction between two pages, either within the same page or to a different web page. There are two types of associations possible in hyperlink-one that connects different parts of same page is called an Intra-Document Hyperlink and the other that connects two different pages is called an Inter-Document Hyperlink.

C. Document structure

The content of web page can also be set in a tree structured format; it relies on the HTML and XML within the page.

D. Web Server Log Data

The web plays an vital role and acts as an interface for extracting logical information. There is a need for data log to trail any transaction of the communications.Log file data can propose valuable information imminent into web site usage. It analyses the action of users over a long period of time.Server logs can be used to offer some technical information related to server load, successful requests, assisting in marketing and site expansion and management activities. Below is an instance of frequent transfer log collected. This data is retrieved from NASA web server log.

EXAMPLE:

199.72.81.55 - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245

unicomp6.unicomp.net - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985

199.120.110.21 - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 200 4085

The server log composes of various attributes-

- **Date:** The date from Greenwich Mean Time (GMT x 100) is tracked for every hit. The date format is YYYY-MM-DD. The instance above shows that the transaction was tracked at 1995-07-01.
- **Time:** Time of transactions. The time format is HH:MM: SS. The instance from above shows that the transaction time was tracked at 00:00:01.
- **Client IP Address:** Client IP is the amount of computer who requested the site.
- **User Authentication:** Some web sites have an authentication provided which requests the user to enter a valid username and password. Each time a new user logs on to a Website, that user's "username" is logged in the log file.

ISSN: 2278 – 1323

- **Server IP Address:** Server IP is a static IP given by Internet Service Provider. This IP will be a reference for access the information from the server.
- **Server Port**: Server Port is a port used for data broadcast. The port used usually is port 80.
- **Server Method (HTTP Request)**: The word demands reference to an image, movie, sound, pdf, .txt, HTML file and more. The above instance indicates that folder.gif was the item accessed.
- **URI:** URI is a pathway from the host. It represents the structure of the websites. For instance:/tutor/images/icons/fold.gif.
- **Agent Log:** The Agent Log produces data on user's browser, browser version, and operating system. This is the noteworthy information, which stores information such as the type of browser and operating system and determines what a user is able to access on a particular site.

E. Pattern Discovery and Pattern analysis

The three main parts of web usage mining are data preprocessing, pattern discovery and pattern analysis.

Creation of Log file :

The quality of the patterns discovered in web usage mining process highly depends on the quality of the data used in the mining processes [1]. When the web browser traces the web pages and stores the Server log file. Web usage data contains information about the Internet addresses of web users with their navigational behavior the basic information source for web usage [2].

1. Web Server Data:

When any user agent (e.g., IE, Mozilla, Netscape, etc) hits an URL in a domain, the information related to that operation is recorded in an access log file. In the data processing task, the web log data can be [3, 4 and 2] preprocessed in order to obtain session information for all users. Access log file on the server side contains log information of user that opened a session [2, 5].

These records have seven common fields, which are:

- User's IP address
- Access date and time
- Request method (GET or POST),
- URL of the page accessed
- Transfer protocol (HTTP 1.0, HTTP 1.1,),
- Success of return code.
- Number of bytes transmitted.

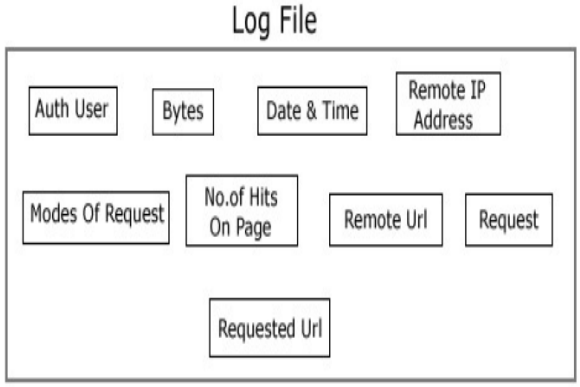

Fig 3. Log File Functional Diagram

2. Preprocessing:

Data Cleaning is also a customized step [6, 7], which includes integrating different usage logs, and parsing data from these usage logs. This process can be performed by detecting file types which have suffixes such as text and hyperlink. The nature of the data to be clustered plays a key role when choosing the right algorithm for clustering.

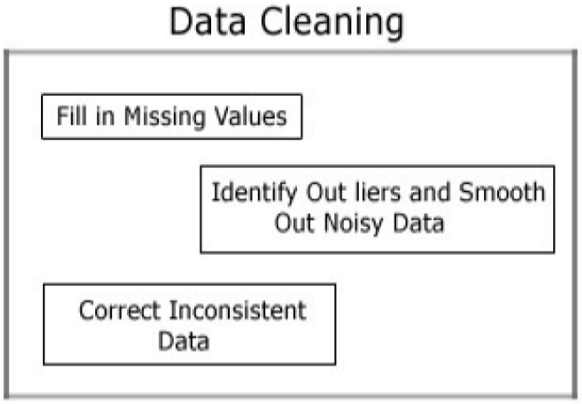

Fig 4. Data Cleaning Diagram

3. Pattern Analysis Phase:

Pattern discovery is the main issue in both web usage mining and data mining [6].The search space increases exponentially as the lengths of patterns to be discovered increase. Also, discovered patterns must be interpreted and logical knowledge must be retrieved from it. Also the comparison of Pattern Discovery on Web Logs Data . Commonly used pattern discovery algorithms that are also suitable for Web Usage Mining are[1,8].

ISSN: 2278 – 1323

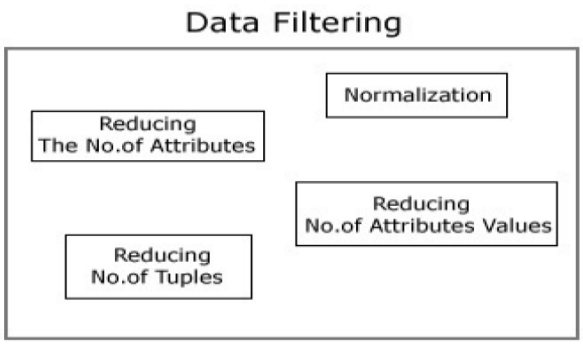

Fig 5. Data Filtering Diagram

Location of a Log file:

A web log is a file to which the Web server writes information each time as a user requests a web site from the server. Places where a log file is located are:

- Web Servers
- Web proxy server
- Client browser.

Following figure shows an example of raw log file. The raw log file contains 546 file where each file contains the information collected during one hour from the activities of the users in a Web store. Each row of log contains the following parts :

• IP address

• URL

• Request Time

• Response Time

• Difference Time

The identification of log files is already done in sessions, the Web page sequences for the same sessions have already been identified in the log file, the Web page sequences for the same the same sessions have to be collected only in the preprocessing step.

F. Apriori Algorithm:

In computer science and data mining, Apriori is a typical algorithm tolearn association rules. Apriori algorithm intends to operate on databases containing transactions (for instance, collections of items bought by customers, or details of a website frequentation). Apriori uses breadth-first search and a tree structure to count candidate item sets resourcefully. It produces candidate item sets of length k from item sets of length k - 1. Then it prunes the candidates which has an uncommon sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it examines the transaction database to determine repeated item sets among the candidates. The key concepts in this algorithm is :-

- **Repeated Item sets**: The sets of item which hasminimum support.
- **Apriori Property**: Any subset of repeated item setmust be frequent.
- **JoinOperation:** To find Ak, a set of candidate k-itemsets is generated by joining Ak-1 with itself.

The advantages of using Apriori algorithm are:-

- Uses large item set property.
- Easily parallelized
- Easy to implement

The Apriori algorithm is an efficient algorithm for finding all repeated item sets. It outfits level-wise search using frequent item property and can be additionally optimized. The Apriori algorithm used is given below.

- *Ak*: Set of repeated item sets of size *k* (with minsupport)
- *Ck*: Set of candidate item set of size *k* (potentially repeated item sets)

L1 = {repeated items};
for(*k* = 1; *Ak*!=; *k*++) **do**
Ck+1 = candidates generated from*Ak*;
for each transaction *t* in database do
increase the count of all candidates in
Ck+1 that are contained in *t*
Ak+1 = candidates in *Ck+1* with min_support
return*k Ak*;

Proposed System and its flow:

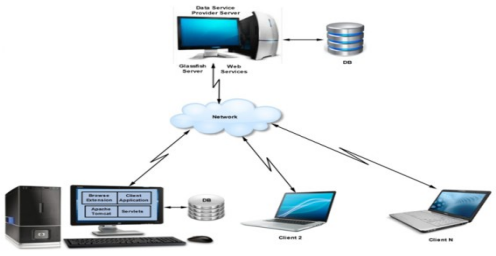

Fig6. Proposed system

ISSN: 2278 – 1323

Steps on client and server side :

Client:

- The client machine has a browser extension that tracks user's browsing behavior and related time details.
- The extension uses servlets on local machine to log data on local database using apache tomcat.
- The client application has features to store local browsing history, filter data and apply user stats to the cloud.
- The user can find the amount of time required to open a particular page or site.

Server:

- Server receives log of user's action from client machines.
- Server applies mining algorithms to find better business intelligence solutions.
- Server application provides a graphical analysis of user's usage patterns.

IV. CONCLUSION

Web usage mining is an application of data mining technique to discover usage patterns from Web data. It helps to understand and serve the need of user. Web usage mining basically has three stages, namely preprocessing, pattern discovery, and pattern analysis. One of the algorithms which is very simple to use and easy to implement is the Apriori algorithm.

ACKNOWLEDGMENT

We would like to express gratitude to our staff,family and friends for guiding us throughout this paper publication process and providing us with excellent support.

REFERENCES

[1] Ai-Bo Song , Zuo-Peng Liang, Mao-Xian Zhao, Yi-Sheng Dong, "Mining Web Log data based on
Key path", Proceedings of the First International Conference on Machine Learning and Cybernetics,
Beijing, 4-5 November 2002.

[2] R. Vaarandi, "A Data Clustering Algorithm for Mining Patterns from Event Logs," in Proceedings of
the 3rd IEEE Workshop on IP Operations and Management. Kansas City, MO, USA: IEEE Press, October2003, pp. 119 – 126.

[3] F. Masseglia, P. Poncelet, and M. Teisseire, "Using data mining techniques on web access logs to
dynamically improve hypertext structure". In ACM SigWeb Letters, 8(3): 13-19, 1999. Web Site Link:
http://portal.acm.org/citation.cfm?id=951440.951443.

4] V elasquez, Bassi J D, YasudaA. "Mining Web data to create online navigation recommendations".
Data Mining, 2004:166-172. Proceedings of the Fourth IEEE International Conference on Data Mining(ICDM'04) 0-7695-2142-8/04 IEEE.

[5] James H. Andrews, Member, IEEE, and Yingjun Zhang, "General Test Result Checking with Log File
Analysis ", 0098-5589/03/ @ 2003 IEEE Published by the IEEE Computer Society.

[6] Junjie Chen and Wei Liu, "Research for Web Usage Mining Model", International Conference on
Computational Intelligence for Modelling Control and Automation,and International Conference on
Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06) 0-7695-2731-0/06 ©2006 IEEE

[7] Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R.W., & Musen, M.A. (2001), "Creating
Semantic Web Contents with Protégé-2000". IEEE Intelligent Systems 16(2), 60-71.

[8] Mike Perkowitz, Oren Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages",
Department of Computer Science and Engineering, Box 352350 University of Washington, Seattle, WA
98195, 1998, American Association for Artificial Intelligence (www.aaai.org).

[9] Sanjay Kumar Malik, Nupur Prakash, S.A.M. Rizvi" Ontology and Web Usage Mining towards an Intelligent Web focusing web logs" 2010 International Conference .

[10] Han J., Pei J., Yin Y. and Mao R., "Mining frequent patterns without candidate generation: A frequent-pattern tree approach" Data Mining and Knowledge Discovery, 2004.

ISSN: 2278 – 1323