Handling Web Log File by using Database in Web Usage Mining

Pragnesh N. Thakkar[1] Prof. Sonal P. Rami[2]
[1]M.E. Student [2]Assistant Professor
[1,2]Department of Information Technology
[1,2]Kalol Institute and Research Center Gujarat, India

Abstract— World Wide Web is a monolithic repository of web pages that provides the Internet users with heaps of information. With the growth in number and complexity of Websites, the size of web has become massively large, and do analysis on that massively large data is being complexes. A Web Usage Mining process comprises of three phases: data pre-processing, patterns discovery and pattern analysis. The data gathered from different sources like client level, server level and proxy level collection. Privacy is a sensitive topic which has been attracting a lot of attention recently due to rapid growth of e-commerce.

Key words: web mining, web usage mining, web usage data sources, Privacy

I. INTRODUCTION

The World Wide Web (WWW) has been influencing both its user & web sites owners. In the past decade, the growth in number of websites & visitors has increased remarkably. As a result large quantity of web data has been generated. Data mining techniques are applicable to mine interesting data. But we cannot apply the data mining techniques directly to the web data since the web data is unstructured or semi -structured. Thus we use web mining which can be applied to web data. [2]

Categories of web mining: Web mining is categorized under three categories:

- Web Content Mining: Web content mining is a process of extracting information from texts, images and other contents.[2]
- Web Structure Mining: Web Structure Mining is a process of extracting information from linkages of web pages.[2]
- Web Usage Mining: Web Usage Mining is a process of extracting information from how to use web sites.[3]

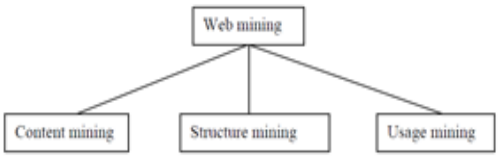

Fig 1 : web mining classification

Web usage mining is an important research area due to following reasons:

Web usage mining can be used for personalization for a user .This can be done by keeping track of previously accessed pages of a user .These pages can be used to identify the typical behavior of the user and to make prediction about desired pages.

- To identify the needed links to improve the overall performance of future accesses, frequent access behaviour for the users can be used. Pre-fetching and caching policies can be made on the basis of frequently accessed pages to improve latency time.
- To improve the actual design of web pages and for making other Modifications to a Web site, common access behaviours of the users can be used.
- Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

II. WEB USAGE MINING

Web usage mining is the application of data mining techniques on large web log repositories in order to extract useful knowledge about user's behavioral patterns. The primary data source in case of web usage mining is a web server log (or web access log). A Web server log is a textual file, independent of server platform, in which a Web server enters a record whenever a user requests for a resource. Do analysis web logs of different sites can help to understand the behavior of users and web structure, thereby improving the design of resources. A sample log entry is shown below in fig 2. [1]

```
213.135.131.79 - - [15/May/2002:19:21:49 -0400]
"GET /features.htm HTTP/1.1" 200 9955
```

Fig. 2: Sample Log Entry

Web usage mining having 3 main steps (see Figure 3): Data Pre-processing, Pattern Discover and Pattern Analysis.

- Data Pre-processing: In this step, data cleaning, user identification, session identification, path completion and transaction identification types of task would be applied.[1]
- Pattern Discover: In this phase, techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition are examined to be applied on data obtained after pre-processing in order to generate identify meaningful patterns.[1]
- Pattern Analysis: In this phase, uninteresting patterns are removed from the patterns identified during pattern discovery phase. There are two most common approaches for the pattern analysis: SQL query mechanism and constructing multi-dimensional data cube to perform OLAP operations.[1]

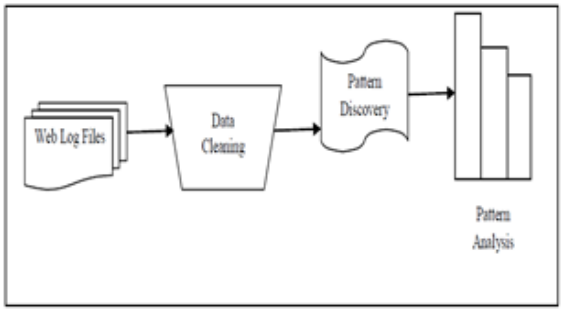

Fig. 3: Phase of Web Usage Mining

Fig. 4: Common Web Log Format

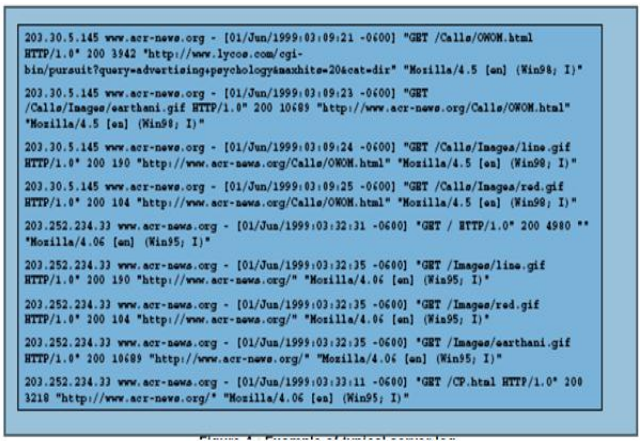

Fig. 5: Example of Server Log

A. Data Sources

1) Server Level Collection

A Web server log is very much important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors.[4] The data recorded in server logs rejects the access of a Web site by multiple users. These log files can be stored in various formats such as Common log or extended log formats. In addition, any important Information passed through the POST method will not be available in a server log. Packet sniffing technology is an alternative method to collecting usage data through server logs. The Web server can also store other kinds of usage information such as cookies and query data in separate logs. Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. Cookies rely on implicit user cooperation and thus have raised growing concerns regarding user privacy. Query data is also typically generated by online visitors while searching for pages relevant to their information needs.[4] Besides usage data, the server side also provides content data, structure information and Web page meta-information (such as the size of a file and its last modified time).[4]

The Web server also relies on other utilities such as CGI scripts to handle data sent back from client browsers. Web servers implementing the CGI standard parse the URI of the requested file to determine if it is an application program.[4] The URI for CGI programs may contain additional parameter values to be passed to the CGI application.[4] Once the CGI program has completed its execution, the Web server sends the output of the CGI application back to the browser.[4]

2) Client Level Collection

Client-side data collection can be implemented by using a re-mote agent or by modifying the source code of an existing browser to enhance its data collection capabilities.[4] The implementation of client-side data collection methods requires user cooperation, either in enabling the functionality of the JavaScript and Java applets, or to voluntarily use the modified browser.[4] Client-side collection has an advantage over server-side collection because it ameliorates both the caching and session identification problems.[4] In fact, it may incur some additional overhead especially when the Java applet is loaded for the first time. JavaScript, on the other hand, consume little interpretation time but cannot capture all user clicks (such as reload or back buttons). These methods will collect only single-user, single-site browsing behavior. [4] A modified browser is much more versatile and will allow data collection about a single user over multiple Websites. The most difficult part of using this method is convincing the users to use the browser for their daily browsing activities. [4] This can be done by offering incentives to users who are willing to use the browser, similar to the incentive programs offered by companies such as NetZero and All Advantage that reward users for clicking on banner advertisements while surfing the Web.[4]

3) Proxy Level Collection

A Web proxy acts as an intermediate level of caching between client browsers and Web servers.[4] 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. The performance of proxy caches depends on their ability to predict future page requests correctly. [4] Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.[4]

B. Data Abstraction

The information provided by the data sources described above can all be used to construct/identify several data abstractions, notably users, server sessions, episodes, click streams, and page views.[4] In order to provide some consistency in the way these terms are defined, the W3C Web Characterization Activity (WCA) has published a draft of Web term definitions relevant to analyzing Web usage.[4] A user is defined as a single individual that is accessing _le from one or more Web servers through a browser. While this definition seems trivial, in practice it is very difficult to uniquely and repeatedly identify users.[4] A user may access the Web through different machines, or use more than one agent on a single machine.[4] A page view consists of every file that contributes to the display on a user's browser at one time. Page views are usually associated with a single user action (such as a mouse-click) and can consist of several files such as frames, graphics, and scripts. When discussing and analyzing user behaviors, it is really the aggregate page view that is of importance. The user does not explicitly ask for \n" frames and \m" graphics to be loaded into his or her browser, the user requests a \Web page." All of the information to determine which files constitute a page view is accessible from the Web server. A click-stream is a sequential series of page view requests. [4] Again, the data available from the server side does not always provide enough information to reconstruct the full click-stream for a site. Any page view accessed through a client or proxy-level cache will not be visible" from the server side. A user session is the click-stream of page views for a singe user

across the entire Web.[4] Typically, only the portion of each user session that is accessing a specific site can be used for analysis, since access information is not publicly available from the vast majority of Web servers. The set of page-views in a user session for a particular Web site is referred to as a server session.[4] A set of server sessions is the necessary input for any Web Usage analysis or data mining tool. The end of a server session is defined as the point when the user's browsing session at that site has ended. Again, this is a simple concept that is very difficult to track reliably.[4] Any semantically meaningful subset of a user or server session is referred to as an episode by the W3C WCA.[4]

III. PRIVACY ISSUES

Privacy is a sensitive topic which has been attracting a lot of attention recently due to rapid growth of e-commerce. It is further complicated by the global and self-regulatory nature of the Web.[4] The issue of privacy revolves around the fact that most users want to maintain strict anonymity on the Web. They are extremely averse to the idea that someone is monitoring the Web sites they visit and the time they spend on those sites. On the other hand, site administrators are interested in finding out the demographics of users as well as the usage statistics of different sections of their Web site. This information would allow them to improve the design of the Web site and would ensure that the content caters to the largest population of users visiting their site. The site administrators also want the ability to identify a user uniquely every time she visits the site, in order to personalize the Web site and improve the browsing experience. The main challenge is to come up with guidelines and rules such that site administrators can perform various analyses on the usage data without compromising the identity of an individual user. Furthermore, there should be strict regulations to prevent the usage data from being exchanged/sold to other sites. The users should be made aware of the privacy policies followed by any given site, so that they can make an informed decision about revealing their personal data. The success of any such guidelines can only be guaranteed if they are backed up by a legal framework.[4] The W3C has an ongoing initiative called Platform for Privacy Preferences (P3P). P3P provides a protocol which allows the site administrators to publish the privacy policies followed by a site in a machine readable format. When the user visits the site for the first time the browser.[4]

Fig. 6: Sample Website Structure

The access data from an IP address (103.4.253.46) recorded on the log are given in Table 1. The paths are found by Heuristics are home.php -- categories.php – clothing.php –categories.php – software.php – software-mobile.php – categories.php – clothing.php – clothing-men.php –categories.php – watches.php – watches-kids.php

Clientip	reqDateTime	URL
103.4.253.46	14/Oct/2013:02 :52:23 -0400	/home.php
103.4.253.46	14/Oct/2013:02 :52:23 -0400	/categories.php
103.4.253.46	14/Oct/2013:03 :52:23 -0400	/clothing.php
103.4.253.46	14/Oct/2013:03 :07:33 -0400	/software.php
103.4.253.46	14/Oct/2013:02 :08:21 -0400	/software-mobile.php
103.4.253.46	14/Oct/2013:02 :08:44 -0400	/categories.php
103.4.253.460	14/Oct/2013:02 :08:51 -0400	/clothing.php
103.4.253.46	14/Oct/2013:02 :08:55 -0400	/clothing-men.php
103.4.253.46	14/Oct/2013:02 :09:18 -0400	/categories.php

Table 1 Sample Log

IV. APPLICATION OF WEB USAGE MINING

The results produced by the mining of web logs can use for various purposes:

- To personalize the delivery of web content [5]
- To improve user navigation through perfecting and caching [5]
- To improve web design; or in e-commerce sites[5]
- To improve the customer satisfaction
- Personalization of Web Content: Web Usage Mining techniques can be used to provide personalized web user experience.
- Support to the Design: Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications.[5]

V. CONCLUSION

Web Usage Mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc.

REFERENCES

[1] Surbhi Anand, Rinkle Rani Aggarwal, 2012, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions", International Journal of Computer Applications (0975 – 888)

[2] Ankita Kusmakar, Sadhna Mishra, 2013, "Web Usage Mining : A Survey on Pattern Extraction from Web Log", International Journal of Advanced Research in Computer Science and Software Engineering

[3] Rajni Pamnani, Pramila Chawan, "Web Usage Mining : A Research Area in Web Mining",

Department of computer technology, VJTI University, Mumbai
[4] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, 2000, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data"
[5] J Vellingiri, S.Chenthur Pandian, 2011, "A Survey Paper on Web Usage Mining", Global Journal of Computer Science and Technology
[6] Sanjeev Dhawan, Swati Goel, 2013, "Web Usage Mining: Finding Usage Patterns from Web Logs", American International Journal of Research in Science, Technology
[7] Mahendra Pratap Singh Dohare, Premnarayan Arya, Aruna Bajpai, 2012, "Novel Web Usage Mining for Web Mining Techniques", International Journal of Emerging Technology and Advanced Engineering
[8] Anupama Prasanth, 2013, "Web Usage Mining - Its Application in E-Services", International Journal of Emerging Technology and Advanced Engineering
[9] Monika Yadav, Mr. Pradeep Mittal, 2013, "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering

