IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. III (Mar – Apr. 2015), PP 17-26
www.iosrjournals.org

Implementation of Intelligent Web Server Monitoring

Rakhi Arya
Department of Information & Technology, Institute of Technology and Management University, Gwalior, India"

Abstract: *Web sites are one of the most important tools for advertisements in international area for universities and other foundation. The quality of a website can be evaluated by analyzing user accesses of the website. To know the quality of a web site user accesses are to be evaluated by web usage mining. The results of mining can be used to improve the website design and increase satisfaction which helps in various applications. Log files are the best source to know user behavior. But the raw log files contains unnecessary details like image access, failed entries etc., which will affect the accuracy of pattern discovery and analysis. So preprocessing stage is an important work in mining to make efficient pattern analysis. To get accurate mining results user's session details are to be known.In our proposed work, user's browsing behavior on a web page is considered with different modes of actions on client side to calculate web page browsing time with more precision and accuracy. It helps us to maintain the accuracy in finding frequent web usage patterns and administrators to evaluate its website usage more effectively. In our proposed work we considered t, t1, t2 three time limits which are subjective to the website management. These values should be selected effectively because if we select small value it will increase load on client side and if we select large value it will not provide good results.*

I. Introduction

Maintaining a website is just as important as building it.To maintain website we need to improve its design. Toimprove the design of website we should find out how it isused by analyzing users' browsing behavior [1].StatisticalAnalysis and Web Usage Mining are two ways to analyzeusers' web browsing behavior. The result of StatisticalAnalysis contains Page Views, Page Browsing Time, and so on. Web usage mining applies data mining methods todiscover web usage pattern through web usage data. Item-SetMining (Fig 1.a), Sequential Pattern Mining (Fig 1.b), andGraph Mining (Fig 1.c), are examples of data miningmethods that can be used to analyze web usage data [2].

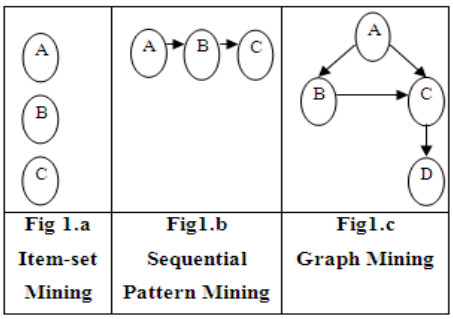


Fig 1.a Item-set Mining	Fig1.b Sequential Pattern Mining	Fig1.c Graph Mining

Web usage data can be collected from three sources: server level, client level, and proxy level [3]. Statisticalanalysis and web usage mining usually use Server Log asmain data source which is server level data source. Server logis a file that automatically created by web server and kept onserver. It contains some data about requests which are sent toweb server. During reconstruction of user's session, server log may not be fully reliable because in some cases such aspage cashing and POST method, data are not recorded inserver log [3]. Moreover, because of it only contains serverside data, cannot be useful when we are interested in user's activities on client side such as hitting back button ofbrowser, switching between tabs or windows and so on. Tocope with such problems we also need to consider client sidedata [4].

The combination of the statistical analysis and web usagemining considering client side data, presents a powerfulmethod to evaluate the usage of website. Among statisticalanalysis results, browsing time is a good scale to evaluatewebsite and users, for example in e-learning systems wherewe expect students spend a minimum duration of time oncertain page, the value less than this minimum may representin attention of student or weakness of the webpage. Amongweb usage mining methods, graph mining can discover user'saccess patterns through complex browsing behavior, for example in parallel browsing where user opens several pagesin new tabs or windows at the same time, we can simulatenavigation path as a graph and apply graph mining todiscover web usage pattern.

The majority of Web usage mining algorithms use Web log files as the main data sources in discovering useful information. Web log records that typically include host name or IP address, remote user name, login name, date stamp, retrieval method, HTTP completion code, and number of bytes in a file retrieved. Therefore, browsing time or duration of stay on a Web page is a key item for Web mining algorithms. However, Web server logs only automatically record the time entering and leaving a certain Web page, without knowing whether the user is continuously working on that page? For example, in e-learning, browsing time within the user profile is used to measure the value of each Web page or performance of a user's learning. With inaccurate browsing time captured, decision makers become conservative to the reliability of the measurement.

II. Web Mining

Web Mining is the use of the data mining techniques to automatically discover and extract information from web documents/services.Web Mining is the extraction of interesting andpotentially useful patterns and implicit information fromartifacts or activity related to the World Wide Web. Webusage mining provides the support for the web site design,providing personalization server and other businessmaking decision, etc. It has quickly becomeone of the most important areas in Computer andInformation Sciences because of its direct applications ine-commerce, CRM, Web analytics, information retrievaland filtering, and Web information systems.

III. Data Sources

Web usage data can be collected from three sources: server level, client level, and proxy level.AWeb server log is an important source for performing WebUsage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. These log files can be stored in various formatssuch as Common log or Extended log formats. An example of Extended log format is given in Figure 2..However, the site usage data recorded by server logs maynot be entirely reliable due to the presence of various levelsof caching within the Web environment. Cached page viewsare not recorded in a server log. In addition, any importantinformation passed through the POST method will not beavailable in a server log. Packet sniffing technology is analternative method to collecting usage data through serverlogs. Packet sniffers monitor network traffic coming to aWeb server and extract usage data directly from TCP/IPpackets. The Web server can also store other kinds of usageinformation such as cookies and query data in separate logs.Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the sitevisitors. Tracking of individual users is not an easy taskdue to the stateless connection model of the HTTP protocol. Cookies rely on implicit user cooperation and thus haveraised growing concerns regarding user privacy. Query data is also typically generated by online visitors while searching for pages relevantto their information needs. Besides usage data, the serverside also provides content data, structure information andWeb page meta-information (such as the size of a file andits last modified time).The Web server also relies on other utilities such as CGIscripts to handle data sent back from client browsers. Webservers implementing the CGI standard parse the URI[1]ofthe requested file to determine if it is an application program. The URI for CGI programs may contain additionalparameter values to be passed to the CGI application. Oncethe CGI program has completed its execution, the Webserver sends the output of the CGI application back to thebrowser.

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	209.456.78.2	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	209.456.78.3	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

Figure 2: Sample Web Server Log

IV. Client Level Collection

Client-side data collection can be implemented by using a remote agent (such as Javascripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.The implementation of client-side data collection methodsrequires user cooperation, either in enabling the functionality of the Javascripts and Java applets, or to voluntarily use the modified browser. Client-side collection has an advantage over server-side collection because it ameliorates both the caching and session identification problems. However,Java applets perform no better than server logs in terms ofdetermining the actual view time of a page. In fact, it mayincur some additional overhead especially when the Java applet is loaded for the first time. Javascripts, 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. A modified browser is much more versatile and will
allow data collection about a single user over multiple Web sites. The most difficult part of using this method is convincing the users to use the browser for their daily browsing activities. This can be done by offering incentives to users who are willing to use the browser.

V. Proxy Level Collection

A Web proxy 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. The performance of proxy caches depends on their ability to predict future page requests correctly. 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.

VI. Log File Format

Currently, there are three formats available to record log files:-

- W3C Extended Log file Format
- Microsoft IIS Log File
- NCSA Common Log file Format

The W3C Extended log file format, Microsoft IIS log file format, and NCSA log file format are all ASCII text formats. The W3C Extended and NCSA formats record logging data in four-digit year format. The Microsoft IIS format uses a two digit year format for years 1999 and earlier and a four-digit format thereafter. The Microsoft IIS log format is provided for backward compatibility with earlier IIS versions.

6.1 Web Access Log Analysis

6.1.1 Statistical Analysis

For web access log analysis, statistical methods, like Google Analytics, are widely used. The results of statistical analysis contain bounce rate (Single page view visits divided by entry pages [5]), page views (The number of times a page was viewed [5]), page browsing time (the time during which users browsed the page), and so on.These observed parameters show features and tendency of web page usage.

Among them, analysis of page browsing time together with knowledge about the page gives

administrators an immediate trigger to reorganize their sites in such a case when users rarely visit or stay at important pages (warning, caution, agreement, etc.) for very short period of time. It is

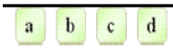
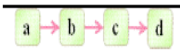
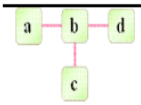
a direct sign that users do not pay much attention to the message which the sites would like to

convey to the users. On the other hand, about the pages, such as sitemap pages etc., should be passed soon, and their browsing time should be short. If the length of browsing time is different from what administrators expect, it suggests to improve the web page. Thus, page browsing time is very useful for administrators to reorganize the web site. Administrators should improve the accessibility or readability so that users will visit at reasonable frequency and stay at such pages for appropriate length of time. We focus on Page browsing time in our Dissertation.

6.1.2Web usage mining (WUM)

Web Usage Mining is an emerging attempt to apply data mining (DM) technique for web access log analyses. WUM can develop a pattern of users' navigation behavior in consideration of combination, order, etc. of the pages accessed by users. Examples of DM technique applied to the analysis are item-set mining, sequential pattern mining, and graph mining, etc. [2]. Particularly, graph mining can extract users 'access patterns as a graph structure like the web site's link structure. It can handle the case where users browse more than one page at the same time. Tab browsers are gaining more popularity, which allow users to open several pages at a time. Graph mining will become an effective analysis means, especially for the web access log created with tab browsers.

Table 1. Example of the methods used in web usage mining (In each pattern, a node intends a web page and an edge intends a users' transition path)

Pattern	Mining method	Features
	Item-set mining	The mined patterns can also be treated as an association rule. Useful to understand the relationships among the pages.
	Sequential pattern mining	This method treats only users' sequential transition, reconstructed by the requested order. It can't consider the *branched* transition to more than one page at the same time.
	Graph mining	This method can extracts patterns, which structures are like the web site's link structure, considering the *branched* transition. Easy to identify the paths frequently used by users in the web site, and useful to reconstruct the web site's link structure etc.

Web usage mining, also called web log mining, includes analysis and prediction of users' behavior in the web site by extracting the access patterns from the page request records like web access logs, applying DM techniques. For example, consider the case where, in some web site, a user browses the page a, and moves to the page b, then opens the page c on another window and moves from b to the page d, and finally browses c and d at the same time. In this case, users' access patterns extracted by WUM are as shown in the Table 2.1.
On shopping sites, which have been growing in recent years, users often browse more than one page at the same time to compare some commodities on the pages. Besides, tab browsers begin to prevail. In such situations, users' transition path becomes branched. Administrators who should improve the web site need to comprehend users' behavior including the branched transition like this. So, in what follows, we adopt graph mining which can treat branched transitions.Web site administrators analyze web access logs by using statistical and WUM method independently. We expect that administrators can analyze users' behavior more in detail with a WUM method considering quantitative information as obtained by a statistical method. In our work, we focus on calculation of page browsing time, andthen use WUM method which takes page browsing time into account, and extracts users' transition patterns by graph mining.

VII. AJAX (Asynchronous Java script And XML)

To calculate page browsing time we have to monitor and record user's web browsing behavior on client side. For this purpose we used remote agent technology AJAX which has following advantages:-

- AJAX is not a new programming language, but a new way to use existing standards.
- AJAX is the art of exchanging data with a server, and update parts of a web page - without reloading the whole page.
- AJAX is a technique for creating fast and dynamic web pages.
- AJAX allows web pages to be updated asynchronously by exchanging small amounts of data with the server behind the scenes. This means that it is possible to update parts of a web page, without reloading the whole page.
- Classic web pages, (which do not use AJAX) must reload the entire page if the content should change.
- Examples of applications using AJAX: Google Maps, Gmail, YouTube, and Face book tabs.
- AJAX applications are browser and platform independent.
- AJAX is based on internet standards, and uses a combination of:
- XMLHttpRequest object (to exchange data asynchronously with a server)
- JavaScript/DOM (to display/interact with the information)
- CSS (to style the data)
- XML (often used as the format for transferring data)

Google Suggest is using AJAX to create a very dynamic web interface: When you start typing in Google's search box, a JavaScript sends the letters off to a server and the server returns a list of suggestions.

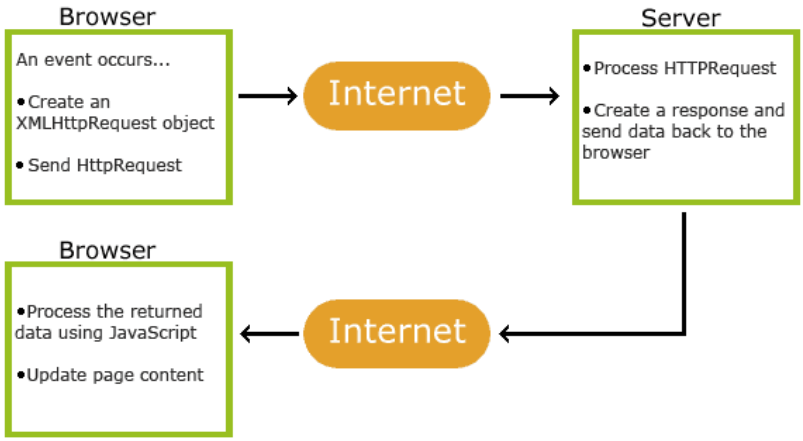


Fig 3: Working of AJAX

7.1 The XMLHttpRequest Object

The keystone of AJAX is the XMLHttpRequest object. All modern browsers support the XMLHttpRequest object (IE5 and IE6 uses an ActiveXObject).The XMLHttpRequest object is used to exchange data with a server behind the scenes. This means that it is possible to update parts of a web page, without reloading the whole page.

VIII. Literature Review

Web usage mining is explored by various researchers and focuses on major research area. In 1996 O. Etzioni explored the question of whether effective Web mining is feasible in practice. He believed that the Web is too unstructured for Web mining to succeed. Indeed, data mining has been applied to database traditionally, yet much of the information on the Web lies buried in documents designed for human consumption such as home pages or product catalogs. Furthermore, much of the information on the Web is presented in natural language text with no machine-readable semantics; HTML annotations structure the display of Web pages, but provide little insight into their content [9]. In 1998 Alex G. Biichner, Maurice D. Mulveena proposed which combined existing online analytical mining as well as web usage mining approaches, and incorporates marketing expertise. The data that is considered not only covers various types of server and web Meta information, but also marketing data and knowledge. Furthermore, heterogeneity resolution thereof and Internet- and electronic commerce-specific pre-processing activities are embedded [10].

Leticia dos Santos Machado, Karin Becker [11] in the year 2000 itself focused on Web Usage mining and discussed about the tools and methodologies for data mining can be used to discover usage patterns from web data. They discussed the three phases of web usage mining namely “preprocessing”, “pattern discovery” and “pattern analysis” in greater detail.

A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period. Once a user was identified, the click stream of each user is partitioned into logical clusters. The method of partitioning into sessions is called as Sessionization or SessionReconstruction.. There are three methods in session reconstruction. Two methods depend on time and one on navigation in web topology.

Time Oriented Heuristics: The simplest methods are time oriented in which one method based on total session time and the other based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes [12] to 24 hours [13]while 30 minutes is the default timeout by R.Cooley[14]. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10minutes then the second entry is assumed as a new session.

Navigation-Oriented Heuristics:uses web topology in graph format. It considers webpage connectivity, however it is not necessary to have hyperlink between two consecutive pages requests. If a web page is not connected with previously visited page in a session, then it is considered as a different session Cooley proposed a referrer based heuristics on the basis of navigation in which referrer URL of a page should exists in the same session. If no referrer is found then it is a first page of a new session. In 2006 Baoyao Zhou [15], devisedasimple algorithm. An access session is created as a pair of URL and the requested time in a sequence of requests with a timestamp. The duration Of an URL is estimated as the difference of request time of successor entry and current entry. For the last URL there is no successor. So the duration is estimated as the average duration of the current

session. The end time of session is the start time and duration. This algorithm is suitable when there are more number of URL's in a session. The default time set by author is 30 minutes per session.

In 2006, NatheerKhasawneh and Chien-Chung Chan [16] presented new techniques for preprocessing web log data including identifying unique users and sessions. They introduced a fast active user-based user identification algorithm with time complexity of O (n). For session identification, we used an ontology based session identification algorithm that uses the website structure to identify users' sessions. Models were introduced for determining three website parameters: number of records per user, inactive user time and number of recorded records per second in the web log to support the active-user-based approach. The output of this preprocessing step can be used as an input for different data mining techniques.

In 2007 an algorithm proposed by Junjie Chen and Wei Liu [17] in which data cleaning and session identification is combined .In this deleting the content foreign to mining algorithms gathered from web logs. User activity record is checked, judges whether the record is spider record or not and judges whether it is embedded object in pages or not according to URL of pages requested and site structure graph. Session record is searched if no session exists, a new session is established. If the present session ends or exceeds the preset time threshold, the pattern will ends it and founds a new one.

In 2008 Murat Ali [18,19] and team devised a new method named Smart Miner. This framework is a part of Web Analytics Software. The sessions constructed by SMART-SRA contains sequential pages accessed from server-side works in two stages and follows Timestamp Ordering Rule and Topology rule. In the first stage the data stream is divided into shorter page sequences called candidate sessions by using session duration time and page stay time rules. In the second stage candidate sessions are divided into maximal sub sessions from sequences generated in the first phase. In the second phase referrer constraints of the topology rule are added by eliminating the need for inserting backward browser moves. The pages without any referrers are determined in the candidate session from the web topology. Then those pages are removed. If a hyperlink exists from the previously constructed session then those pages are appended to the previous sessions. In this sessions are formed one after another. An agent simulator is developed by authors to simulate an actual web user. It randomly generates a typical web site topology and a user agent to accesses the same from its client side and acts like a real user. An important feature of the agent simulator is its ability to model dynamic behaviors of a web agent. Time constraint is also considered as the difference between two consecutive pages is smaller than 10 minutes.

In 2008 another method using Integer Programming was proposed by Robert F.Dell [20]. The advantage of this method is construction of all sessions simultaneously. He suggests that each web log is considered as a register. Registers from the same IP address and agent as well as linked are grouped to form a session. A binary variable is used and a value of 1 or 0is assigned depending on whether register is assigned a position in a particular session or not. Constraints such as each register is used at most only once and in only one session for each ordered position. A maximization problem is formulated. To improve the solution time the subset of binary variables is set to zero. An experiment is conducted to show how objective function is varied and results are obtained with raw registers and filtered for MM objects and errors. Unique pages and links between pages are counted. Chunks with same IP address and agent within a time limit is formed such that no register in one chunk could ever be part of a session in another. Experiment is focused with IP address with high diversity and a higher number of registers. Sessions produced better match an expected empirical distribution.

Graphs are also used for session identification. It gives more accurate results for session identification. Web pages are represented as vertices and hyperlinks are represented as edges in a graph. User navigations are modeled as traversals from which frequent patterns can be discovered. i.e., the sub traversals that are contained in a large ratio of traversals [21].

In 2009, Mehdi Heydari and team [22] proposed a method where they considered client side data for reconstruction of user's session. There are three phases in this method. In first phase an AJAX interface is designed to monitor user's browsing behavior. Events such as Session start, end, on page request, on page load, on page focus are created along with user's interaction and are recorded in session. In the second phase a base graph is constructed using web usage data. Browsing time of web pages in indicated as vertices. Traversal is a sequence of consecutive web pages on a base graph [21]. A database is created with traversals. In phase three graph mining method is applied to the database to discover weighted frequent pattern. Weighted frequent pattern is the pattern when weight of traversal is greater than or equal to a given Minimum Browsing Time.

Maintaining a website is just as important as building it. Tomaintain website we need to improve its design. To improve the design of website we should find out how it is used by analyzing users' browsing behavior [1].

In 2002 Kurt D. Fenstermacher Mark Ginsburg proposed framework encompasses client side applications beyond the Web browser. Expanding monitoring beyond the browser to incorporate standard office productivity tools enables analysts to derive a much richer and more accurate picture of user behavior on the Web. To ease the burden of applying the framework, they are investigating the development of a toolkit that

would enable non-programmers to select applications and events they wish to monitor and choose from a menu of actions to take in each case.

In 2003 Juan Vel'asquez, Hiroshi Yasuda and Terumasa Aoki proposed a way to study the visitor behavior in a Web site, based in web content and usage mining. A web site is a semi structured collection of different kinds of data, whose motivation is show relevant information to visitor and by this way capture her/his attention. Understand the specifics preferences that define the visitor behavior in a web site, is a complex task. An approximation is supposed that it depend the content, navigation sequence and time spent in each page visited. These variables can be extracted from the web log files and the web site itself, using web usage and content mining respectively. Combining the describe variables, a similarity measure among visitor sessions is introduced and used in a clustering algorithm, which identifies group of similar sessions, allowing the analysis of visitors behavior. In order to prove the methodology's effectiveness, it was applied in a certain web site, showing the benefits of the described approach [24].

In 2005, they described a framework for a recommender system that predicts the user's next requests based on their behavior discovered from Web Logs data. They have compared results from three usage mining approaches: association rules, sequential rules and generalized sequential rules. They have used two selection rules criteria: highest confidence and last subsequence. Experiments are performed on three collections of real usage data: one from an Intranet Web site and two from an Internet Web site.

In Dec 2006 Yu-Hui Tao [25] and team proposed a method that uses intentional browsing data (IBD) collected online from the user's interaction with the web pages, for adjusting the estimation of the browsing time for potential better application results. According to them as IBD is a web browser component such as scroll bar, copy, save-as, and so on, the browsing idle time can be better adjusted with any event of IBD.

In 2006 Maximilian Viermetz[26] and team introduced a generic browsing model extending the traditional serial or single window model to cover the use of multiple tabs. It is crucial to understand how the use of multiple tabs impacts on web usage mining, especially on the understanding of a session and its reconstruction. In their model, they present and analyze an approach to detect use of multiple tabs within sessions. The existence and increasing prominence of the use of multiple tabs is shown by this approach to be of relevance to business analysis as well as research results.

In 2007 KoichiroMihara[27] and team proposed a novel web usage mining method to combine the statistical analysis of page browsing time and the graph based data mining technique in order to exact users' typical browsing behavior. For web access log analysis, statistical methods like Google Analytics are widely used. The results of statistical analysis contains bounce rate, page views, page browsing time etc.Among them, analysis of page browsing time together with Knowledge about the page gives administrators an immediate trigger to reorganize their sites in such a case when users rarely visit or stay at important pages. Graph mining can extract users' access patterns as a graph structure like the web site's link structure. It can handle the case where users browse more than one page at the same time.

In July 2007**,** I-Hsien Ting, Chris Kimble and Daniel Kudenko proposed a users' browsing behavior analysis approach which is based on applying web usage mining techniques. Two web usage mining techniques in the approach are introduced, including Automatic Pattern Discovery (APD) and Co-occurrence Pattern Mining with Distance Measurement (CPMDM). A combination method is also discussed to show how potential browsing problems can be identified [1].

In Dec 2007**,** they focused on extraction of Sequential Patterns (SPs) with very low support from a large preprocessed Web usage data, to discover the behaviors of minority users of a Web site. Due to the sequential nature of the Web user's activity, Sequential Pattern Mining (SPM) is particularly well adapted for the study of Web usage data. Traditional SPM techniques with very low support produce large number of SPs. They are unsuitable for extraction of knowledge about the minority users because of large diversified user's behaviors and difficult to locate. Here, they proposed a novel approach called Cluster and Extract Sequential Patterns (CESP) that works based on divisive principle, where initial large Web log data split into smaller clusters (sub-logs) through ART1 neural network based clustering, and then Apriori like SPM technique is applied on each Cluster to extract SPs which reveal the behaviors of minority users. Several experiments were conducted on diversified Web log files, enabled us to discover interesting SPs having very low support (0.06 %). The study reveals that discovery of such SPs by a traditional SPM algorithms were impractical.

In Feb 2008, they proposed a complete framework and findings in mining Web usage patterns from Web log files of a real Web site that has all the challenging aspects of real-life Web usage mining, including evolving user profiles and external data describing ontology of the Web content. Even though the Web site under study is part of a nonprofit organization that does not "sell" any products, it was crucial to understand "who" the users were, "what" they looked at, and "how their interests changed with time," all of which are important questions in Customer Relationship Management (CRM). Hence, they presented an approach for discovering and tracking evolving user profiles. They also described how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from Web log data.

Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior.

IX. Justification Of Need

In previous work, done by Mehdi Heydari[22] the browsing time of page is calculated on the basis of session timeout but there are possibilities that the user may have lost focus from the browsing page and would have returned after some time or till the session is time out. For e.g. the user opens a page and after some time user is not at all interested in browsing, then the browsing time of that page should be less than the (session timeout time - Focus time of page) or user may left the page to do some other work like using MS-Word for some time and then returns back to the web page. The time spent on other applications should also be deducted from the browsing time of a page to get corrected results.

Thus actual browsing time calculation will become changed and it needs to be calculated as total actual time spent on the page. This work, hence, focuses on more precise and approximately accurate browsing time calculation of a page for a given session of the user.

X. Framework

We have used Net Beans IDE version 6.9.1 for implementation of our proposed work and MySQL Server version 5.0 for storing data. In Proposed work ,first we monitor and record web browsing behavior of user on client side by checking whether user is continuously working on web page or not and calculates browsing time accordingly and then record it into a log file. To record user's web browsing behavior, we use remote agent technology. To do this, we design an AJAX interface. The interface is a customized web application server which is able to monitor user's browsing behaviors on client side. AJAX allows web pages to be updated asynchronously by exchanging small amounts of data with the server behind the scenes. This means that it is possible to update parts of a web page, without reloading the whole page. Classic web pages, (which do not use AJAX) must reload the entire page if the content should change. AJAX applications are browser and platform independent.

In our proposed work we send information of client like Page, Page Id, Event, current Date and Time to server using AJAX interface. Sending asynchronously requests is a huge improvement for web developers. We used AJAX because many of the tasks performed on the server are very time consuming and before AJAX, this operation could cause the application to hang or stop. With AJAX, the JavaScript does not have to wait for the server response, but can instead execute other scripts while waiting for server response and deal with the response when the response is ready.

All browsing behaviors are monitored and kept in the user session until session timeout. After session timeout data is recorded in log file.

SID	PAGE	PID	RID	BT	E	T	D
S1	A	A1	NULL	6	L	39884	D1
S1	A	A1	NULL	0	G	39890	D1
S1	B	B1	A1	5	L	39890	D1
S1	B	B1	NULL	0	G	39895	D1
S1	A	A1	B1	5	F	39895	D1
S1	A	A1	NULL	0	G	39900	D1
S1	C	C1	A1	6	L	39901	D1
S1	C	C1	NULL	0	G	39907	D1
S1	C	C1	C1	18	F	39926	D1
S1	C	C1	NULL	0	G	39944	D1
S1	E	E1	C1	7	L	39945	D1
S1	E	E1	NULL	0	G	39952	D1
S1	C	C1	E1	36	F	39954	D1
S1	C	C1	NULL	0	G	39990	D1
S1	F	F1	C1	8	L	39990	D1
S1	F	F1	NULL	0	G	39998	D1
S1	E	E1	F1	13	F	40000	D1
S1	G	G1	E1	5	L	40013	D1
S1	G	G1	NULL	0	G	40018	D1
S1	B	B1	G1	15	F	40019	D1
S1	B	B1	NULL	0	G	40034	D1
S1	D	D1	B1	6	L	40034	D1
S1	D	D1	NULL	0	G	40040	D1

S1	B	B1	D1	24	F	40042	D1
S1	B	B1	NULL	0	G	40066	D1
S1	C	C2	B1	6	L	40066	D1
S1	C	C2	NULL	0	G	40072	D1
S1	F	F2	C2	5	L	40073	D1
S1	F	F2	NULL	0	G	40078	D1
S1	C	C2	F2	22	F	40080	D1
S1	E	E2	C2	4	L	40102	D1
S1	E	E2	NULL	0	G	40106	D1
S1	D	D1	E2	10	F	40108	D1
S1	G	G2	D1	60	L	40118	D1
S1	G	G2	NULL	0	G	40178	D1
S1	G	G2	G2	48	F	40190	D1
S1	G	G2	NULL	0	G	40238	D1

XI. Assumptions

To explain above example, we had taken a session Timeout time (t)of 100 seconds and after every 30(t1) seconds of load time a check function will be called which checks the difference between current time(when check function called) and last active time(Which may be equal to load time or focus time or key press time or mouse move time).If this difference is greater than 30 (t2) seconds, we assumed that user is not interested in browsing on that web page or user is busy in another work. Here also we recorded the gone time of page which helps us in estimating browsing time which is lesser as compared to previous work.

XII. Summary

In general most of the users have tendency to open several pages simultaneously and in between, use some non browsing applications such as Ms-word, Excel etc for their own personal work, in such cases data recorded in server log only shows the requested time of the web pages and cannot help us to find out which web page and for how long has been really browsed on client machine.

In this paper, we are calculating browsing time of web pages considering possible events of user's behavior like when the user is not navigating the web page throughout the time, is not attentive on the page during the navigation or navigate to some other application. Here the calculation is done only on the client side by facilitating the user with simultaneous opening of several web pages in new tabs or windows. It helps us to reconstruct user's session as it has occurred including navigation paths and pages browsing time. As we have calculated the browsing time with more precision, it helps us to find web usage patterns with more accuracy and which in turn helps administrators to evaluate its website usage more effectively.

Our research in future is to create more efficient session reconstructions through graphs and mining the sessions using graph mining as quality sessions gives more accurate patterns for analysis of users. Web Usage Mining and its algorithms have a bigger scope as far as research is concerned. Web mining and its application area is still in its infancy and requires more research. Besides Web content and Web Link, the Web Usage Mining is one of the most important areas of web mining research. We have got different application areas like Business Intelligence, E-Commerce or alike. These application areas have got more research interest. The kind of data we can recently have from Web Log is not adequate. So, this research area is also highly promising. This research area canalso use Content and/or structure mining as a tool. Web Mining and specifically Web Usage Mining can give rise to different application areas, which will really be beneficial for Web Users, society, and obviously for the governments. Specifically the defense professionals can have very useful information by mining usage data. Data Security and Privacy of data should be taken into consideration for Web Miner or Web Usage miners.

References

[1]. I-Hsien Ting, Chris Kimble and Daniel Kudenko, Applying Web Usage Mining Techniques to Discover Potential Browsing Problems of Users, Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007)

[2]. R. Iváncy and I.Vajk.Frequent Pattern Mining in Web Log Data.ActaPolytechnica Hungaria, Journal of Applied Sciences at Budapest Tech Hungary, Special Issue onComputational Intelligence. Vol.4, No.1, pp.77-99, 2006.

[3]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, Jan 2000, Volume 1, Issue 2 – page 12

[4]. Maximilian Viermetz, CarstenStolz, Relevance and Impact of Tabbed Browsing Behavior on Web Usage Mining, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence

[5]. The Web Analytics Association (WAA). Web Analytics Definitions – Version 4.0.http://www.webanalyticsassociation.org/, 2007.

[6]. J.Srivastava, R.Cooley, M.Deshpande, P-N,Web usage mining:discovery and applications of usage patterns from web data,SIGKDD Explorations 1 (2) (2000) 12-23.

[7]. Demin Dong, "Exploration on Web Usage Mining and its Application, ",IEEE,2009.

[8]. Raju G.T. and Sathyanarayana P. "Knowledge discovery from WebUsageData : Complete Preprocessing Methodology, ", IJCSNS 2008.
[9]. O. Etzioni, The world -wide Web: Quagmire or gold mine ? Communications of the ACM 39 (11) (1996) 65-68.
[10]. Alex G. Biichner, Maurice D. Mulveena" Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", (1998).
[11]. Leticia dos Santos Machado, Karin Becker, " Distance Education: a Web Usage Mining Case Study for the Evaluation of Learning Sites".2003.
[12]. Catlegde L. and Pitkow J., "Characterising browsing behaviours in the world wide Web,", Computer Networks and ISDN systems, 1995.
[13]. SpilipoulouM.andMobasher B, Berendt B.,"A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," INFORMS Journal on Computing Spring ,2003.
[14]. Robert.Cooley,BamshedMobasher, and JaideepSrinivastava, "Webmining:Information and Pattern Discovery on the World Wide Web,",InInternational conference on Tools with Artificial Intelligence, pages558-567, Newport Beach, IEEE,1997.
[15]. Baoyao Zhou, Siu Cheung Hui and AlvisC.M.Fong,"AnEffective Approach for Periodic Web Personalization,", Proceedings of theIEEE/ACM International Conference on Web Intelligence. IEEE,2006.
[16]. Natheer Khasawneh , Chien-Chung Chan "Active User-Based and Ontogy-Based Web Log Data Preprocessing for Web Usage Mining" ,Dec 2006.
[17]. Jungie Chen and Wei Liu, "Research for Web Usage Mining Model," International Conference on Computational Intelligence for ModellingControl and Automation, IEEE, 2006.
[18]. Murat Ali Bayir, Ismail HakkiToroslu, AhmetCosar and GuvenFidan"Discovering more accurate Frequent Web Usage Patterns, ",arXiv0804.1409v1, 2008
[19]. Murat Ali Bayir, Ismail HakkiToroslu, AhmetCosar and GuvenFidan"SmartMiner:A new Framework for Mining Large Scale Web UsageData," , International World Wide Web Conference Committee, ACM,2009.
[20]. Robert F.Dell ,PabloE.Roman, and Juan D.Velasquez, "Web UserSession Reconstruction Using Integer Programming," , IEEE/ACMInternational Conference on Web Intelligence and IntelligentAgent,2008.
[21]. SeongDae Lee, and Hyu Chan Park, "Mining Weighted FrequentPatterns from Path Traversals on Weighted Graph, ",InternationalJournal of Computer Science and Network Security, VOL.7 No.4, April2007.
[22]. Mehdi Heydari, Raed Ali Helal, and Khairil Imran Ghauth, A Graph- Based Web Usage Mining Method Considering Client Side Data, ", International Conference on Electrical Engineering and Informatics, IEEE, 2009.
[23]. Juan Vel'asquez Hiroshi Yasuda and Terumasa Aoki "Combining the web content and usage mining to understand the visitor behavior in a web site", 2003.
[24]. Yu-Haitao , Tsung-Pei Hong andYu-Ming Su, "Web usage mining withintentional browsing data, ", Expert Systems with Applications , ScienceDirect,2008.
[25]. Maximilian Viermetz, CarstenStolz, Relevance andImpact of Tabbed Browsing Behavior on Web UsageMining, Proceedings of the 2006 IEEE/WIC/ACMInternational Conference on Web Intelligence.
[26]. K. Mihara, M. Terabe, K. Hashimoto, "A Graph-BasedWeb Usage Mining Considering Page Browsing Time", inProc.of the Second International Conference on Knowledge,Information and Creativity Support Systems (KICSS'07),p.199-206, Ishikawa, Japan, Nov. 2007.
[27]. Sebastian A. Rios and Juan D.Velasquez, "Semantic Web Usage Mining by a Concept-based Approach for Off-line Website Enhancement", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
[28]. Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path CompletionTechnique in Web Usage Mining,,", International Symposium onComputer Science and Computational Technology, IEEE,2008.
[29].]Saud R. Aghabozorgi and The Ying Wah,"Dynamic Modelling by Usage Data for Personalization" 2009 13th International Conference Information Visualization.