Pre-Processing and Analyzing Web Logs for Web Analytics to Improve Web Organization

1S. Uma Maheswari and 2Prof. S.K. Srivatsa

1Research Scholar, SCSVMV University, Department of IT, Kanchipuram, India.
2Senior Professor, Prathyusha Institute of Tech & Mgmt, Department of ECE, Chennai, India.

ARTICLE INFO

ABSTRACT

The web access logs are the best repositories for the information source. It maintains the entire record of even a tiny low event. Most of existing websites have a stratified content organisation. The way of organizing could also be slightly different from the visitor’s expectation of organizing the web site. Particularly, it’s typically unclear to the visitant at which location a specific document is found. First, these log files are pre-processed and regenerate into specified formats. Therefore web usage mining techniques will apply on these web logs. This paper reviews the method of Preprocessing that is helpful to take clear web log data from the online server log file of an educational institute. The preprocessed and analyzed results are used in many areas such as net traffic analysis, economical web site administration, website modifications, system improvement and customization and business intelligence etc.

INTRODUCTION

With the technological advancements, businesses have gone online. Web seems to be too huge for effective data warehousing and data mining. World Wide Web (WWW) serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, e-commerce etc. Information is arranged in proper hierarchy in the form of websites. The web also contains a rich and dynamic collection of hyperlink information. Collection of web pages named websites are accessed via hyperlinks. Nowadays internet plays a vital role for providing information to all kinds of users to obtain their needs. Day by day the usage and accessibility of internet is increasing tremendously. Websites are highly helpful for providing any kind of information to any kind of users at any time. A web server usually registers a weblog entry for every access of webpage. When the user communicates with the website, the interaction details are automatically recorded in web server as in form of web logs.

Interesting information extracted from the visitors browsing data such as Analysis of who browsed, what can give important insight into, for example, what are the buying patterns of existing customers help analysts to predict. Mining the required data is a significant challenge in web data mining. Web data mining is the process to extract the interesting knowledge from huge amount of data. Stream data grows rapidly, so there is an augmented need to perform pre-processing on stream data.

Taxonomy of Web Mining:

There are enormous wealth of information on web such as Financial information like stock quotes, Book/CD/Video stores, Restaurant information and Car prices. Even though it has many sort of information, the web posses great challenges for effective resource and knowledge discovery(Umamaheswari,2014). The web seems to be two huge for effective data warehousing and mining. Also the complexity of web pages is far greater than that of any old text documents.

Only a small portion of the information on the web is truly relevant (Micheline Kamber, 2010). It is possible to get lots of data on user access patterns and also possible to mine interesting nuggets of information. The process of Searching the Web is illustrated in the following “Fig. 1”.

Corresponding Author: S. Uma Maheswari, Research Scholar, SCSVMV University, Department of IT, Kanchipuram, India.
Web has recently become a powerful platform for retrieving information and discovering knowledge from web data. The idea of discovering useful patterns in data may have many names such as data mining, knowledge extraction, Information discovery, Information Harvesting, Data Archeology, and Data pattern processing (Navin Kumar Tyagi, 2010).

**Basic Structure of Web Log Mining:**

The figure given below shows the basic structure of web log mining. Access Log is the input for preprocessing.

![Web Log Mining Diagram](image-url)

**Fig. 2:** Basic structure of web log mining.
Access log is the input to pre-processing block. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site preprocessing a web usage mining model aims to reformat the original web logs to identify user’s access sessions, pre-processing a web usage mining model aims to reformat the original web logs to identify user’s access sessions. The web logs are maintained as in form of line of text (Umamaheswari, 2014).

1. Web Server – Here the log files gives more complete and accurate user’s interaction information along with the website. The World Wide Web Consortium retains a standard format to use for the web server log files. It stores client IP address, request date, request time, page requested, Hyper Text Transfer Protocol (HTTP) code, bytes transferred, user agent and referrer.

2. Proxy Server – The proxy servers behave like a mediator between the browser and the web server. It gets HTTP request from user and it passes them to the web server.

3. Browser – Web Logs for the particular user are kept in the browser machine. The browsers are programmed and scripting languages are employed to collect the client side data.

Pre-Processing and Analyzing Web Log Files:

Nowadays there is no quality data and no quality mining results. So the Quality decisions must be based on quality data. Generally the duplicate or missing data may cause incorrect or even misleading statistics. Also the Data warehouse needs consistent integration of quality data. Moreover the Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse. That is why it is important to preprocess the data. Preprocessing (Chitraa, 2012) is an important activity in web usage mining and treated as a key to success. In Web Usage Mining the data preprocessing consumes more time and is considered as a complex process in terms of calculations. This process involves gathering information from different sources and grouping them and finally translating into a form adaptable to preprocessing technique. Most of the researchers (Yinghui Yang, 2005; Mehrdad Mahdaviand, 2009) done their research on web usage mining as preprocessing is the initial step of their work. Later many techniques are applied on the preprocessed web logs to make a group for better processing of the web logs. The familiar clustering algorithms like k-means, modified k-means and Harmony k-means algorithms are used by some of the authors (Yinghui Yang, 2005; Mehrdad Mahdaviand, 2009; Sudipto Guha, 1998).

Generally, it is easily determined that usage mining has valuable responses to the marketing of businesses and a direct impact to the success of their promotional strategies and internet traffic. This information is gathered on a daily basis and continues to be analyzed consistently. Analysis of this pertinent information will help companies to develop promotions that are more effective, internet accessibility, inter-company communication and structure and productive marketing skills through web usage mining.

Related Work:

This author (Yinghui Yang, 2005) focused on grouping the customer transactions by using the clustering technique. The set of transactions in a group has some similarities, so we can easily identified the customer behaviour and the web site analyst can able to understand the customer expectation and make the website customer friendly. In other point of view, making the website like more personalized and more user friendly is very important. The researcher used the pattern based clustering approach to group the similar type of transactions. Some measure is followed to group the transactions, for example {starting_time = morning, avg_time_page < 2, category = 3, total_time < 10 min} may be the behavioural pattern for grouping the transactions. The result may be the webpages of news, finance, share or email.

The author (Mehrdad Mahdaviand, 2009) dealt with two types of groups one is Web Clustering Groups which groups the relative pages from the web server log files, the second is User Clustering Groups which groups the user who refers the same type of web pages. Divisive Hierarchical Clustering Algorithm is used to group the Web Log files and User of similar type.

The author (Sudipto Guha, 1998) mainly focused on the data preprocessing step to remove the unnecessary data such as images, extra click events. Pattern discovery algorithms are used to eliminate the unwanted data from the web server log files.

They have taken the data from NASA server web log files and remove the unwanted data to improve the efficiency of the web log data analysing process. No specific data mining techniques are applied to web log files after pre-processing.

Kavita Das and O P Vyas (Kavita Sharma, 2011) have presented a model for web personalization approach using web mining. The server side and browser side details are taken for consideration. The author proposed bottom-up approach for achieving web personalization from personalized websites. The websites are personalized for individual users by analyzing the user’s browsing history.

K.R. Suneetha and R Krishnamoorti (Suneetha, 2011) have developed an Intelligent Recommendation System to determine pages that are most likely to be visited by the user in future. IRS assists site owners in optimization, improving user satisfaction etc.
The web usage pattern analysis is a method of distinguishing browsing patterns by analyzing the user’s navigation and behaviour. The internet server log files that store the knowledge concerning the guests of internet sites is employed as input for the web usage pattern analysis method.

**Proposed Method:**

**Data cleaning:**

Data collection (Malavizhi, 2012) is the initial step in web log preprocessing. Irrelevant records are eliminated during data cleaning. Data cleaning (Vijayashri Losarwar, 2012) is the process of removing noisy and irrelevant data that are not helpful for mining the knowledge from the web logs.

**User and Session Identification:**

The task of user and session identification is to find out the different user sessions from the original web access log. Session identification (Sheetal, 2012) is the process of dividing the individual user access logs into sessions. A referrer-based method is used for identifying sessions.

Whenever a user interacts with the website, they spend some time in each web page. Session is time duration spending on each web page by the single user. Session identification (Vijayashri Losarwar, 2012) is the process of dividing the individual user access logs into sessions. The login and logout time is considered for identifying the starting and ending time of each session. Some steps are followed to find the user session.

1. If the user is identified as a new user then it is taken as a new session.
2. For the same session, when the refer page is null then it is considered as a new session.
3. If the time between page requests exceeds 30 minutes then it is treated as one new session.

The web logs are updates each time a user starts a new session. Initially the log file contains each and every detail regarding the user, the IP address, website name, time stamp and other details. But these details are generated based on each and every second, so to make the log files light which we obtained from different sources, some preprocessing steps are first taken into action. There are many techniques by which we can reduce the density of log content in a log file. In this paper we are considering only five entities and they are IP Address, user name, website name, session and frequency. The Extraction process of the session timing and the frequency is calculated by taking the time difference and the total number of clicks on a particular web site given in a log file (Suguna, 2012).

**Path Completion:**

There are chances of missing pages after constructing transactions due to proxy servers and caching problems in web server logs. In such condition it becomes mandatory for identifying the user’s access path, and adding the missing paths. Because of local buffers existence, some requested pages are not recorded in access log. The goal of path completion is to fill all the missing references that are not recorded. The solution for path completion is if a requested page is reachable by a hyperlink from any of the visited pages by the user, it is assumed that it is added in the session.

**Data Formatting:**

The data formatting is the final step in preprocessing. The preprocessed web log information is properly formatted suitable for applying the pattern discovery algorithms.

**Analysis of Web Log Files:**

Analyzing log data which is being used as the significant basic data in webservice research. The paper introduces an idea for website analysis by discovering web users’ behaviour patterns as well as discovering knowledge from the website structure. The analyzed web log data can be used in many applications such as Web personalization, Web recommendation, Tracing Visitor’s online Behaviours for web site organization and E-commerce applications, etc.

**Web personalization:**

Web personalization is defined as any action that adapts the information or services provided by a Web site to the needs of a user or a set of users, taking advantage of the knowledge gained from the users’ navigational behavior and individual interests, in combination with the content and the structure of the Web site.

The steps of a Web personalization process includes the collection of Web data, the modeling and categorization of these data, the analysis of the collected data and the determination of the actions that should be performed. The site is personalized through the highlighting of existing hyperlinks, the dynamic insertion of new hyperlinks that seem to be of interest for the current user, or even the creation of new index pages. Most of the research efforts in Web personalization correspond to the evolution of extensive research in Web usage mining (Becchetti, 2010; Kavita Das, 2011).
Web recommendation:
This paper guides the web site developer for better Web Recommendation. Web recommendation systems are considered as an important role to understand customers’ behavior, interest, improving customer convenience, increasing service provider profits and future needs.

Tracing Visitor’s online Behaviours for web site organization:
It must to trace the visitors’ on-line behaviors for website usage analysis. Actually it is an analysis to get Knowledge about how visitors use Website which could provide guidelines to web site reorganization and helps to prevent disorientation.

E-commerce Applications:
E-commerce means more than just build up a web site, then sit back and relax. Web Mining systems need to be implemented to Understand visitors’ profiles, Identify company’s strengths and weaknesses and Measure the effectiveness of online marketing efforts. Web Mining support on-going, continuous improvements for E-businesses. Web Usage Mining techniques are used to find out hidden patterns to improve business ideas. In addition, employee as well as online user satisfaction is important for an e-commerce based enterprise.

The trend of online shopping is growing in India, Mainly the urban city users are doing online shopping through various e-commerce websites like ebay.com. With the extensive expansion in the number of E-commerce websites, a lot of online user data is available on web logs of these sites.

Experimental Results:
The web log files are gathered from the college web server. Then it is preprocessed and analyzed by using Deep Log analyzer tool to know about visitor’s online behaviours and Page details, Browsing History, Number of visits, Popular web pages and User’s Preferences etc. The sample data set and some of the resultant feature data set such as visitor’s details, Page details and Web Site details are given here.

Sample Dataset:

<table>
<thead>
<tr>
<th>FileName</th>
<th>Page Views</th>
<th>Data Transferred(Kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/Default.asp</td>
<td>18365.0</td>
<td>385123.0</td>
</tr>
<tr>
<td>/download.asp</td>
<td>1938.0</td>
<td>27168.0</td>
</tr>
<tr>
<td>/robots.txt</td>
<td>1532.0</td>
<td>2231.0</td>
</tr>
<tr>
<td>/download/dltrial.exe</td>
<td>1246.0</td>
<td>6221606.0</td>
</tr>
<tr>
<td>/dl.asp</td>
<td>738.0</td>
<td>16744.0</td>
</tr>
<tr>
<td>/buy.asp</td>
<td>645.0</td>
<td>12063.0</td>
</tr>
<tr>
<td>/screenshots.asp</td>
<td>591.0</td>
<td>7132.0</td>
</tr>
<tr>
<td>/pop.aspx</td>
<td>507.0</td>
<td>319.0</td>
</tr>
<tr>
<td>/dla.xml</td>
<td>292.0</td>
<td>3983.0</td>
</tr>
<tr>
<td>/contacts.asp</td>
<td>265.0</td>
<td>2948.0</td>
</tr>
<tr>
<td>/clients.asp</td>
<td>239.0</td>
<td>5144.0</td>
</tr>
<tr>
<td>/compare.asp</td>
<td>237.0</td>
<td>4568.0</td>
</tr>
<tr>
<td>/submit-feedback.aspx</td>
<td>221.0</td>
<td>3115.0</td>
</tr>
<tr>
<td>/affiliates.asp</td>
<td>208.0</td>
<td>401.0</td>
</tr>
<tr>
<td>/buynow.asp</td>
<td>202.0</td>
<td>120.0</td>
</tr>
</tbody>
</table>

Unprocessed Web Log File:
Processed & Resultant Data Set after analysing the Log files:
The visitor details, country and Number of visits they made have been taken from analysed web log files which support the web site organization.

<table>
<thead>
<tr>
<th>Visitor</th>
<th>Country</th>
<th>Number of Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.14.141.98</td>
<td>United States</td>
<td>6.0</td>
</tr>
<tr>
<td>12.14.141.98</td>
<td>United States</td>
<td>2.0</td>
</tr>
<tr>
<td>12.151.36.11</td>
<td>United States</td>
<td>3.0</td>
</tr>
<tr>
<td>125.17.7.231</td>
<td>India</td>
<td>4.0</td>
</tr>
<tr>
<td>125.238.50.102</td>
<td>New Zealand</td>
<td>6.0</td>
</tr>
<tr>
<td>128.175.23.125</td>
<td>United States</td>
<td>2.0</td>
</tr>
<tr>
<td>129.132.124.7</td>
<td>Switzerland</td>
<td>2.0</td>
</tr>
<tr>
<td>129.2.17.234</td>
<td>United States</td>
<td>3.0</td>
</tr>
<tr>
<td>13.16.137.11</td>
<td>United States</td>
<td>5.0</td>
</tr>
<tr>
<td>130.225.206.55</td>
<td>Denmark</td>
<td>3.0</td>
</tr>
<tr>
<td>130.235.4.70</td>
<td>Sweden</td>
<td>4.0</td>
</tr>
<tr>
<td>130.80.28.26</td>
<td>United States</td>
<td>3.0</td>
</tr>
</tbody>
</table>

The following diagram depicts the number of visits made by many visitors from which top visitors and their relevant information can be gathered for enhancing the web sites and so they may give best service to various customers according to their needs.

**Conclusion:**

The accuracy & quality of pattern mining algorithms are improved with the help of preprocessing techniques. Analysis of this usage data will provide the companies with the information needed to provide an effective presence to their customers. These present some of the benefits for external marketing of the company’s products, services and overall management. Internally, usage mining effectively gives information to improvement of communication through intranet communications. Still some more research is needed to improve the efficiency of the algorithms to facilitate the website visitors, website analyst and website personalization. Association Rule Mining can be applied to Web Mining to understand the website visitor’s behavior, attracting the website visitors, personalizing the websites and enhancing the websites for business point of view.

**REFERENCES**


