Empirical Analysis of Web usage mining

Dr. B.V.Ramana Murthy*1  B.Mallikarjun Rao*2
*1Department of Computer Science and Engineering, Stanley College of Engineering and technology for Women, Hyderabad.
*2Research scholar, Rayalaseema University.

Abstract—WWW and Data mining are the dynamic areas of contemporary research. A combination of these areas is called as Web mining. The Web mining word can be used in different ways and has many applications. The Web mining has been divided into areas like Content mining and Usage mining. The Content mining describes the discovery process from lakhs of web sites and its pages across the WWW. The usage mining in this research paper, is the process of mining server side log files or the user information from server, user browsing habits and browsing patterns on one or more web site pages & different sources. In this paper we define mining of the web pages and log-file patterns mining, which presents an overview of the different research issues, its techniques, and their developmental efforts. Here the study is mainly on the problems and we examine the techniques associated with Usage mining.

Keywords— Web mining, Usage mining, pattern mining, LOGML.

I. INTRODUCTION

With the growth of information across the World Wide Web, it has become necessary for users to utilize tools to find, to extract, to filter, and to evaluate the wanted information from the hoard of resources. It is crucial for organizations and companies to track and analyze user browsing behavior patterns to market their products. These factors raise the necessity of generating server-side and client-side systems which can mine for information across the Internet and in particular World Wide Web.

Website mining [4] can be defined as the discovery and analysis of useful statistics provided from the WWW. This information is broadly defined and describes the search and retrieval of information and these resources are available from millions of web sites and different web databases and is referred as Web Content mining, and the retrieval of information and analysis of browsing patterns from more than one servers or services is referred as Web Usage mining[30].

The expansion of the WWW has resulted in a large amount of data that freely available across the Web for user access. Web technology is not evolving in incremental steps, but it is turbulent, erratic, and often rather unmanageable. It is estimated that the internet has expanded and is doubling in size every six to ten months. In the years to come, the advancement of computer related technologies and their decrease in the costs have expanded the means available to collect and store the data. As a consequence the amount of meaningful information stored has been increasing at a very fast rate. Traditional information analysis techniques are useful for creating reports from data and to confirm a predefined hypothesis about the data. However large volumes of data being collected and created new challenges for storage and such techniques, as the data collected over a period of time contains hidden knowledge about the business or character of a customer and their behavior. With the rapid growth of WWW, the information, its modeling and prediction [22] of user behavior, restructuring a web site and placing a particular link has become an important part of knowledge discovery process. The various types of data has to be managed and efficiently organized in a way that they can be accessed by different users. Therefore, the approach of data mining techniques on the WWW is now the focus of our research. Many methods are used to discover the concealed information in the World Wide Web, which web mining has been developed into an autonomous research area. From the application point of view, information and knowledge obtained from the web can be applied to manage activities related to electronic commerce, digital libraries[23], electronic governance[42], electronic business[41], electronic services[24].

Since the web technology feeds on ideas, by which many efficient search engines and companies are built. As many companies use internet to give and procure information, knowledge discovery on the web has become an important research area. As the growth of resources started available on the web, it has become mandatory for companies to discover the usage patterns and analyze the discovered patterns. The pattern mining is an area with an assortment applications, in which repeated pattern mining is an area where patterns are discovered in web log data.

The server log report is to be analyzed using a Log Markup language (LOGML), an XML[1]
application developed to describe the activity report of server log files, and it is in LOGML format for a web site. The data to be mined is collected from server log files, which are useful for studying the possibilities and to collect the links and structures. The focus is to find the frequent patterns in server log data, which is the information about the user actions, which can be used for promotions & publicity purposes, and also for generating powerful user profiles. In this paper the recurrent patterns can be mined [18] and enquiries are done from the Web usage mining point of view. The divergent patterns in web log mining [34] are item-sets, sequences and graphs. The resulting report can be website, browser and usage statistics. The log database is mined using LOGML[34] for the information to be extracted from web logs. By using LOGML[34] the preprocessing steps are easier to implement. The frequent pattern mining model is used to extract interesting patterns from the Log Markup Language database.

The structure of this paper is as follows. The Section 2 institutes the basic tasks of web mining and its approaches. In Section 3 the usage mining is given in detail. Section 4 describes the basics of Log Markup Language. The algorithms used in the pattern extraction phase of the mining process are described briefly in Section 5.

II. WEB MINING PROCESS

WWW mining [3] involves applications that focus at finding and taking out hidden information that of data stored on the WWW. Thus, web mining [9] can be placed into three different classes based on which part of the WWW is to be mined. These three categories are content mining, structure mining and usage mining.

The Content mining[43] cover the problems of automatic data filtering, classes of smart search agents and to customized web agents.

The Usage mining [30] depends on the structure of the site, and extracts information from user’s access behavior which is stored in server logs.

The different classes of data are multimedia data, the XML documents, and semi-structured HTML documents and unstructured text. The aim of content mining is to make available a well-organized mechanism to help the users to find the information they request. Content mining[43] includes the task of arranging and clustering [25] the documents and understanding search engines for accessing the variety of documents by keywords, classes and contents.

The structure mining is the process of discovering the constitutional design of a site it discovers the link structures at the Inter-document level.

The focus of the structure mining is to recognize the root pages, having useful information and which are supported by diverse links pointing to the root page. Nodes are Web pages having a lot of referencing hyperlinks means that the content of the page is useful, preferable and maybe reliable, which contain many links to parent pages, thus they help in clustering the different root pages. Mining the structure of the Web supports the task of Web content mining, the document retrieval, and the reliability and relevance of the documents. The graph structure can be investigated by Web structure mining to improve the performance of the information retrieval and classification of the records.

Usage mining [19][30], the main topic of this paper focuses on information discovery from the usage of different sites and the task of finding the actions of the users while they are going through the sites. The taste of the visitors are used to enhance the quality of electronic commerce and its services, to customize the web sites and to enhance the constitution and server performance.

The process of usage mining [19][20] is to discover the knowledge which can be extracted from the actions of users, which are stored in server logs.

Applications of Web Mining

Usage mining is used:

1. To analyze the user clicks as a data and combine data from different sources and use analytics for customized interactive publicity programs.
2. To indicate browsing actions of a user and to forecast the required web pages.
3. By determining repeated actions of users and links can be established to improve the overall performance for future actions.
4. By storing of web pages so that better search engines are built.
5. The mining of patterns is used to gather perceptive patterns to improve customer relationship management.
6. In the study of how browsers are used and the users interactiveness with the interface.
7. Usage mining of patterns gives a key to comprehend the traffic behavior and subsequently used to deal with policies on web caching.
transmission, and load comparison and data distribution.

8. In identifying the intrusions, frauds and penetration into the systems.


10. In the areas of Medicine, Sciences and Engineering.

11. To know the how browsers work.

12. In network traffic analysis to determine the requirements and different loads.

More-ever the requests from the client side stored in the database are missing since most of the algorithms work based only the server side data. Some commonly used data mining algorithms for usage mining is association rule mining [32], sequence mining [28] and clustering [29].

### III. WEB USAGE MINING

Usage mining is the task of applying the techniques of data mining to find usage patterns from page data and to understand and serve the need of the user as the user browses the web. Usage mining is used in many areas such as digital business, CRM, Services, digital learning, electronic governance, Digital Libraries [23], promotion and marketing, biotechnology and so on.

The process of usage mining consists of three main steps: (i) pre-processing (ii) finding patterns and (iii) analysis.

The usage mining process is presented in below Figure 1.

There are number of data pre-processing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data by removing irrelevant information like image files, CGI scripts and other information for this reason the data have to be converted in the preprocessing phase such that the output of the conversion can be used as the input of the algorithms. In the pattern matching different algorithms are used for matching patterns obtained for discovering interesting patterns. The resulting patterns are analyzed for the results obtained by the algorithms. In classifying threats and fight against terrorism and fraud detection.

The server side does not contain records of page accesses that are cached on the proxy servers or on the client. Besides the log files on the server, proxy server files provides additional information.

The log file data is given as input to the pre-processing stage, where it changes the data contained in the various sources into the ideas necessary for sequence discovery. This stage needs much effort in the mining process because of incomplete data. The client and server side sessions are identified by IP address, agent and server side click stream data, which may not be sufficient in identifying the different sessions.

Some of the problems encountered are:

- Isolated address but many sessions – The Service providers have multiple proxy servers by which users able to access the Internet. Many number of user’s access a site which are connected to a single proxy server over a time period.

- Multiple address but one Session – Some ISPs dynamically assign an IP address over a request from the user, by which a session can have multiple IP addresses.

- Many addresses but a single user – Tracking a user is difficult when he uses different machines to access same web site, as each session may have different IP addresses.

- Different browsers but a single user – When a server is accessed by a user with different browsers from same machine and the user may appear as a multiple user.

Let each user may be identified by using path, cookie or login sessions. The click sessions of each user has to be divided as requests from other servers are not available.

The users click stream are divided into sessions by time out methods. The timeout session is based on the results of browsing behavior [5]. The extracted information of the user action is available from the server log’s request field.

![Figure 1. Usage mining](image-url)
Figure 2 shows a general architecture [4] of web usage mining proposed by B. Mobasher et al. The frequent patterns discovery phase needs only the pages visited by a user, the sequences of the pages are irrelevant and the pages are ordered in a predefined order. The Sequence mining [28] needs the ordering of the pages, and if a page was visited more than once by a user in a time interval, then it is also taken into account. Sequence mining can be used to discover the pages which are accessed immediately one after another. Using this knowledge the activity of the users can be determined and the next to be visited pages can be predicted.

For sub-tree or web site graph mining the sequences are needed but also the structure of the web pages visited by a user. The forward navigations of the users are taken, which form a tree for each user.

IV. LOG Markup Language (LOGML)

The web log files are the server log files which are generated by using LOGML[34] are compressed form of Log files. The site master store log file data in several files. The log file has information like user id, user timings and statistics for a single day. As much more data will be accumulated in log files, therefore the old data is deleted to make room for the new log files. All the reports are combined and generated by site master to get a final report.

LOGML[2] is based on extensible markup and modeling language (XGML)[1], an XML based application makes the task of generating reports easier and can generate reports of log files and can combine them with XSLT[6] to produce HTML reports. The reports and graphs of web server to be analyzed are generated by using LOGML[34] with the combination of other XML applications. The log file information can be structurally expressed by using LOGML[2] and by XML. The root elements of the web log file is used to describe the graph and its structure. The LOGML[27] file has three sections under root element. The first narrates the graph generated by the visit of users to pages and links. The second is the information of host, agents and others. The third is the report of the user sessions [21], which is a sub graph of graph generated on user visits with a time stamp. The time stamp compute the total time of the user session.

LOGML for web data mining

The problems that arise in web data mining are solved using LOGML [2]. One of the problem is referred to as frequent patterns, it encircle mining approach and it derives to include more intricate patterns like tree mining and graph mining. From the previous studies of association and frequent pattern rules, information is retrieved from intricate structures from the same database. Such intricate patterns contain sub-trees, acyclic graphs, and frequent directed & undirected sub-graphs.

The LOGML file has the graph and user sessions, which are sub-graphs of the net graph and can be used to extract intricate and factual patterns. Given a LOGML document of a web site, one can perform different mining tasks. A site can be shaped as a digraph and can have forward links as well as back references, which create a recurring series of states. If the user actions are given in a file that stored in a database, we can discover normally occurring sub-graphs.

The ever expanding web results in huge data that is accessible for users across the frontiers. The web is the single largest source of data available. According
to the NetCraft Survey the total web sites in usage are given below in the Fig3 [33].

Fig 3. Total web sites in usage

NetCraft is a service provider devoted to tracking technology on the Internet, which provides security services, including anti-fraud and anti-phishing services, research data and analysis on many aspects of the Internet. In June 2010 Netcraft survey received responses from 206,956,723 sites. Total no of sites across all domains is found to be 240,749,695.

V. OVERVIEW OF THE MINING ALGORITHMS

Before carrying the research on usage mining, understand the mechanism of mining algorithms to analyze the results. The methods and algorithms developed from statistics, data mining, machine learning and pattern recognition are used in Pattern Matching. Efficient discovery of frequent patterns from large data sets plays an important role in data mining. It is the topic of investigation in data mining, with wide uses in industry and huge inferences in other topics of data warehousing and data mining.

The range of frequent-pattern mining algorithms covers mining associations, correlations, sequential patterns and closed-patterns, partial periodicity, etc. and their applications are in classification, data warehousing, spatial databases, multimedia databases, time-series databases, text databases, and World Wide Web [3].

Torben B Pedersen et al [29] proposed a nearer way to examine click data that overcome large storage requirements, I/O cost, and scalability problems. An integration of the existing Hypertext Probabilistic Grammar (HPG) and Click sequence Table approach is used to mine web logs.

K.A.Smith [26] proposed an approach by utilizing R. Srivastava et al [13] proposed an approach to usage based web personalization [36], the offline tasks of usage data, and the online process of page customization based on the mined knowledge.

Self-Organizing map (SOM) organize pages into a two-dimensional map, based on the user behavior, rather than content of the pages. The resultant tool can be used to know the user behavior with their characteristics. This tool can be used for navigation by users and also by web masters.

John R. Punin et al proposed XGMML[1] and LOGML [27] in discovering curious information from browsing actions as stored in web access logs. XGMML[1] is an application based on XML format and it describes a graph structure and LOGML is an application to describe log reports of web servers. A website graph is generated in XGMML[1] format for a web site. The web-log report is generated in LOGML[34] format for a web site from web log files and the web graph.

The focus of this paper is to provide an overview of how to use mining techniques for discovering different patterns in a Web log database using the LOGML[34] format. As with web mining techniques, it covers Apriori[31], the most commonly used Association Rule algorithm[31] that utilizes the frequent itemset strategy. Apriori [32] and its various kinds of improvements and extensions, projection and frequent-pattern growth techniques, constraint-based mining, sequential pattern mining methods. Apriori[32] was the first scalable algorithm designed for association-rule mining algorithm. Apriori [32] is an improvement over the AIS[38] and SETM[37] algorithms. The Apriori[32] algorithm searches for large itemsets during its initial pass and uses its result as the seed for discovering other large datasets during subsequent passes. Many association rule algorithms have been used, such as Apriori[32] and Partition.

Comparative studies of Web Usage Mining[30]

<table>
<thead>
<tr>
<th>S.No</th>
<th>Description</th>
<th>Observation</th>
<th>Algorithm</th>
<th>Application</th>
<th>Area Used</th>
<th>Visibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Click Sequences</td>
<td>Overview large itemsets management, resource efficiency and scalability issues</td>
<td>CAN &amp; PRM</td>
<td>Web page behavior</td>
<td>frequent pattern mining</td>
<td>Predict user navigation preferences based on page views and page ordering time</td>
</tr>
<tr>
<td>2</td>
<td>Web Mining</td>
<td>User behavior based</td>
<td>Efficient clustering algorithms</td>
<td>Recommendation systems</td>
<td>Frequent pattern mining</td>
<td>Predict user navigation preferences based on page views and page ordering time</td>
</tr>
<tr>
<td>3</td>
<td>Self Organizing Map</td>
<td>User log data</td>
<td>Log Mining</td>
<td>Sparse Clustering</td>
<td>Predict user navigation preferences based on page views and page ordering time</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Log Analysis</td>
<td>User log data</td>
<td>Effective clustering algorithms</td>
<td>Recommendation systems</td>
<td>Predict user navigation preferences based on page views and page ordering time</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Web Access Patterns (WAP)</td>
<td>Frequent WAP</td>
<td>TAI-VAP, CATCH</td>
<td>Web access patterns</td>
<td>Predict user navigation preferences based on page views and page ordering time</td>
<td></td>
</tr>
</tbody>
</table>

The navigational action of a user is registered in the server logs in sequence, on analyzing the traits of
these navigational actions with the browsing habits, we can forecast user’s next action and offer customized website content and service based on prediction of web pages.

Web Personalization [36] is any action that makes the Web experience of a user personalized to the user’s taste.

LOGSIM utilizes self organizing map to organize web pages into a two dimensional map based on the users browsing behavior.

LOGML[34] is used to describe web log reports and combine them with XSLT to produce HTML reports.

Web access pattern, which is the sequence of accesses taken by users frequently, is a kind of interesting and useful knowledge.

An Analytic Hierarchy Process[39](AHP) technique is used for designing and evaluating the responses to the questionnaires, a mathematical decision making technique that allows consideration of both qualitative and quantitative aspects of decisions. It reduces complex decisions to a series of pair wise comparisons, then synthesizes the results. It is often used in business risk management.

The AHP[39] technique gives the weights of the attributes considered in the questionnaire, which helps the researcher as an indicator for the further process. The attributes are known group-wise, the task that follows is testing their applicability in practices and relevance of the attributes identified. For the purpose, the questionnaire has been applied for five attributes. These attributes are employed on an underlying scale with values ranging from 1 to 9 to rate the relative preferences for two criteria (see Table 1), by generating pair-wise comparison matrices using AHP techniques[39].

<table>
<thead>
<tr>
<th>Intensity of Importance</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Equal importance</td>
</tr>
<tr>
<td>2</td>
<td>Equal to moderate importance</td>
</tr>
<tr>
<td>3</td>
<td>Moderate importance</td>
</tr>
<tr>
<td>4</td>
<td>Moderate to strong importance</td>
</tr>
<tr>
<td>5</td>
<td>Strong importance</td>
</tr>
<tr>
<td>6</td>
<td>Strong to very strong importance</td>
</tr>
<tr>
<td>7</td>
<td>Very strong importance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criteria</th>
<th>CS</th>
<th>WP</th>
<th>SOM</th>
<th>LOGML</th>
<th>WAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Click Sequence</td>
<td>1.5</td>
<td>1.0</td>
<td>2.0</td>
<td>0.25</td>
<td>0.23</td>
</tr>
<tr>
<td>Web Personalization</td>
<td>3.0</td>
<td>1.0</td>
<td>3.0</td>
<td>0.32</td>
<td>0.23</td>
</tr>
<tr>
<td>Self Organizing Map</td>
<td>2.0</td>
<td>0.25</td>
<td>2.0</td>
<td>0.32</td>
<td>0.14</td>
</tr>
<tr>
<td>Log Markup Language</td>
<td>1.5</td>
<td>0.25</td>
<td>1.5</td>
<td>0.876</td>
<td>0.25</td>
</tr>
<tr>
<td>Web Access Pattern</td>
<td>3.0</td>
<td>0.25</td>
<td>3.0</td>
<td>0.25</td>
<td>0.23</td>
</tr>
<tr>
<td>Total</td>
<td>1.25</td>
<td>0.25</td>
<td>1.25</td>
<td>0.25</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table1. Scale for pair wise comparison[40]

VI. CONCLUSIONS

In this paper we survey the research area of web usage log mining. For the survey, we focus on the representation issues and also on the process. Web logs are being used by academic and industry researchers, to study, to understand and to improve the interaction between the user and web services. This introduces the process of web log mining, and shows how frequent patterns are discovered and they applied on the web log data in order to obtain useful information about the user navigational behavior. We have seen that some of the traditional data mining algorithms have been extended or new algorithms have been used to work on the web data. With the growth of web based applications there is a significant interest in analyzing the web data to better understand web usage and apply the knowledge to better services to the user. However, web usage mining raises some questions that must be answered before such robust tools can be developed.

REFERENCES


