Web Usage Mining: Architecture, tools& application

Somanjoli Mohapatra[1]Chinmaya Dash[2] Prakash Chandra Behera[3]

[1]Assistant Professor, Department of Sciences, St. Claret College, Bangaloe-560013
[2]Assistant Professor, Department of Sciences, St. Claret College, Bangaloe-560013
[3]Assistant Professor, Department of Sciences, St. Claret College, Bangaloe-560013

Abstract:
A huge amount of data is available in form of web documents over the World Wide Web and is increasing day by day. Web mining is used to extract useful information from web documents which is categorized into three types, namely, web content mining, web structure mining and web usage mining. Web Usage Mining mainly deals with discovery and analyzing of usage patterns in order to serve the needs of web based applications. The process of Web Usage Mining mainly consists of three inter-dependent stages: data preprocessing, pattern discovery and pattern analysis. Web usage mining useful for the applications like e-commerce to do personalized marketing, fight against terrorism, fraud detection, to identify criminal activities, web design etc.
Keywords*-Web Usage Mining, Web usage mining process, Web usage mining applications and Tools.*

1 INTRODUCTION

Web Mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining is the application of data mining techniques to extract knowledge from Web data including Web documents, hyperlinks between documents, usage logs of web sites etc. A common taxonomy of web mining defines three main research lines: Web Content Mining, Web Structure Mining and Web Usage Mining.Web Mining can be categorized in to three broad areas of mining.

Web Content Mining: Web Content Mining (WCM) is responsible for exploring the proper and relevant information from the contents of web. It focuses mainly inner document level.

Web Structure Mining: Web Structure Mining (WSM) is the process by which we discover the model of link structure of the web pages.

Web Usage Mining: Web usage mining is a research field that focuses on the development of techniques and tools to study users web navigation behavior. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web based applications. It also called as Web log mining.

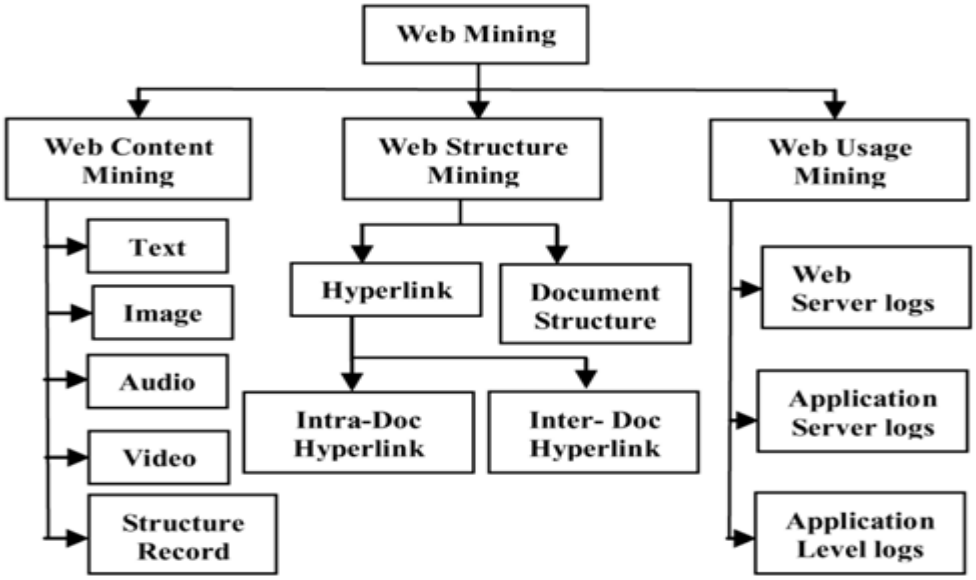

Fig 1: Types of Web Mining

A Web log file records activity information when a Web user submits a request to a Web Server. A log file resides at three different places in system:

Web Server Log: It resides in web server and notes activity of the user browsing website. There are four types such as access logs, agent logs, error logs and referrer logs.

Web Proxy Server Log: It contains information about the proxy server from which user request came to the web server.

Client browser Log Files: It resides in client's browser and to store them special software are used.

2 WEB USAGE MINING:

The web usage mining related to the application of data mining tools and technique. The web usage mining is used to discover usage patterns from web data in order to understand the user's need for navigating on the web. Web usage mining is used to discover the navigation patterns from web data, predicts the behavior of user while the user interacts with the web and also it helps to improve large collection of resources.Mainly there are four types of data sources present in which usage data is recorded at different levels they are: client level collection, browser level collection, server level collection and proxy level collection.

Client Level collection: This data shows the behavior of a single user on single site. Client side data collection requires user participation for enabling java scripts or java applets. The advantage of data collection at client side is that it can capture all clicks including pressing of back or reload button [1].

Browser Level Collection: Second method of data collection is by modifying the browser. It shows the behavior of single user over multiple sites. The data collection capabilities are enhanced by modifying the source code of existing browser.

Server Level Collection: Web server log stores the behavior of multiple users over single site. These log files can be stored in common log format or extended log format. Serv-er logs are not able to store cached page views. Another technique used for usage data collection at server level is TCP/IP packet sniffing.

Proxy Level Collection: Proxy servers are used by internet service provider to provide World Wide Web access to customers. By predicting the usage pattern of the visitor Web Usage Mining improves the quality of e- commerce services and enhances the performance of web structure and web server.

3 WEB USAGE MINING PROCEDURE AND TECHNIQUES

The whole procedure of using Web usage mining for Web recommendation consists of three steps, i.e. data collection and pre-processing, pattern mining (or knowledge discovery) as well as knowledge application.The steps involved in Web Usage Mining are as follows:

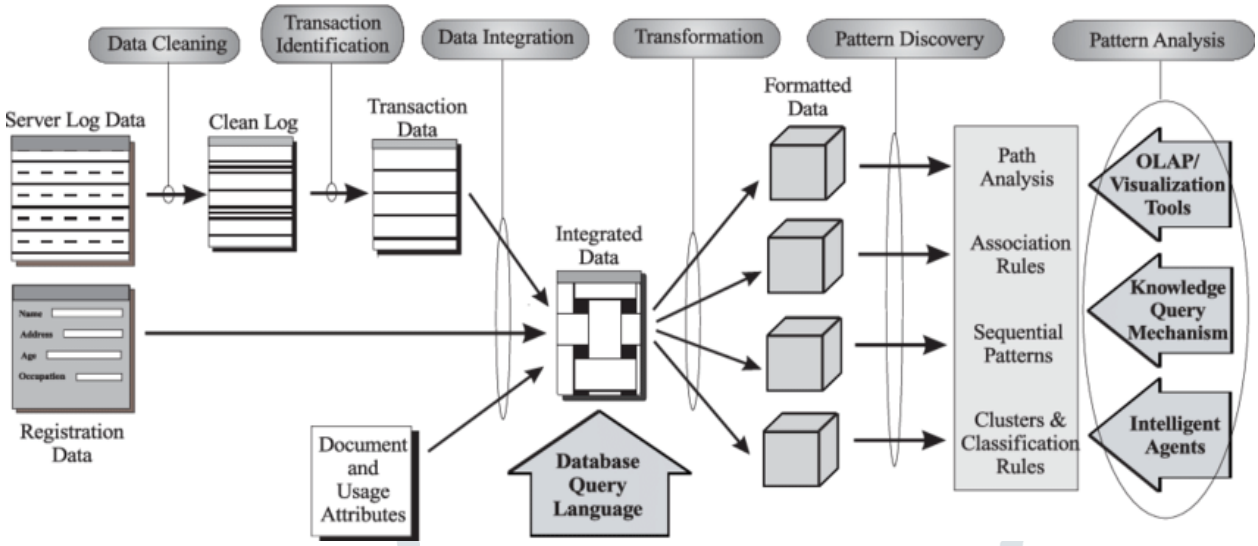

Fig 1: Architecture of Web Usage Mining

3.1 Data collection : It is the very first step of Web usage mining. It involves extraction of log data from server log files. Data can be basically collected from three sources[1,25,26]:

The server side: These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format.

The Proxy Side: Many Internet Service Providers give to their customer proxy server services to improve navigation speed through caching. The main difference with the server side is that proxy servers collect data of groups of users accessing groups of web servers.

The Client Side: Access data can be tracked also on the client side by using JavaScript or applets, or even modified browsers. These techniques avoid the problems of session identification.

3.2 Data Integration: Integrate multiple log files into a single file is defined as data integration.

3.3 Data preprocessing : Real world data may be noisy or inconsistent so we have to preprocess them to make them consistent and reliable. So preprocessing phase is very important step of web usage mining [9]. The main steps of preprocessing are:

Data Cleaning/ Data Reduction: The purpose of data reduction process is to remove unwanted data that may affect the overall mining process[4]. Use of this algorithm, these types of useless data is removed and the mining process gets be evaluated results comparatively fast.

User Identification: User identification refers to identify unique users.

Session Identification: Session identification refers to differentiate the web log entries into different user sessions by a session timeout. Once a user was identified then click stream is divided in to clusters. This method of division is called Session Reconstruction or Sessionization.

Path Completion: This step is used to check the missing pages after constructing transactions. The missing page problem is due to proxy servers and caching problem of clients.

3.4 Pattern discovery : In Pattern Discovery phase, data mining techniques like association rule mining and clustering applied on web log files after preprocessing to discover the useful pattern.

Association rule mining: Association rule mining is one of the data mining technique which is used to discover useful pattern. It works on generating frequent pattern and rules. In web log file number of URL

visit by number of users so we can identify frequently accessed web pages by users which can help to understand user needs.

Clustering Analysis: Clustering Analysis is used to group the data or items which have similar attributes or characteristics. Clustering analysis defined as similar characteristics users are group together without knowledge of group definition.

Sequential pattern analysis: Sequential pattern analysis is used to find that a suspected user visit a particular link A followed by link B in a time ordered set of sessions. By using this approach we can predict the suspected user psychology which is useful in crime detection.

Classification: In this method web server data is classified according to some common attributes like hour of the day in which data accessed. Classification is a mapping method of data that could be one or several predefined data.

3.5 Pattern Analysis: The main purpose of pattern analysis is to analyze the pattern which is identified during pattern discovery phase. Its main purpose is to find out a valuable model or standard pattern for specific web usage mining application.

OLAP (Online Analytical Processing Technique) is a powerful paradigm for strategic analysis of relational database which is very useful in business systems. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management, budgeting and forecasting, financial reporting and similar areas, with new applications coming up, such as agriculture.

Data and Knowledge Querying: Query mechanism such as SQL is the most common method of pattern analysis. When a user makes a lot of errors during login on any e-commerce site, it may be a malicious user that wants to guess the password.

Usability analysis: Itis a modeling technique to accessing the behavior of user on the web site. The main reason of SQL injection attack is an inefficient input validation in the database.

Visualization Technique: Visualization Technique is a method that used to understanding the behavior of web users by graphical method.

4 TOOLS USED IN WEB USAGE MINING

Some tools used to explore Web Usage Mining are:

Web Utilization Miner (WUM): WUM uses mining language MINT which is the mining language serving as interface between the user and the miner. MINT supports the specification of criteria of statistical, structural and textual nature. To discover the navigation patterns satisfying the expert's criteria, WUM exploits an innovative aggregated storage representation for the information in the web server log.

KOINOTITES*:* KOINOTITES, is a software system that exploits Web Usage Mining and user modeling techniques for the customization of information to the needs of individual users. KOINOTITES processes the Web server log files, and organizes the information of a Web site into groups, which reflect common navigational behavior of the Web site visitors.

Web miner: A general and flexible framework for web usage mining to extract relationship from data collected in large web data repositories.

Web Site Information Filter System: Web SIFT system uses content and structure information from the web site in order to identify potentially interesting results from the mining usage data.

WebViz: WebViz provides that selectively filtered a web server log, control bindings to graph attributes and also reissue of logged sequence of requests.

Web log miner: It uses data mining and OLAP on treated and transformed web access files. Mining a web server log files.

Web mate: The user profile is inferred training examples, proxy agent provides effective browsing and searching help.

Web usage miner: It exploits an innovative aggregated storage representation for the information in the web server log.

i-miner*:* To optimize the concurrent architecture of fuzzy clustering algorithm and fuzzy inference system to analyze the trends, pattern discovery and trend analysis from web usage data mining.

5. APPLICATIONS OF WEB USAGE MINING

The main applications of Web Usage Mining are :

Personalization of web content: Web Personalizationis Web based information systems adaptive to the needs and interests of individual users, or groups of users. Typically, a personalized Web site recognizes its users, collects information about their preferences and adapts its services, in order to match the user's' needs.

Web design: This helps in the designing the web structure based on the user's query and can help to retrieve only relevant documents.

E-commerce*:* Web Usage Mining retrieves the user information from the various web logs. This information may be related to their personal information like age, qualification, their interests, their economy, their usage time, frequently accessing sites, their buying behavior etc.

Web Advertising/Marketing: Web advertising also referred to as an online advertisement. The use of popular websites can be an effective way of introducing new a product to the customer.

Pre-fetching and caching: Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time.

Transaction Analysis: Analysis of e-commerce uses clickstream data to determine the marketing effectiveness of the site by quantifying user behavior while actually visiting the site visitor browsing the site recording the translation in a sales transaction.

Modification of web site: For successful website, modification according to user need is essential. Required modification is successfully determined by the web usage mining of the server log data.

Fraud detection: Unauthorized users can be traced using search results of web log data. A user unsuccessfully trying the access to any web site may be an intruder tries to break the password of restricted area of website.

Customer Relationship Management: It focuses creating value for the customer and company over the long term and the relationships are built with the customers, which provide value for services.

Product/Site recommendation: Web site and various products can be recommended to users according to the user interest using web usage mining.

Identify Web Robots: Web Robots are software programs behaves like human for target website. These programs are very harmful for websites because they may crack a password or may breakdown the site by continuous fake requests.

To improve web server program's performance: Web usage mining is very useful for improvement of performance of the of web server.

6 CONCLUSION

This paper has discussed about the Web Mining and its types, and also discuss the issues related to log files. In this paper we conclude that WUM is mining process to extract useful pattern from log files and enhance the performance of web pages using personalization.Future scope of Web Usage Mining resides on Digital forensics investigations, Crime investigation, Automated data cleaning, Robot detection and filtering, Transaction identification etc.

7 REFERENCES

[1] Bharti Joshi, Ph.D., SuhasiniParvatikar, "Analysis of User Behavior through Web Usage Mining", International Journal of Computer Applications, International Conference on Advances in Science and Technology (ICAST-2014).
[2] Aditi Shrivastava, Nitin Shukla, "Extracting Knowledge from User Access Logs", International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012
[3] Amit Pratap Singh , Dr. R. C. Jain, "A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May – June 2014.
[4] G.D.Praveenkumar,R.Gayathri, "A Process of Web Usage Mining and Its Tools", International Journal of Advanced Research in Science, Engineering and Technology, Vol. 2, Issue 11 , November 2015.
[5] NanhaySingh ,Achin Jain , Ram Shringar Raw, "COMPARISON ANALYSIS OF WEB USAGE MINING USING PATTERN RECOGNITION TECHNIQUES", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013.
[6] C. Sakthipriya , G. Srinaganya , Dr. J. G. R. Sathiaseelan, "An Analysis of Recent Trends and Challenges in Web Usage Mining Applications", International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 4, April 2015, pg.41 – 48.
[7] Aarti M. Parekh, Anjali S. Patel, Sonal J. Parmar, Prof.Vaishali R. Patel, "Web usage Mining:Frequent Pattern Generation using Association Rule Mining and Clustering", International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 04, April-2015.
[8] S. E. Salama, M. I. Marie, L. M. El-Fangary and Y. K. Helmy, "Web Anomaly Misuse Intrusion Detection Framework for SQL Injection Detection", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No. 3, 2012, pp. 123-129
[9] Dr.S. Vijiyarani and Ms. E. Suganya, "RESEARCH ISSUES IN WEB MINING", International Journal of Computer-Aided Technologies (IJCAx), Vol.2, No.3, July 2015.