Web Usage Mining: A Survey on User's Navigation Pattern from Web Logs

Richa Patel[1] Akshay Kansara[2]
[1]P.G. Student [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]Mehasana, India

Abstract— With an expontial growth of World Wide Web, there are so many information overloaded and it became hard to find out data according to need. Web usage mining is a part of web mining, which deal with automatic discovery of user navigation pattern from web log. This paper presents an overview of web mining and also provide navigation pattern from classification and clustering algorithm for web usage mining. Web usage mining contain three important task namely data preprocessing, pattern discovery and pattern analysis based on discovered pattern. And also contain the comparative study of web mining techniques.

Key words: web mining, web usage mining, web log data, classification, clustering

I. INTRODUCTION

In today's world internet has become extremely popular and its growth is very rapid. The information is available on internet for people whom they are using for their different intend. The resources for using internet are growing fast, it is necessary for users to use automatic tools for discover desired information. People require systems at client side and server side for finding out the desired information. Above system used to mine data and extract information from that source. So for a specific user only interesting information of web is useful and rest of the information not important. Several users are interested in content of web and they can browse by using search engines. Using web log files, we can find out information related to web access pattern. These web log files provide the information related to behavior of user. In business area, user's behavior plays an important role for extracting information.

In this area, user navigation patterns are described as the common browsing behaviors along with a group of users. During navigation, many users may have common interests .so navigation patterns should capture the overloaded information or user's need. In addition, navigation patterns should also be able to differentiate among web pages based on their different meaning to each pattern.

The rest of paper is as follows: Section II presents the overview of web mining and types of web mining. Section III related to literature work. Section IV explains the comparisons of WCM, WSM and WUM. Section V represent conclusion.

II. WEB MINING

A. Overview

Web mining is a part of data mining techniques to automatically discover and extract knowledge from the web. It can be broadly divided into three tasks: web content mining, web usage mining, web structure mining.

Classification of Web Mining can be understood using the given Fig .1

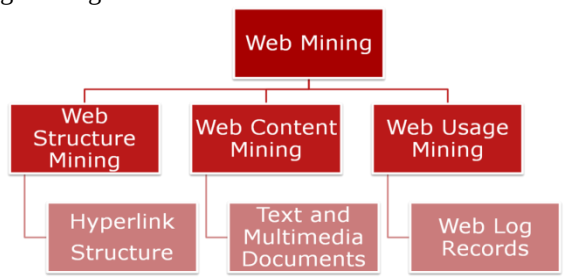

Fig.1: Web Mining Classification

Web Mining is an application of Data Mining which deals with the extraction of interesting pattern from the World Wide Web. In above figure web mining can be classified into three categories: web structure mining, web content mining and web usage mining. Web Structure Mining is the process of inferring knowledge from the World Wide Web and links between web pages. The structure of a typical web graph contains web pages as nodes and hyperlinks as edges between related web pages. It is the process of using graph theory to examine the node and connection structure of a web site. Web content mining is the process of extracting useful information from the available web pages. Generally, the web content mining contain the several types of source data such as textual, image, audio, video, metadata as well as hyperlinks. Web Usage mining is a part of data mining techniques to discover useful navigation patterns from web log. Resource data is usually collected when user interact with web server such as web server log, proxy server logs, user queries, registration data. In short, Web usage mining is a process of extracts information from user how to use web sites. Web content mining is a process of extracts information from texts, images and other contents. Web structure mining is a process of extracts information from hyperlinks of web pages.

B. Web Usage Mining

Web usage mining is an part of web mining. It contain the two categories, first is general access pattern tracking and second is customized usage tracking. Now, general access pattern is a mining process using the history of the web page visited by user. And customized usage tracking is targeted on specific user. Web usage mining, the art of analyzing user interactions with a web page, has been dealt by several researchers using different approaches.[1] there are many research area in data mining techniques like association rule mining, classification ,clustering and Sequential-pattern-mining-based.

C. Web Usage Mining Techniques

These techniques given below:

1) Sequential-pattern-mining-based: Allows the discovery of temporally ordered Web access patterns,
2) Association-rule-mining-based: Finds correlations among Web pages,
3) Clustering-based: Groups users with similar characteristics,
4) Classification-based: Groups users into predefined classes based on their characteristics [2].

Here all are given briefly:

1) Association Rules Mining:

Association rule mining can be used for related pages that are most often referenced together in a single server sessions. These rules refer to sets of pages that are accessed together with a support value more than some predefined threshold. Association rule is a concept of data mining that is used for basket transaction data analysis. There are several association rule algorithms have been used, such as Apriori, Partition. This rule is being applicable for business intelligence and marketing applications, e-Commerce and also it can help web designers to restructure their web site.

2) Sequential Pattern Mining:

A database consists of sequences of events or values that change with time, is called a time-series database. Time-series database is widely used to store historical data in a variety of areas such as, scientific data, financial data, medical data, and so on. Different mining techniques have been designed for mining time-series data, basically there are four kinds of patterns we can get from various types of time series data:1) Trend analysis, 2) Sequential pattern, 3)Similarity search, and 4) Periodical patterns. This technique of sequential pattern mining attempts to find the relationships between occurrences of sequential events, and also to find if there exists any specific order of the occurrences We can find the sequential patterns of cross different items and also find the specific individual items. Generally, Sequential pattern mining is used in analyzing of DNA sequence. An example of sequential patterns is that every time IBM stock drops at least 4%, Microsoft stock will drops 5% within three days.

3) Classification:

Classification is a supervised learning technique, to automatically generate a model that can be classify as a class of objects so that to predict the classification or missing attribute value of future objects whose class may not be known. Classification is a two-step process. In the first step is model construction, it is based on the collection of training data set, a model is constructed to describe the predetermine characteristics of a set of data classes or concepts. Class label of training data is known is called supervised learning (i.e., which class the training sample belongs to is provided). In the second step, model usage is used to predict the classes of future objects or data. . Classification can be done by using several learning algorithms such as decision tree classifiers, k-nearest neighbor classifiers, naïve Bayesian classifiers, Support Vector Machines etc. The decision tree induction Classification algorithms such as C4.5, ID3, CART, and Hunt's algorithm can be used to predict if page is of interest to the user.

4) Clustering:

Classification is a supervised learning technique, clustering is another mining technique similar to classification. However clustering is an unsupervised learning technique. Clustering is a technique in which set of data item or object having similar characteristics they are group together. Objects characteristics are same within a cluster but dissimilar to characteristics of object in other clusters. Clustering can be performed on either the page views or the user views. Clustering analysis in web usage mining intends to find the cluster of page, user, or sessions from web log file, where each cluster represents a group of objects with common interesting or characteristic. The clustering method are model-based, partition based, grid based, hierarchical based. To construct partition of data, we use partition method like k-mean, k-medoid. Each partition represent cluster.

5) Web Log Data

Log files can be classified into three categories:

- Web Server Log Files: Web server log files located in web server and notes activity of the user browsing website. There are four types of web server logs i.e., transfer logs, error logs, agent logs, and referrer logs.
- Web Proxy Server Log Files**:** web proxy server log files contain information about the proxy server from which user request came to the web server.
- Client browser Log Files**:** client browser log files reside in client's browser and to store them special software are used.

6) Log Files Parameters

To recognizing user browsing patterns, they are use log files parameters which are given below.

- User Name: To identify the user who has visited the website and this identification normally is IP address.
- Path Traversed**:** it is the path taken by the user within the website.
- Visiting Path: in which path taken by the user while visiting the website.
- Time Stamp: It is the time spends by user on each page and is normally known as session.
- Page Last Visited: It is the last page visited by the user while leaving the website.
- Success Rate: It is measured by copying and downloads activity carried out on the website.
- User Agent: It is the browser that user send request to server.
- URL: It is the resource that is accessed by the user and it may be of any format like CGI, HTML etc.
- Request Type: It is the method that is used by user to send request to the server and it can be either GET or POST method.

a) Types of Log File Format

There are mainly three kinds of log file formats that are used by majority of the servers.

Common Log File Format: It is the standardized text file format that is used by most of the web servers to generate the log files.[16]

Format is given below:

- Log Format "%h %l %u %t \"%r\" %>s %b" common CustomLog logs/access_log common eg: 127.0.0.1 RFC 1413 frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326[16]
- Combined Log Format: It is same as the common log file format but with three additional fields i.e., referral field, the user_agent field, and the cookie field.[16]

Format is given below:

Log Format "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{Useragent i\"" combined CustomLog log/access_log combined eg**:** 127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I;Nav)"[16]

- Multiple Access Logs: It is the combination of common log format and combined log file format but in this format multiple directories can be created for access logs.[16]

Format is given below:

LogFormat "%h %l %u %t \"%r\" %>s %b" common CustomLog logs/access_log common CustomLog logs/referer_log "%{Referer}i -> %U" CustomLog logs/agent_log "%{User-agent}i"[16]

The above techniques help in the development of effective marketing strategies as well as design of better Web sites. Generally, users are not ready to disclose personal information and may tend to give false information to the websites. Hence, it is assumed that the anonymity of users for WUM in general, particularly for non-commercial web sites where user registration is not required, and disregards clustering and classification.

D. Web Usage Mining Steps

There are several steps in web usage mining given in fig.2

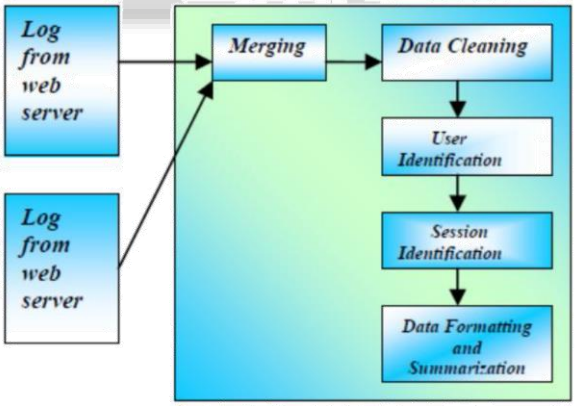

Fig. 2: stages of preprocessing [3]

The main objectives of preprocessing are to reduce the size of data and at a same time quality being enhanced. Preprocessing contain the following steps – Merging of Log files from Different Web Servers, Data cleaning, User Identification, Session Identification, Data formatting and Summarization as shown in Fig. 2.

1) Merging

At the beginning of the data preprocessing is merging, the request from the all web log files, merging all log files with web server name and synchronization of web server clocks, including time zone differences.

2) Data Cleaning

The second step of data preprocessing is data cleaning, it consists of remove unnecessary request from web server log files. We need to eliminate the irrelevant entries if all web log entries are not valid. Usually, this process remove request which contain non-analyzed data such as images, multimedia files and page style files.

3) User Identification

In generally, the web log file contains only the computer address (name or IP address) and the user agent (for ECLF log files). The web log file also contains the user login that can be used for the user identification when user login or register in particular web sites.

4) Session Identification

To group the activities of a single user from the web log files is called a session [4]. A user session is called when as longer as user connected to particular web sites. Default session taken as a 30 minutes time-out. A session is a set of reference pages from one web site during one logical time period. Historically session identification is done by a user logging into a computer, performing the task and then logging off. The login and logoff represent the logical time period of the start session and end of the session.

5) Data Formatting & Summarization

This is the last step of data preprocessing, the structured or formatted file containing user sessions and visits are transformed into a relational database model.

III. LITERATURE REVIEW

The focus of literature review is to study and analyze about available technique to predict the users' navigation pattern with the web or web site. . There are many Web log analysis tools presented for mine the data from log record on Web page. Log record provides useful information likes IP address, URL and time and so on. Analyzing and discovering of log could help organizations to discovered more potential customers, pages popularity number of times a page has been visited etc. that can help in reorganizing the Web site for fast and easy customer access, , attracting more advertisement capital by intelligent adverts improving links and navigation, turning viewers into customers by better site architecture.

Internet is a gold mine, but only for those companies who realize the importance of Web mining and adopt a Web mining strategy now.[5] Generally, Companies have to implement Web mining systems to identify their own strength and weakness , and to understand their customers' profiles of their E-marketing efforts on the web through continuous improvements. In this paper author describe the overview of web mining and types and application of web mining. From the data analysis and graphing workspace tool, we are helpful in providing useful information related to the user access patterns, which could not be possible by using traditional approaches.[6] in this paper they considered university's peak working time for analysis of web traffic. Traffic data was selected based on daily, monthly and hourly including request volume and page volume to generate cluster models. In web usage mining grouping of web access sequences can be used to determine the behavior or intent of a set of users.[7] The task of grouping web sessions based on similarity and consists of maximizing the intra-group similarity while

minimizing the inter-group similarity is done using sequence alignment method.[7] In this paper author describe a new method to group web sessions, which considers the global and local alignment techniques of similarity measurement.. Length of session also plays an important role for session. Sequence alignment method can be improved by affine gaps in computing global alignment value .we can also find the web session similarity based on Fuzzy Rough Set Theory. In that, rough set theory defines the characteristics of each cluster. Web site graph contain vertices of web pages and edges of link between web pages. Using Table Filling Algorithm to generate the cluster Traditional clustering methods create clusters by describing the members of each cluster whereas the rough set based clustering techniques create clusters describing the main characteristics of each cluster.[9] [8], There is an attempt to provide an overview of the state of the art in the research of web usage mining, while discussing the most relevant tools available in the sphere as well as the niche requirements that the current variety of tools lack. All tools are related to statistical analysis techniques, likes WEBMINER, Web Tool, Web Mate and so on. The main uses of web content mining are to gather, categorize, organize and provide the best possible information available on the Web to the user requesting the information [17]. In this paper web content mining tool scan the HTML documents, text and images. They compare the web content mining tools like screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, and Web Content Extractor. Screen-scrapper needs prior knowledge of proxy server and some knowledge of HTML and HTTP where as other tools do not require any such knowledge and it need Internet connection to run [17]. Web contains noisy data, redundant information and which mirrored web pages in and abundance [18]. They use web content extraction method to remove noisy data present in web documents. They use three phases to perform web content extraction method. First phase list of documents are selected, second phase documents are preprocessed, and in third phase result are presented to user. Proposed method shows better performance when compared with existing methods [18]. Web content mining tool and techniques and analyzed the method for retrieving information. This method focuses on the following objectives: Focusing on the role of web content extraction and identifying list problems when mining list of documents. Studying the solutions to this problem, presenting the method which is used to identify required patterns in an effective manner [18] Examining a Web content data consist of structured, semi – structured and un- structured data. Web content mining techniques such as text based clustering, partition clustering, hierarchical clustering, graph based clustering, neural network based clustering , fuzzy clustering and probabilistic clustering. And they also describe the research issue on web content mining such as data/information extraction in that product or result search extraction, web information integration and schema matching, opinion extraction from online sources in that customer review of products, forums, blogs, and chat rooms .knowledge synthesis, segmenting web pages and detecting noise in that like content of web page without advertisement, navigation links and copyrights notices. An EM algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.[10]This algorithm is a type of partitioning clustering method and it is a mixture of Markov chain used for clustering user session. In this algorithm there are two steps, first is expectation step in which value is expected from cluster probabilities and second is maximization step in which they compute distribution parameters and their like hood. This algorithm work same as k-mean in that set of parameters are re-computed until desired value is achieved. This algorithm used for improve cluster quality. Machine learning research area have two activity first is clustering and second is classification. Both the classification and clustering algorithm are used to find the maximum and minimum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.[11]In this paper ,author use graph partitioned clustering algorithm to discover the navigation pattern and LCS classification algorithm to predict future requests. After applying clustering and classification three performance parameters accuracy, coverage and F1 Measure are measured. They provide accuracy 0.87% when threshold was 0.9.To predict future request and finding navigation pattern we use clustering and classification techniques. In this paper there are two phase, one is online phase and second is offline phase. The offline phase takes care of preprocessing and clustering, while the classification and prediction is performed during the online phase [12].Ant-based clustering method used to discover or extract user's navigational pattern from web log files. The LCS classification algorithm uses the knowledge from offline stage and predicts the users' next request [12]. After applying clustering and classification they measure maximum accuracy 74% when threshold value increases. In this paper we deal with classification algorithms for studying the user/client behavior and for the generation of interested user patterns [13]. In this paper they give the comparison of classification algorithms based on some factor like accuracy, precision, session based timing and recall. Classification algorithms are to be applied on the web log data and then performance of these algorithms can be measured. Here classification algorithm likes Decision Tree Classifier, Naive Byes Classifier, Support Vector Machine, Neural Networks, Rule Based Classifier, and K-Nearest Neighbor Classifier are analyzed. After analyzing above algorithms, author say that the Decision Tree classifier and SVM algorithm are perform well as compared to others. Clustering is useful in characterizing customer groups based on purchasing patterns, categorizing web documents that have similar functionality [14]. In this paper they uses graph based clustering algorithm which is Chameleon clustering algorithm, it takes data point as an input and forms clusters. They uses adjacency matrix in which session as rows and pages as column. This matrix generates the data point which is input to algorithm. This algorithm works same as KNN graph. It has two phases. In first phase, they use graph partition algorithm to partition cluster into small sub-clusters. And in second phase, they merge the cluster until relative number of cluster is achieved. This algorithm provides both interconnectivity and closeness to identifying the most similar pairs of cluster.

IV. COMPARATIVE STUDY OF WCM, WSM AND WUM

Here, given table represent the comparison of WCM, WSM and WUM:

Specification	WCM	WSM	WUM
Tasks	information from the web content / documents	It discover the model underlying the link structure of the web	It tries to make sense of data generated by web surfer's session or behavior
Input Data	primary data	It uses primary data	It uses secondary data
Data View	structured, semi-structured, unstructured	Linking of structure	Interactive Data
Method	Machine Learning, statistical Method, Association Rule, Proprietary algorithm	Proprietary algorithm	Statistical Method, Machine Learning
Application Categories	Finding Extraction Rules, Finding Pattern in text, Finding Frequent Sub-structure, Categorization, Clustering	Categorization, Clustering	Site Construction, Adaption and Management, Marketing User Modeling
Outcome	Bag of word, Phrases, Edge labeled graph	Graph	Relational Table, Graph

V. CONCLUSION

This paper provide a more current evolution and of Web Usage Mining. It describes the user navigation pattern from classification and clustering algorithm. Web Usage Mining is a fast rising research area for generating interested log information. And also provide the comparison table of Web Content Mining, Web Structure Mining and Web Usage Mining.

REFERENCES

[1] Niranjana. Kannan and Dr. Elizabeth Shanthi, "Classification and Clustering of Web Log Data to Analyze User Navigation Patterns", JGRCS, Volume 1, No. 1, August 2010
[2] Yew-Kwong Woon ,Wee-Keong Ng, Ee-Peng Lim,"Web Usage Mining: Algorithms and Results"
[3] Ramya C, Kavitha G, "An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network", Fifth International Conference on Information Processing, Augus-2011,Bangalore, INDIA
[4] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", ICAIES'2012,July 15-16, 2012
[5] Monika Yadav, Pradeep Mittal, "Web Mining: An Introduction", IJARCSSE, Volume 3, Issue 3, March 2013
[6] Shakti Kundu, "An Intelligent Approach Of Web Data Mining", IJCSE, Vol. 4 No. 05 May 2012
[7] Bhupendra S Chordia, Krishnakant P Adhiya, "Grouping Web Access Sequences Using Sequence Alignment Method", IJCSE, Vol. 2 No. 3 Jun-Jul 2011
[8] Chhavi Rana, "A Study of Web Usage Mining Research Tools", Int. J. Advanced Networking and Applications,Volume:03 Issue:06 Pages:1422-1429 (2012)
[9] T. Vijaya kumar & H. S. Guruprasad, "Clustering Of Web Usage Data Using Fuzzy Tolerance Rough Set Similarity and Table Filling Algorithm", IJCSEITR, Vol. 3, Issue 2, Jun 2013, 143-152
[10] Norwati Mustapha, Manijeh Jalali, Abolghasem Bozorgniya, Mehrdad Jalali, "Navigation Patterns Mining Approach based on Expectation Maximization Algorithm", World Academy of Science, Engineering and Technology Vol:3 2009-02-26
[11] V.Sujatha, Punithavalla, , "Improved User Navigation Pattern Prediction Technique From Web Log Data", International Conference on Communication Technology and System Design 2011
[12] K. Devipriyaa and Dr. B.Kalpana, "Users' Navigation Pattern Discovery using Ant Based Clustering and LCS Classification", JGRCS, Volume 1, No. 1, August 2010
[13] Supreet Dhillon, Kamaljit Kaur , " Comparative Study of Classification Algorithms for Web Usage Mining", IJARCSSE, Volume 4, Issue 7, July 2014
[14] T.Vijaya Kumar, Dr. H.S.Guruprasad2, "Clustering Of Web Usage Data Using Chameleon Algorithm", IJIRCCE, Vol. 2, Issue 6, June 2014
[15] Naresh Barsagade, **"**Web Usage Mining and Pattern Discovery: A Survey Paper"
[16] Nanhay Singh1, Achin Jain1, Ram Shringar Raw1. "Comparison Analysis Of Web Usage Mining Using Pattern Recognisation Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013
[17] Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi, " Overview Of Web Content Mining Tools", The International Journal Of Engineering And Science (IJES) ,Volume 2 ,Issue 6 ,2013
[18] Shanmuga Priya, S. Sakthivel, "An Implementation Of Web Personalization Using Web Mining Techniques", IJCSMC, Vol. 2, Issue. 6, June 2013, pg.145 – 150

