ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

Group Movement Pattern Mining in Web Access Log for Navigation Behavior

K.Anitha

Assistant Professor, Dept. of CSE, Cauvery College of Engineering and Technology, Trichy, India

ABSTRACT: Web usage mining (WUM) is the technique of extracting useful and interesting patterns from web access log. Web access log is the log-file that contains registered user access by the server. WUM is otherwise called as web log mining that is used to understand visitor's behavior and navigation or routing preferences of the visitor mainly for evaluating the effectiveness of the website and for enhancing the quality of the website respectively. The aim of discovering frequent patterns in web log data is to obtain information about the navigational behavior of the user. This can be used for advertising purpose, for creating dynamic user profiles, etc. The click-streams generated by various users often follow distinct patterns. The proposed distributed mining algorithm is Group Movement Pattern Mining that identifies a group of visitors with similar patterns. The proposed algorithm comprises of two phases that are local mining phase and cluster ensembling phase. In local mining phase, the system finds navigation pattern of each user and compute the similarity between the users to identify the local groups. In cluster ensembling phase, the global group can be identified by clustering local groups. Distributed mining algorithm achieves good grouping quality.

KEYWORDS: web usage mining, pre-processing, log file, pattern discovery, web recommendation

I. INTRODUCTION

The recent and massive growth of web site over the 'World Wide Web' (WWW) increases the necessity of doing the projects in web usage mining. By this, the lifetime value of the customers can be identified, and also improve their cross marketing strategies. Website affects on the size and depth level of web site structures. This results in negative impression of presented content and weaker usability. These problems can be solved by recommendation and personalization. Recommendation is narrow image of personalization and defined as suggestion— pointing specific information objects to the user.

People use the internet to get information in field of interest, do research works, increase awareness about the world, get currently updated news, etc. Getting the most appropriate information has become complicated for web users. It is easy to make offered prioritized area of interest if and only if interest of user from browsing history can be recognized.

The data over the web is collected in the form of server access logs that are generated by the interaction of clients with the web site and stored in the form of transaction logs. The web servers generally store this information in the form of server logs or access logs. Different genres of organizations can make use of this data by analyzing for respective purposes. The session of analyzing server logs providing information on how to better structuring of web site for effective use and benefit of the organization.

Web Usage Mining involves determining the frequency of the page access by the clients and then finding the common traversal paths of the users. With the help of this analysis, web site owner can re-structure the web site with the navigation results.

Web Usage Mining is a four-step process. The first step is data collection, the second step is data pre-processing, the further step is pattern discovery and the last step is pattern analysis. In the pre-processing stage, the system involves cleaning of the click stream data and the data is partitioned into a set of user transactions with their respective visits to the website. During the stage of pattern discovery, the use of machine learning algorithms are performed on the transaction logs to find hidden patterns and the behaviour of the users. In the final stage, the discovered patterns from the prior stage are further processed and filtered producing models that can be served as an

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

input to different visualization tools and report generation tools. The process called as pattern analysis. The last stage performs the filtering, aggregation and characterization on the discovered patterns.

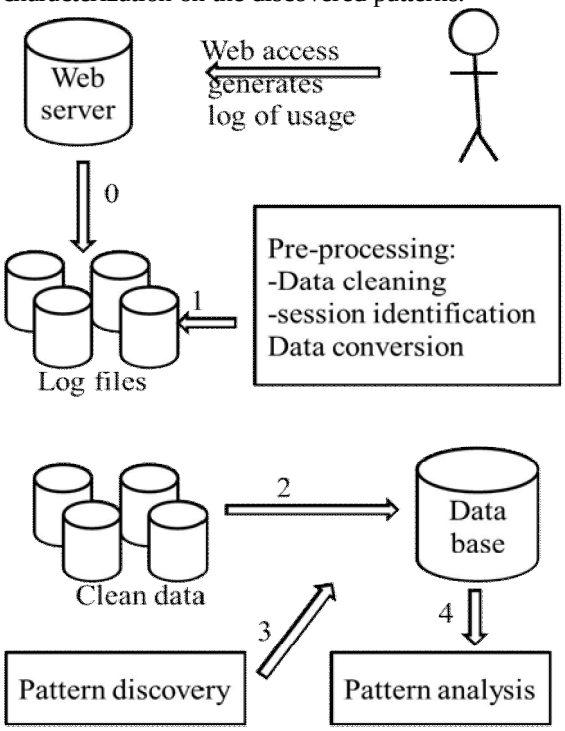

Fig.1. A logical view on web usage mining

Identification of user intentions helps to personalize the experience of web browsing and also has some other benefits. The web pages likely to browse frequently by a user can be cached to reduce the time of retrieval and loading the same web page on network considerably. In e-commerce, awareness about user intention plays the vital role to become very about the targeted clients for reduced marketing advertisement and for increased clients count. This results to propose a system for interest of user according to browsing pattern and the time duration at each web page by the user.

Recommendation and personalization can be seen as taking any actions adapting objects' presentation on the Web site to the user needs. The consequence of conducting such actions may be higher user retention calculated by the number of opened pages and total session duration.

Five major steps followed in web usage mining are

1. Data collection – Web log files, which keeps track of access of users i.e., visits of all the visitors
2. Data Integration – Integrate multiple log files (different log file in different format) into a single file
3. Data pre-processing – Cleaning and structuring data to prepare for pattern extraction
4. Pattern extraction – Extracting interesting patterns
5. Pattern analysis and visualization – Analyze the extracted pattern i.e., knowledge presentation
6. Pattern applications – Apply the pattern in real world problems

II. RELATED WORK

In recent years, there has been an increase in research works done with regard to web usage mining i.e., Future user request prediction.

WANG Tong et al., [1] offered an improved AprioriAll algorithm based on the original AprioriAll algorithm. The new algorithm includes the property of the UserID during the every step of generating the candidate set and each

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

step of scanning the database by which to decide whether an item in the candidate set must be put into the large set which will be used to generate next candidate set.

Ming-Syan et al., [2] presented a new data mining technique that involves mining path traversal patterns in a distributed information i.e., documents or objects are connected together to make possible interactive access. There is two steps procedure. First, derive an Algorithm to convert original log data sequence into a set of maximal forward references. Second, derive another Algorithm to determine the frequent traversal patterns i.e., large reference sequences from the maximal forward references obtained.

Hengshan Wang et al., [3] proposed two common data mining algorithms – Fpgrowth and PrefixSpan into Web Usage Mining. In WUM model, sequential pattern mining and PrefixSpan is used. Maximum Forward Path is combined with WUM model to trim down the nosiness of "pseudo visit" done by browser cache as well as to enhance the extracting of frequent traversal paths.

Kobra Etminani et al., [4] proposed ant-based clustering algorithm to pre-processed web log data to extract frequent patterns and then it is exhibited in an understandable format.

N. Sujatha et al., [5] have proposed a new method to develop the web sessions' cluster quality from k-means clustering using Genetic Algorithm (GA). First, a customized k-means algorithm is used to cluster the user sessions. The refined initial condition allows the iterative algorithm to converge to a "better" local minimum. Second, they have proposed a GA based refinement algorithm to improve the cluster quality.

Paola Britos et al., [6] described Self Organized Maps, a kind of artificial neural network which is used in the process of WUM to detect user's patterns. It converts the data storage in the Web Servers Log files to an input of Self Organized Maps.

III. PROPOSED SYSTEM

The web log tracker is the proposed system that personalizes and recommends the web access patterns. It consists of

- Data collection
- Data pre-processing
- GMP mining
- Web recommendation

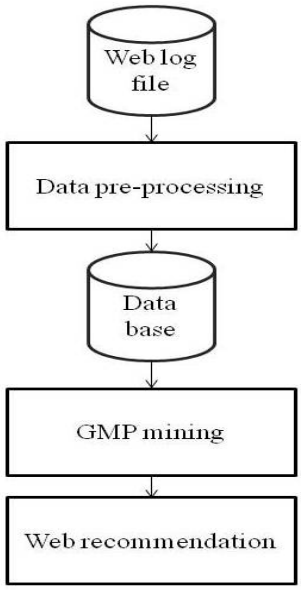

Fig.2. Architecture of web log tracker

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

DATA COLLECTION

Data Collection is the first step in web usage mining process. It consists of gathering the relevant web data. Data source can be collected from the Apache server access logs of Periyar Maniammai University, Vallam, Thanjavur. The log file is in the form of text file.

DATA PREPROCESSING

The information available in the web is heterogeneous and unstructured. Therefore, the pre-processing phase is a prerequisite for discovering patterns. The goal of preprocessing is to transform the raw click stream data into a set of user profiles. Data preprocessing presents a number of unique challenges which led to a variety of algorithms and heuristic techniques for pre-processing tasks such as merging and cleaning, user and session identification etc.

DATA CLEANING

Data Cleaning is a process of removing irrelevant items such as jpeg, gif files or sound files and references due to spider navigations. Improved data quality improves the analysis on it. The Http protocol requires a separate connection for every request from the web server. If a user request to view a particular page along with server log entries graphics and scripts are download in addition to the HTML file. An exception case is Art gallery site where images are more important. Check the Status codes in log entries for successful codes. The status code less than 200 and greater than 299 were removed.

USER IDENTIFICATION

In web usages mining does not require knowledge about a user history because the users visit or request given more than one time to the server. If client visit more than one time, then it generate multiple sessions for each user. It is also known as User activity Records. User Identify by using IP address and User Agent in log files. Client request to server then it generate log files at that time client also send user agent to server. The combination of IP + AGENT we can find out users.

SESSION IDENTIFICATION

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 portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. Time oriented is depends on the Time stamps or date and time of request in the server log file. Sessionize by means of the difference between First request and last request is < =30 minutes. Structure oriented capture in the referrer fields of the server logs. Structure oriented depends on Referrer fields is currently open or that user currently login referrer. Means it's belonging to more than one "open" constructed session.

PATH COMPLETION

The path completion confirms that the user how visited the web pages. Path completion is depends on mostly URL and REFF fields in server log file. It is also graph model. Graph model represents some relation defined on Web pages (or web), and each tree of the graph represents a web site. Each node in the tree represents a web page (html document), and edges between trees represent the links between web sites, while the edges between nodes inside a same tree represent links between documents at a web site.In the path completion Missing Reference this method also used. Missing Reference means the user backtrack should not be stored in server log file. It cached in client side.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

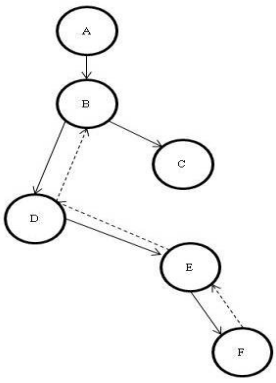

Fig.3. Web site structure

GMP MINING

GMPMine algorithm identifies groups of objects and determines their movement patterns. the GMPMine algorithm, where s represents the location sequence data set and N denotes the number of objects of interest. The minimal similarity threshold (SIMmin) is the lower limit of the similarity between two objects belonging to the same group.

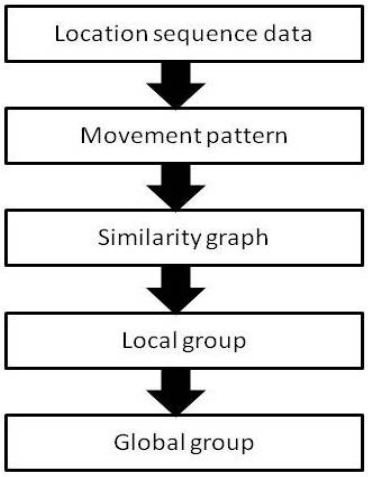

Fig.4. Functionality of GMP mining

WEB RECOMMENDATION

This can be achieved by the suggesting clients with the observed knowledge through the user interpretation. The personalization and recommendation are the main theme of doing the web usage mining.

IV. APPLICATIONS

The main aim of WUM is to gather interesting and useful information about users routing or movement patterns (i.e., to characterize web users). Later, this information can be used to improve the web site from the perception of user. The outcome produced by the web log mining can used for a variety of purposes: (i) to personalize

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

the delivery of web content; (ii) to improve user navigation through pre-fetching and caching; (iii) to improve web design; or in e-commerce sites (iv) to improve the customer satisfaction.

PERSONALIZATION OF WEB CONTENT:

Web Usage Mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behaviour by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users. Personalized Site Maps are an example of recommendation system for links. There is an adaptive technique to reorganize the product catalogue of the products according to the forecasted user profile. A survey on existing commercial recommendation systems, implemented in e-commerce web sites.

PRE-FETCHING AND CACHING:

The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. Typically, Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time.

SUPPORT TO THE DESIGN:

Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications. Stratograms are used to evaluate the organization and the efficiency of web sites from the users' viewpoint. Web Usage Mining techniques are used to suggest proper modifications to web sites. Adaptive Web sites represents a further step. In this case, the content and the structure of the web site can be dynamically reorganized according to the data mined from the users' behavior.

E-COMMERCE:

Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure.

V. CONCLUSION

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 behaviour. 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 that the pre-processing stage is an important work in mining to facilitate efficient pattern analysis. To get accurate mining results user's session details are to be known.This project implements the new method of web usage mining by using group movement pattern mining algorithm called web log tracker.

REFERENCES

[1] WANG Tong, HE Pi-lian, "Web Log Mining by an Improved AprioriAll Algorithm", World Academy of Science, Engineering and Technology 4 2005

[2] Ming-Syan Chen, Jong Soo Park, Philip S. Yu, "Efficient Data Miningfor Path Traversal Patterns", IEEE Transactions On Knowledge And Data Engineering, Vol. 10, No. 2, March/April 1998.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 4, April 2016

[3] Hengshan Wang, Cheng Yang, Hua Zeng, " Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan", Communications of the IIMA 2006 Volume 6 Issue 2
[4] Kobra Etminani, Mohammad-R. Akbarzadeh-T., Noorali Raeeji Yanehsari, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", IFSA-EUSFLAT 2009
[5] N. Sujatha, K. Iyakutty, "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.464-476
[6] Paola Britos, Damián Martinelli, Hernán Merlino, Ramón García-Martínez, "Web Usage Mining Using Self Organized Maps", International Journal of Computer Science and Network Security, VOL.7 No.6, June 2007
[7] Qiankun Zhao, Sourav S. Bhowmick, "Sequential Pattern Mining: ASurvey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003118 , 2003.

BIOGRAPHY

Anitha is the Assistant Professor, Dept. of CSE in Cauvery College of Engineering and Technology, Trichy, Tamil Nadu, India. She has completed her B.E from Kings College of Engineering and Technology, Anna University and M.Tech from Periyar Maniammai University, India. She can be contacted by her e-mail id: anitha.kaliyarasu@gmail.com