Recommend Websites through Weblog Files using Association Rule

Dhruva Mistry
Post Graduate Student
C.E Dept, BVM Engineering College, Gujarat, India

Kirti J. Sharma
Assistant Professor
C.E Dept, BVM Engineering College, Gujarat, India

Samip A. Patel
Assistant Professor
I.T Dept, BVM Engineering College, Gujarat, India

ABSTRACT

In recent years netizens prefers that web accessing is fast in nature and give appropriate results without any confusion. Recommendation is one of the most useful system based on analysing web log files to be applied for web personalization in future. Web Usage Mining(WUM) is also applicable for online marketing and site modification. This system consists of three main interdependent tasks of WUM which are Data Preprocessing, Pattern Discovery and Pattern Analysis. Association rules are used for relate pages that are most often reference together in a single server session. The main goal of recommendation system in this research is to improve web site usability and to gives recommendations of websites to users of their use.

General Terms

A. Web Usage Mining

Web Usage Mining is an 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 emphasize on finding user access patterns from web browsing data stored in Web server log, proxy server logs or browser logs. The usage data records the user's behavior when the user browses or makes transactions on the web site.

B. Association Rule Mining

Association rule are used for prediction of next event or discovery of associated event. An association rule X-->Y is a relationship between two itemsets X and Y such that X and Y are disjoint and are not empty. A valid rule is a rule having a support higher or equals to minsup and a confidence higher or equal to minconf. The support is defined as sup(x-->Y) = sup (X U Y) / (number of transactions). The confidence is defined as sup(x-->Y) = sup (X U Y) / sup (X).

In the web log files, the transaction consists of the number of URL visits by the client, to the web site. Applying different association rule mining algorithm we can predict which are web pages frequently accessed together by users of website. The discovery of such rules from the access log can be of tremendous help in reorganizing the structure of the web site. The frequently accessed web pages should be organized in their order of importance and be easily accessible to the users[2].

Keywords

Association Rule Mining,Web Usage Mining,Frequent Pattern(FP) Tree.

1. INTRODUCTION

There exists huge amount of data on the internet and so it becomes much more difficult for users to access information effectively. Analysing and modelling web navigation behaviour is helpful in understanding demands of onlineusers. Application of mining of web log files of a web server is consists of three main tasks: 1)In Preprocessing we remove all irrelevant user's request from web server log files leading to log reduction followed by users' identification and sessions' identification.2) In pattern discovery we extract association rules and find co-occurrence patterns from web log files.3)In pattern analysis we analyse the set of generated rules independent of website's topology to extract valid set of rules that achieves highest coverage of dataset. This paper's experimental results provided relative recommendation pages that can help web designer to restructure websites, web application or portals to better serve web customers.

2. RELATED WORK

Web log files are the best source to predict user's behavior. Along with the useful information, the raw log files also contain entries for unnecessary details like image access, failed entries etc. which are of no use from the perspective of the WUM[3].

Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta[2] proposed a new approach of pattern discovery using association rule mining. They defines types and format of web log files. Also proposed algorithms for data cleaning. user identification and session identification as this includes in data preprocessing step. Using association rule for pattern discovery they found some limitation in declaration of minimum support and minimum count values. As solution of this first apply clustering technique in which grouping objects with similar behaviour in different clusters so that reduces the input data set for association rule mining. For that they use DBSCAN algorithm with two parameters: minEps which is data area to be cover and minPts to form a clusters. In the next step to find user's access patterns apply association rule mining with Apriori algorithm on clustered data. They use min_confidence =85% and min_support=30%.For pattern analysis use OLAP/visualization tools.This research implemented on website log of JACPC for 2months collections of data. Using graphical representation they show comparison of raw log data and preprocessed log data. Further they also differentiate results of simple ARM and clustered ARM. At last step they providing to users the websites of their choice and it gives better accuracy in result with their proposed method.

Bamshad Mobasher, Robert Cooley, Jaideep Srivastava [1] apply usage based clustering technique for web personalization. That is consist of two main task:1) offline process mining of data and 2)online process automated web page customization based on mined data. The offline part gives user transaction files by performing data preprocessing steps on raw log files.For online part frequent item sets and URL clusters are used with active session of users and so provide dynamic recommendation to user. Also provide generalized architecture. Data Preprocessing includes data cleaning, user identification, session identification, transaction identification and support filtering. In data cleaning removing all redundant references. The goal of session identification is to divide the page accesses of each user into individual sessions. For that they proposed simple heuristics using the referrer and agent log in server log. Trnsaction identification to dynamically create clusters of reference of each user. Based on user's browsing behaviour each page reference can be categories as content reference, auxiliary reference ,hybrid reference. Given data preprocess task results in unique URLs U and set of user transaction T. For usage based clustering they compute overlapping clusters of URL references based on co-occcurence patterns across user transaction. Association rule used for capture relationship among items based on co-occurrence of patterns. Association rule Hypergraph partitioning is successfully run in variety of domains like content based categorization of web documents. Once URL cluster have been computed the recommendation engine will apply with current user session. The recommendation engine used fixed size sliding window over active user session history and added to user session before that page is sent to the browser. So that adaptive nature of websites will occur resulting from recommendation engine.

M. Maged M. Deghaidy , Khaled Mahmoud Badran, Gouda Ismail Mohamed[3] introduce Web recommendation framework applicable to online business and marketing. In the first step includes sequential eight steps: 1) load unstructured web log data into RDBMS,2)Page aliasing, 3)create new fields RecID, UserID, Timestamp, Session for each transaction,4) User Identification, 5) Session Identification, 6)Data/Time format adjustment ,7) removing irrelevant entries ,8) Data aggregation into single session excluding page hits <2.In the pattern discovery they also defines steps includes: Programmatically divide preprocessed log randomly into two datasets DS1 and DS2,Using WEKA software apply Apriori algorithm for knowledge flow, Generate ARFF file for DS1 and DS2 based on click stream field in each datasetand finally extract association rule with different values of support and confidence. Pattern analysis task is also containing steps that: Load generated association rule in RDBMS, measure similarity between generated association rules from DS1 and DS2,calculate coverage percent and distinct coverage percentage of click stream data of both dataset, Represent sets of rules into graph and in final step to generate recommendation display association rule with descending order according to confidence. After this implementation they compare Generated Rules Similarity. Rules Coverage Percentage for both datasets and for no. of rules with table and graphical representation. At final they concluded that proposed recommendation frame work consistent with the Collaborative filtering approach. It is based on usage patterns discovery and analysis describing users' behaviors and predicting what users will like based on their similarity to other users without relying on machine analyzable content and also capable of accurately recommending items without requiring an understanding of the item itself.

Amit Dipchandji Kasliwal, Dr. Girish S. Katkar[7]introduce Rapid Miner data mining tool for data preprocessing, pattern discovery on NASA weblog files. Data cleaning the first step in knowledge discovery process is done by Matlab code in sequential execution. In the first step web log file exists in text file is loaded into matlab. In the second step matlab convert into matrix format and then after extracted required information from it. User identification is done by easily assign different IP address as individual users. Time threshold 30 min is taken as expiration time for session identification and at last it generates ARFF file which is input file for RapidMiner for further association rule mining. RapidMiner is open source and freeware under GNU AGPL.It provides well suited user interface and do analytical process through pipeline. In pattern discovery phase Rapidminer first READ ARFF file. In the next step it converted in to matrix of numeric, non numeric, discrete and binomial values. After that it matrix is accepted by Fp-Growth and then association rule block for final execution.Implementation is done on web log files of archives.org. Also display generated association rule after applying support and confidence in RapidMiner. This research is very useful to find frequent user visits websites during specific period.

Rahul Neve, K.P Adhiya[7] have done comparison of web mining algorithms for web page prediction in recommendation system. The first algorithm to compare is Generalized Sequential Patterns(GSP) that follows multi-pass and candidate generate-and test approach. In the first scan of database it finds all frequent items with >=min_support value. Each subsequent pass start with seed set of sequential patterns found in previous pass. The second algorithm is Prefix Span Algorithm is work as pattern growth method for mining sequential patterns. This algorithm differ with pattern growth in a way that instead projecting sequential database, it projecting only frequent prefixes. It examines prefix subsequence's and projects only their corresponding postfix subsequence's into projected database. After study of both algorithm they proposed a module to generate frequent access pattern of web pages that is combination of give algorithms. Input to the module is Ip-address with token of URL's sequences. The proposed method mainly give some token to the URL's in sequence and then group into by using Ip-address/domain name. To store module is used to convert back the token into URLs which further used by online web page recommendation system. At last they concluded that Prefix-span algorithm works faster compare to GSP for low min_support.

3. PROPOSED SYSTEM

The proposed recommendation system is mainly useful to find which web pages are more likely to be accessed next page by the users in near future. Relations between given web log are determine by association rule technique with some value of support and confidence.

Figure-1 shows architecture of web page recommendation system. First, all user's web access activities of website is recorded by web server and stored in web server logs[7].The proposed model consists of five major steps:

3.1 Data Collection and Preprocessing

In the initial step web log files of different web servers and website files are combined in single log file for process. Data filtering is essential when large dataset is used for extracting knowledge in different application.

In this research irrelevant entries based on user requirement for different applications are to be removed from the log.User and session identification is also done in this step. Data Preporcessing consists of three substeps:

3.1.1 Remove the entries with extension of .jpg,.png,.gif etc.

3.1.2 Assign each unique ip address as unique userId.

3.1.3 Assign each session different sessionId.Subsubsections

3.2 URL Cluster

This research mainly apply association rule for finding frequent patterns from web log data.But association rule mining has some limitation said by Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta[2].

Limitation of ARM

The association rule mining there is some limitation that too many rules are generated and no guarantee for all generated rules to be relavant. Minimum support and minimum confidence parameters are set in such a way to eliminate false discoveries. When minimum support is too small, every rule will get a chance to be true, leading to wrong recommendation and when minimum support is too large, for small data set, wrong predictions may occur [2].

Solution to avoid this limitation it is suggested to apply clustering process to grouping of object with similar behaviour in different clusters[2].

In this research we had created clusters in a way that outliers will be removed and reduces the input data set to be small for Association rule mining. Consequently the numbers of rules are reduced and the extracted rules are highly relevant and meaningful. K-mean clustering technique is used for grouping URLs ,we call clustering obtain in this way, URL clusters.

3.3 Association Rule Generation

Association rule mining is technique to find relations among data. It also finds hidden patterns in stored data. A valid rules are generated which having support is higher or equals to minsup and confidence is higher or equals to minconf. We are giving minsup, minconf values 30%,75% respectively and transaction database to the input in this research. Generated rules are displayed in sequence in which they were accessed.

Here we had calculate each URL sequence based on their support and confidence value.Also display rules according to calculated score with higher values comes first.

3.4 Frequent Itemset Generation

Frequent items are web pages that are more likely to be accessed next page by the users in the log files. In this step transaction database is processed with minsup value. It is defined according to the dataset. The set of all frequent items are displayed in at least minsup transactions. An itemset is just a set of items that is unordered.

Frequent item set is generated by applying FP-growth algorithm. It works efficiently when it is used with large datasets with almost similar in nature.

3.5 Recommendation Engine

Colleborative filtering (recommendation) is based on usage patterns discovery and analysis describing users' behaviors and predicting what users will like based on their similarity to other users without relying on machine analyzable content.This research observe below steps:

3.5.1 Retrieve Ip address and client request for resource

3.5.2 Search the particular User's past history or user profile from frequent item sets stored in database.

3.5.3 Build recommendation set.

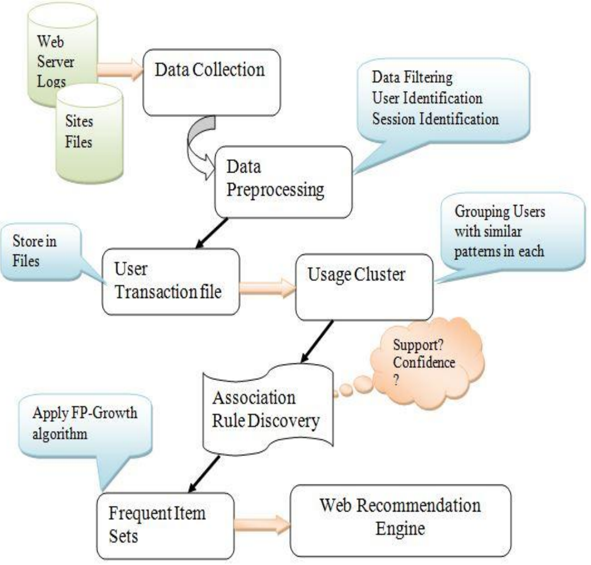

Figure-1 Web Page Recommendation System

4. DATASET DESCRIPTION

In our proposed framework for experimental purposes, we have used the log file contains seven month's worth of all HTTP requests to the University of Saskatchewan's WWW server. The University of Saskatchewan is located in Saskatoon, Saskatchewan, Canada. It consists of 47000 records.

Data Source	Input data	Records after cleaning	Websites	Application
Single site, Multi user	Web server Log	35090	1	Association Rules

Table-1 Taxonomy Dimensions

5. EXPERIMENT RESULTS

For the study of web mining algorithms a system is developed with FP-Growth algorithm which saves memory.The experiments were conducted on Intel Core i3 with 2GB RAM is used to run this project on Windows 7 OS.

Experimental was conducted on Saska College web log files containing 47000 records in raw log file. Following table shows the experimental result proposed method with existing method.

	Recommend websites using ARM	WUM for Predicting User Access Behavior[6]	Web page Prediction in Recommendation system[7]
Dataset	NASA	NASA	NASA
No of records	10k	-	10k
Tool	-	Rapid miner	-
Records after cleaning	6981	-	1986
Mining Algorithm	Fp-growth	Fp-growth in Rapid miner	GSP & prefix span
No. of USER	108	-	645
No. of URLs	369	-	278
Rules generated	query2.lycos.cs.cmu.edu -→ tanuki.twics.com , wpbfl2-45.gate.net , wpbfl2-45.gate.net	[uplherc.upl.com] --> [mission-sts-69.html, mission-sts-71.html]	-
Accuracy	Promising	Good	Best

Table-2 Comparison of Proposed system with existing

Applicable to students, employees in their related field of search for knowledge gain in routine.

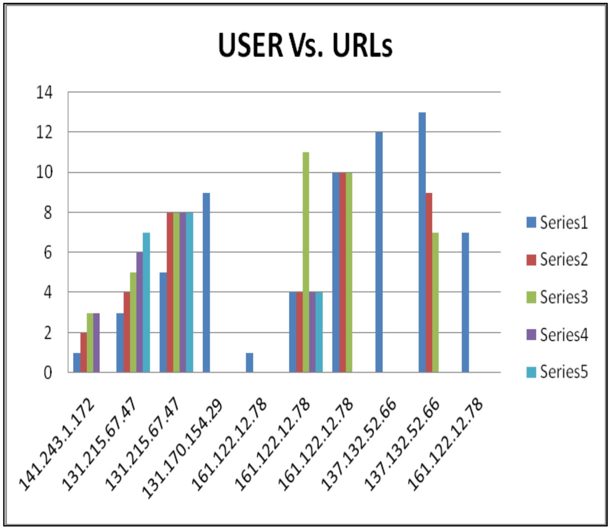

Figure-2 No. of Users accessing URLs

6. CONCLUSION & FUTURE SCOPE

This research useful to society in a way that it extracting knowledge from web usage data and predict user behavior over internet. Due to the navigation of many users and the change of their login time or interests, the web navigation profiles should be extracted again which is a time consuming. Incremental and adaptive navigation profiles will be more suitable for the prediction engine and is a key feature for the future.

7. REFERENCES

[1] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava,"Creating Adaptive WebSites Through Usage-Based Clustering of URLs", Journal of knowledge and information.

[2] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta,"Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery",International Journal of Data Mining Techniques and Applications,2013.

[3] M. Maged M. Deghaidy , Khaled Mahmoud Badran, Gouda Ismail Mohamed,"Web Recommendation Framework based on Association Rules Coverage to be Applied for Site Modification", IJCA,2014.

[4] Prateek Gupta,Surendra Mishra "Improved FP Tree algorithm with customized web log preprocessing",IJCST,2011.

[5] S.Revathi, Dr.T.Nalini"Performance Comparison of Various Clustering Algorithm",IJARCSSE,2013.

[6] Amit Dipchandji Kasliwal, Dr. Girish S. Katkar,"Web Usage mining for Predicting User Access Behaviour", IJCSIT,2015.

[7] Rahul Neve 1, K.P Adhiya "Comparative Study of Web Mining Algorithms for Web Page Prediction in Recommendation System",IJARCCE,2013.

[8] V.Chitraa, Dr.Antony Selvadoss Thanamani" A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications, 2011.

[9] Ms. Dipa Dixit, Ms. M Kiruthika,"PREPROCESSING OF WEB LOGS", International Journal on Computer Science and Engineering, 2010.

[10] Heidar Mamosian, Amir Masoud Rahmani, Mashalla Abbasi Dezfouli,"A New Clustering Approach based on Page's Path Similarity for Navigation Patterns Mining", IJCSIS, 2010.