A Review – Clustering and Preprocessing For Web Log Mining

Shital S. Kontamwar,
M.TechStudent CSE,
Nuva College of Engg. & Technology
Nagpur , India ,
skontamwar@gmail.com

Prof. Anil Warbhe
Assistant Professor,
MIET,
Gondia. , India
mtech2008@rediffmail.com

Prof. Shyam. Dubey
HOD. CSE Department
Nuva college of Engg & Technology,
Nagpur, India,
shyam.nuva@rediffmail.com

Abstract- World Wide Web is consist of large amount of information and provides it different kinds users. Everyday number of users use log on internet.. Internet information growing enormously. Users accesses are documented in web logs. As huge storage log files are growing rapidly .One of the application of Data Mining is Web Usage Mining works on users logs. It consist of various steps such as user identification ,session identification and clustering. Again removing robot entries. In previous years data preprocessing analysis system algorithm on web usage mining has been used buts algorithm lacks on scalability problem. This proposes session identification process and building transaction preprocessing ,data cleaning by using efficient data mining algorithm . The experimental results may show considerable performance of proposed algorithm.

Keywords:-*Data Mining, Clustering, preprocessing, K-medoid, Apriori.*

*****

I. INTRODUCTION

World Wide Web (WWW) is growing rapidly everyday in the number of websites and also the population of users. The basic purpose of website is to provide useful information to its users efficiently and timely. In competition with other websites every site at the same time websites are competing to acquire their own shares of visitors. Many websites are providing number of services to users according to users needs. Again providing the facilities such as online discussions, questionnaires online quiz and finding their interest and again proving the data according to their search on different websites. One of subtle approach is to predetermin into web log files for revealing patterns of users' interests on the websites. It is well know that users' online interactions with the website are recorded in server web log files that serve as a valuable pool of information[2]. As user visits any website their information, logs are stored. .Eventually this log data is growing tremendously. Log data differs from one and another datasets used during data mining, and there are various problems which must be addressed in preparation for data mining. The main problem is to get a reliable dataset for mining. Therefore the data should be pretreated and users' accessing behavior is to be constructed as transactions. These transactions are to be reliable.

The Common log formats or Extended Log Formats only records the visitors browsing activities rather than the details of the visitor's identity. This means that different visitors sharing the same host cannot be differentiated. If there are proxy servers the problem became much severe. Users are identified easily by using Cookies or authentication mechanism. But users are not attracted by these types of sites due to privacy concerns. Web usage mining includes three main steps: Data Preprocessing, Knowledge Extraction and analysis of extracted results. Preprocessing plays a vital role because of the complex nature of the Web architecture as most of mining is done on this step, 80 % mining is done here. The raw data is pretreated to get reliable sessions for efficient mining dependent tasks of data cleaning, user identification, session identification, and clustering and construction of transactions. Data cleaning is the task of removing irrelevant records that are not necessary for mining. User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user sessions. Path completion is used to fill missing page references in a session. Classifications of transactions are used to know the users interest ad navigational behavior. The next step in web usage mining is knowledge extraction in which different data mining algorithms are applied on preprocessed data these techniques such as association mining, clustering, classification etc. The third step is pattern analysis in which tools are provided to facilitate the transformation of information into knowledge. Here we are using clustering algorithm k-medoid. K-medoids algorithm is sensitive to outliners because an object with an extremely large value may sustainably distort distribution of data. it measure distance from cluster center. After getting output from k-medoid, it is applied as input for apriori. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by

identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

II. OVERVIEW ON EXISTING SYSTEM

A. *A New Clustering and Preprocessing for Web Log Mining:*

Web mining is the application of data mining, chart technology, artificial intelligence etc, to the web data and identifies user's visiting behaviors and getting out the interests using patterns. Web mining has become one of the important areas in computer and information science. The purpose of Web Usage Mining techniques in log data to find out the behavior of users which is used in various function like pre fetching, creating attractive web sites, personalized services, adaptive web sites, customer profiling etc. Web server's gathers information about user's interactions in log files. Such as whenever server requests for resources and receives resources.

Log files records maintain information such as client IP address, URL requested etc. By using the theory of distribution in Dempster-Shafer's [1] theory, the belief function similarity measure in this algorithm adds to the clustering task. The ability to extract the anxiety from number. Web user's navigation performance. Log data differs from other datasets used in data mining and there are various problems which must be addressed in preparation for data mining. The main problem is to get a reliable dataset for mining. Therefore the data should be pretreated and users' accessing behavior is to be constructed as transactions. These transactions are to be reliable.

B. . *A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining :*

In Web Usage Mining (WUM), [2] Web session clustering plays a key role to classify web visitors on the basis of user click history and similarity measure. Swarm based web session clustering helps in many ways to manage the web resources effectively such as web personalization, schema modification, website modification and web server performance. In this paper, we propose a framework for web session clustering at preprocessing level of web usage mining. The framework will cover the data preprocessing steps to prepare the web log data and convert the categorical web log data into numerical data. A session vector is obtained, so that appropriate similarity and swarm optimization could be applied to cluster the web log data. The hierarchical cluster based approach will enhance the existing web session techniques for more structured information about user session.

C: A Survey On Parallelization Of Data Mining Techniques

The overview of various parallelization techniques to improve the performance of existing data mining algorithms and make the capable of handling large amount of data. There is variety of techniques to achieve the parallelization in data mining field, in this paper a brief introduction to few of the popular techniques is presented. The second part of this paper contains information regarding various data algorithms that are proposed by various authors based on these techniques Web cache and the IP address Misinterpretation are the two drawbacks in the server log.

D: An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a dynamic ART1 Neural Network **:**

This paper presented an effective methodology for preprocessing required for WUM process. The experimental results illustrate the importance of the data preprocessing step and the effectiveness of our methodology. Next ART1 clustering algorithm to group hosts according to their Web request patterns. a complete preprocessing methodology for discovering patterns in web usage mining process to improve the quality of data by reducing the quantity of data has been proposed.

III. PROPOSED SYSTEM

The proposed system focuses on data cleaning, session Identification process and building the transactions in preprocessing stage. In this research a referrer-based method is given for effectively constructing the reliable transactions in data preprocessing. This gives us the clustering process which uses k-mediod algorithm. We want the resultant groups of pages with common user profiles and this can be achieved by using apriori algorithm for association.

1) *Clustering of data (k-medoid), and Association (apriori)*

Clustering of datameans arranging the data in a similar type as a form database. All similar type of data in single cluster ,different type of data in a different cluster in a database and user searching the data easily also user searching all time. K-medoid algorithm is used for clustering of common type of data element from n elements The *k*-medoids algorithm is a clustering Algorithm related to the *k*-means algorithm and the chooses data points as centers. Basically the *k*-means and *k*-medoids algorithms are partitions data into groups and both used to minimize the distance between points labeled in as cluster to a point as the center of that cluster. The algorithm as follows:

Algorithm:

- K: the number of cluster,
- D : a data set containing n objects.

Output : A set of k clusters

Method

1) Arbitrarily choose k objects in D as the initial representative objects or seeds
2) Repeat
3) Assign each remaining objects to the cluster with the nearest representative objects
4) Randomly select a non representative object , o_{random} ;
5) compute the total cost ,S, of swapping representative object ,$o_{j,}$ with o_{random};

6) if S<0 then swap o_j with o_{random} to form the new set of k representative objects;
7) until no change;

Then A**priori** algorithm is used for store a link as cookies in a database which user search. Apriori is designed to operate on databases containing transactions. Apriori an algorithm provide itemsets and association rule for set mining and confidence in frequent item sets association rule learning over transactional databases. It proceeds further by identifying the frequent of each item from the database again and again up to larger item set. And this continuous until those item sets appear through in the database. And terminates when no further successful extensions are found.

Association rule[3] generation is usually split up into two separate steps:
1. First, minimum support is applied to find all frequent item sets in a database.
2. Second, these frequent item sets and the minimum confidence constraint are used to form rules. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations). The set of possible item sets is the power set over and has size 2n− 1 (excluding the empty set which is not a valid item set). Although the size of the power set grows exponentially in the number of items n in I, efficient each is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent item set, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent. Exploiting this property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent itemsets.

Apriori Algorithm Pseudocode
Procedure

Apriori
(T, minSupport) { //T is the database and minSupportis the minimum support
L1= {frequent items};
for(k= 2; Lk-1!=∅; k++) {
Ck= candidates generated from Lk-1
// that is Cartesian product Lk-1x Lk-1and eliminating any k-1 size item set that is not //frequent
foreach transaction **t** in database **do**{
#increment the count of all candidates in Ck that are contained in t
Lk= candidates in Ck with min Support
}//end for each
}//end for
return Uk, Lk;
}

2) *Clean Similar Data :*

We process on data and clean those similar data entries from database. All data store in a database. Repeated data, repeated link also stored and large amount of data in a database will be generated if one single link suppose user is repeatedly open and repeatedly store in a database.

CONCLUSION

Algorithm used in existing system is lacks in scalability problem. Usage data collection on the Web is incremental. Therefore, there is a need for mining algorithms to be scalable. This can be focused in proposed system. Our research in is to create more efficient session reconstructions more accurate patterns for analysis of users. As internet users are increasing by using clustering algorithm we can have more efficient data and algorithm here are used on large data and we can work on large data and data cleaning saved the disk space. Authentication by admin side makes system more secure than existing solution.

REFERENCES

[1] A New Clustering and Preprocessing for Web Log Mining 1B.Uma Maheswari, 2 Dr. P.Sumathi ,1Doctoral student in Bharathiyar University, Coimbatore ,Tamil Nadu, India2Asst. Professor, Govt. Arts College, Coimbatore, Tamil Nadu, 2014 World Congress on Computing and Communication Technologies
[2] "Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining" Tasawar Hussain, Dr. Sohail AsgharCenter of Research in Data Engineering(CORDE)Muhammad Ali Jinnah University (MAJU),Islamabad, Pakistan
[3] "A Survey On Parallelization Of Data Mining Techniques"- International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 4, Jul-Aug 2013, pp. 520-526
[4] "An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a dynamic ART1 Neural Network" Ramya C*, Kavitha G***Department of Studies in Computer Science and Engineering U.B.D.T. College of Engineering, Davangere-04 Fifth International Conference on Information Processing, August-2011,Bangalore, INDIA
[5] BamshadMobasher "Data Mining for Web Personalization," LCNS, Springer-Verleg Berlin Heidelberg, 2007. [6] Catlegde L. and Pitkow J., "Characterising browsing behaviours in the world wide Web," Computer Networks and ISDN systems, 1995.
[6] Chungsheng Zhang and LiyanZhuang , "New Path Filling Method onData Preprocessing in Web Mining ," Computer and InformationScienceJournal , August 2008.
[7] Cyrus Shahabi, Amir M.Zarkessh, JafarAbidi and Vishal Shah "Knowledge discovery from users Web page navigation, " In.Workshop on Research Issues in Data Engineering,Birmingham,England,1997.
[8] Istvan K. Nagy and Csaba Gaspar-Papanek "User Behaviour Analysis Based on Time Spenton Web Pages,"Web Mining Applications in E-commercce and E-Services, Studies in Computational Intelligence, 2s009, Volume 172/2009, 117-136, DOI: 10.1007/978-3-540-88081-3_7 -Springer