Chapter 18

Improved Pre-Processing Stages in Web Usage Mining Using Web Log

J.Umarani[1] and S.Kavitha[2]

Abstract
Enormous growth in the web persists both in number of web sites and number of users. The growth generated large volume of data in during user's interaction with the web site and recorded in web logs. Web site owners need to understand about their users by accessing these web logs. Web mining perks up to comprehend range of concepts of diverse fields. Web Usage Mining (WUM) is the recent research field that it corresponds to the process of Knowledge Discovery in Databases (KDD). It comprises three main categories: Pre-Processing, Pattern Analysis, Pattern Discovery. WUM extracts behavioral data from web users data and if possible from web site information (structure and content). In this paper, we propose a customized application specific methodology for preprocessing the Web logs and combining WUM with Association Rule Mining.

Keywords: Steganography, Least Significant Bit, Data hiding, digital images.

Introduction

As in standard data mining, the aim in web mining is to determine and recover useful and attractive patterns from a huge dataset. There has been enormous interest towards web mining. In web mining, this dataset is the massive web data. Web data contains different kinds of information, including, web documents data, web structure data, web log data, and user profiles data. Two different approaches are projected on the definition of web mining. One approach is process-based and the other is data-based. Data-based definition is more widely accepted today. In this perspective, web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure or usage data is used in the mining process. There are no differences between web mining and data mining compared in general. All of web data can be mined mainly in three different dimensions, which are web content mining, web structure mining, and web usage mining.

[1]*Assistant Professor in Computer Applications, Thanthai Hans Roever College, (Autonomous), Perambalur, Tamilnadu, India.*
[2]*Assistant Professor in Computer Applications, D.G.Vaishnav College (Autonomous), Chennai, Tamilnadu, India.*

S. Vignesh and A. Philip Arokiadoss (ed.), *Statistical Approaches on Multidisciplinary Research*, Volume I

There are several reasons for the appearance of web mining [1]. First of all the World Wide Web is huge and efficient source for data mining and data ware housing. The size of the web is very large on the orders of terabytes and it is still growing rapidly. Many organizations, individuals or societies offer their public information through web. Also, the content of the web pages are much more composite than any other conventional text documents. Today, web pages lack standard structure; they contain more complex style than standardized formats. Web Mining can be broadly divided into three categories as shown in Fig 1 according to the kinds of data to be mined:-

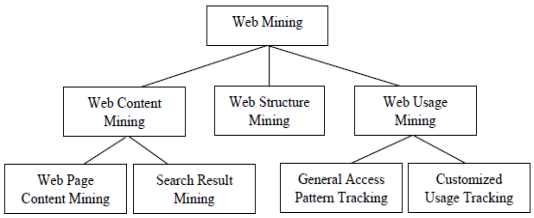

Fig 18.1 Taxonomy of Web mining

The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the user/consumers' needs. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the problems encountered on the Web. Therefore, Web mining becomes a popular active area and is taken as the research topic for this investigation. Web Usage Mining [2], [3] 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. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server. The remainder of this paper is organized as follows: Section 2 provides a brief review of Web Usage Mining. In Section 3, we explain the related work in this area. In Section 4, we introduce our proposed algorithm and an illustration of the algorithm. Finally, we concluded our work.

Web Usage Mining (WUM)

Web usage mining (WUM) [4] [5] can be defined as the application of data mining techniques to weblog data in order to discover user access pattern. Web usage mining has various application areas such as user behavior prediction, site-reorganization and web personalization. Web usage Mining comprises of three phases:

- Preprocessing
- Pattern discovery
- Pattern Analysis

The data stored in the log files don't present an accurate picture of the user's accesses to the web server [6]. Data preprocessing is the process to convert the raw data available in log files into the database tables for making it suitable for applying

the data mining algorithm. Hence preprocessing of web log data is most essential and a pre-requisite phase before it can be used for the pattern discovery task. Due to large amount of irrelevant entries in the web log file, the original log files cannot be directly used in the web usage mining process. Therefore the preprocessing of web log file becomes significant and important. A log file contains information related to the user queries on a website. Web usage mining may be used to improve the website structure or giving recommendations to visitors. The research on data preprocessing of Web Usage Mining is a field in focus nowadays.

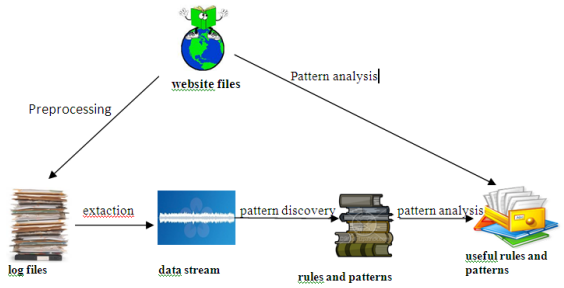

Fig 18.2 Web Usage Mining Process

Web Log File: An Overview

It is a simple text file which keeps record of the requests that are submitted by the user to the server while accessing a website. Key fields contained in log file are: username, IP address, user agent, date, time, number of bytes transferred etc.[7] Example of a log file entry stored on IIS server taken from www.uietkuk.org is shown in figure 3:

2013-03-10 12:57:40 208.91.198.202 GET /downloads.html- 80 - 37.228.106.69 HTTP/1.1 Opera/9.80 200 5444 31

Fig 18.3 sample log file entry

Various fields are explained below:

Date: The date of access is recorded for each hit. The format is YYYY MM-DD [8]. In fig. 2, date is 2013-03-10.
Time: It denotes the time at which access was made. The time format is HH:MM: SS. In fig. 2 time is 12:57:40.
Server IP: It is the server IP address. Its value in fig.3 is 208.91.198.202.
Server method: It is the method by which servers send information to the client. It can be: GET, POST, or Head. In fig. 3, GET method is used.
Uri-stem: URI-Stem is a path of the page accessed from the host. Its value in fig. 3 is /downloads.html.
Server Port number: It is a communication port used for transmission of data between client and server. Usually port number 80 is used.
Client IP: It is the IP address of the client. Its value in fig. 3 is 37.228.106.69.
Version and User-Agent: It denotes internet protocol/version and Browser type/version respectively. HTTP is the protocol with version 1.1and browser used is Opera with version 9.80 in the above figure.

Status: This indicates the status code returned by the web server to the client to indicate the success or failure of the transaction. 200 are returned by the server in above figure, denoting success.
Number of Bytes transferred: It indicates the number of bytes transferred by the server to the client. 5444 bytes are sent by the server.
Time taken: It indicates the time taken for the complete request and response cycle.

Types of Web Server Logs
Web server logs keeps track of pages visited by a user as well as details related to the accesses such as IP address of client, request date and time, page that is requested, HTTP code, number of bytes transferred, user -agent, referrer etc. This data can be saved into a single or multiple files. These files are usually not be easily retrieved by the general internet users. Different types of server log files are:

1. Access log: It keeps track of all the requests that are sent by the client and hence processed by the server [7]. Information about the user is then processed to determine user behavior [8]. Commonly indicated fields in this log are: Date and time of access, client IP, User authentication, Server name, its IP address and port [3]. Its format is same as in fig 2.

2. Error log: Whenever an error occurs in accessing a page, which is being requested by client to web server, the entry is made in this log [2]. For example, error log file records an entry when a user clicks on a particular link that does not locate the promised page or web site, as a result of which a message is displayed ‘‘Error 404 File Not Found’’. It is beneficial for the web page designers to optimize the links of web site [9]. Error code 404 is clearly shown in Figure 4 below.

Fig 18.4 Sample Error Log entry

3. Agent log file: It contains the information about user’s browser, its version and operating system and its version [7]. This information is used by the web site designer and administrator for the analysis of which specific browser is used by users. Also, this can be used to find out the most popular browsers and operating systems among the users. In figure 5 the browser used is Mozilla with version 5.0 and Operating system used is Windows 7.

Fig 18.5 Sample Agent Log Entry

4. Referrer log file: It contains the detail about the referrer. For example, as someone jumps from www.google.com to any site by clicking the link, referrer log file of that web server will record a referrer entry that a user came from www.google.com [9]. The Referrer Log tells what web sites link to a server [8]. In figure6 referrer is www.google.com.

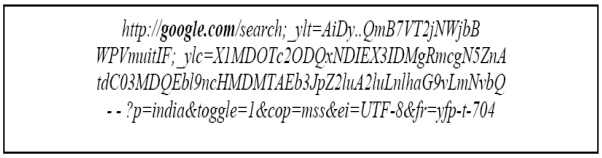

Fig 18.6 Sample Referrer log Entry

Related Work

In Web usage mining several data mining techniques can be used. Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between groups of users with specific interest. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together. Web usage mining is elaborated in many aspects. Besides applying data mining techniques also other approaches are used for discovering information. For example [10] has introduced a web usage mining intelligent system to provide taxonomy on user information based on transactional data by applying data mining algorithm, and also offers a public service which enables direct access of website functionalities to the third party.

Santosh Kumar et. al. [11] concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. The comparison of memory usage and time usage is compared using Apriori algorithm and Frequent Pattern Growth algorithm. *Patel et al [12]* discusses the process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented Web Usage Mining applications and some Web Mining software. *In paper [13]* provides an introduction to the field of Web mining and examines existing as well as potential Web mining applications applicable for different business function, like marketing, human resources, and fiscal administration. Suggestions for improving information technology infrastructure are made, which can help businesses interested in Web mining hit the ground running.

In [14] an overview of the web mining concept has been presented and how it can be useful and beneficial to the business improvement by facilitating its applications in various areas over the internet. The contribution of this paper is towards the various areas containing web sites on internet, which can make best use of different web mining techniques to improve their business decisions based on the user behavior analysis which can ultimately help in improving the relevance of their web site to suit their user needs and adding value to their business growth. *Kharwar et al [11]* implements the high level process of Web Usage Mining using basic Association Rules algorithm call Apriori Algorithm. Here, Web Usage Mining, approach has been combining with the basic Association Rules, Apriori Algorithm to optimize the content of the serve log data. Finally, this paper will present a finding association Rule from server log which are useful in many application like cache for web page, Marketing, Targeted Advertising etc. Sheetal A. Raiyani et. al. [15], proposed the algorithm called DUI (Distinct User Identification) as per author. It analyses more factors, such as user's IP address, Web site's topology, browser's edition, operating system and referrer page.

Wasavand et al. identified user's navigation pattern by data cleaning on web log file. They also used classification algorithms to identify user's interested website [16].

Manisha conducted data cleaning and distinct user identification technique which enhance the preprocessing steps of web log usage data. Using user identification

they found out the distinct user based on their attended session time. This will help in personalizing the websites. [17] Pankaj M. Meshram, Prof. Gauri A. Chaudhary used clustering techniques to complete the path and improve the websites performance. [18]. T. Vijay Kumar, H. S. Guruprasad, Bharath Kumar K. M., Irfan Baig, and Kiran Babu S introduced a new idea of incorporating available website knowledge for better session construction which would eventually lead to betterpatterns during pattern discovery. By using concept based approach they captured the actual intuition of the user which is sole purpose of any mining process. By identifying user's navigation between concepts, they have generated user profiles which will be useful for administrator to predict user behavior for a particular group of users. Recommendation models based only on usage information are inherently incomplete because they neglect domain knowledge. Field Extraction. [19]

Proposed System

The main goal of the proposed system is to identify usage pattern from web log files of a website. collections of items bought by customers, or details of a website frequentation).In this paper we proposed a new algorithm which combines the concept of association mining and Clustering instead of mining association rules from the web log data directly we have mined the clusters selected by user. Figure 7 represents proposed approach.

Algorithm Description

Input: A web log database.
Output: Frequent item sets
Method:

1) Scan the database D and partition the transaction table into clusters using K-means algorithm. Apply the method from step 2 to 6 on user selected cluster.
2) The set of frequent 1 item sets say L1, can then be determined. It consists of candidate 1 item Sets which satisfy minimum support
3) To discover the set of frequent 2- item sets
4) The algorithm iterates to find upto n- frequent item sets
5) From user selected cluster find out the n-frequent item sets.

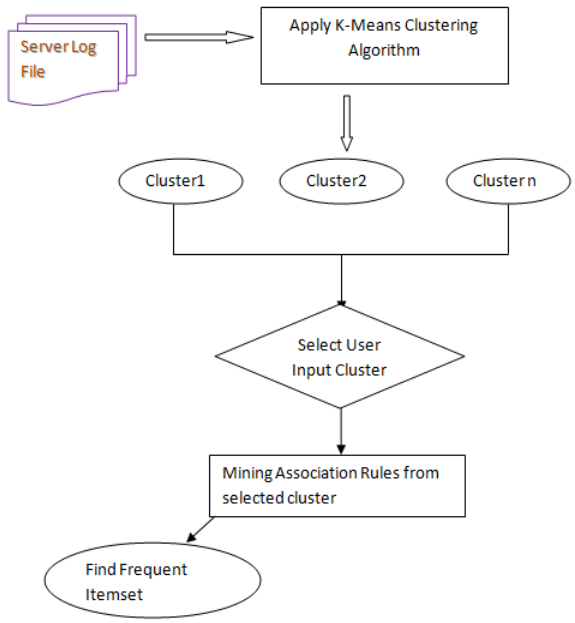

Fig 18.7 Proposed System

Clustering Algorithm

Clustering is a technique to search hidden patterns that exists in datasets. It is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the other clusters. A popular clustering method that minimizes the clustering error is the k-means algorithm. It partitions the input dataset into k clusters. First select k initial centers based on desired number of clusters. The user can specify k parameter value. Each data point is assigned to nearest centroid and the set of points assigned to the centroid is called a cluster. Each cluster centroid is updated based on the points assigned to the cluster. The process will be repeated until the centroids remain the same or no point changes clusters. In this algorithm mostly Euclidean distance is used to find distance between data points and centroid.

Algorithm:

The k-means clustering algorithm

Input: D:{d1,d2....dn}\\set of n items
K //Number of desired clusters

Output: A set of k-clusters.

Steps:

1. Arbitrarily choose k-data items from D as initial centroids;
2. **Repeat** assigns each item di to the cluster which has the closest centroid, Calculate new mean for each cluster;

until convergence criteria are met.

Association Rule Mining

Given a server log files that represent user activities, the main purpose of Association Rules is to generate all Association Rules that have support and confidence greater than the user specified minimum support (called min_sup) and minimum confidence (called min_conf) respectively. An algorithm for finding all Association Rules, henceforth, referred to as the Apriori algorithm. In Apriori

algorithm, discovery of association rules require repeated passes over the entire database to determine the commonly occurring set of data items. Therefore, if the size of disk and database is large, then the rate of input/output (I/O) overhead to scan the entire database may be very high. We have proposed a new Algorithm, which improves the Apriori algorithm for repeated scanning of large databases for frequent item sets generation. In our algorithm, transaction dataset will be used in the transposed form.

Algorithm:
Association Rule Mining for each cluster
1. Read the database to count the support of C1 to determine L1 using sum of rows.
2. L1= Frequent 1- itemsets and k: = 2
3. While (k-1 ≠ NULL set) do
Begin Ck: = Call Gen_candidate_itemsets (Lk-1)
Call Prune (Ck)
for all itemsets i l€ I do
Calculate the support values using dot-multiplication of array;
Lk : = All candidates in Ck with a minimum support;
k:=k+1
End
4. End of step-3
End Procedure
Procedure Gen_candidate_itemsets (Lk-1)
Ck = Φ
for all itemsets I1 € Lk-1 do
for all itemsets l2 € Lk-1 do
if I1[1] = I2 [1] ^ I1 [2] = I2 [2] ^ … ^ I1 [k-1] < I2 [k-1] then
c = I1 [1], I1 [2] … I1 [k-1], I2 [k-1]
Ck = Ck U{c}
End Procedure
Procedure Prune (Ck)
for all c € Ck
for all (k-1)-subsets d of c do
if d Lk-1 then
Ck = Ck – {c}
End Procedure

Conclusion
The use the internet has made repeated knowledge extraction from Web log files a necessity. Information provided are interested in techniques that could learn Web users' information needs and preferences. This can improve the effectiveness of their Web sites by adapting the information structure of the sites to the users' behavior. The aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. A way to evaluate the effectiveness of a Web site and its information access tools is through the mining of web log files. Proposed algorithm is used to generate association rules that associate the usage pattern of the clients for an website. In the proposed work we have combined the association mining with the clustering instead of mining association rules from the web log data directly we have mined the clusters. The goal of clustering is to organize data circulated over the Web into groups / collections in order to facilitate data availability

and accessing, and at the same time meet user preferences. Therefore, the main benefits include: increasing Web information accessibility, understanding user's navigation behavior, improving information retrieval and content delivery on the Web.

References

B.Naveena Devia, Y.Rama Devib, B.Padmaja Ranic, R.Rajeshwar Raod, Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce International Conference on Communication Technology and System Design 2011.

B.Santhosh Kumar,K.V.Rukmani, Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms Int. J. of Advanced Networking and Applications Volume:01, Issue:06, Pages: 400-404 (2010)

Han J., Pei J., Yin Y. and Mao R., "Mining frequent patterns without candidate generation: A frequent-pattern tree approach" Data Mining and Knowledge Discovery (2004).

Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang -Ning Tan."*Web usage Mining: Discovery and applications of usage patterns from web data*". ACM SIGKDD. Jan 2000.

Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques.

K Sudheer Reddy, Dr. G. Partha Saradhi Varma, Dr. I. Ramesh Babu. " Preprocessing the Web Server Logs – An illustrative approach for effective usage mining*".* ACM SIGSOFT. Volume 37 Number 3. May 2012.

Ketul B. Patel,Dr. A.R. Patel, Process of Web Usage Mining to find Interesting Patterns from Web Usage Data International Journal of Computers & Technology www.ijctonline.com ISSN: 2277-3061 Volume 3, No. 1, AUG, 2012.

Manisha V. "A Step up in Data Cleaning and User identification of Preprocessing on Web Usage data". International Journal of Advanced Research in Computer Engineering and Technology IJARCET, 2014

Morzy T, Wojcie M, and Zakrazewicz M. "Web Use Clustering" International Symposium On Computer and Information Sciences, 2000.

Pankaj, M. & Gauri, A. "Mining of Web Logs Using Preprocessing and Clustering". International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 12,December 2014.

Pradnya Purandare, WEB MINING: A KEY TO IMPROVE BUSINESS ON WEB IADIS European Conference Data Mining 2008.

Priyanka Patil and Ujwala Patil. "*Preprocessing of web server log file for web mining*". Proceedings of National Conference on Emerging Trends in Computer Technology. April 21, 2012.

Rahi, Priyanka. "Business Intelligence: A Rapidly Growing Option through Web Mining." arXiv preprint arXiv:1208.5875 (2012).

Robert F, Dell P, Roman E, and Juan D., "Web User Session Reconstruction Using Integer Programming," IEEE/ACM International Conference on Web Intelligence and Intelligent Agent, 2008.

Sheetal A. Raiyani, Shailendra Jain and Ashwin G. Raiyani, "Advanced Preprocessing using Distinct User Identification in web log usage data", ISSN : 2278 – 1021, IJARCCE, Vol. 1, Issue 6, August 2012.

Surbhi Anand, and Rinkle Rani Aggarwal,"An efficient Algortihm for Data Cleaning of Log file Using File Extensions,"*International Journal of Computer Application* (0975-88), 48(8): (2012).

Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood," *Web Usage Mining: A Survey on Preprocessing of Web Log File*", In the proceedings of International Conference on Information and Emerging Technologies, 2010

Vijay Kumar, T. & Guruprasad, H.S. & Bharath Kumar, K.M. &Irfan, B. & Kiran, B. "A New Web Usage Mining Approach for Website Recommendations Using Concept Hierarchy and Website Graph". International Journal of Computer and Electrical Engineering, Vol. 6, No. 1, February 2014.

Wasavand, C. & Devale, P.R & Ravindra, M. "Data PreprocessingMethod of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern". IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, 2014.