PATTERN DETECTION WITH IMPROVED PREPROCESSING IN WEB LOG

Priyanka Dubey*

Prof. Roshni Dubey**

ABSTRACT

The past fifteen years are characterized by an exponential growth of the Web both in the number of Web sites available and in the number of their users. This growth generated huge quantities of data related to the user's interaction with the Web sites, recorded in Web log files. Moreover, the Web sites owners expressed the need to better understand their visitors in order to better serve them. The Web Use Mining (WUM) is a rather recent research field and it corresponds to the process of knowledge discovery from databases (KDD) applied to the Web usage data. It comprises three main stages: the preprocessing of raw data, the discovery of schemas and the analysis (or interpretation) of results. A WUM process extracts behavioral patterns from the Web usage data and, if available, from the Web site information (structure and content) and on the Web site users (user profiles). In this thesis, we bring two significant contributions for a Web Use Mining process. We propose a customized application specific methodology for preprocessing the Web logs and a modified frequent pattern tree for the discovery of patterns efficiently.

* Department of Computer Sc. & Engineering, SRIT, RGPV University, Jabalpur, India

** Asst Prof, Department of Computer Sc. & Engineering, SRIT, RGPV University, Jabalpur, India

1. Introduction

The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view, users, web service providers, business analysts. The users want to have the effective search tools to find relevant information easily and precisely. 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 [4], [5] 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 process of Web usage mining also consists of three main steps: (i) pre-processing, (ii) pattern discovery and (iii) pattern analysis. In this work pattern discovery means applying the introduced frequent pattern discovery methods to the log data. For this reason the data have to be converted in the pre-processing phase such that the output of the conversion can be used as the input of the algorithms. Log files are stored on the server side, on the client side and on the proxy servers. Logs are processed in Common Log Format. Pattern analysis means understanding the results obtained by the algorithms and drawing conclusions. In pattern discovery phase methods and algorithms used have been developed from several fields such as statistics, machine learning, and databases. This phase of Web usage mining has three main operations of interest: association (i.e. which pages tend to be accessed together), clustering (i.e. finding groups of users, transactions, pages, etc.), and sequential analysis (the order in which web pages tend to be accessed). Pattern analysis is the last phase in the overall process of Web usage mining. In this phase the motivation is to filter out uninteresting rules or patterns found in the previous phase.

2. Introduction of WUM

Web Usage Mining can be used to make search relevant by determining frequent access behavior for users, needed links can be identified to improve the overall performance of future accesses. Web Usage mining has been defined as the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of Web based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Web Usage Mining may be applied to data such as contained in logs files. 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 [2]. The aim in web usage mining is to discover and retrieve useful and interesting patterns from a large dataset. In web mining, this dataset is the huge web data. Web data contains different kinds of information, including, web structure data, web log data, and user profiles data. 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. Web usage mining has various application areas such as web pre-fetching, link prediction, site reorganization and web personalization. Most important phases of web usage mining is discovering useful patterns from web log data by using pattern discovery techniques such as Apriori, FP-Growth algorithm[4]. Figure below represent WUM process.

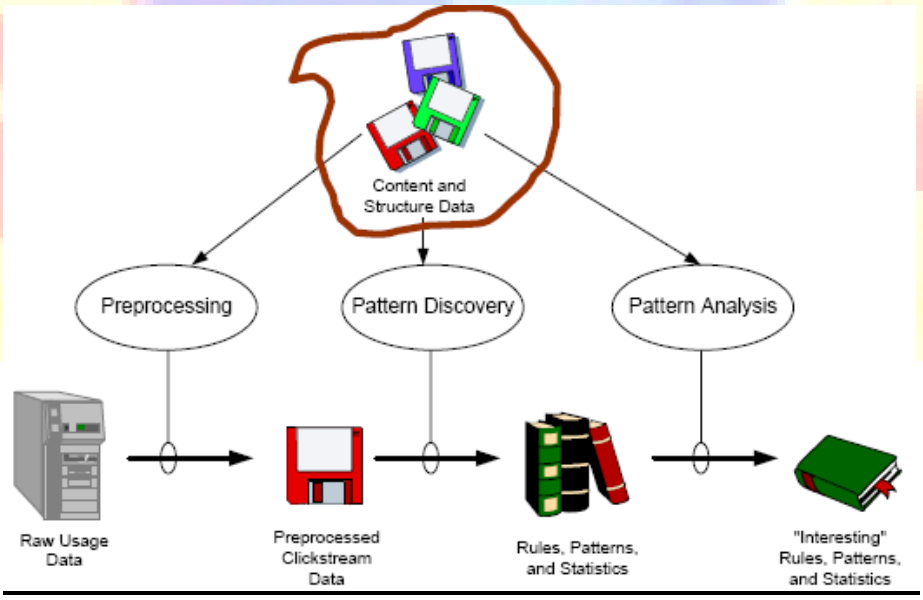

Figure 1: WUM Process

3. Introduction of WebLogs

Web log files are files that contain information about website users and are created by web servers automatically. Log file contain information about like user name, IP address, date, time, byte transfer, access request. Web log file is a simple plain text file which record information about each web users and display log files data in different formats.[5][6]

I. W3C extended log file format.

II. IIS log format.

III. NCSA common log format.

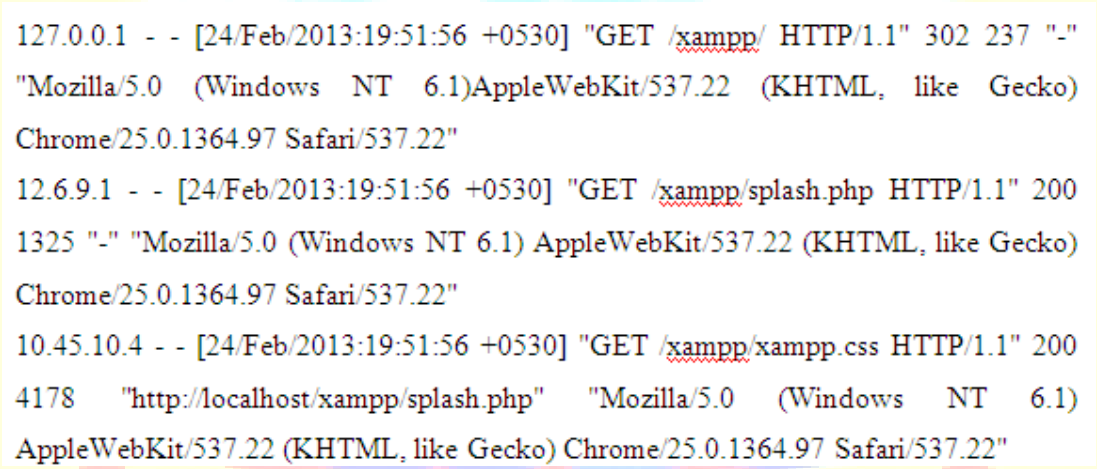

Figure 2: Sample Web Log

4. Web Log Preprocessing

The inputs to the pre-processing phase are the log and site files. The outputs are the user session file, transaction file Web servers register a Web log entry for every single access they get, in which important pieces of information about accessing are recorded, including the URL requested, the IP address from which the request originated, and a timestamp. A log file can be located in three different places-Web Servers, Web proxy Servers, and Client browsers [8]. Steps of Web log preprocessing are shown below. We have add op one more step before data cleaning that is cleaning according to criteria given by user as per need.

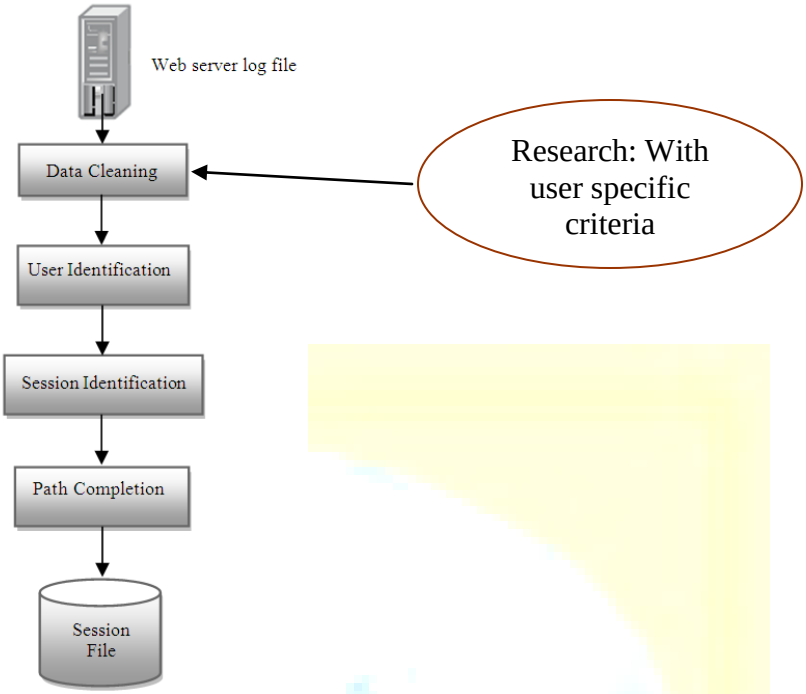

Figure 3: Web Log Preprocessing

We introduced one more step in traditional pre-processing steps, before data cleaning, Customization. In this step we clean log on the basis of user requirement for application. User selects choice of application for which he wants to perform usage mining according to that our pre-processing approach works.

Steps for proposed web log pre-processing are as follows:

1. Input raw web accessing log file.
2. Take user choice for normal, multimedia, graphics or e-commerce applications.
3. Read raw web log file and remove logs according to user selection to make intermediate file.
4. Identify users & resources uniquely & assign a unique id to them.
5. Identify resource accessed by users according to id.
6. Create preprocessed file by mapping of user id and resource ids accessed by them.

5. Pattern detection using Frequent Pattern tree

Apriori algorithm searches for large itemsets during its initial database pass and use its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent itemsets and those below are called small

itemsets. The algorithm is based on the large itemset property which states: Any subset of a large itemset is large and any subset of frequent item set must be frequent. The FP-tree method is efficient and scalable for mining both long and short frequent patterns and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent-pattern mining methods. The FP-tree algorithm avoids candidate generation steps [7]. The main idea of the algorithm is to maintain a frequent pattern tree (FP-Tree) of the database [5]. It is an extended prefix-tree structure, storing crucial quantitative information about frequent sets. The tree nodes are frequent items and are arranged in such a way that more frequently occurring nodes will have a better chances of sharing nodes than the less frequently occurring ones. The method starts from frequent 1-itemsets as an initial suffix pattern and examines only its conditional pattern base (a subset of the database), which consists of the set of frequent items co-occurring with the suffix pattern. The algorithm constructs the conditional FP-tree and performs mining on this tree. A hash-based technique is used to reduce the size of the candidate k- patterns. Another variation is to reduce the number of transactions to be scanned at higher values of k. Since a transaction that does not contain any frequent k-pattern cannot contain any frequent (k+1) - pattern, these types of transactions can be marked during the Kth scanning and are not considered in the subsequent scanning . All existing Algorithms have their own advantages and drawbacks. If all the transactions are different then Apriori algorithm is good otherwise FP-tree Algorithm is good. In FP-tree Algorithm, the tree will generate one branch for each

transaction. The tree size will become complex for storing in the memory. The FP-tree Algorithm will take more time for recursive calls in the algorithm. For generating Frequent Patterns the pointer have to traverse all the nodes. So it will take more time, so to remove these drawbacks we proposed a new method which is nonrecursive. Algorithm will take less execution time for access paths which are

not having uncommon items. The main idea of the algorithm is to maintain a frequent pattern tree of the database. It is an extended prefix-tree structure, storing crucial quantitative information about frequent patterns. This algorithm scans the data base once for generating page table. This table stores the information about web pages, the number of times the user accessed that web page and the pointer field that stores the reference of that webpage in the pattern base tree.

6. Conclusion

In order to make a website popular among its visitors, System administrator and web designer should try to increase its effectiveness because web pages are one of the most important advertisement tools in international market for business. The obtained results of the study can be used by system administrator or web designer and can arrange their system by determining occurred system errors, corrupted and broken links. In this study, analysis of web server log files of smart sync software has done by using web log expert program. Other web sites can be used for similar kind of studies to increase their effectiveness. With the growth of web-based applications our research work can be utilized in industry and application oriented system. Our proposed method of customized web log preprocessing, rather than traditional approach, may reduces size of raw web log file. Improved Frequent Pattern Tree structure uses iterative approach with support count restriction to reduce execution time and memory rather than traditional Frequent Pattern Tree. In future research, work is carried out for developing a web usage mining tool with customized web log preprocessing and combined pattern analysis approaches according to different application.

Acknowledgements

I feel pleasure in conveying my profound thanks to my supervisor Prof. Roshni Dubey, Computer Science & Engineering, SRIT, Jabalpur, for his constant Support, valuable Guidance and Encouragement. I also thankful to all staff members of Computer Engg for their time to time supports. I am also very thankful to my family members.

References

[1] Rahul Mishra, Abha Choubey "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining ", International Journal of Advanced Research in Computer Science and Software Engineering 2012

[2] Sanjay Kumar Malik, Nupur Prakash, S.A.M. Rizvi" Ontology and Web Usage Mining towards an Intelligent Web focusing web logs" 2010 International Conference .

[3] Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei "Web usage mining based on WAN users' behaviours" 2010 International Conference.

[4] 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.

[5] 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), Vol-48, No-8, Jan 2012.

[6] T. Subha Mastan Rao, Thinley Lhendup, Thinley Wangdi, Sujata Pradhan, "Hybrid Model for Preprocessing and Clustering of Web Server Log" 2013

[7] J. Han, J. Pei, and Y. Yin. "Mining frequent patterns with out candidate generation". IEEE, Sept.1998 pp-365-378.