ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

An Empirical Proposal towards the Algorithmic Approach and Pattern in Web Mining for Assorted Applications

Harleen Puri[1], Arvind Selwal[2], Anuradha Sharma[3]

M. Tech Scholar, Dept. of CSE, Ambala College of Engineering and Applied Research, Devsthali, Ambala, India[1]

Associate Professor,Dept. of CSE, Ambala College of Engineering and Applied Research, Devsthali, Ambala, India[2]

Assistant Professor, Dept. of CSE, Ambala College of Engineering and Applied Research, Devsthali, Ambala, India[3]

ABSTRACT: Data mining or the analysis phase of the knowledge discovery process is the computational process of discovering patterns in large data sets that involves methods at the intersection of artificial intelligence, machine learning, statistics, and database system. The classical goal of the data mining and machine learning process is to fetch and extract information from a data set and transform it into an understandable structure for further use. Besides raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Web Usage Mining is the type of data mining technique to discover interesting usage patterns from web data, in order to discover useful pattern 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 may 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. Some of the typical data collected and saved at a web server include IP addresses, page references, and access time of the users. In this paper a new technique is proposed to discover the web usage patterns of websites from the server log files with the foundation of clustering and improved Apriori algorithm.

Keywords: Apriori algorithm, association rule mining, clustering, rule learning, web server log data, web usage mining

I. INTRODUCTION

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. It is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. It also allows looking for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and web spiders. Structure mining is used to examine data related to the structure of a particular Web site and Web Usage Mining is applied to many real world problems to discover interesting user navigation patterns for improvement of web site design by making additional topic or recommendations observing user or customer behaviour [7].

Web Usage Mining 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. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users and is the main input to the present Research. This work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files [6].

II. LITERATURE REVIEW

B.Santhosh Kumar et al implements three phases of Web usage mining namely preprocessing, pattern discovery, and pattern analysis. Apriori algorithm is used to generate an association rule that associates the usage pattern of the clients for a particular website. The output of the system was in terms of memory usage and speed of producing association rules. Pooja Sharma et al proposed a clustering algorithm to find out data clusters for both numerical and nominal data by calculating the average and log values of data set. This algorithm improves the techniques of Web Usage Mining by first discover the log files of individual users at one place. Martinez-Romo et al have analyzed different information

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

retrieval methods for both, the selection of terms used to construct the queries submitted to the search engine, and the ranking of the candidate pages that it provides, in order to help the user to find the best replacement for a broken link. To test the sources, they have also defined an evaluation methodology which does not require the user judgments, what increases the objectivity of the results. MahendraPratap Singh Dohare et al proposed a new reactive session reconstruction method. This algorithm is better than previously developed both time and navigation oriented heuristics as it does not allow page sequences with any unrelated consecutive requests to be in the same session. They have also implemented agent simulator for generating real user sessions. Resul Das et al analyzed the web server user access logs of Firat University to help system administrator and Web designer to improve their system by determining occurred system errors, corrupted and broken links by using web using mining. PriyankaPatil et al have focused on web log file format, its type and location. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log file. Data preprocessing is an important step to filter and organize appropriate information before using to web mining algorithm. They have also proposed two algorithms for field extraction and data cleaning. Preprocessing web log file is used in data mining techniques, also used in intrusion detection system as input to detect intrusion.

III. WEB USAGE MINING PROCESS

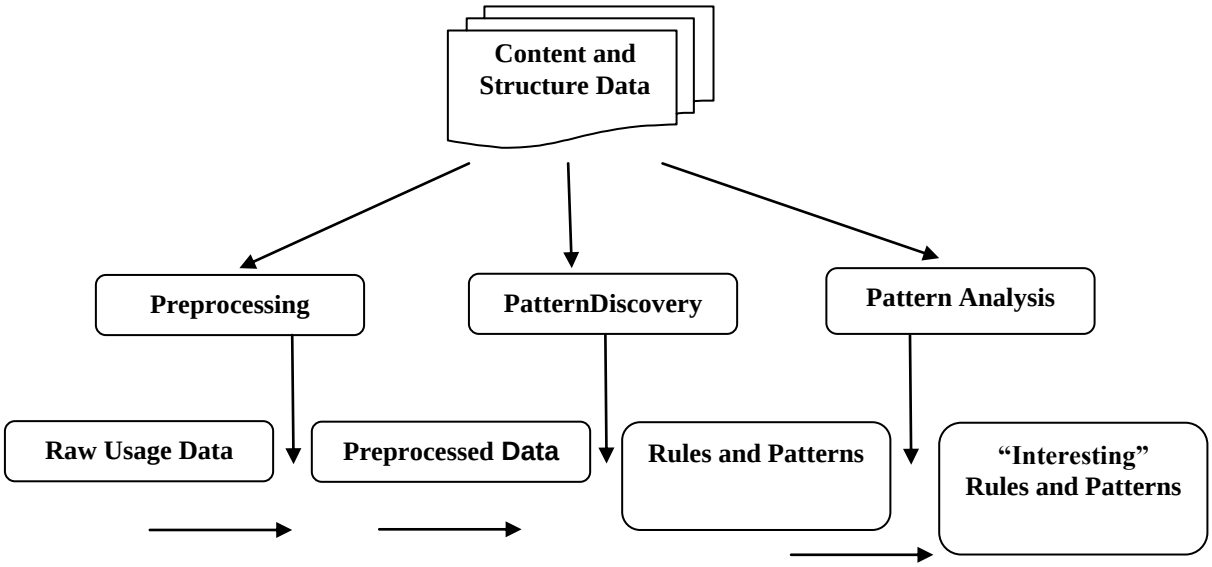

Fig. 1Web usage mining process

The server log consists of several attributes. The attributes are as follows:-

1. *Date:* The date from Greenwich Mean Time (GMT x 100) is recorded for each hit. The date format is YYYY-MM-DD [1].
2. *Time:* Time of transactions. The time format is HH:MM: SS [1].
3. Client IP Address: Client IP is the number of computer who access or request the site [1].
4. *User Authentication:* Some web sites are set up with a security feature that requires a user to enter username and password. Once a user logs on to a Website, that user's "username" is logged in the log file [1].
5. *Server IP Address:* Server IP is a static IP provided by Internet Service Provider. This IP will be a reference for access the information from the server [1].
6. *Server Port:* Server Port is a port used for data transmission. Usually, the port used is port 80 [1].
7. *Server Method (HTTP Request):* The word request refers to an image, movie, sound, pdf, .txt, HTML file and more [1].
8. *URL:* URL is path from the host. It represents the structure of the websites. For examples:/tutor/images/icons/fold.gif [1].
9. *Agent Log:* The Agent Log provides data on a user's browser, browser version, and operating system. This is the significant information, as the type of browser and operating system determines what a user is able to access on a site [1].

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

A. Pattern Discovery and Pattern Analysis

The three main stages of web usage mining are data preprocessing, pattern discovery and pattern analysis. Data preprocessing involves removal of unnecessary data. Pattern discovery data mining techniques are used in order to extract patterns of usage from Web data. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the web access log. Pattern Analysis is the final stage of the Web usage mining. The aim of this process is to extract the interesting rules or patterns from the output of the pattern discovery process by eliminating the irrelative rules or patterns [1].

IV. PROBLEM STATEMENT

The classical classical Apriori Algorithm makes use of bottom up approach for pattern extract and rule mining whereby the frequent subsets are extended one item at a time. This step is known as candidate generation and finally groups of candidates are tested against the data. This algorithm terminates as soon as no further successful extensions are extracted. The proposed technique will overcome the problems and complexities associated with the apriori algorithm by using specialized search and navigation based on the top down approach.

V. PROPOSED TECHNIQUE AND THE FLOW

1. An effective web usage mining algorithm shall be designed with the foundation of clustering and improved apriori algorithm.
2. The algorithmic approach shall be applied on the server log files for analysis and reports generation based on the usage patterns in the log files.
3. The log files and the results obtained will be used as a forensic database as well as associative rule mining.

With the advancements in the information and communications technology, the research is going on in the stream of data mining and machine learning. Thus, novel and efficient methods are required to mine the knowledge from large and huge sets of databases. Data mining refers to the extraction of core from the knowledge discovery in databases. It is the procedure for finding the useful and potential knowledge in database. Association rules are associated with the prominent knowledge of data mining and results that can be defined as the relations and dependency between the data items with the usage of support andconfidence. In the existing algorithms of the association rules mining and machine learning, apriori is the ancestor that was offered and discovered in 1993. The core idea of the apriori is scanning the database repeatedly. With the paradigm that the subset of the frequent data items are frequent patterns that can be gained with the length of frequent (k+1)-itemsets L_{k+1} from the frequent k-itemsets L_k. At the k time it scans the database only the candidate items C_{k+1} that generates from the L_k was concerned. Further, the appearance time of the C_{k+1} can be verified by another scanning database. There are lots of improved algorithm for apriori such as AprioriTID, Apriori Hybri, Multiple oins, Reorder and Direct etc. The main idea of these algorithms is according the theory that the subset of frequent items is a frequent set and the superset of an infrequent set is an infrequent itemset. These are used to scan the database repeatedly for mining the association rules.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

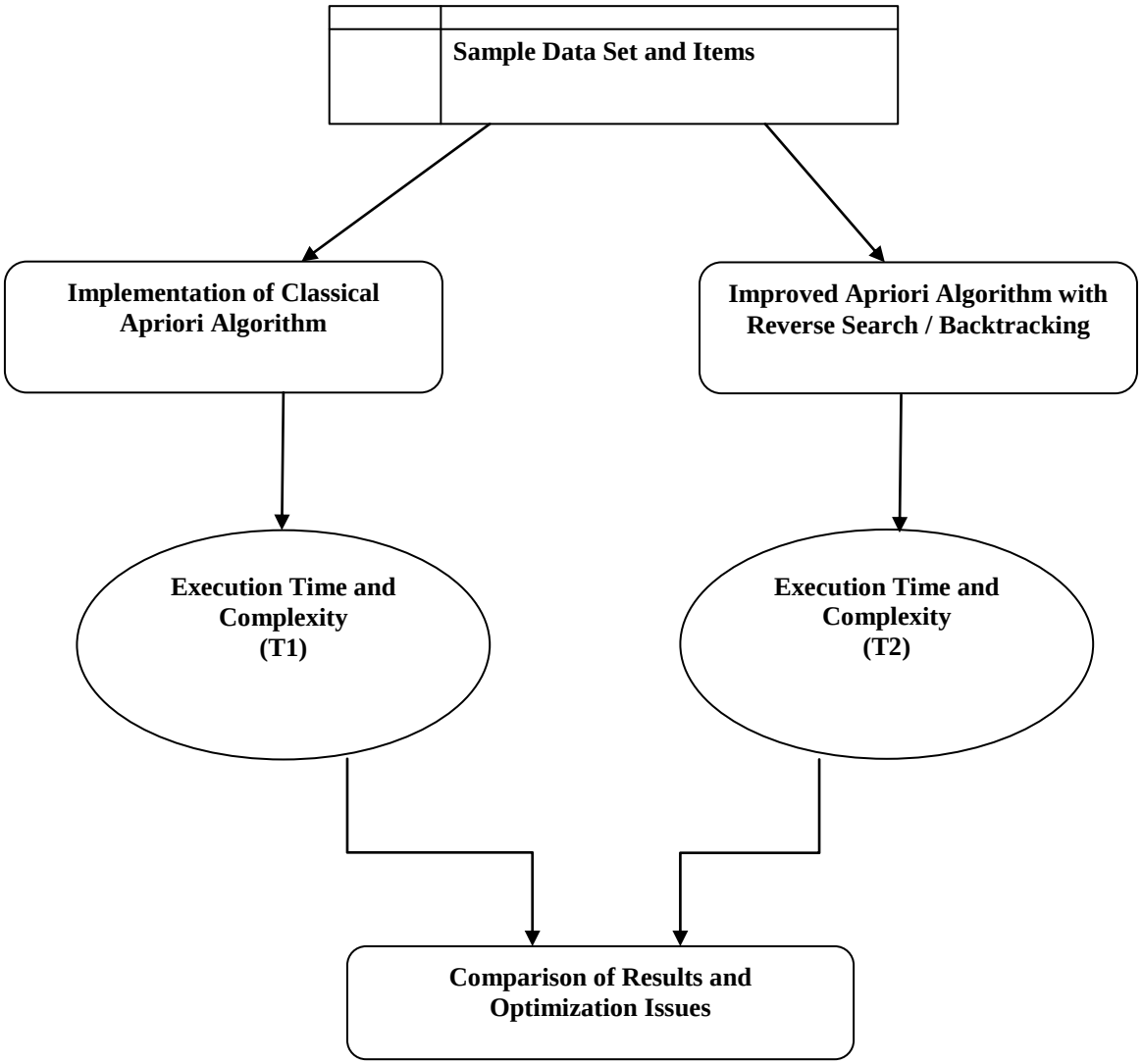

Fig.1: Proposed implementation of the Apriori Algorithm and comparison flow

VI. ANTICIPATED TECHNICAL IMPLEMENTATION AND APPROACH

1. The proportional analysis on various association rule mining using on the clustering and its applications in the web log files
2. Design of a new algorithmic approach towards apriori algorithm in log files analysis and forensic information
3. Implementation of the improved and efficient algorithm on the sample log record fetched from the live server
4. Relative study of the results with the existing techniques and generation of the detailed reports
5. Acceptance and trust level of the hypothesis and objectives specified in the research proposal
6. Framing out the conclusion and future work from the implementation performed and results fetched

VII. CONCLUSION

Web usage mining is 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. One of the algorithms which is very simple to use and easy to implement is the Apriori algorithm. In this paper a new technique is proposed to discover the web usage patterns of websites from the server log files with the foundation of clustering and improved Apriori algorithm. The effective algorithm will be proposed with the improvements as well as the implementation of Apriori Algorithm. The forthcoming step in the research work shall be to design the improved version of the Apriori Algorithm that shall be implemented on the Server Log Files for Association Rule Mining.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

REFERENCES

[1]. Mishra Rahul, ChoubeyAbha, "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, Vol.2, pp.311-318,2012.
[2]. PatilPriyanka, PatilUjwala, "Preprocessing of Web Server Log File for Web Mining", World Journal of Science and Technology, pp.14-18,2012.
[3]. DohareMahendraPratap Singh, Arya Premnarayan, Bajpai Aruna, "Novel Web Usage Mining for Web Mining Techniques", International Journal of Emerging Technology and Advanced Engineering, Vol.2, Issue 1, pp.253-262, 2012.
[4]. Sharma Pooja, BhartiyaRupali, "An efficient Algorithm for Improved Web Usage Mining", International Journal of Computer Technology and Applications, Vol.3 (2), pp.766-769, 2011.
[5]. Romo Juan Martinez, Araujo Lourdes, "Analyzing Information Retrieval Methods to Recover Broken Web Links", ECIR, pp.26-37,2010.
[6]. Kumar B. Santhosh, Rukmani K. V., "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms", International Journal of Advanced Networking and Applications, Vol.1, Issue 6, pp.400-404, 2010.
[7]. Suneetha K. R., Krishnamoorthi, R., "Identifying User Behavior by Analyzing Web Server Access Log File", International Journal of Computer Science and Network Security, Vol.9 No. 4, pp.327-332, 2009.
[8]. Das Resul, Turkoglu Ibrahim, Poyraz Mustafa, "Analyzing of System Errors for Increasing A Web Server Performance by Using Web Usage Mining", Journal of Electrical and Electronics Engineering, Vol.7 No. 2, pp.379-386, 2007.
[9]. Pingxiang Li, Jiangping Chen, FulingBian, "A Developed Algorithm Of Apriori Based On Association Analysis", Geo-spatial Information Science (Quarterly), Vol.7, Issue 2, pp.108-112, 2004.