ISSN (Online) : 2249 - 054X

International Journal of Computing and Corporate Research

Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners

http://www.ijccr.com

International Manuscript ID : ISSN2249054X-V3I2M3-032013

VOLUME 3 ISSUE 2 March 2013

A PRAGMATIC ALGORITHMIC APPROACH AND PROPOSAL FOR WEB MINING

Pooja Rani

M.Tech. Scholar

Patiala Institute of Engineering and Technology

Punjab, India

Abstract

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. 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 : Web Mining, Knowledge Discovery, Apriori Algorithm, Server Log Files

INTRODUCTION

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V3I2M3-032013
VOLUME 3 ISSUE 2 March 2013

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].

LITERATURE SURVEY

B.Santhosh Kumar et al [6] 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.

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V3I2M3-032013
VOLUME 3 ISSUE 2 March 2013

The output of the system was in terms of memory usage and speed of producing association rules.

Pooja Sharma et al [4] 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 [5] have analyzed different information 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.

Mahendra Pratap Singh Dohare et al [3] 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 [8] 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.

Priyanka Patil et al [2] 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

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V3I2M3-032013
VOLUME 3 ISSUE 2 March 2013

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.

WEB USAGE MINING PROCESS

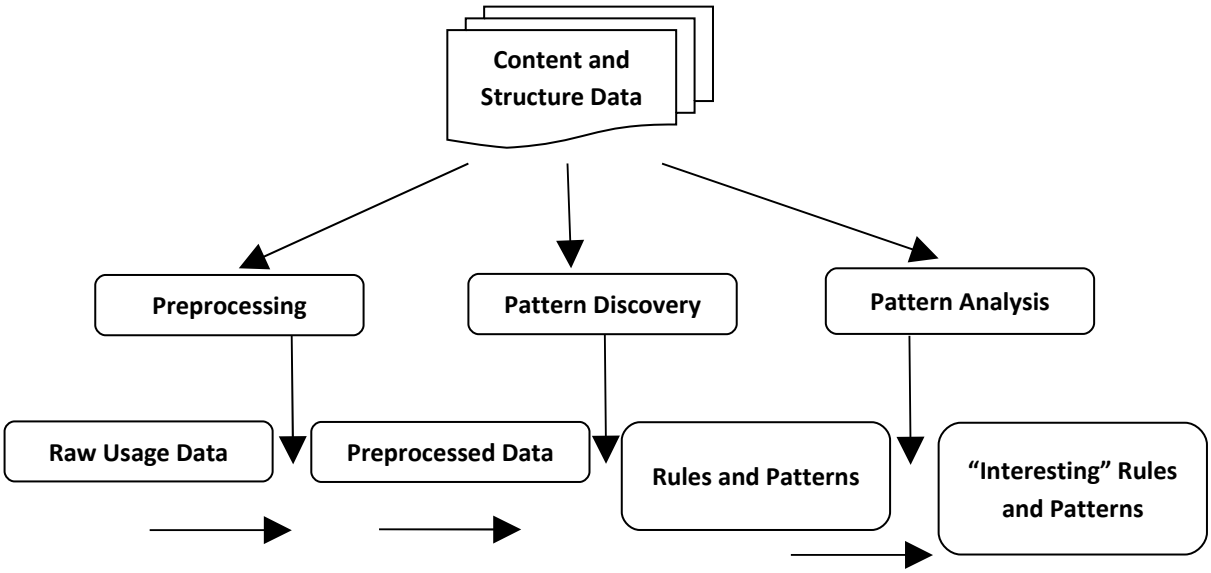

Figure 1 : Web Usage Mining Process

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].

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V3I2M3-032013
VOLUME 3 ISSUE 2 March 2013

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) URI: URI 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].

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

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers
International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V3I2M3-032013
VOLUME 3 ISSUE 2 March 2013

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].

PROPOSED TECHNIQUE AND THE FLOW

- An Effective Web Usage Mining Algorithm shall be designed with the foundation of Clustering and Improved Apriori Algorithm
- 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.
- The Log Files and the Results Obtained will be used as a Forensic Database as well as Associative Rule Mining

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

ISSN (Online) : 2249 - 054X
International Journal of Computing and Corporate Research
Multi Disciplinary Journal for Publication of Review and Research Papers

International Refereed and Indexed Journal for Research Scholars and Practitioners
http://www.ijccr.com
International Manuscript ID : ISSN2249054X-V3I2M3-032013
VOLUME 3 ISSUE 2 March 2013

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.

REFERENCES

[1] Rahul Mishra, Abha Choubey, 2012. Discovery of Frequent Patterns from Web Log Data by using FP-Growth Algorithm for Web Usage Mining

[2] Priyanka Patil, Ujwala Patil, 2012. Preprocessing of Web Server Log File for Web Mining

[3] Mahendra Pratap Singh Dohare, Premnarayan Arya, Aruna Bajpai, 2012. Novel Web Usage Mining for Web Mining Techniques

[4] Pooja Sharma, Rupali Bhartiya, 2011. An efficient Algorithm for Improved Web Usage Mining

[5] Juan Martinez-Romo, Lourdes Araujo, 2010. Analyzing Information Retrieval Methods to Recover Broken Web Links

[6] B. Santhosh Kumar, K.V.Rukmani, 2010. Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms

[7] K.R.Suneetha, Dr. R. Krishnamoorthi, 2009. Identifying User Behavior by Analyzing Web Server Access Log File

[8] Resul DAS, Ibrahim TURKOGLU, Mustafa POYRAZ, 2007. Analyzing of System Errors for Increasing A Web Server Performance by Using Web Usage Mining