THE SURVEY REPORT ON INTEGRATION BETWEEN WEB USAGE MINING AND DATA MINING TECHNIQUES

Zalak Prajapati[1] , Prof. Riya Parmar[2]

[1] *Research Scholar, CE-LDRP-ITR, KSV University Gandhinagar, Gujarat, India*
[2] *Lecturer, CE-LDRP-ITR, KSV University Gandhinagar, Gujarat, India*

ABSTRACT

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Web Usage Mining can be described as the discovery and analysis of user accessibility pattern, during the mining of log files and associated data from a particular Web site, in order to realize and better serve the needs of Web-based applications. Web usage mining itself can be categorized further depending on the kind of usage data considered they are web server, application server and application level data. Discovering hidden information from Web log data is called Web usage mining. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behaviour of the users. This can be used for advertising purposes, for creating dynamic user profiles etc.

Keyword: - *Web usage mining, Apriori algorithm, improved Frequent Pattern Tree algorithm, Web log mining*

1. INTRODUCTION

The Web is a vast, volatile, diverse, dynamic and mostly amorphous data repository, which stores incredible amount of information/data, and also enhance the complexity of how to deal with the information from the different opinion of view, users, web service providers and business analyst. The users wish for the effective search tools/engine to locate related information easily and accurately [1].Web usage mining is the process of finding out what users are looking for on the internet. Few users might be looking at only documented data, whereas some others might be interested in multimedia data. It is the submission of facts and figures mining techniques to find out interesting usage patterns from World Wide Web facts and figures in alignment to realize and better serve the desires of Web颅 based applications[3].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..[5]

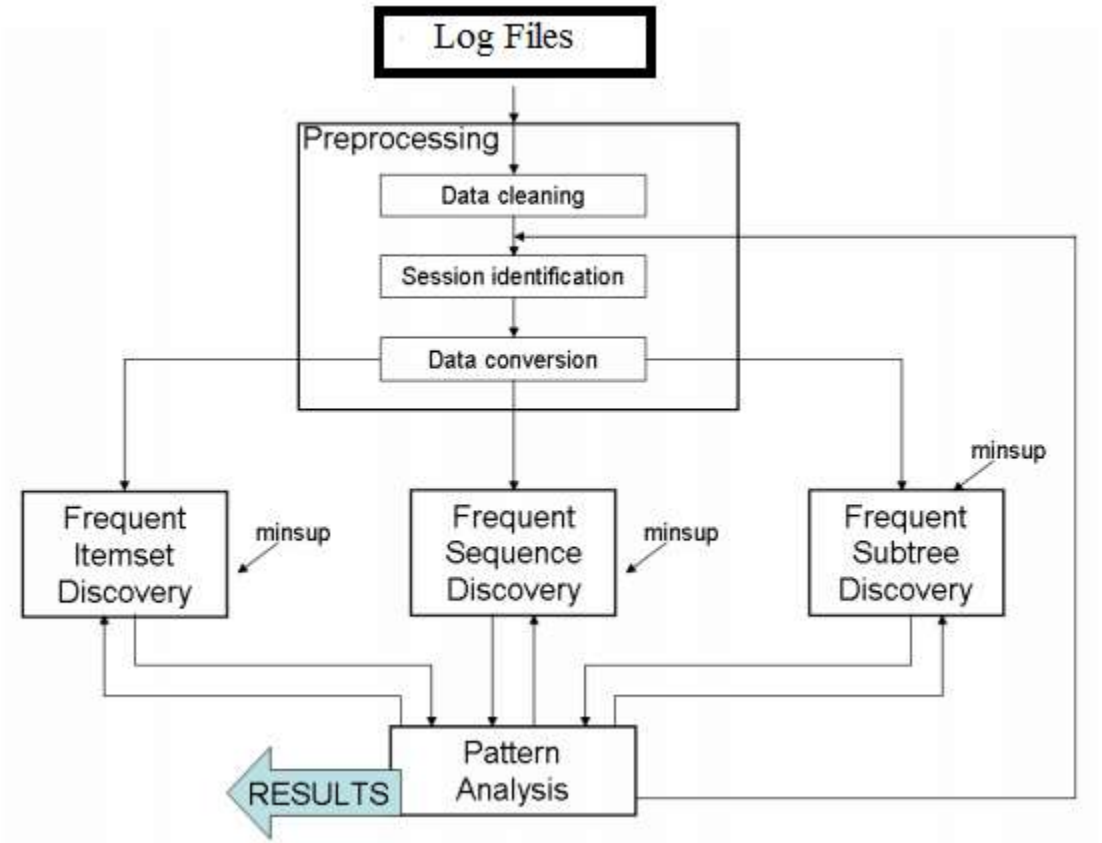

Figure 1: Process of web usage mining [4]

Figure 1 shows the process of Web usage mining realized as a case study in this work. As can be seen, the input of the process is the log data. The data has to be preprocessed in order to have the appropriate input for the mining algorithms. The different methods need different input formats, thus the preprocessing phase can provide three types of output data[4].

2. 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 group of users with specific interest [5].Application Level Data: New types of events can be characterized in an application, and logging can be twisted on for them therefore generating histories of these particularly characterized events. It should be noted however, that numerous end submissions need a combination of one or more of the methods directed in the classes above.[7]

3. PROPOSED ALGORITHM

3.1 APRIORY ALGORITHM

- PREPROCESSING MODULE
- APRIORI OR FP GROWTH ALGORITHM MODULE
- ASSOCIATION RULE GENERATION
- RESULTS

The current Research work is planned to work on log files. Apriori is a typical algorithm for frequent item set mining and association rule learning over transactional databases. It is proceed by recognize the frequent individual items in the database and extend them to big and big item sets as long as those item sets appear sufficiently often in the database [5].The frequent item sets find out by Apriori can be used to find out association rules which highlight general trends in the database: this has applications in domains such as market basket analysis[7].

3.2 FP-TREE ALGORITHM

FP tree is a solid data architecture that retained important, absolutely vital and quantitative intonations considering common patterns [8].

The main attributes of Frequent Pattern tree are:

- It comprises of one root marked as "root", a set of piece prefix sub-trees as the child of the root, and a frequent-item header chart.
- one-by-one node in the piece prefix sub-tree comprises of three areas

4. CONCLUSIONS

Web usage mining is the procedure of finding out which users are looking for the internet. It can be described as the sighting and scrutiny of user ease of access pattern, during mining of files and its connected data from a Web site. The main drawback of Apriori algorithm is that the candidate set creation is costly, especially if a large number of patterns and/or long patterns exist. The main drawback of FP-growth algorithm is the explosive quantity of lacks a good candidate generation method. Future research can combine FP-Tree with Apriori candidate generation method to solve the disadvantages of both apriori and FP-growth. In future the algorithm can be extended to web content mining, web structure mining, etc. The work can also be extended to extract information from image files.

6. REFERENCES

[1] K .S .K .D. Association Rules Mining: A Recent Overview, GTS International Tran on Computer Science, Vol.65 (1), 2006,pp.45-65

[2]A R "Fast Algorithms for Mining Association Rules", Sep12-15 1994, Chile, 487-99, pdf, 1-55860- 153-9.

[3]Mannila H,"Efficient algorithms for discovering
association rules mining." conference Knowledge Discovery in Databases (SIGKDD). 181-83.

[4]Tan, P. N., M. St., V. Kumar, "Introduction to web Mining",Addison-Wesley, 2013, 769pp.

[5]I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation, 2nd ed. San Mateo.

[6] P. Becuzzi, M. Coppola, and M. Vanneschi, ìMining of Association Rules in Very Large Databases: A Structured Parallel Approach," Proc. Europar-99, vol. 1685, pp. 1441-1450, Aug. 1999.

[7] R. Jin and G. Agrawal, "An Efficient Implementation of Apriori Association Mining on Cluster of SMPs," Proc. Workshop High Performance Data Mining (IPDPS 2001), Apr. 2001.

[8] J. Han and M. Kamber, "Data Mining: Concepts and Techniques" .Morgan Kaufmann Publishers, 2000.

[9] E.-H. Han, G. Karypis, and V. Kumar, "Scalable Parallel Data mining for Association Rules," Proc. ACM SIGMOD 1997, May 1997.
[10] E-H. Han, G. Karypis, and V. Kumar, "Scalable Parallel Data mining for Association Rules," IEEE Trans. Data and Knowledge Eng.,vol. 12, no. 3, May/June 2000.

BIOGRAPHIES

	**Zalak Prajapati** is a Student in the Master of Computer Engineering Department. She Studying Master of Computer Engineering (ME-CE) degree from LDRP-ITR,KSV University,Gandhinagar,Gujarat,India. Her research interests are Data Mining and Web Usage Mining
	Prof. Riya Parmar is Lecturer in Computer Department in LDRP-ITR,KSV University,Gandhingar,Gujarat,India. Her Research interest are Data mining,Computer Network