AN OVERVIEW ON RARE ITEMSETS MINING IN WEBLOG DATA

B.Amsapriya[1], S. Gunasekaran [2],

[1]*M.Phil Research Scholar*
[2] *Asst. Professor*
Department of Computer Science,
Thanthai Hans Roever College, Perambalur
Bharathidasan University, Trichy.

Abstract: *With the fast also volatile growth of data available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by web servers. Weblog is unstructured data and therefore mining rare information from weblog is a very challenging task. The weblog is unformed data and contains information about User Name, IP Address, Time Stamp, Access-Request, number of Bytes Transferred, etc. The log files are maintained by the web servers. These weblog files give details about the user. Frequent itemset mining in weblog data is a heavily researched area in the field of data mining with wide range of applications but rare Itemset mining differs from frequent itemset mining where it locates the uninteresting patterns, i.e., it detects the data items that arise very rarely. This paper presents a literature review on different techniques for mining rare (ie) infrequent itemsets in weblog data.*

Keywords: **Infrequent, Web log mining, Itemsets, Data Mining**

I. INTRODUCTION

Data mining is the discovery of hidden information found in databases and can be viewed as a step in the knowledge discovery process. Data mining functions include clustering, classification, prediction, and link analysis (associations). One of the most important data mining applications is that of mining web data. This mining is helpful for analyzing customer behavior in retail trade, banking system etc., Web data mining is an emerging research area where mining data is an important task and various algorithms has been proposed in order to solve the various issues related to the web mining in existing dataset [1].

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. [2] It is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. 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 behavior. 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. There are three general classes of information that can be discovered by web mining:

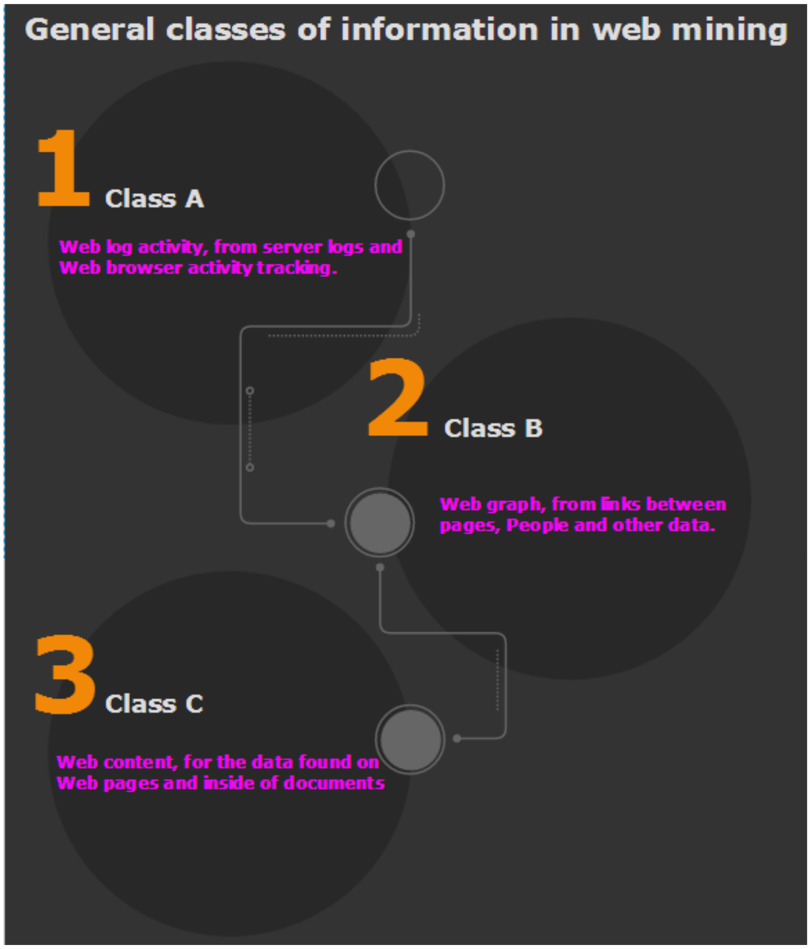
General classes of information in web mining
1 Class A
Web log activity, from server logs and Web browser activity tracking.
2 Class B
Web graph, from links between pages, People and other data.
3 Class C
Web content, for the data found on Web pages and inside of documents

Figure1. General Classes of information in web mining

Rare itemset mining is the process of mining data in a set of items. The resulted infrequent set data supports the minimum support threshold. An infrequent pattern is a pattern that occurs rarely in a web log dataset. Some infrequent itemset may also suggest the occurrence of interesting rare events. For example, if {Fire = Yes} is frequent other than {Fire = Yes, Alarm = On} is infrequent, then second one is an interesting infrequent pattern since it can indicate faulty alarm systems. To find such infrequent situations, the expected support of a pattern must be determined, so that, if a pattern turns out to have a considerably lower support than expected, it is declared as an interesting infrequent item [3]. Infrequent itemset deserve special attention because they represent major difficulties for data mining algorithms. This paper analyzes the rare Itemset mining for weblog data algorithm to conquer such problem.

II. Literature Review

Adda et al. [4] approach and algorithm name as Apriori for Infrequent And Non-present Item-set Mining (ARANIM), ARANIM discover of non-present patterns and infrequent patterns using infrequent item-set mining and they have also proposed a framework to represent the different categories of patterns based on the frequency constraint which by means of an instantiation process leads to the representation of frequent, infrequent and non-present pattern mining problems.

Troiano and Scibelli [5] propose Rarity, a top-down breadth-first level-wise algorithm; they explore the power set lattice from the top, reaching the border line of non-infrequent itemsets, this approach is applied in Rarity.

R. Agrawal et al [6]. proposed the Apriori algorithm in association rule mining, to identify the frequent itemsets in the large transactional database. Apriori works in two phases. During the first phase it generates all possible Itemsets combinations. These combinations will act as possible candidates. The candidates will be used in subsequent phases. In Apriori algorithm, first the minimum support is applied to find all frequent itemsets in a database and Second, these frequent itemsets and the minimum confidence constraint are used to form rules. The main drawback of Apriori is the generation of large number of candidate sets. The efficiency of apriori can be improved by Monotoni city property, hash based technique, Partioning methods and so on.

Luca Cagliero et al [7]. developed a frequent pattern growth algorithm. The drawback of Apriori can be improved by this algorithm. It is implemented without generating the candidate sets. This algorithm proposes a tree structure called FP tree structure, going to collect information from the database and creates an optimized data structure as Conditional pattern. Initially it Scans the transaction database DB once and Collects the set of frequent items F and their supports and then Sort the frequent itemsets in descending order as L, based on the support count. This algorithm reduces the number of candidate set generation, number of transactions, number of comparisons.

Sakthi Nathiarasan et al [8]. proposed the Minimal Infrequent Weighted Itemsets Miner. The MIWI Mining procedure is similar to IWI Mining. However, since MIWI Miner focuses on generating only minimal infrequent patterns, the recursive extraction in the MIWI Mining procedure is stopped as soon as an infrequent itemset occurs. It finds both the infrequent itemsets and minimal infrequent itemset mining. The advantage of MIWI algorithm is reduction in generating the candidate sets, reducing the computational Time, improved the efficiency of algorithm performance compared to FP-Growth algorithm.

FP-Growth algorithm can be improved by Broglet’s FP growth algorithm [9]. Initially it scans the frequencies of the items and all infrequent items, that is, all items that appear in fewer transactions than a user-specified minimum number are discarded from the transactions, since, they can never be part of a frequent item set. The items in each transaction are sorted, so that they are in descending order with respect to their frequency in the database. It reduces the computational cost in FP-Growth.

IFP min algorithm [10] that uses a recursive approach to mine minimally infrequent Item sets (MIIs). The infrequent item sets are then reported and it gets pruned from the database. The items presented in the modified database are individually frequent. This algorithm then selects the MIIs and it divides into two non-disjoint sets as residual database and projected database. First the IFP-min algorithm is applied to residual database, where the MIIs are reported, if the database has the single item then it is considered to be the item itself or as an empty set. Then IFP-min algorithm is applied to projected database. The itemsets in the projected database share the lf-item as a prefix. The MIIs obtained from the projected database by recursively applying the algorithm are compared with those obtained from residual database. If an itemset is found to occur in the second set, it is not reported; otherwise, the lf-item is included in the itemset and is reported as an MII of the original database. The use of residual tree is to reduce the computational time.

III. Web Mining

It is the term of applying data mining techniques to automatically discovery and extract useful information from the World Wide Web documents and services. Data mining efforts associated with the web called web mining. Web data mining can be broadly categorized into three areas of interest based on which part of the web to mine.

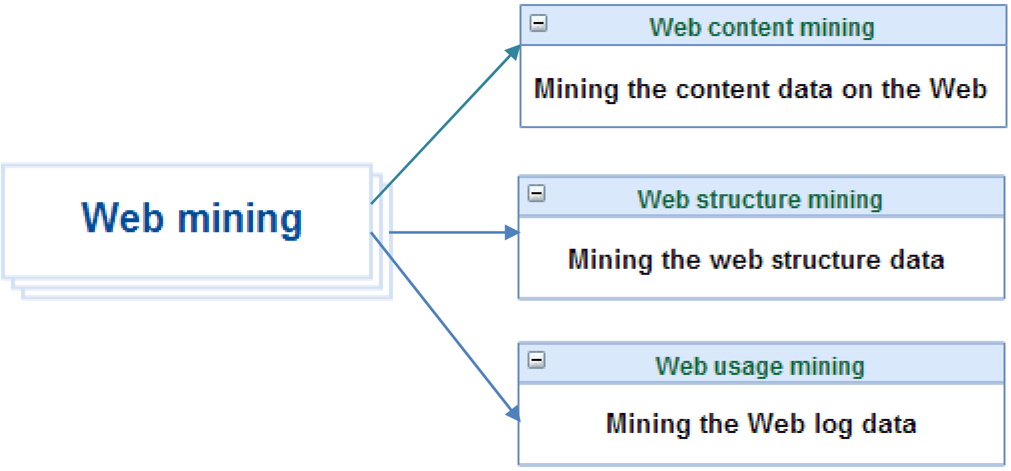

Figure2. Types of Web Mining

Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services [11]. In Web Mining, data can be collected at the server side, client-side, proxy servers, or obtained from an organization's database. Each type of data collection differs not only in terms of the location of the data source, but also the kinds of data available, the segment of population from which the data was collected, and its method of implementation.

IV. WEB LOG MINING

Web servers are register a log entry for every single access they get. A huge number of accesses are registered and collected in an ever growing web log [12].

a. Log files

Log files are files that list the actions that have been occurred. These log files reside in the web server. Computers that deliver the web pages are called as web servers. The Web server stores all of the files necessary to display the Web pages on the user’s computer. All the individual web pages combines together to form the completeness of a Web site. Images/graphic files and any scripts that make dynamic elements of the site function. The browser requests the data from the Web server, and using HTTP, the server delivers the data back to the browser that had requested the web page. The browser in turn converts, or formats, the files into a user viewable page. This gets displayed in the browser. In the same way the server can send the files to many client computers at the same time, allowing multiple clients to view the same page simultaneously. The Log files in different web servers maintain different types of information. [14] The basic information present in the log file are listed below,

- ***User name:*** This identifies who had visited the web site. The identification of the user mostly would be the IP address that is assigned by the Internet Service provider (ISP). This may be a temporary address that has been assigned. Therefore here the unique identification of the user is lagging. In some web sites the user identification is made by getting the user profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified.

- ***Visiting Path:*** The path taken by the user while visiting the web site. This may be by using the URL directly or by clicking on a link or trough a search engine.

- ***Path Traversed:*** This identifies the path taken by the user within the web site using the various links.

- ***Time stamp:*** The time spent by the user in each web page while surfing through the web site. This is identified as the session.

- ***Page last visited:*** The page that was visited by the user before he or she leaves the web site.

- ***Success rate:*** The success rate of the web site can be determined by the number of downloads made and the number copying activity under gone by the user. If any purchase of things or software made, this would also add up the success rate.

- ***User Agent:*** This is nothing but the browser from where the user sends the request to the web server. It's just a string describing the type and version of browser software being used.

- ***URL:*** The resource accessed by the user. It may be an HTML page, a CGI program, or a script.

- ***Request type:*** The method used for information transfer is noted. The methods like GET, POST.

These are the contents present in the log file. This log file details are used in case of web usage mining process.

b. Location of a Log File

A Web log is a file to which the Web server writes information each time a user requests a web site from that particular server. [15] A log file can be located in three different places:

- ***Web Servers:*** The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser. The contents of the file will be the same as it is discussed in the previous topic. In the server which collects the personal information of the user must have a secured transfer.

- ***Web proxy Servers:*** A Proxy server is said to be an intermediate server that exist between the client and the Web server. Therefore if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers maintain a separate log file for gathering the information of the user.

- ***Client browsers:*** This kind of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window the entries to the log file is done only by the Web server.

c. Types of Web Server Logs

Web Server logs are plain text (ASCII) files and are Independent from the server. [10]There are some Distinctions between server software, but traditionally there are four types of server logs:

i. Transfer Log
ii. Agent Log
iii. Error Log
iv. Referrer Log

The first two types of log files are standard. The referrer and agent logs may or may not be "turned on" at the server or may be added to the transfer log file to create an "extended" log file format.

d. Tools of web log mining

Different types of tools [16] used in various stages of web log mining are described in below table,

S.NO	Tools	Description
1	AWUSA	A framework based on combination of information architecture, automated usability evaluation and web mining techniques for data gathering and analysis
2	Web Quilt	Web logging and visualization system that helps web design teams capture usage traces which can be aggregated and visualized in a zooming interface that shows the web pages people viewed.
3	KOINOTITES	A system which uses data mining techniques for the construction of user communities on the Web
4	Web Tool	It uses sequential pattern mining which relies on PSP an algorithm developed by the authors
5	Web Mate	The user profile is inferred from training examples
6	Clementine	To browse data using interactive graphics to find important features and relationships.
7	WEBMINER	A general and flexible framework for Web usage mining, the application of data mining techniques, such as the discovery of association rules and sequential patterns, to extract relationships from data collected in large Web data repositories

VI. INFREQUENT ITEMSETS

Infrequent itemsets are rarely found in database. They are frequently considered to be uninteresting and eliminated using the support measure. Such itemsets are known as infrequent itemsets. Itemset mining is an exploratory data mining technique widely used for discovering valuable correlations among data. Frequent itemsets mining is a core component of data mining and variations of association analysis, like association rule mining. Infrequent itemsets are produced from very big or huge data sets by applying some rules or association rule mining algorithms like Apriori technique, that take larger computing time to compute all the frequent itemsets.

In this section, we provide definitions of key terms that explain the concepts frequent and infrequent itemset, let A be the collection or set of items entailed by database records, e.g. the set of items a consumer collects in a shopping complex, according to market basket analysis it is referred to as an itemset. Moreover, let I = {i1, i2... in} be a set of n different elements called items. Let the database DB is a collection of transactions over I, T is associated with each and every transaction and Tid is a unique index for each transaction. The subset of itemset X= $\{i_a, i_b, i_c... i_z\} \in$ I and its length consist a number of itemset in X. A z-length item set consist transaction in DB with different itemset and length z. The frequency (Number of occurrences) of an itemset X called support count of X, and it is denoted by Supp (X).

Frequent and infrequent itemset are depend on f_s and r_s where f_s is a frequent support count threshold and r_s is a infrequent support count threshold and $f_s< r_s$. Moreover a particular itemsets are said to be frequent if and only if $\text{Supp}(X) \geq f_s$ and infrequent if and only if $\text{Supp}(X) \leq r_s$. The support count of superset of an itemset is related to its subsets itemset. Let we take two itemset A and B such that $A\subset B$, the frequency of A itemset is at least B frequency, or we can say A is part of B then $\text{Supp}(A) \geq \text{Supp}(B)$, $\forall A\subset B$ [17].

a. Infrequent Itemset Mining

Various types of algorithm have been proposed for infrequent itemset mining and it is different from frequent pattern mining algorithms. Infrequent itemset consist those itemset that do not occur frequently but it may also generate interesting rules, so if a infrequent pattern consists high confidence rule then it should not be discarded completely. Koh and Rountree [18] proposed a more efficient algorithm name as Apriori-Inverse, which finds perfectly sporadic rules and imperfectly sporadic rules (irrelevant) without generating all the unnecessarily frequent items. They use three parameters in Apriori-Inverse such as Fixed Threshold, Adaptive Threshold, and Hill Climbing. In Apriori-Inverse finds all perfectly sporadic rules much more quickly than Apriori. Szathmary et al. [18] proposes two algorithms Minimal Infrequent Generators (MRG)-Exp and A Infrequent Itemset Miner Algorithm (ARIMA). First is MRG-Exp used to finding minimal infrequent generators, we focus on frequent itemsets generators in lattice. Second ARIMA is to get all infrequent itemsets from minimum rate itemset. Tsang et al. [18] propose a tree structure approach RP-Tree for mining a subset of infrequent association rules and information get component that helps to identify the more interesting association rules. RP-Tree, examines all infrequent-item nodes in the initial tree, and all nodes that have less support than an infrequent-item are infrequent items themselves, RP-Tree must find all infrequent item itemsets.

V. Conclusion

Web usage mining and data mining to find patterns is a growing area with the growth of Web-based applications. Application of web usage data can be used to better understand web usage, and apply this specific knowledge to better serve users. This paper has attempted to give an overview to the process of rare itemset mining from a Web server data and gives a detailed look about the web log file, its contents and its types etc., Added to these information it also gives a detailed description of how the file is being processed in web usage mining process. The various mechanisms that perform each step in mining the log file is being discussed in literature review. The contribution of the paper is to introduce the process of web log mining, and to show how infrequent itemset mining tasks can be applied on the web log data in order to obtain useful information about the user’s sessions.

REFERENCE:

1. *Tan, P. N., M. Steinbach, V. Kumar, .Introduction to Data Mining”, Addison-Wesley, 2005, 769pp.*

2. *Faustina Johnson, Santosh Kumar Gupta, “Web Content Mining Techniques: A Survey”, International Journal of Computer Applications, vol. 47, no.11, June 2012.*

3. *Preeti Chopra, Md. Ataullah, “A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms”, International Journal of Engineering and Advanced Technology (IJEAT), vol. 02, issue 03, Feb. 2013.*

4. *Adda M., Wu L., White S., and Fengr Y., “Pattern Detection with Rare Itemset Mining,” International Journal on Soft Computing, Artificial Intelligence and Applications, vol. 1, no. 1, pp. 1-17, 2012.*

5. *Troiano L. and Scibelli G., “A Time-Efficient Breadth-First Level-Wise Lattice-Traversal Algorithm To Discover Infrequent Itemsets,” Data Mining and Knowledge Discovery, vol. 28, no. 3, pp. 773-807, 2014.*

6. *R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 487-499, 1994.*

7. *Luca Cagliero and Paolo Garza “Infrequent Weighted Itemset Mining using Frequent Pattern Growth”, IEEE Transactions on Knowledge and Data Engineering, pp. 1- 14, 2013.*

8. *Han, J., Pei, J., & Yin, Y. “Mining frequent patterns without candidate generation”. In Proc. ACM-SIGMOD Int. Conf. Management of Data (SIGMOD ’96), Page 205-216, 2000.*

9. *Grahne O. and Zhu J. “Efficiently Using Prefix-trees in Mining Frequent Itemsets”, In Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining, 2004.*

10. *A.Gupta, A. Mittal, and A. Bhattacharya, “Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees,” Proc. Int’l Conf. Management of Data (COMAD), pp. 57- 68, 2011.*

11. *Wahab, M.H.A., Mohd, M.N.H., Hanafi, H.F., Mohsin, M.F.M.: Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science, Engineering and Technology 48 (2008).*

12. *L.K. Joshila Grace,V. Maheswari,Dhinaharan Nagamalai "Web Log Data Analysis and Mining" N. Meghanathan et al. (Eds.): CCSIT 2011, Part III, CCIS 133, pp. 459–469, © Springer-Verlag Berlin Heidelberg 2011.*

13. *Vijayalakshmi, S., Mohan, V., Suresh Raja, S.: Mining Constraint-based Multidimensional Frequent Sequential Pattern in Web Logs. European Journal of Scientific Research 36, 480–490 (2009).*

14. *Jain, R.K., Kasana1, D.R.S., Suresh Jain, D.: Efficient Web Log Mining using Doubly Linked Tree. International Journal of Computer Science and Information Security, IJCSIS 3 (July 2009).*

15. *Suneetha, K.R., Krishnamoorthi, R.: Identifying User Behavior by Analyzing Web Server Access Log File. IJCSNS International Journal of Computer Science and Network Security 9, 327–332 (2009).*

16. *D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, Web usage mining as a tool for personalization: A survey, User Modeling and User Adapted Interaction, 13(4), 2003, 311-372. Kluwer Academic Publishers.*

17. *A.Gupta, A. Mittal, and A. Bhattacharya, "Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees," Proc. Int'l Conf. Management of Data (COMAD), pp. 57- 68, 2011.*

18. *Brijesh Bakariya, Ghanshyam Thakur "An Efficient Algorithm for Extracting Infrequent Itemsets from Weblog" The International Arab Journal of Information Technology, Vol. 16, No. 2, March 2019.*