An Efficient System on High Utility Infrequent Itemsets Mining over Weblog Data

A. Kalaiselvi[1] P. Jayapriya[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Maharaja Arts & Science College, Coimbatore, India

Abstract— In recent years, mining infrequent itemsets over weblog databases have attracted much attention and their significance in many applications like fraud detection, web portals, information access and retrieval tools, giving information on problems occurred to the users, etc. 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. It gives details about the user. Infrequent Itemset mining differs from frequent itemset mining where it locates the uninteresting patterns, i.e., it detects the data items that arise very rarely. Itemsets which do not occur frequently in the database. All the itemsets which has value lesser than the support, will be considered as infrequent item sets. Data mining techniques like association rules, sequential patterns, clustering and classification can be used to discover frequent patterns. This paper aims to develop a novel dynamic algorithm for infrequent itemset over weblog data.

Key words: Infrequent Itemsets, Weblog Data, Web Usage Mining, Association Rule

I. INTRODUCTION

Uncommon Itemset mining shifts from continuous itemset mining wherever it calls attention to the exhausting examples, i.e., it recognizes the information things which happen infrequently. Rare cases are worthy of special concern since they express significant difficulties for data mining algorithms. But there must be concentration on the rare items set also. Because infrequent itemset may gain extra profit than frequent ones. It can be a business strategy to be followed for earns profit in large amount [1].

The rapid as well explosive development of data existing over the Internet, the World Wide Web has turn into a great stage to store, distribute also recover data and mine valuable information. Web Usage Mining is the area of Web Mining that deals with the removal of attractive information from classification data formed by web servers. Due to the possessions of the vast, various, active and formless nature of Web data, Web data research has met a many challenges like extensibility, multimedia and sequential problems etc. There are three common classes of data which can be discovered by web mining [2]:

S. NO	Class	Description
1	A	Web log activity, from server logs and Web browser activity tracking.
2	B	Web graph, from links between pages, People and other data.
3	C	Web content, for the data found on Web pages and inside of documents

Table 1: Classes in Web Mining

II. INFREQUENT ITEMSETS

It is an investigative information digging system generally utilized for finding profitable relationships among information. Visit itemsets mining is a middle segment of information mining and additionally deviations of affiliation examination, for example, affiliation lead mining. They are shaped from amazingly colossal informational collections by applying a few standards generally affiliation manage mining strategies like Apriori method, which gain part of processing time to figure all the incessant itemsets [3].

A few itemset are once in a while found in web log database that are frequently viewed as boring as well as are killed utilizing the help calculate. That information is known as rare itemsets. While tremendous prominent of uncommon information is undesirable, a couple of them may be useful to the examination, essentially those that relate to negative relationships in information. Some rare itemset may likewise propose the event of intriguing uncommon occasions [4].

To discover such unprecedented conditions, the foreseen keep up of a precedent should be recognized, all together that, if an example ends up having a fundamentally lesser manage than evaluated; this is affirmed as an appealing inconsistent thing. Rare itemset merit extraordinary consideration since they speak to real troubles for information mining calculations. Mining rare examples is a testing endeavor on the grounds that there are huge quantities of such examples which can be gotten from a given informational index [5].

III. WEB USAGE MINING

It is the way toward implementing information mining systems to find intriguing examples from web use information. Web utilization digging gives better comprehension to serving the necessities of Web-based applications. Web Usage Mining focuses on the methods which can calculate the navigational model of the user although the users interact with the web. This is mostly divided into two categories, they are general access pattern tracking and customized usage tracking. In common way pattern tracking data is exposed by using the narration of web page visited by user although in modified perform tracking mining is targeted on definite user [6]. Generally there are four types of data sources there in that handling information is recorded at various levels they are:

- Client level collection
- Browser level collection
- Server level collection
- Proxy level collection

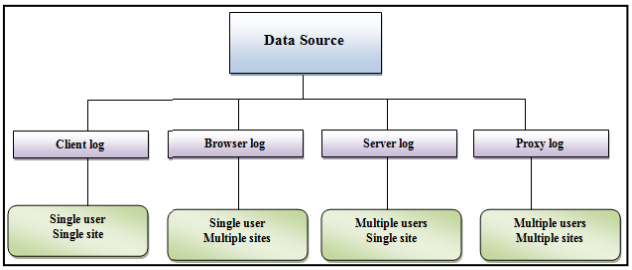

Fig. 1: Types of Data Sources

A. Client Level Collection

Client level collection is means of java scripts or java applets. This data show the behavior of a single user on single site. Client side data collection requires user participation for enabling java scripts or java applets [7].

B. Browser Level Collection

Browser level collection of the data collection is by modifying the browser. It shows the behavior of single user over multiple sites. The data collection capabilities are extend by modifying the source code of existing browser. Browser provides much more versatile data as they consider the behavior of single user on multiple sites.

C. Server Level Collection

Server level collection behavior of multiple users over single site. Server log files can be stored in common log format or extended log format. Server logs are not able to store cached page views. Another technique used for usage data collection at server level is TCP/IP packet sniffing [8].

D. Proxy Level Collection

Proxy log servers are used by internet service provider to provide World Wide Web access to customers. These server stores the behavior of multiple user at multiple site. Proxy log functions like cache server and they are able to produce cached page views [9].

There are many popular programs for usage pattern mining. Different types of tools used in all the three stages of web usage mining are described in following table [10].

S.No	Tool	Feature	Function
1	AWUSA	A framework depends on grouping of data architecture, automatic usability evaluation also web mining methods for data assembly as well as investigation.	Automated website usability evaluation.
2	Web Quilt	Web logging as well as visualization scheme which helps web design teams confine usage traces which can be aggregated also visualized in a zooming boundary that confirms the web pages people viewed.	To run usability tests and analyze the collected data from web logs.
3	KOINOTITES	A system which uses data mining techniques for the construction of user communities on the Web.	Personalization.
4	Web Tool	It uses sequential pattern mining which relies on PSP an algorithm developed by the authors.	Usage profiling.
5	Web Mate	The user profile is inferred from training examples.	As Proxy agent provides effective browsing and searching Help.
6	Clementine	To browse data using interactive graphics to find important features and relationships.	CRM
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.	Restructure a Web site, and in analyzing user access patterns to dynamically present information tailored to specific groups of users.

Table 2: Tools in Web usage Mining

IV. WEB LOG DATA ANALYSIS & MINING

Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analyzing these log files gives a neat idea about the user.

These files are listing the actions that have been occurred 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 [11]. 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 [12].

A. Log File content

The Log files in different web servers maintain different types of information. The basic information present in the log file is

- User Name
- Visiting Path
- Path Traversed
- Time Stamp
- Page Last Visited
- Success Rate
- URL
- Required Type

These are the substance there in the log record. These are used if there should arise an occurrence of web utilization mining strategy [13].

B. Log Files Location

A Web log is a folder to that the web server writes data every time a user needs a web site from which exacting server [14]. A log file can be positioned in three dissimilar places:

- Web Servers
- Web proxy Servers
- Client browsers

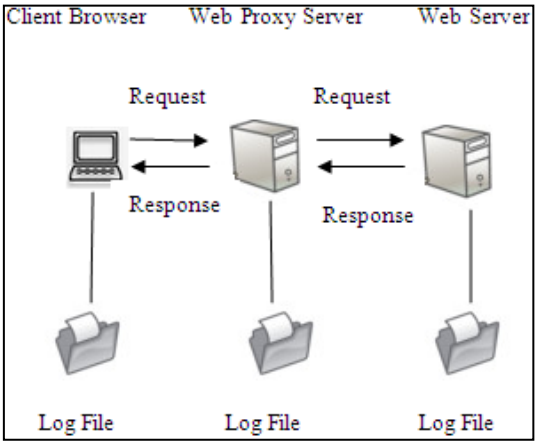

Fig. 2: Web Proxy Server Log files

C. Types of Web Server Logs

These are basic content documents in addition to they are independent of the attendant. Which is few contrasts between attendant programming, aside from more often than not there are 4 sorts [15]:

– Server Logs
– Transfer Log
– Agent Log
– Error Log
– Referrer Log

V. PROPOSED SYSTEM

This proposed system improves extracting infrequent itemsets from weblog data effectively. This system consists of major five steps named as [16]

– Joining

Generate Candidate Sets

– Pruning

Identify Infrequent Itemsets

– Binary Search

Identify the count for generating candidate set

– Verification

Extract Frequent Itemsets based on minimum support

– Hiding

Extract Sensitive Rules based on minimum Confidence

1) Step 1: Binary Search

Low = min (SOW) → SOW = Size of weblog in length of shortest item in D
High = max (SOW) → length of longest item in D
Mid = (low + high) / 2
While (low <= high)
K = mid

2) Step 2: Joining

C_k is the collection of K-frequent itemsets
While ($F_{k-1} \neq 0$)
Do $C_k = \emptyset$;
For (each pair of itemset)
$\{X_1, X_2 \ldots X_{k-2}, X_{k-1}\}$ F_{k-1} and $\{X1, X_2 \ldots X_{k-2}, Y_{k-1}\}$
Do if ($X_{k-1} < Y_{k-1}$)
Then Z = {X1, X2… Xk-2, Xk-1, Y_{k-1}};
// Z is the new itemset of size K [17].

3) Step 3: Filtration

P_k is the potential itemset with size K
IF_k is the collection of K-infrequent itemset
$P_k = C_k$;
IF_k =;
For(ach infrequent itemset $Z \in \cup_{i=2}^{k-1} IF_i$)
Do if ($Z \subset \{Y \mid \exists\ Y \in P_k\}$)
Then $P_k = P_k - Y$;
$IF_k = IF_k\ Y$;

4) Step 4: Verification

// F_k is the K-frequent itemset collection
$F_k = \emptyset$
For (each potential itemset Z P_k)
Do if (support (Z) ≥ minsup)
Then $F_k = F_k \cup Z$;
Else IFk = $IF_k \cup Z$ [18];

5) Step 5: Hiding Sensitive Rule

// S_k is the collection of K-sensitive itemset
Sk =∅;
For (each infrequent itemset $Z \in F_k$)
Do if (confidence (Z) ≥ minconf)
Then $S_k = S_k \cup Z$;
Else $NS_k = NS_k \cup Z$;
K = K + 1;
Print $\cup_{i=1}^{k-2} S_i$; [19]

VI. RESULT & DISCUSSIONS

This approach integrates the concept of two algorithms such as Infrequent Itemset Mining for Weblog (IIMW) association rule mining. In our experiments, our proposed algorithm compared to IIMW. We presented a new approach for mining rare itemsets in large databases. The described algorithm is differing from existing implementations. In this section, we compare IIMW algorithm to propose association rule mining based IIMW.

Minimum Support (%)	Number of Candidate Count	
	Proposed system	IIMWD
10	1.8	2.7
20	1.3	2.2
30	0.4	0.4
40	0.2	0.2
50	0.18	0.18
60	0.1	0.12
70	0.08	0.04
80	0.03	0.01
90	0	0
100	0	0

Table 3: Runtime Comparisons on Dataset

In Table3 ascertain the hopeful check and IIMW calculation contrast and the proposed framework. In this examination the quantity of competitor check produced with various help edge. This procedure is hopeful age process after this procedure the continuous and rare itemset could be extricated.

Minimum Support (%)	Number of Infrequent itemset	
	Proposed system	IIMWD
10	0.2	0
20	0.18	0
30	0.09	0
40	0.07	0
50	0.04	0
60	0.03	0
70	0.02	0
80	0.01	0
90	0	0
100	0	0

Table 4: Infrequent Itemset versus Minimum Support

In Table4 proposed system extract number of infrequent itemset with different minimum support (10% to 100%) and compare with IIMW algorithm. In our proposed algorithm generate candidate count, frequent patterns and infrequent patterns. In our system, we don't count those infrequent itemset which is having 0 frequencies because in the log file or web data; the web page is not visited. We only accept those infrequent itemset which length is 1 or greater

than 1. All experimental results show that the proposed algorithm improves better than existing approach.

VII. CONCLUSION

The proposed system developed to reduce the operation and find out infrequent itemset. Reduce the transaction and input/output cost. Also find the infrequent itemset from largest weblog itemset to smallest weblog itemset at the minimum level of scanning original database. The rare itemsets are presenting rarely in the database. Occasionally rare itemsets are more significant as they take useful data that regular patterns may not provide. Rare itemset mining is a not easy task. This research present an efficient approach for mining rare item set for time variant dynamic data set. It has an idea for removing infrequent itemsets and hiding sensitive rules. An integrated method for pruning frequent itemset and hiding sensitive rules of association rule mining is suggested. One of these methods expressed as filtration for joining operation and the rest is hiding the sensitive rule. It is observed that the proposed approach generates frequent rules effectively based on Filtration and Hiding methods.

REFERENCE

[1] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. SIGKDD Explor. Newsl., 2(2):66–75, 2000.
[2] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino. On maximal frequent and minimal infrequent sets in binary matrices. Annals of Mathematics and Artificial Intelligence, 39:211–221, 2003.
[3] L. Szathmary. Symbolic Data Mining Methods with the Coron Platform (M´ethodes symboliques de fouille de donn´ees avec la plate-forme Coron). PhD in Computer Sciences, Univ. Henri Poincar´e Nancy 1, France, Nov 2006.
[4] L. Szathmary and A. Napoli. CORON: A Framework for Level wise Itemset Mining Algorithms. In Suppl. Proc. Of the 3rd Intl. Conf. on Formal Concept Analysis (ICFCA ’05), Lens, France, pages 110–113, Feb 2005.
[5] H. Yun, D. Ha, B. Hwang, and K. Ryu. Mining association rules on significant rare data using relative support. Journal of Systems and Software, 67(3):181–191, 2003.
[6] C. Gianella, J. Han, J. Pei, X. Yan, and P. Yu. Mining frequent patterns in data streams at multiple time granularities. In Proceedings of the NFS Workshop on Next Generation Data Mining, 2002.
[7] Wang, C.Y., Tseng, S.S., and Hong T.P. Flexible online association rule mining based on multidimensional pattern relations, Information Sciences, vol. 176, pp. 1752–1780, 2006.
[8] Luca Cagliero and Paolo Garza , ”Infrequent Weighted Itemset Mining Using Frequent Pattern Growth”, IEEE Trans. Vol. 26, No. 4, April 2014.
[9] Mehdi Adda, Lei Wu, Sharon White, Yi Feng, ”Pattern Detection with Rare item-set Mining” International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.
[10] J. Srivastava and R. Cooley. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, 1:12–23, 2000.
[11] J. Yang and J. Logan. A data mining and survey study On diseases associated with paraesophageal hernia. In AMIA Annual Symposium Proceedings, pages 829–833, 2006.
[12] Jiawei Han · Hong Cheng · Dong Xin · Xifeng Yan “Frequent pattern mining: current status and future Directions” Data Min Knowl Disc (2007) 15:55–86 DOI 10.1007/s10618-006-0059-1.
[13] K. Beyer and R. Ramakrishnan. Bottom-up computation of sparse and iceberg cube.SIGMOD Rec., 28:359–370, 1999.
[14] Mehdi Adda, Lei Wu, Sharon White(2012), Yi Feng ” pattern detection with rare item-set mining” International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.
[15] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB, pages 487– 499, 1994.
[16] R. Agrawal, T. Imieinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, Proceedings of the ACM SIGMOD International Conference on the Management of Data, pages 207– 216, Washington DC, 1993. ACM Press.
[17] W. Shi, F.K. Ngok, and D.R. Zusman. Cell density regulates cellular reversal frequency in Myxococcus xanthus. In Proceedings of the National Academy of Sciences of the United States of America, volume 93(9), pages 4142–4146,1996.
[18] X. Dong, Z. Niu, X. Shi, X. Zhang, and D. Zhu. Mining both positive and negative association rules from frequent and infrequent itemsets. In ADMA, pages 122–133, 2007.
[19] Luca Cagliero and Paolo Garza “Infrequent Weighted Itemset Mining using Frequent Pattern Growth”, IEEE Transactions on Knowledge and Data Engineering, pp. 1- 14, 2013.

