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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

A Study on Association Rule Mining Algorithms Used in Web Usage Mining

Dhana Praveena A [1], Selvi V [2]

PhD Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, India[1]

Assistant Professor, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, India[2]

ABSTRACT: Web Usage Mining is an application of Data Mining which is used to identify the user needs from web log. It does so by discovering interesting and most frequent patterns based on users' navigational behaviors. Source data mainly consist of the logs that are collected when users access web servers and might be represented in standard format. Web server log files act as storage for frequent word sequences. The word sequence comprises of IP address, page reference and access time. The study focuses on comparison of Apriori, AprioriTID and AprioriHybrid algorithms.

KEYWORDS: Web Usage Mining, Association Rule Mining, Frequent Pattern, Apriori

I. INTRODUCTION

Web is a vast and dynamic repository which comprises of mostly raw data which is a source to the enormous supply of information and also raises the complexity of how to deal with the information excavated from this repository. Hence the web users need an effective search tool to find relevant information easily and to learn users' needs. Web usage mining is one of the applications of data mining technique which discovers the interesting usage patterns from web data. The main purpose of discovering such patterns is to understand and better serve the needs of the web based applications. Web usage is divided into three tasks: Preprocessing, Pattern analysis and Pattern Discovery. Preprocessing – includes the fusion, synchronization identification, user identification and sessionization. Pattern Analysis – pull outs interesting knowledge from frequent patterns and used for website modification. . Pattern Discovery- applies pattern discovery algorithms on raw data.

II.RELATED WORK

In [2], authors compared the time complexity of four association rule mining algorithms. The authors have proposed an improved version of Apriori algorithm which reduces the time consumption to find the frequent itemset. The speed of the algorithms is calculated, compared and concluded that all the algorithms are efficient in certain areas [3]. The accurateness of the association rule mining algorithms is compared by the authors in [5]. **Sequential Patterns** are used to discover frequent subsequences among large amount of sequential data. In web usage mining, sequential patterns are exploited to find *sequential* navigation patterns that appear in users' sessions frequently[10]. Association Rules are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining. When applied to Web Usage Mining, association rules are used to discover associations among web pages that frequently appear together in users' sessions. The typical result has the form "X.html, Y.html ⇒ Z.html" which states that if a user has visited page X.html and page Y.html, it is very likely that in the same session, the same user has also visited page Z.html. Mining association rules problems from large database has become the most advanced, important and dynamic research contents. The selection of association rule is based on support and confidence. The confidence factor indicates the strength of the implication rules, i.e. the confidence for an association rule is the ratio of the number of transactions that contain X U Y to the number of transactions that contain X; whereas the support factor indicates the frequencies of the occurring patterns in the rule. i.e., the support for an association rule is the percentage of transactions in the database that contain X U Y. Given the database DB, the problem of mining association rules involves the

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

generation of all association rules among all items in the given database DB that have support and confidence greater than or equal to the user specified minimum support and minimum confidence.

III. ASSOCIATION RULES

Association Rules are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining. When applied to Web Usage Mining, association rules are used to discover associations among web pages that frequently appear together in users' sessions. The typical result has the form "X.html, Y.html ⇒ Z.html" which states that if a user has visited page X.html and page Y.html, it is very likely that in the same session, the same user has also visited page Z.html. Mining association rules problems from large database has become the most advanced, important and dynamic research contents. The selection of association rule is based on support and confidence. The confidence factor indicates the strength of the implication rules, i.e. the confidence for an association rule is the ratio of the number of transactions that contain X U Y to the number of transactions that contain X; whereas the support factor indicates the frequencies of the occurring patterns in the rule. i.e., the support for an association rule is the percentage of transactions in the database that contain X U Y. Given the database DB, the problem of mining association rules involves the generation of all association rules among all items in the given database DB that have support and confidence greater than or equal to the user specified minimum support and minimum confidence.

Support: The percentage of task-relevant data transactions for which the pattern is true.

$$\text{Support}(XY) = \frac{\text{No. of Transactions containing X and Y}}{\text{Total No. of Transactions in D}}$$

$$\text{Confidence}(XY) = \frac{\text{No. of Transaction containing X and Y}}{\text{No. of transaction containing X}}$$

Confidence: The measure of certainty or trustworthiness associated with each discovered pattern.

APRIORI ALGORITHM

The traditional algorithm used for mining all frequent item sets and strong association rules was AIS algorithm. After a period of time, AIS algorithm was modified and renamed as Apriori. Apriori was initially proposed by R. Agrawal. Apriori is the most supervised and important algorithm for mining frequent item sets. It captures the large dataset at the time of its initial database passes and that dataset is used as the base for finding out other large datasets during the subsequent passes. This algorithm is based on the large item set property. It uses pruning techniques to avoid measure bound items. There are several key concepts used in Apriori algorithm such as Frequent Itemsets, Apriori Property and Join Operation. It identifies the frequent individual things within the information and extends them to larger and bigger item sets as long as those item sets seem sufficiently typically within the information. Apriori algorithmic rule confirms frequent item sets that may be used to determine association rules that highlight general trends within the information.

APRIORITID ALGORITHM

AprioriTID algorithm uses the Geneartion operation to generate the candidate itemsets. The difference between Apriori and AprioriTID algorithms is that the database is not referred for counting the support after the primary pass. Instead, a group of candidate itemsets is used for this purpose for k>1. If a group does not have any

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

candidate k_itemset, then the candidate k_itemset won't have any entry for that transaction. This can reduce the number of transactions within the set containing the candidate itemsets as compared to the database. Since the value of k increases each entry will be smaller than the corresponding transactions because the variety of candidates within the transaction will continue decreasing. Apriori exclusively performs higher than AprioriTID during its initial passes however in later passes AprioriTID certainly have higher performance than Apriori.

APRIORI HYBRID ALGORITHM

Apriori examines the database for every transaction. On the other hand, AprioriTID scans the candidate itemset for obtaining support count. Based on these observations, the Apriori Hybrid algorithm has been proposed. In the earlier passes, Apriori does better than AprioriTid. In later passes, AprioriTID performs better than Apriori. So Apriori Hybrid uses Apriori in the initial passes and switches to AprioriTid in the later passes.

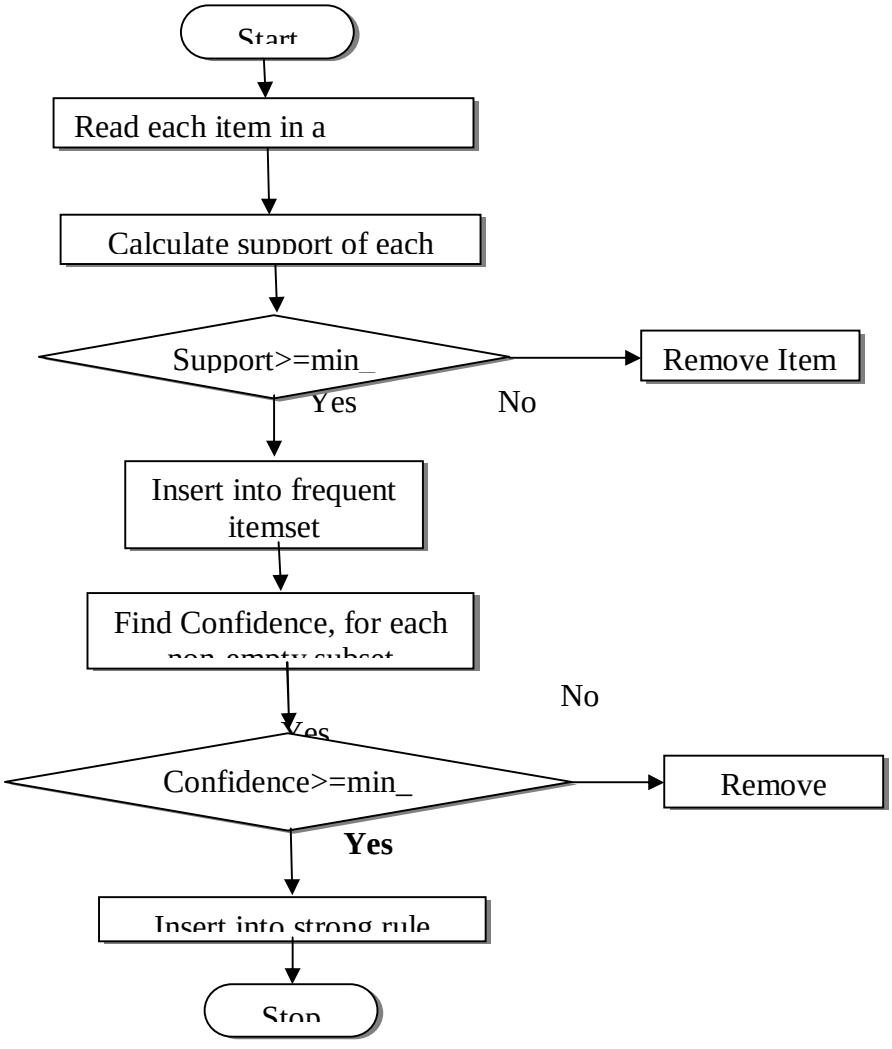

Fig. 1 : Process Flow of Apriori Algorithm

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

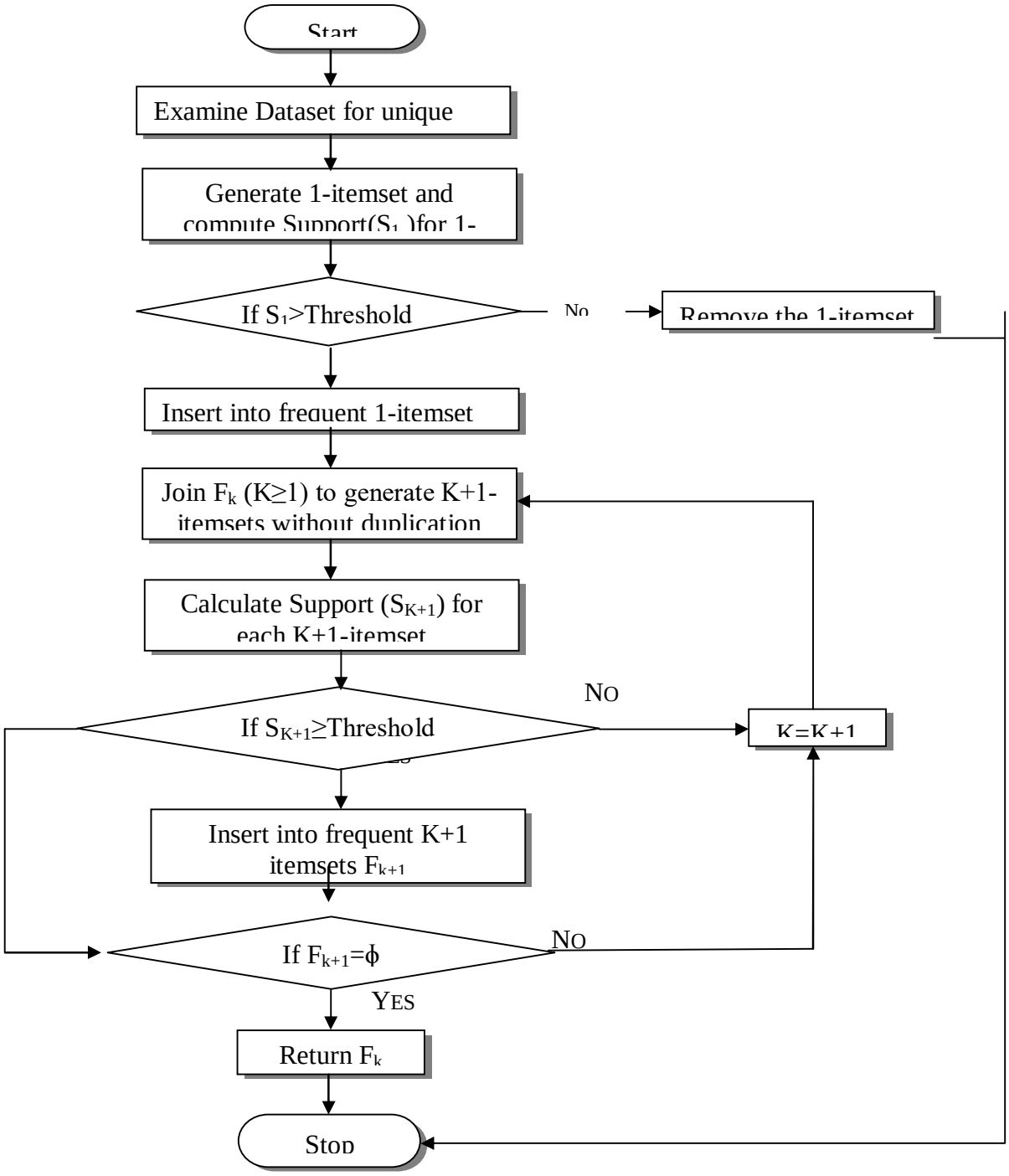

Fig:2 Process Flow of AprioriTID Algorithm

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

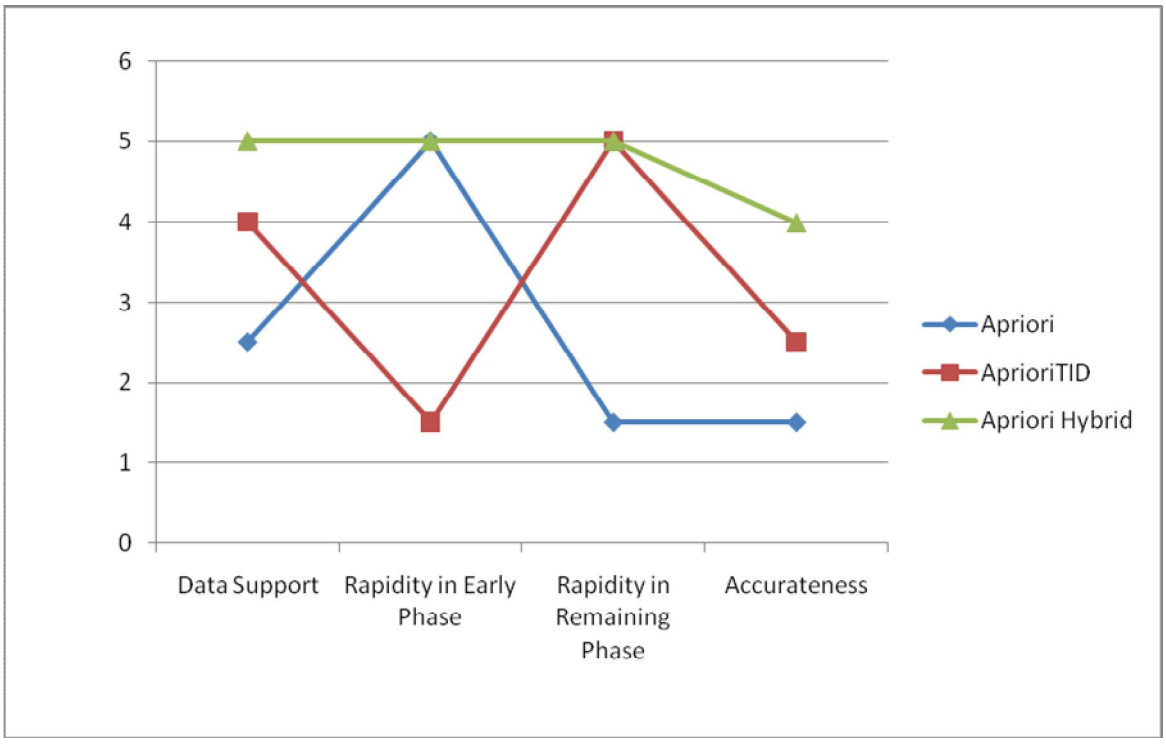

Fig 3: Graphical Representation of the performance of Association Rule Mining Algorithms

This figure shows the comparison of the three association rule mining algorithms namely Apriori, Apriori Tid and Apriori Hybrid. It shows that Apriori Hybrid algorithm is most efficient in all the aspects like Data Support, Speed in initial phase and in remaining phase and accuracy.

TABLE 1: COMPARISON OF ASSOCIATION RULE MINING ALGORITHMS

ATTRIBUTES	APRIORI	APRIORITID	APRIORIHYBRID
DATA SUPPORT	AVERAGE	HUGE	VERY BIG
RAPIDITY IN EARLY PHASE	HIGH	SLOW	HIGH
RAPIDITY IN REMAINING PHASE	SLOW	HIGH	HIGH
ACCURATENESS	A SMALLER AMOUNT	AVERAGE, BUT HIGHER THAN APRIORI	MORE ACCURATE

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

IV. CONCLUSION

This paper presents the extensive of study of various Association Rule Mining algorithms in data mining which are really useful and very much needed to obtain useful facts or associations among data items in large data sets to take some important decision making in any kind of problems. This paper gives the outline of three Association Rule Mining algorithms namely Apriori, AprioriTid, and AprioriHybrid in which all algorithms are evaluated and the merits and demerits are reported. In comparative study, all three algorithms have been compared with respect to three important criteria such as Data Support, Rapidity and accurateness. Based on rapidity, the Apriori hybrid algorithm is good However, the Apriori and AprioriTID algorithms outperform well than the Apriori Hybrid with respect to Accurateness. The comparative result be evidence for that the Apriori Hybrid algorithm is more suitable for obtaining significant associations from very large datasets in a speedy and accurate manner.

REFERENCES

1. Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the international conference on very large data bases (pp. 407–419).
2. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association between sets of items in massive database. International proceedings of the ACM-SIGMOD international conference on management of data (pp. 207–216).
3. Attila Gyenesei, A Fuzzy approach for mining quantitative association rules, Technical Report: TUCS-TR-336 Year of Publication, 2000.
4. Borgelt, C. “Efficient Implementations of Apriori and Eclat”. Workshop of frequent item set mining implementations (FIMI 2003, Melbourne, FL, USA).
5. Hipp, J., G¨untzer, U., and Nakhaeizadeh, G. “Algorithms for Association Rule Mining – A General Survey and Comparison”, SIGKDD Explorations ACM, JULY 2000.
6. S. Rao, P. Gupta 2012. “Implementing improved algorithm over Apriori data mining association rule algorithm”, *IJCST*, vol. 3, pp.489-493, ISSN: 2229-4333.
7. Pratima Gautam and K.R. Pardasani, “Algorithm for Efficient Multilevel Association Rule Mining” *In (IJCSE) International Journal on Computer Science and Engineering*, Volume 02, No. 05, 1700-1704, 2010.
8. Qiankun Zhao, Sourav S. Bhowmick, Association Rule Mining: A Survey, Technical Report, CAIS, Nanyang Technological University, Singapore, 2003.
9. Irina Tudor, Association rule mining as a data mining technique, BULETINUL universitatii Petrol-Gaze din Ploiesti, Vol. LX No1/ 2008, page 49-56.
10. J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In:Proc. Conf. on the Management of Data (SIGMOD’00, Dallas, TX). ACM Press, New York, NY, USA 2000.
11. R. Porkodi, Dr. B.L Shivakumar, An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm, pp.1510 – 1514, March 2009.
12. Srikant, R. & Agrawal, R., "Mining quantitative association rules in large relational tables", SfGMOD Rec., ACM, 1996, 25.
13. P. Prithiviraj and R. Porkodi, “A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study”, Open J. Comput. Sci. Eng. Surv, 2015
14. A Modi, R Krishnan “An Improved Method for Frequent Itemset Mining- International Journal of Emerging Technology and Advanced Engineering”, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 5, May 2013)
15. JR Jeba, Dr.S.P.Victor, Comparison of Frequent Item Set Mining Algorithms, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (6) , 2011, 2838-2841.
16. Khurana, K., and Sharma, S. “A comparative analysis of association rule mining algorithms”. *International Journal of Scientific and Research Publications*, Volume 3, Issue 5, May 2013