Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.299 – 306

RESEARCH ARTICLE

APRIORI ALGORITHM AND FILTERED ASSOCIATOR IN ASSOCIATION RULE MINING

Himani Bathla, Ms. Kavita Kathuria
M.Tech student, Department of CSE, Shri baba masth nath engineering college
Assistant professor, Department of CSE, Shri baba masth nath engineering college
himanibathla989@gmail.com, kavita.kathuria1990@gmail.com

Abstract: ***Most algorithms used to identify large itemsets can be classified as either sequential or parallel. In most cases, it is assumed that the itemsets are identified and stored in lexicographic order (based on item name). This ordering provides a logical manner in which itemsets can be generated and counted. This is the normal approach with sequential algorithms. On the other hand, parallel algorithms focus on how to parallelize the task of finding large itemsets. Mining Associations is one of the techniques involved in the process mentioned in chapter 1 and among the data mining problems it might be the most studied ones. Discovering association rules is at the heart of data mining. Mining for association rules between items in large database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making. Today, research work on association rules is motivated by an extensive range of application areas, such as banking, manufacturing, health care, and telecommunications. It is also used for building statistical thesaurus from the text databases, finding web access patterns from web log files, and also discovering associated images from huge sized image databases.***

Keywords: ***KDD, WWW, CAR, CHAID, AIS***

I. INTRODUCTION

Mining for association rules between items in large database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making. Today, research work on association rules is motivated by an extensive range of application areas, such as banking, manufacturing, health care, and telecommunications. It is also used for building statistical thesaurus from the text databases, finding web access patterns from web log files, and also discovering associated images from huge sized

image databases. For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule below:

Computer => antivirus software [support = 2%; confidence = 60%] ….(1)

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule (1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. The rest of this chapter is organised as follows: Section II summarizes related researches. Section III gives introduction of various data mining tools, Section IV describes Apriori association mining algorithms, Section V describes Experimental results and analysis. In Section VI, we draw conclusion and give future work.

II. RELATED WORK

Jaishree Singh, Dr. J.S. Sodhi[1], has explained Classical Apriori algorithm generates large number of candidate sets if database is large. And due to large number of records in database results in much more I/O cost. In this project, we proposed an optimized method for Apriori algorithm which reduces the size of database. In our proposed method, we introduced an attribute Size_Of_Transaction (SOT), containing number of items in individual transaction in database. **Chang-Hung Lee, Ming-Syan Chen[2],** the discovery of association relationships among a huge database has been known to be useful in selective marketing, decision analysis, and business management. A popular area of applications is the market basket analysis, which studies the buying behaviors of customers by searching for sets of items that are frequently purchased together (or in sequence). **Farah Hanna AL-Zawaidah, Yosef Hasan Jbara[3]**, in this paper we present a novel association rule mining approach that can efficiently discover the association rules in large databases. The proposed approach is derived from the conventional Apriori approach with features added to improve data mining performance. We have performed extensive experiments and compared the performance of our algorithm with existing algorithms found in the literature. Experimental results show that our approach outperforms other approaches and show that our approach can quickly discover frequent itemsets and effectively mine potential association rules. **Du Ping, Gao Yongping[5]**, In this paper, they have explained an enhance algorithm associating which is based on the user interest and the importance of itemsets is put forward by the paper, incorporate item that user is interested in into the itemsets as a seed item, then scan the database, incorporate all other items which are in the same transaction into item sets, Construct user interest itemsets, reduce unnecessary itemsets; through the design of the support functions algorithm not only considered the frequency of itemsets, but also consider different importance between different itemsets. The new algorithm reduces the storage space, improves the efficiency and accuracy of the algorithm. **Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu[23]**, In this paper, they have analyzed the basic ideas and the shortcomings of Apriori algorithm, studies the current major improvement strategies of it. In order to solve the low performance and efficiency of the algorithm caused by its generating lots of candidate sets and scanning the transaction database repeatedly, it studies the pruning optimization and transaction reduction strategies, and on this basis, the improved Apriori algorithm based on pruning optimization and transaction reduction is put forward. According to the performance comparison in the simulation experiment, by using the improved algorithm, the number of frequent item sets is much less and the running time is significantly shortened as well as the performance is enhanced then finally the algorithm is improved.

III. DATA MINING TOOLS

Data mining , a set of techniques used for the purpose of obtaining information from the data. Statistical analysis of data using a combination of techniques and artificial intelligence algorithms and data quality information in the disclosure of confidential information, a process of transformation. In this context, SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB, commercial and **RapidMiner** (YALE), **Weka**, **R**, **C4.5**, **Orange**, **KNIME** developed several programs, including open source.

i) Open Source Programs Data Mining

Data mining applications is necessary to use a computer program to do. In this context, most software is developed. In this section, the Open Source Data Mining Programs and **RapidMiner** (YALE), **Weka** and **R** programs mentioned.

i) RapidMiner (YALE)

By scientists from Yale University in the United States was developed using Java language. RapidMiner (previously: Rapid-I, YALE) is a mature, Java-based, general DM tool currently in development by the company RapidMiner, Germany. Previous versions (v. 5 or lower) were open source. RapidMiner also offers the option of application wizards that construct the process automatically based on the required project goals (e.g. direct marketing, churn analysis, sentiment analysis).

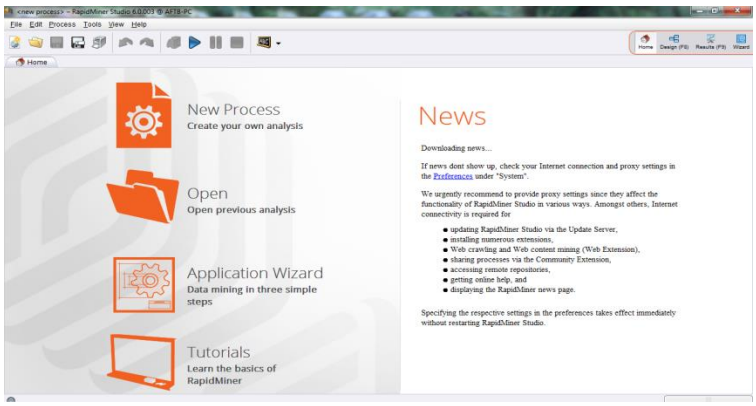

Figure 1 : RapidMiner Application Menu

ii) WEKA

Weka is a Java-based, open-source DM platform developed at the University of Waikato, New Zealand. The software is free under GNU GPL 3 for noncommercial purposes. Weka has had mostly stable popularity over the years, which is mainly due to its user friendliness and the availability of a large number of implemented DM algorithms. It is still not as popular as RapidMiner or R, both in business and academic circles, mostly because of some slow and more resource demanding implementations of DM algorithms. Although it is not a single tool of choice in DM, Weka is still quite powerful and versatile, and has a large community support.

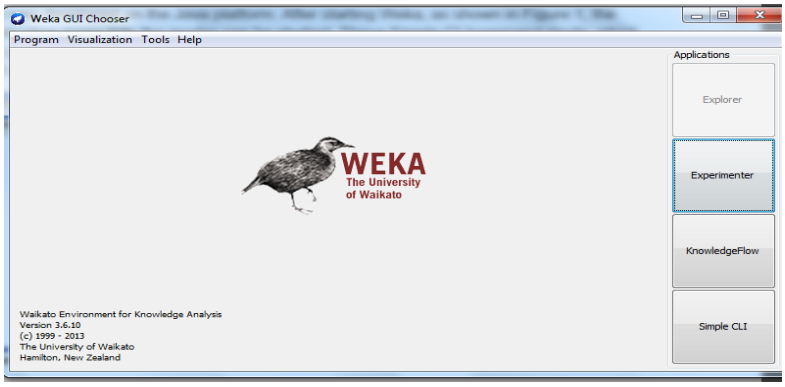

Figure 2 : WEKA Application Menu

IV. APRIORI ASSOCIATION ALGORITHMS

The Apriori algorithm developed by[1] is a great achievement in the history of mining association rules. It is by far the most well-known association rule algorithm. This technique uses the property that any subset of a large itemset must be a large itemset. Also, it is assumed that items within an itemset are kept in lexicographic order. The Apriori generates the candidate itemsets by joining the large itemsets of the previous pass and deleting those subsets which are small in the previous pass without considering the transactions in the database. By only considering large itemsets of the previous pass, the number of candidate large itemsets is significantly reduced.

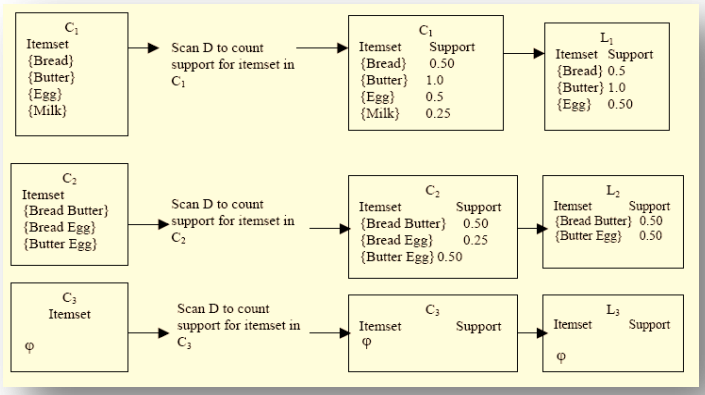

Figure 3 : Discovering Large Itemsets using the Apriori Algorithm

Apriori incorporates buffer management to handle the fact that all the large itemsets Lk-1 and the candidate itemsets Ck need to be stored in the candidate generation phase of a pass k may not fit in the memory. A similar problem may arise during the counting phase where storage for Ck and at least one page to buffer the database transactions are needed[1]. [1] considered two approaches to handle these issues. At first they assumed that Lk-1 fits in memory but Ck does not. The authors resolve this problem by modifying apriori_gen() so that it generates a number of candidate sets Ck′ which fits in the memory. Large itemsets Lk resulting from Ck′ are written to disk, while small itemsets are deleted. This process continues until all of Ck has been measured. The second scenario is that Lk-1 does not fit in the memory. This problem is handled by sorting Lk-1 externally. A block of Lk-1 is brought into the memory in which the first (k-2) items are the same. Blocks of Lk-1 are read and candidates are generated until the memory fills up. This process continues until all Ck has been counted.


```
Function count(C: a set of itemsets, D: database)
begin
for each transaction T  D="Di do begin
forall subsets x  T do
if x    C then
x.count++;
end
end
```

Algorithm . Apriori

Input:
I, D, s
Output:
L
Algorithm:
//Apriori Algorithm proposed by Agrawal R., Srikant, R. [**1**]
//procedure LargeItemsets
1) C 1: = I; //Candidate 1-itemsets
2) Generate L1 by traversing database and counting each occurrence of an attribute in a transaction;
3) **for** (k = 2; $L_{k-1} \neq \varphi$; k++) **do begin**
//Candidate Itemset generation
//New k-candidate itemsets are generated from (k-1)-large itemsets
4) Ck = apriori-gen(L_{k-1});
//Counting support of Ck
5) Count (Ck, D)
6) Lk = {c C_k | c.count $\geq$ minsup}
7) **end**
9) L = kL_k

i) Apriori-TID: Apriori-TID uses the Apriori's candidate generating function to determine candidate itemsets before the beginning of a pass. The main difference from Apriori is that it does not use the database for counting support after the first pass. Rather, it uses an encoding of the candidate itemsets used in the previous pass denoted by *Ck* . In Apriori-TID, the candidate itemsets in Ck are stored in an array indexed by TIDs of the itemsets in Ck. Each Ck is stored in a sequential structure. In the kth pass, Apriori-TID needs memory space for Lk-1 and Ck during candidate generation. It was also found that Apriori-TID outperforms Apriori when there is a smaller number of *Ck* sets, which can fit in the memory and the distribution of the large itemsets has a long tail. That means the distribution of entries in large itemsets is high at early stage.

ii) Apriori-Hybrid: The Apriori-Hybrid technique was developed which uses Apriori in the initial passes and switches to Apriori-TID when it expects that the set *Ck* at the end of the pass will fit in memory. Therefore, an estimation of *Ck* at the end of each pass is necessary. Also, there is a cost involvement of switching from Apriori to Apriori-TID. The performance of this technique was also evaluated by conducting experiments for large datasets. It was observed that Apriori- Hybrid performs better than Apriori except in the case when the switching occurs at the very end of the passes.

V. IMPLEMENTATION AND PERFORMANCE EVALUATION

While implementing Apriori algorithm such as spend a large overhead to deal with large candidate set, many repeat comparison of itemset in join step and repeatedly scanning the transaction database requires a lot of I/O load etc. Hence to analyze the algorithm in depth, we have used WEKA tool, which is build for various kind of data mining algorithms and in respected research area.

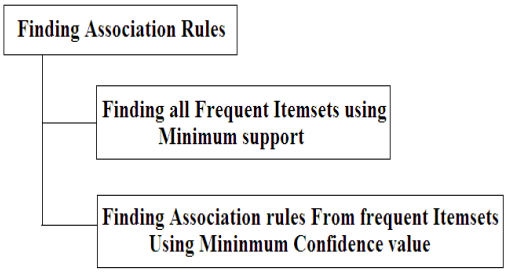

Figure 4: Finding the association Rule

i) Implementation of Apriori Algorithm: To perform the Apriori algorithm, the best open source data mining tool is Weka, which is developed at the University of Waikato, New Zealand, first we retrieve the dataset that is already exist in weka tool, by which we could perform the algorithms and analyze the objectives.

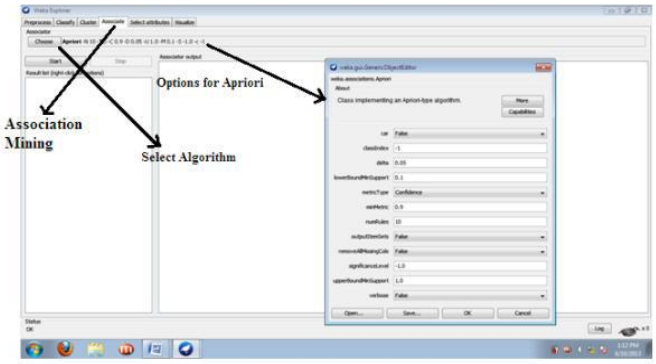

Figure 5: Apriori Algorithm with different properties in Weka

ii) Implementation of Filter Associator

To perform the Filter Associator, we have to do same procedure as Apriori algorithm i.e. just select the Filter Associator in place of Apriori algorithm. After taking the value of support and confidence, the execution of Filter Associator is done by clicking the "Start" button and according to that it generates the best association rules.

iii) Performance Evaluation of Apriori algorithm and Filter Associator

After performing the execution of both algorithms: Apriori algorithm and Filter Associator in the weka tool, we found that Apriori algorithm takes more number of cycle performed and for specific value of support it also generates extra large itemsets compare to the Filter Associator.

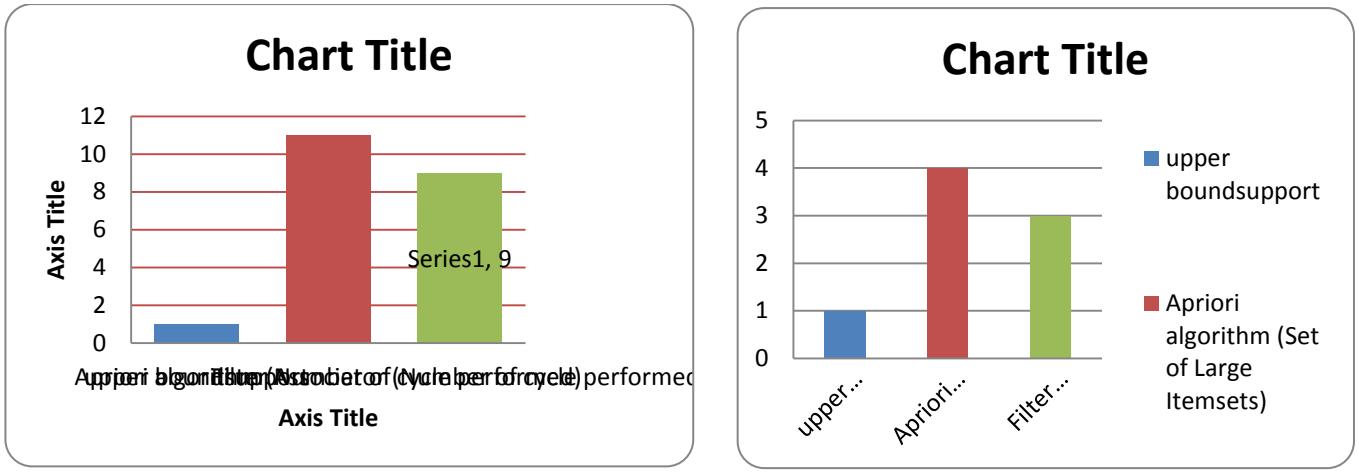

Figure 6: Performance of Apriori and Filtered Associator algorithms

VI. CONCLUSION

Conclusion

In this paper we have discussed various association rule algorithms and compared two algorithms: Apriori algorithm and Filter Associator. We have analyzed the frequent itemsets generation and number of cycle performed over the Apriori algorithm and Filter Associator in the context of association analysis. According to the comparison of above two algorithms on weka tool, we conclude that Filter Associator is efficient algorithm than Apriori algorithm based on above two factors (Number of cycle performed, large itemsets) because the Apriori algorithm generates more number of cycle performed and generate extra large itemsets which degrades the performance of algorithm.

Future Recommendations

Some of the future enhancements of the thesis are presented below:

- The work presented in the thesis can be extended for multi-level association rule mining.
- The work can be enhanced to generate multi-dimensional association rules.
- A tool for generating association rules can be developed. This tool can choose the approach for frequent itemsets mining according to the properties of the dataset to be mined.

REFERENCES

1. Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, 487-499.

2. A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. Proceedings of the 21st International Conference on Very large Database,1995.

3. Anurag Choubey, Ravindra Patel, J.L.Rana, "A Survey of Efficient Algorithms and New Approach for Fast Discovery of Frequent itemset for Association Rule Mining", IJSCE ,ISSN: 2231-2307, vol. 1, issue 2,May 2011.

4. Dr. Varun Kumar, Anupama Chadha, **"**Mining Association Rules in Student's Assessment Data", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

5. Du Ping, Gao Yongping, " A New Improvement of Apriori Algorithm for Mining Association Rules", International Conference on Computer Application and System Modeling (ICCASM 2010), Vol. 2, 529-532.

6. Farah Hanna AL-Zawaidah, Yosef Hasan Jbara, "An Improved Algorithm for Mining Association Rules in Large Databases", World of Computer Science and Information Technology Journal (WCSIT), ISSN: 2221-0741 Vol. 1, No. 7, 311-316, 2011.

7. Hassan M. Najadat, Mohammed Al-Maolegi, Bassam Arkok, "An Improved Apriori Algorithm for Association Rules", International Research Journal of Computer Science and Application Vol. 1, No. 1, June 2013, PP: 01 – 08.

8. H.Toivonen, "Sampling large databases for association rules". In Proc. 2006 Int. Conf. Very Large Data Bases(VLDB'06),pages 134-145, Bombay, India, Sep.2006.

9. Hu Ji-ming, Xian Xue-feng. " The Research and Improvement of Apriori for association rules mining", Computer Technology and Development 2006 16(4) pp. 99-104.

10. Jiao Yabing, "Research of an Improved Apriori Algorithm in Data Mining Association Rules", International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.

11. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

12. Jaishree Singh, Hari Ram, Dr. J.S. Sodhi, "Improving Efficiency of Apriori Algorithm Using Transaction Reduction ", International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013.

13. Li Qingzhong , Wang Haiyang, Yan Zhongmin, "Efficient mining of association rules by reducing the number of passes over the database", Computer Science and Technology, 2008, pp. 182-188.

14. L. Klemetinen, H. Mannila, P. Ronkainen, et al. (1994) "Finding interesting rules from large sets of discovered association rules". Third International Conference on Information and Knowledge Management pp. 401-407.Gaithersburg, USA.

15. Li Yang, Mustafa Sanver; Mining Short Association Rules with One Database Scan; Int'l Conf. on Information and Knowledge Engineering; June 2004.

16. Mohammed M. Mazid, A.B.M. Shawkat Ali and Kevin S. Tickle(2008), "Finding a Unique Association Rule Mining Algorithm Based on Data Characteristics", 5th International Conference on Electrical and Computer Engineering ICECE.

