ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Survey on various improved Apriori Algorithms

Shilpi Singla[1], Arun Malik[2]

M.Tech Scholar, Dept. of Computer Science, Lovely Professional University, Jalandhar, India [1]

Assistant Professor, Dept. of Computer Science, Lovely Professional University, Jalandhar, India [2]

Abstract: Data Mining is a way of obtaining undetected patterns or facts from massive amount of data in a database. Association rule mining is a major technique in the area of data mining. Association rule mining finds frequent itemsets from a set of transactional databases. Apriori algorithm is one of the earliest algorithm of association rule mining. Apriori employs an iterative approach known as levelwise search. In this paper, we have presented a survey of most recent work that has been done by researchers in Association rule based mining using Apriori algorithm.

Keywords: Association rule mining, Data mining, Apriori algorithm.

1.INTRODUCTION

Data Mining is a way of obtaining undetected patterns or facts from massive amount of data in a database. Data Mining is also known as knowledge discovery in databases (KDD). Data mining is more in demand because it helps to reduce cost and increases the revenues [1]. The various applications of data mining are customer retention, market analysis, production control and fraud detection. Data Mining is designed for different databases such as object-relational databases, relational databases, data warehouses and multimedia databases. Data mining methods can be categorized into classification, clustering, association rule mining, sequential pattern discovery, regression etc. Amongst these methods, association rule mining is very important which results in generating strong association rules.

Association rules was first proposed by R.Agrawal which aims at finding frequent itemsets from a set of transactional databases. The various algorithms in associations rule mining are Apriori, FP-Growth, Direct hashing and pruning (DHP), Apriori Tid. It is based on support and confidence values. The definitions of probability are:

Support (A->B) =P (AUB)
Confidence (A->B) =P (B|A)

The rules that meet the condition of minimum support (min_supp) and minimum confidence (min_conf) values are known as strong association rules. The itemsets which appears in the data set frequently are known as frequent itemsets. If the support value of itemsets A is greater than or equal to minimum support threshold value, then itemsets A is called frequent itemsets. If the support value of itemsets A is smaller than the minimum support threshold value, then itemsets A is called infrequent itemsets.

The process of association rule mining is categorized into two steps which are shown in figure 1.

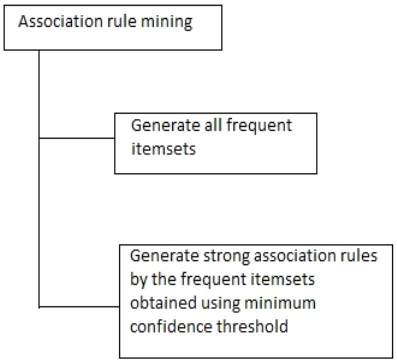

Figure 1: Association rule generation

2. APRIORI ALGORITHM

Apriori algorithm is the fundamental algorithm of association rule mining proposed by R.Agrawal and S.Srikant in 1994 [5]. Apriori employs an iterative approach known as levelwise search. In Apriori, (k+1) itemsets are generated from k-itemsets. First, scan the database for count of each candidate and compare candidate support count with minimum support count to generate set of frequent 1-itemsets. The set is denoted as L1. Then, L1 is used to find L2, set of frequent 2-itemsets, which is further used to find L3 and so on, until no more frequent k-itemsets can be found [7]. After finding set of frequent k-itemsets, it is easy to generate strong association rules. The process of finding each Lk requires the database to be scanned completely once. An important Apriori property is used to improve efficiency of level-wise generation of frequent itemsets which tells that any subset of frequent itemset must be frequent. A two-step process is followed:

a) Join step

In join step, Ck is generated by joining Lk-1 with itself.

b) Prune step

In prune step, if any (k-1) itemset that is not frequent cannot be a subset of a frequent k-itemset [2].

Pseudo Code[2]-

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Enter: a database D, the minimum threshold of support min-sup
Output: the database of all the frequent
a) L1 = (frequent 1 - Items);
b) for (k = 2; Lk-1; k + +) do begin
c) Ck = apriori_gen (Lk-1); / / after two-step connection and pruning operations generate a new candidate frequent itemsets
d) for all transactions t €D do begin
e) Ct = subset (Ck, t); / / in the database to scan t included in the candidate frequent itemsets Ck
f) for all candidates c € Ct do
g) c.count + +;
h) end;
i) Lk = (c € C | c.count ≥ min-sup)
j) end;
k) Answer = UkLk ;

Advantages of Apriori algorithm:
- Apriori algorithm is easy to understand.
- It is simple to implement.
- It uses large itemset property.
- It is easily parallelized.

Disadvantages of Apriori algorithm:
- It requires many database scans.
- It is less efficient and accurate.
- It takes more time and consumes more memory.

3. REVIEW ON SEVERAL IMPROVEMENTS OF APRIORI ALGORITHM

Various improved algorithms have been proposed to vanquish the weaknesses of Apriori algorithm in various ways. Here presents eight different approaches that face the common drawback.

3.1 Improvement by reducing redundant pruning operation and increasing efficiency of support calculation

3.1.1 Enlightenment

To weed out obstructions of frequent itemsets mining in Apriori algorithm the authors **Wanjun Yu, Xiaochun Wang and Fangyi Wang, Erkang Wang and Bowen Chen (2008)** [2] has given a new algorithm named as Reduced Apriori Algorithm with Tag (RAAT). In Apriori algorithm, there is large number of candidate 2-itemsets and less tendency to determine support value. So, it takes lot of time to scan the database repeatedly and decreases the efficiency. To improve this, RAAT uses Apriori-gen operation to form candidate 2-itemsets which results in diminishing the pruning operation. RAAT also follows the concept of tag to increase the speed of support calculation. As a result, RAAT shortens the time and improves efficiency. The experimental results shown that RAAT performs well when it is compared with Apriori algorithm in a number of times.

3.2 Improvement by reducing transactions and memory utilization

3.2.1 Enlightenment

The authors **Jaishree Singh, Hari Ram, Dr.J.S.Sodhi (2013)** [3] have introduced a modified Apriori Algorithm called an improved Apriori Algorithm(IAA) to conquer the limitations of classical Apriori Algorithm. The classical Apriori algorithm scans the database many times. If database contains ample number of records, it takes huge time to scan the database which results in increasing I/O cost. The improved Apriori Algorithm reduces the scanning time by eliminating the transactions containing irrelevant records. It uses the concept of attribute named as Size_Of_Transaction (SOT) which contains the number of items exists in specific transaction. It also decreases the I/O cost. By comparing improved Apriori Algorithm with classical apriori algorithm, it was shown that improved Apriori Algorithm is better on the basis of efficiency and optimization. This algorithm has certain drawback also as it has to deal with new database after every generation of frequent itemset. In future, we can divide the database among processors to remove this drawback.

3.3 Improvement based on customer habits

3.3.1 Enlightenment

Shuo Yang et al (2012) [4] proposed a theorem to improve the traditional Apriori Algorithm. The traditional Apriori Algorithm takes more time to scan the database in order to find out the frequent itemsets. This increases the complexity and decreases efficiency. The proposed algorithm decreases the database access on the basis of customer habits. It uses relative theorems to find frequent itemsets. For applying improved Apriori algorithm to E-commerce, there will be a need to develop a shopping site because when customers visit the shopping site the system will automatically find out their next purchasing goods based on goods already available in their shopping basket. So, it will save time and increases the efficiency and provides more benefit. The customers could easily generate association rules with the help of improved Apriori algorithm and suggests useful products to customers within a reasonable time. The tool used for developing the site is Macromedia Dreamweaver 8; database management tool is Microsoft SQL Server 2000 and Web server is Tomcat 6.0. According to experimental results, it was shown that improved Apriori Algorithm when compared with traditional Apriori algorithm is more efficient.

3.4 Algorithm based on grid computing

3.4.1 Enlightenment

The authors **Mrs. R. Sumithra and Dr (Mrs). Sujni Paul (2010)** [5], have developed a new method distributed apriori association rule to recover the limitations of classical apriori algorithm. In this paper, main focus is on eliciting the knowledge. The implementation of both algorithms is shown using concept of grid computing. Grid computing is a form of distributed computing that enables the developers to work together on a single task at same time. Grid computing has capability to increase the efficiency and decreases the cost of computing networks by optimizing the resources. It is best for large workloads. By comparing, it was shown that distributed apriori association rule on grid based environment is better than classical apriori algorithm. The future scope is that

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

knowledge could be extracted in parallel to produce more optimized result.

3.5 Improvement based on count-based method and generation record

3.5.1 Enlightenment

In this paper, on the basis of analysis and study of previous efforts that researcher have applied, an improved Apriori algorithm (IAA) for association rule mining is proposed by **Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu and Wenbao Jiang (2009)** [6]. The IAA conquer the limitations of original apriori algorithm. The IAA introduces a new count based approach which is used to elicit the redundant candidate itemsets and uses generation record to reduce the scanning time of database. The IAA meets the various challenges correlates with association rule mining such as reducing I/O cost, improving efficiency and increasing processing speed. From the experimental results, it was proved that IAA is better than original Apriori algorithm because IAA counts each candidate itemsets once. But C-R problem exists in IAA which could not be solved where C represents condition item sets and R represents result item sets.

3.6 Algorithm based on concept lattice

3.6.1 Enlightenment

The authors **XUE-GANG HU, DE-XING WANG, XIAO-PING LIU, JUN GUO and HAO WANG (2004)** [7] introduced a new Quantitative extended concept lattice (QECL) method which is based on the concept of lattice to mine association rule .Concept lattice is a type of induced lattice which comes into being based on the partial ordering relation between O, D, R where O represents set of objects, D represents set of attributes and R is the binary relationship between O and D. This method is better than existing Apriori algorithm because we can mine association rules easily without finding frequent itemsets and easily obtained strong association rules within a reasonable amount of time. It eliminates the burden of scanning database repeatedly. As a result, the efficiency and accuracy of mining association rules are improved.

3.7 Algorithms based on function interest

3.7.1 Enlightenment

The classical Apriori algorithm generally focuses on only two aspects: minimum support and minimum confidence to generate strong association rule. There may be a chance that sometimes it is necessary to determine strong association rules for making decisions and sometimes less strong rules are required. To fulfill this condition, the authors **WEI-MIN MA and ZHU-PING LIU (2008)** [8], proposed two revised algorithms based on Apriori: AMS (Algorithm for mining stronger association rules) and AMLS (Algorithm for mining less strong association rules) which focus on three aspects: minimum support, minimum confidence and minimum interest. These algorithms works in the form of matrix to decrease the scanning time of database. On the basis of comparison of classical Apriori algorithm with AMS and AMLS, it was proved that AMS and AMLS are better than classical Apriori algorithm.

3.8 Improvement by reducing redundant operation

3.8.1 Enlightenment

In this paper, the classical Apriori algorithm is discussed. The faults that are present in classical Apriori algorithm such as consuming more time to generate candidate itemsets, scanning the database repeatedly are also discussed. In order to remove all these faults, the authors **Yanfei Zhou, Wanggen Wan, Junwei Liu and Long Cai (2010)** [9], described an improved Apriori algorithm. This improved Apriori algorithm consists of three segments: First is decreasing number of judgements during the time of generating frequent candidate itemsets. Secondly, pruning frequent itemsets. Finally, optimize the database. The improved Apriori algorithm was compared with classical Apriori algorithm on the basis on different support degree, different number of trading services and different number of items. From this comparison, it was proved that improved Apriori algorithm improves performance, increases efficiency, and reduces the redundant operation while producing frequent itemsets and strong association rules.

4. CONCLUSION

Association rule mining is a major technique in the area of data mining. Apriori algorithm is one of the earliest algorithm of association rule mining. In this paper, we have presented a survey of most recent work that has been done in Association rule based mining using Apriori algorithm. After doing survey of various algorithms, we can make a conclusion that in improved Apriori algorithm the main focus is on generating less candidate sets which contains all frequent items within a reasonable amount of time.

ACKNOWLEDGMENT

The author is extremely grateful to Asst. Professor Arun Malik for his encouragement to complete this paper.

REFERENCES

[1] "Introduction to Data Mining and Knowledge Discovery" Third Edition by Two Crows Corporation ISBN: 1-892095-02-5

[2] Wanjun Yu, Xiaochun Wang and Fangyi Wang, Erkang Wang, Bowen Chen, "The Research of Improved Apriori Algorithm for Mining Association Rules" 2008 11th IEEE International Conference on Communication Technology Proceedings, 978-1-4244-2251-7/08/$25.00 ©2008 IEEE.

[3] 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 ISSN 2250-3153.

[4] Shuo Yang, "Research and Application of Improved Apriori Algorithm to Electronic Commerce" 2012 11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science, 978-0-7695-4818-0/12 $26.00 © 2012 IEEE DOI 10.1109/DCABES.2012.51

[5] Mrs. R. Sumithra, Dr (Mrs). Sujni Paul, "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery" 2010 Second International conference on Computing, Communication and Networking Technologies, 978-1-4244-6589-7/10/$26.00 ©2010 IEEE.

[6] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang, "An Improved Apriori-based Algorithm for Association Rules Mining"

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 978-0-7695-3735-1/09 $25.00 © 2009 IEEE DOI10.1109/FSKD.2009.193.

[7] XUE-GANG HU, DE-XING WANG, XIAO-PING LIU, JUN GUO, HAO WANG, "THE ANALYSIS ON MODEL OF ASSOCIATION RULES MINING BASED ON CONCEPT LATTICE AND APRIORI ALGORITHM" Proceedings of the Third International Conference on Machine Leaming and Cybernetics, Shanghai, 26-29 August 2004, 0-7603-8403-2/04/$20.00 @2W4 IEEE.

[8] WEI-MIN MA, ZHU-PING LIU, "TWO REVISED ALGORITHMS BASED ON APRIORI FOR MINING ASSOCIATION RULES", Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008, -1-4244-2096-4/08/$25.00 ©2008 IEEE.

[9] Yanfei Zhou, Wanggen Wan, Junwei Liu, Long Cai, "Mining Association Rules Based on an Improved Apriori Algorithm", 978-1-4244-585 8- 5/10/$26.00 ©2010 IEEE.

BIOGRAPHIES

Shilpi Singla, received her B.Tech degree in Computer Engineering and currently doing Mtech from Lovely Professional University, Jalandhar in department of Computer Science and Technology. Currently, doing research on Apriori Algorithm to generate frequent itemsets within a reasonable amount of time. Her area of interest is in Data Mining.

Arun Malik, Assistant Professor, Dept. of Computer Science and Technology, Lovely Professional University, Jalandhar, India