



The Research on Top Down Apriori Algorithm using Association Rule

Shikha Maheshwari

Department of Computer Science & Engineering,
Shri Vaishnav Institute of Technology & Sciences,
Indore (M.P.), India

Pooja Jain

Asst.Prof
Shri Vaishnav Institute of Technology & Sciences,
Indore (M.P.), India

Abstract: Association rule mining finds interesting association or correlation relationships among a large set of data items, which is an essential task of data mining. Apriori is an important algorithm for mining frequent itemsets for Boolean association rules. Firstly, the concept of association rules is introduced and the classic algorithms of association rule are analyzed. Classic Apriori algorithm not used for large dataset because number of iteration is more. And improved top down approach is useful for large amount of data set. This algorithm uses less space, less number of iteration.

Keywords: Data mining, Association rules, Apriori algorithm, frequent item

I. INTRODUCTION

Data mining is the main part of KDD. Data mining generally involves four classes of task; regression, classification, clustering, and association rule learning. Data mining refers to discover knowledge in huge amounts of data. It is a precise discipline that is concerned with analyzing observational data sets with the objective of finding unsuspected relationships and produces a summary of the data in novel ways that the owner can understand and use. Data mining is a kind of process of decision support. It gets the potential and useful information and acknowledges from practical application data which is large, incomplete, noisy, ambiguous and random. Data mining relates to extracting a large of data from database, transforming, analyzing and modeling handling these data, and withdrawing the critical data to aid decision making[1]. Association rule mining has been well studied in data mining, mainly for basket transaction data analysis.

Association rule mining finds interesting association or correlation relations among a large set of data items[2,3]. In this paper Association rule mining algorithm- Apriori algorithm which is commonly used in data mining is mainly discussed.

II. ASSOCIATION RULE MINING

This section introduces the basic concepts of frequent pattern mining for discovery of interesting associations and correlations between itemsets in transactional and relational database. Association rule mining can be defined formally as follows:

$I = \{i_1, i_2, i_3, \dots, i_n\}$ is a set of items, such as products like (computer, CD, printer, papers ...and so on). Let DB be a set of database transactions where each transaction T is a set of items that $T \subseteq I$. Each transaction is associated with unique identifier, transaction identifier (TID). Let X, Y be a set of items an association rule has the form $X \rightarrow Y$ where $X \cap Y = \emptyset$. X is called the antecedent and Y is called the consequent of the rule where, X, Y is a set of items called as an itemset or a pattern[4]. Let $\text{freq}(X)$ be the number of rows (transactions) containing X itemset in the given database. The support of an itemset X is defined as the fraction of all rows containing the itemset.

The support of an association rule is the support of union of X and Y , [5] i.e.

$$\text{Support}(X \rightarrow Y) = (X \cup Y)$$

The confidence of an association rule is defined as the percentage of rows in D containing itemset X that also contain itemset Y [6], i.e.

$$\text{Confidence}(X \rightarrow Y) = P(X/Y)$$

III. CLASSICAL APRIORI ALGORITHM USING BOTTOM UP APPROACH

Apriori is a classic algorithm for learning association rules in data mining. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules[7,8]. The Apriori algorithm is a classical data mining method for association rule discovery typically applied to market basket data, such as the study of what products tend to be purchased together in an on-line market place (e.g. Amazon etc).

There are two properties: Generate all frequent item sets: One frequent itemsets is an itemsets which support higher than minimum support (min_sup);

Generate all confident association rule from frequent itemsets: One confident association rule is a rule which confidence level greater than minimum confidence (minconf)[9] The properties are used in Apriori algorithm to scanning the database, resulting in Boolean association rules frequent itemsets. Specifically, Apriori uses an iterative search method layer by layer, where k -dimensional itemsets are used to explore $(k+1)$ -dimensional itemsets. First, the set of frequent 1-

dimensional itemsets is found and denoted L_1 , Next, L_1 is used to find L_2 , the set of L_2 frequent 2-itemsets, which is used to find L_3 , and so on until no more frequent k -dimensional itemsets can be found. Finally, getting the rules from large set of data items. How L_{i-1} is used to find L_i is consisting of two step process, join and prune actions as followed

1. **The join step:** Join L_{k-1} with itself, than combine the same extension item appeared to generate a possible candidate k -dimensional itemsets, this set of candidates is denoted C_k , $C_k \supseteq L_k$.
2. **The prune step:** Scan the database to determine the count of each candidate in C_k . When the count is less than the minimum support count, it should be delete from the candidate itemsets. If any $(k-1)$ dimensional subset of a candidate k -dimensional itemsets is not in L_{k-1} , the candidate cannot be frequent either, after this we can get the k dimensional itemsets, which is denoted L_k .

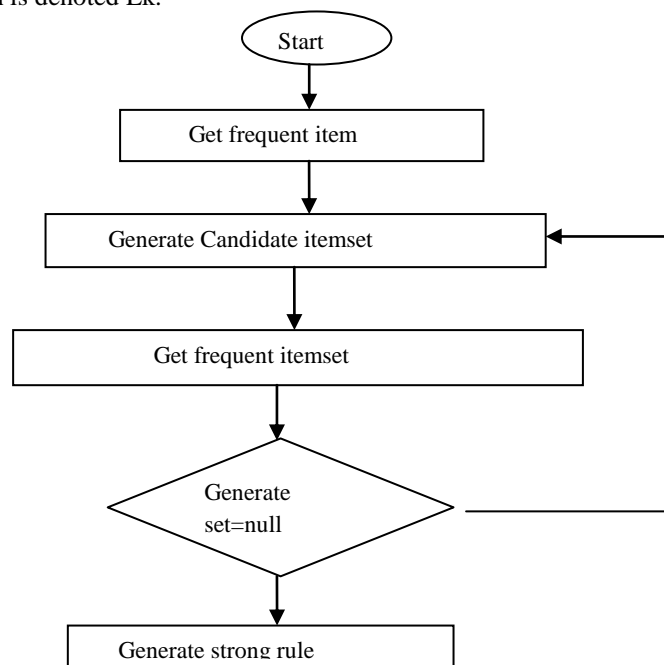


Figure 1 Flow chart of classical algorithm

IV. IMPROVED APRIORI ALGORITHM USING TOP DOWN APPROACH

The classical Apriori algorithm follows bottom up approach. The improved algorithm uses top down approach, where in the rules are generated by avoiding generation of un necessary patterns. The major advantage of this approach is that, the number of data base scan reduces and times also reduce.

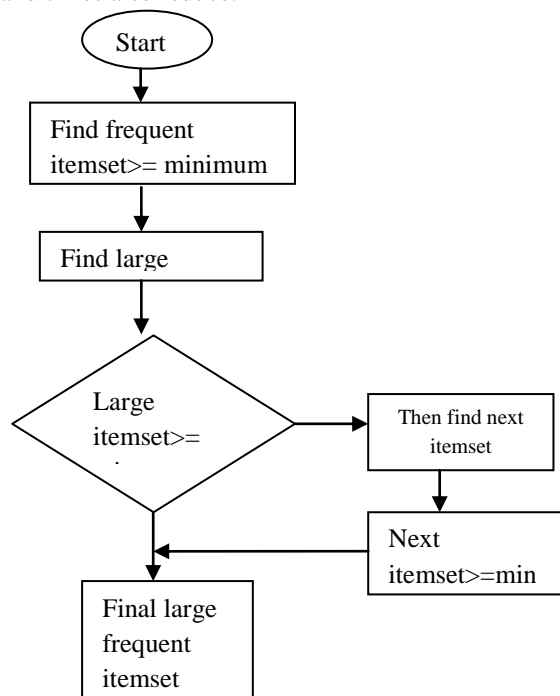


Figure 2 Flow chart of improved algorithm

A. Improved Apriori algorithm illustration

Generate frequent item sets for example, a transaction database as shown in table 1, in which there are nine affairs, that is, $|D|=9$. Apriori assumes that the items of the affair are stored by the order of dictionary. Minimum support threshold $\text{min support} = 2/9 = 22\%$, minimum support count is 2.

Step1 In this algorithm firstly count Minimum support of all data item set.

T id	Item set
T1	A,B,E
T2	B,D
T3	B,C
T4	A,B,D
T5	A,C
T6	B,C
T7	A,C
T8	A,B,C,E,
T9	A,B,C
T10	F

Item Set	Support count
{A}	6
{B}	7
{C}	6
{D}	2
{E}	2
{F}	1

Step2

In this step transaction T 10 delete because itemset F is not belong to min support. Next In new database table search the large data item set.

T id	Item set
T1	A,B,E
T2	B,D
T3	B,C
T4	A,B,D
T5	A,C
T6	B,C
T7	A,C
T8	A,B,C,E,
T9	A,B,C

itemset	Support count
{A,B,C,E}	1

Step 3 :

The large data itemset {A, B, C, D} item set not belong to support count. So it is not frequent item .then next large itemset scan in database table.

Three large itemset scan but {A, B, D} are not frequent item set because it not belong min support. Finally {A, B, C},{A,B,E} are frequent itemset.

Item set	Support count
{A,B,E}	2
{A,B,D}	1
{A,B,C}	2

Item set	Support Count
{A,B,C}	2
{A,B,E}	2

Generation of frequent item set

V. EXPERIMENT RESULT

We compared the performance of our improved Apriori algorithm with the classical Apriori algorithm. The experimental platform is Intel Core(TM) i3 processor. Minimum requirements 1 GB RAM, Windows Operating System7. . The algorithm develop in java development kit version 1.7 .Integrated development environment net beans 7.3.1.required library jcommon 1.0.20.jar, jfreechart-1.0.16.jar

The first experiment is done under the environment that the count of transaction is different but the same support count 40% (show in Figure 3).From Figure 3, it is easy to get the conclusion the improved Apriori algorithm use less time and memory.

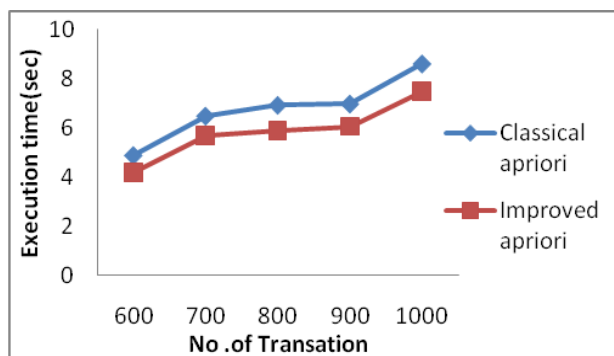


Figure 3 Relationship of the Number of Transaction with time consumption

Another experiment test is relationship of support count with time consumption

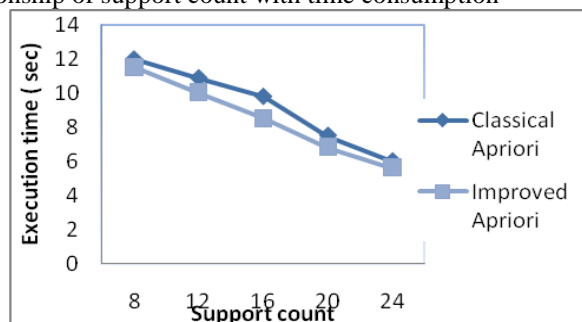


Figure 4 Relationship of support count with time consumption.

VI. CONCLUSION

In this paper, the improved version of Apriori algorithm is proposed to overcome the deficiency of the basic Apriori algorithm. The basic Apriori algorithm follows bottom up approach which suffers from increased number of database scan. The novel proposed method follows top down approach which reduces the number of database scans. The improved version Apriori algorithm is more efficient which takes less time.

REFERENCES

- [1] Lanfang Lou, Qingxian Pan, Xiuqin Qiu, New Application of Association Rules in Teaching Evaluation System, International Conference on Computer and Information Application IEEE2010
- [2] Luo Fang, Qiu Qizhi, The Study on the Application of Data Mining Based on Association Rules, International Conference on Communication Systems and Network Technologies IEEE 2012
- [3] A. Savasere, E. Omiecinski, and S. Navathe. "An efficient algorithm for mining association rules in large databases". In Proc. Int'l Conf. Very Large Data Bases (VLDB), Sept. 1995, pages 432-44[4] Karthiya Banu.R, Dr. Ravanan.R, Gopal.J, Analysis an implementation of association rule mining 978-1-4244-8594-9/10, 2010 IEEE
- [5] Yanfei Zhou, Wanggen Wan*, Junwei Liu, Long Cai, Mining Association Rules Based on an Improved Apriori Algorithm, 978-1-4244-585 8- 5/10 ©2010 IEEE
- [6] Huiying Wang, Xiangwei Liu, The research of improved association rules mining Apriori algorithm, international conference on Fuzzy System and Knowledge Discovery IEEE 2011
- [7] Mining Association Rules Based on an Improved Apriori Algorithm Yanfei Zhou, Wanggen Wan*, Junwei Liu, Long Cai, 978-1-4244-585 8- 5/10/ ©2010 IEEE
- [8] Agrawal.R and Srikant.R. "Fast algorithms for mining association rules". In Proc Int'l Conf. Very Large Data Bases (VLDB), Sept. 1994, pages 487-499
- [9] RuPeng Luan*, SuFen Sun, JunFeng Zhang, Feng Yu, Qian Zhang, A Dynamic Improved Apriori Algorithm and Its Experiments in Web Log Mining, 9th International Conference on sFuzzy Systems and Knowledge Discovery 978-1-4673-0024-7/10 2012 IEEE pp1261-64
- [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] Ya-Han Hu, Yen-Liang Chen. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism Decision Support System, 2006, pp.1-2
- [12] Agrawal, R., Imielinski, T., and Swami, A. N. Mining Association Rules Between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD, International Conference on Management of Data, pp.207-216, 1993.
- [13] Agrawal. R. and Srikant. R., Fast Algorithms for Mining Association Rules, Proceedings of 20th International Conference of Very Large Data Bases. pp.487-499, 1994.