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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Comparison of Data Mining Algorithms for Effective Performance

Nikhil Belwate

Software Consultant, Dept. Mobility, Quinnox Consultancy Services, India

ABSTRACT: Data mining on large relational databases has gained popularity and its significance iswell recognized. However, the performance of SQL based data mining is known to fall behind specialized implementation since the prohibitive nature of the cost associated with extracting knowledge, as well as the lack of suitable declarative query language support. Frequent pattern mining is a foundation of several essential data mining tasks. These facts motivated us to develop original SQL-based approaches for mining frequent patterns.This paper compares three algorithm based on the parameter assize of the database, efficiency and speed

KEYWORDS: Database, Data mining, FP Growth algorithm, Apriori algorithm, sorting

I. INTRODUCTION

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is Market Basket Analysis.

Data mining is the automated discovery of non-trivial, implicit, previously un-known, and potentially useful information or patterns embedded in databases Brieflystate, it refers to extracting or mining knowledge from large amounts of data The motivation for data mining is a suspicion that there might be nuggets of useful information hiding in the masses of unanalyzed or under analyzed data, and there-fore methods for locating interesting information from data would be useful. From the beginning, data mining research has been driven by its applications. While and industries have long recognized the benefits of data mining, data mining techniques can be effectively applied in many areas and can be performed on a variety of data stores, including relational databases, transaction databases and data warehouses.

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Data mining (DM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering

Data mining techniques are the result of a long process of research and product development. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

II. LITERATURE SURVEY

1. Mining Efficient Association Rules Through Apriori Algorithm Using Attributes
   Authors MamtaDhanda, SonaliGuglani, Gaurav Gupta, RIMT – IET have define Apriori algorithm is not an efficient algorithm as it is a time consuming algorithm in case of large dataset. This paper illustrate the apriori algorithm disadvantages and utilization of attributes which can improve the efficiency of apriori algorithm. The working of algorithm is well explained with respect to example.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

2. Research on the FP Growth Algorithm about Association Rule Mining
This is web based research article written by Wei Zhang and published on www.researchgate.net. The article describe research on improving the mining performance and precision is necessary, so many focuses of today on association rule mining are about new mining theories, algorithms and improvement to old methods Research on the FP Growth Algorithm about Association Rule Mining.

3. Association rule mining using improved FP-Growth Algorithm (IJFTRE)
This paper is presented by Palak Patel and Purnima Gandhi and the describe FP-Growth algorithm construct conditional frequent pattern tree and conditional pattern based from database which satisfies the minimum support. This information is also use in mining algorithms comparison

4. The Research of Data Mining Algorithm Based on Association Rules (International conf. 2012)
Author Lei Chen, china prepare document for in-depth study of the existing data mining and association rule mining algorithms. The brief details on Association rules provide direction to work in mining technologies.

III. PURPOSE OF SYSTEM

It has become increasingly necessary for users to utilize automated tools in find the desired information resources, and to track and analyse their usage patterns**.** Association rule mining is an active data mining research area. However, most ARM algorithms cater to a centralized environment. Distributed Association Rule Mining (D-ARM) algorithms have been developed. These algorithms, however, assume that the databases are either horizontally or vertically distributed.

IV. APRIORI ALGORITHM AND ADVANCE APRIORI

Apriorialgorithm is the most classical and important algorithm for mining frequent itemsets. Apriori is used to find all frequent itemsets in a given database DB. The key idea of Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-itemsets are used to explore (k+1)-itemsets. In the beginning, the set of frequent 1-itemsets is found. The set of that contains one item, which satisfy the support threshold, is denoted by L1. In each subsequent pass, we begin with a seed set of itemsets found to be large in the previous pass. This seed set is used for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually large (frequent), and they become the seed for the next pass.

1. Apriori Algorithm:

• Item: article in the basket.

• Itemset**:** a group of items purchased together in a single transaction.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 2}	1
{1 3}*	2
{1 5}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2

C_3

Itemset	Support
{2 3 5}*	2

C'_2

100	{1 3}
200	{2 3}, {2 5}, {3 5}
300	{1 2}, {1 3}, {1 5}, {2 3}, {2 5}, {3 5}
400	{2 5}

C'_3

200	{2 3 5}
300	{2 3 5}

Fig 1.1: Table Example

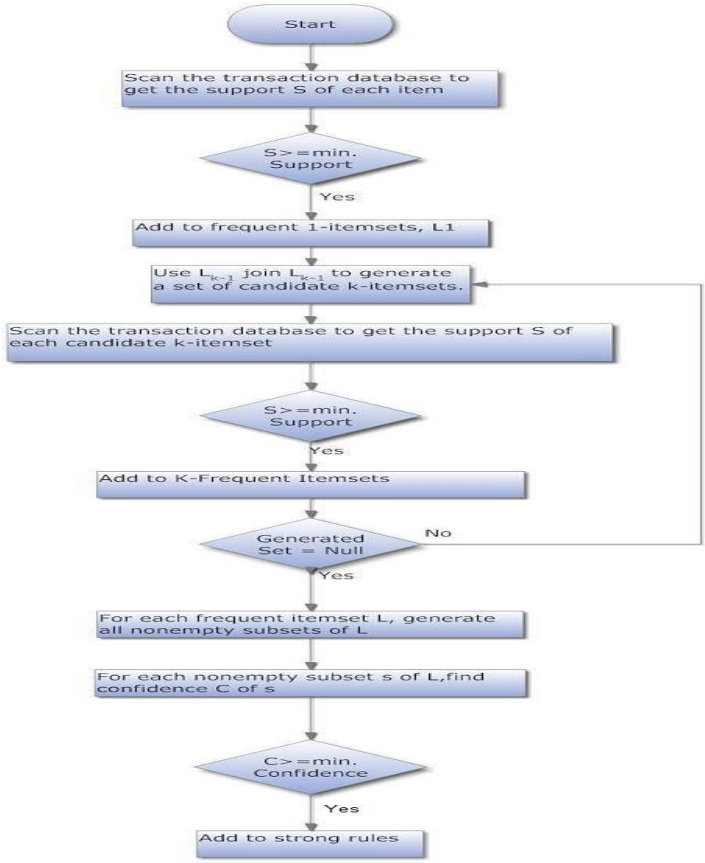


Fig 1.2: Flow chart

2. **Improved Algorithm**:

The improved algorithm is described in following steps:

Input:

- D, a database of transaction
- Min_sup, the minimum support count threshold

1. In the first iteration of the algorithm, each item is a member of the set of candidate 1-itemset C1.The algorithm simply scans all of the transaction to count the number of occurrences of each item.
2. The set of frequent item sets, L1, is determine by Comparing the candidate count with minimum support count which contains candidate 1-itemsets satisfying minimum support.
3. To generate the set of frequent 2-itemsets, L2, the algorithm generate a candidate set of 2-itemset and then the transactions in D are scanned and the support count of each candidate item set in C2 is accumulated and then repeating the step 2.
4. Then D2 was determined from L2.
5. Generate C3 candidates from L2 and scan D2 for count of each candidate and then repeating step 2.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

6. At the end of the pass, determining which of the candidate item sets are actually large, and they become the seed for the next pass.

7. This process continues until no new large item sets are found.

The improved Apriori algorithm reduce the number of database scanning and the redundancy while generating subtests and verifying them in the database. Because of which this algorithm takes less time for generating frequent item set as compared to classical Apriori algorithm.

V. FP GROWTH ALGORITHM

FP-growth algorithm is an efficient method of mining all frequent item sets without candidate‘s generation. FP-growth utilizes a combination of the vertical and horizontal database layout to store the database in main memory. Instead of storing the cover for every item in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item. This new data structure is denoted by FP-tree (Frequent-Pattern tree) (Hanet al 2000). Every node additionally stores a counter, which keeps track of the number of transactions that share the branch through that node. Also a link is stored, pointing to the next occurrence of the respective item in the FP-tree, such that all occurrences of an item in the FP-tree are linked together. Additionally, a header table is stored containing each separate item together with its support and a link to the first occurrence of the item in the FPtree. In the FP-tree, all items are ordered in support descending order, because in this way, it is hoped that this representation of the database is kept as small as possible since all more frequently occurring items are arranged closer to the root of the FP-tree andthus are more likely to be shared.

The algorithm mine the frequent item sets by using a dividend- conquer strategy as follows: FP-growth first compresses the database representing frequent item set into a frequent-pattern tree, or FP-tree, which retains the item set association information aswell. The next step is to divide a compressed database into set of conditional databases (a special kind of projected database), each associated with one frequent item. Finally, mine each such database separately. Particularly, the construction of FP-tree and the mining of FP-tree are the main steps in FP-growth algorithm

Algorithm:

Input: DB: transaction database;

Min_sup: the minimum support threshold

Output: frequent itemsets

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Example

TID	Items bought
100	{f, a, c, d, g, i, m, p}
200	{a, b, c, f, l, m, o}
300	{b, f, h, j, o}
400	{b, c, k, s, p}
500	{a, f, c, e, l, p, m, n}

Item	frequency
f	4
c	4
a	3
b	3
m	3
p	3

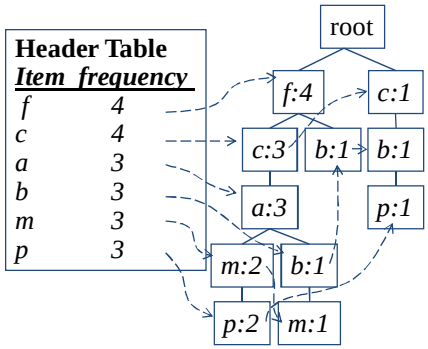
- For *support = 50%*
- Step 1: Scan DB find frequent 1-itemsets
- Step 2: Order them in descending order
- Step 3: Scan DB and construct FP-Tree

May 21, 2002

Fig 2.0: Example

Example – Conditional Pattern Base

- Start with the frequent header table
- Traverse tree by following links from frequent items
- Accumulate prefix paths to form a conditional pattern base



***Conditional* pattern bases**

item	cond. pattern base
c	f:3
a	fc:3
b	fca:1, f:1, c:1
m	fca:2, fcab:1
p	fcam:2, cb:1

May 21, 2002

Fig 3.0: Condition Pattern

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

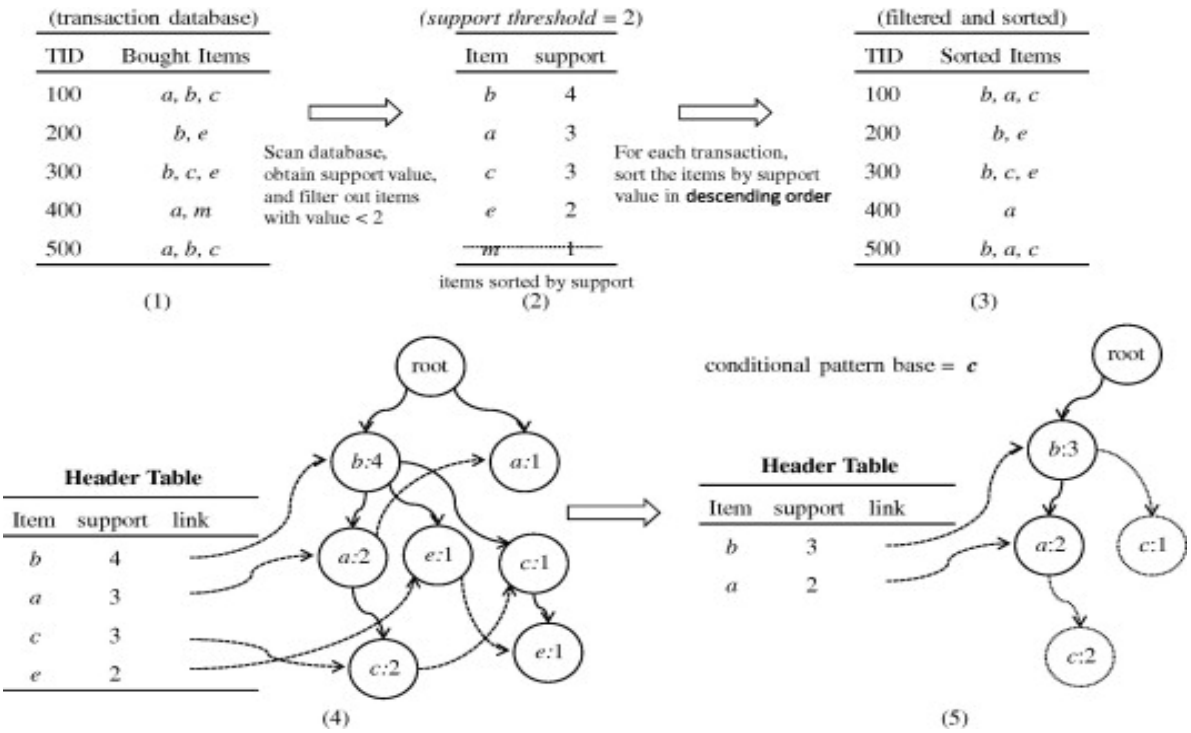


Fig 4.0: Calculation

One of the advantages of *FP-growth* over other approaches is that it constructs a highly compact *FP-tree*, which is usually substantially smaller than the original database and thus saves the costly database scans in the subsequent mining processes. Although an *FP-tree* is rather compact, it is unrealistic to construct a main memory-based *FP-tree* when the database is large. The *FP-tree* consists of a there data structure in which each node stores an item as well as a counter, also a link pointing to the next occurrence of the respective item in the *FP-tree.* Additionally a header table is stored containing each separate item together with its support and a link to the ¯rest occurrence of the item in the *FP-tree.* In *FP-growth*, the cover of an item is compressed using the linked list starting from its node-link in the header table, but every node in this linked list needs to store its label, a counter, a pointer to the next node, a pointer to its branches and a pointer to its parent. Therefore, the size of such a complex *FP-tree* should be large. Table shows for the total number of nodes in *FP-growth* and the compression rate of the *FP-tree*. However using RDBMSs provides us the benefits of using their buyer management systems specially developed for freeing the user applications from the size consider-actions of the data. And moreover, there are several potential advantages of building

VI. COMPARING ALGORITHMS

We have discussed different algorithms for association rule mining on different size of database. In the first algorithm we have seen the improved Apriori algorithm which takes less time for generating frequent item set. In the second algorithm we have seen the 'Feature Based Association Rule Mining Algorithm' (FARMA) which is efficient than other algorithm and it speed up the data mining process. Third algorithm we have seen it is better and faster with respect to all parameters.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

No.	Parameters	Apriori	Improved Apriori Algorithm	Fp-growth
1.	Database Size	Small	Large	Very Large(Distributed)
2.	Database Scan	N times	N times	Only twice.
3.	Candidate set generation	Large number of candidate sets are generated	.candidate sets are generated but less than aprori	Candidate set is not generated..
4.	Memory requirement	Large	Large	Less
5.	Speed	Slow	Faster than apriori	Fastest
6.	Cost	More i/o cost	More i/o cost	Less i/o cost
7.	Output	Candidate set	Candidate set	Fp-tree

VII. CONCLUSION AND FUTURE WORK

The association rules play a major role in many data mining applications, trying to find interesting patterns in data bases. In order to obtain these association rules the frequent sets must be previously generated. The most common algorithms which are used for this type of actions are the Apriori and FP-Growth. The performance analysis is done by varying number of instances and confidence level.

The efficiency of both algorithms is evaluated based on time to generate the association rules. From the experimental data presented it can be concluded that the FP-growth algorithm behaves better than the Apriorialgorithm. There are several ways to improve the database access of Apriori algorithm thereby improving also the efficiency of the execution. Based on the modified code, set size and set size frequency were introduced.

These factors helped in a more rapid generation of possible association of frequent items. In terms of database passes, the modified apriori provides less database access compared with the original one that makes its execution faster. Suggestions in finding other ways to generate combinations are encourage. Currently, further research in finding a faster way of pruning candidate keys is undergoing in finding the ideal starting size of combination size.

REFERENCES

1. Agrawal R.,Imielinski T., Swami A.N.,"Mining association rules between sets of items in large databases", In Proceedings ACM SIGMOD International Conference on Management of Data, Vol. 22, No. 2, of SIGMOD Record, Washington, pp. 207–216
2. Agrawal R., SrikantR.,"Fast algorithms for mining association rules"
3. Fayyad U. M., Piatetsky-Shapiro G., Smyth, P.,"From data mining to knowledge discovery in databases", AI MagazineVol. 17, No. 3, pp. 37-54
4. D.W. Cheung, et al., "A Fast Distributed Algorithm for Mining Association Rules," Proc. Parallel and Distributed Information Systems, IEEE CS Press, 1996,pp. 31-42
5. M.J. Zaki and Y. Pin , "Introduction: Recent Developments in Parallel and Distributed Data Mining," J. Distributed and Parallel Databases , vol. 11, no. 2, 2002,pp. 123-127

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

6. D.W. Cheung , et al., "Efficient Mining of Association Rules in Distributed Databases, "IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, 1996,pp. 911-922
7. Agrawal R.,Imielinski T., Swami A.N.,"Mining association rules between sets of items in large databases", In Proceedings ACM SIGMOD International Conference on Management of Data, Vol. 22, No. 2, of SIGMOD Record, Washington, pp. 207–216
8. James R. Groff's "SQL, the complete reference" pp. 189-255
9. A. Schuster and R. Wolff, "Communication-Efficient Distributed Mining of Association Rules," Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, 2001, pp. 473-484.
10. Han, J., Kamber, M.,"Data mining concepts and techniques"
11. Agrawal R.,Imielinski T., Swami A.N.,"Mining association rules between sets of items in large databases", pp. 207–216

BIOGRAPHY

Nikhil Sharad Belwateis a consultant and Research Assistant in the working in mobility department of Quinnox consultancy services. He received Master of Management Studies(MMS) and B.E. (Information technology) from University of Mumbai, India. His research interests are Computer programming, IoT, and artificial intelligence etc.