International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 01 | Jan-2018 www.irjet.net p-ISSN: 2395-0072

Comparative analysis of Apriori and Apriori with hashing algorithm

Aesha J Doshi[1], Barkha Joshi[2]

[1,2] *Department of Computer Engineering , Sardar Vallabhbhai patel Inst. of Tech, Vasad, India*

Abstract: Data mining is a powerful technology to discover information within the large amount of the data. It is considered as an important subfield in knowledge management. Research in data mining continues growing in various fields of organization such as Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition, business, education, medical, scientific etc. Data mining algorithms are used to retrieval data from large database very efficiently. Association rule mining is used to find the frequent item set from the large database based on the relation. In this paper we are try to compare apriori and apriori with hashing algorithm and try to find which algorithm is better to provide accurate result in less amount of time.

Key-word – Association rule, Apriori algorithm, Apriori with hashing algorithm.

INTRODUCTION:

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. it involves database and data management aspects, datapreprocessin, model and inference considerations,interestingnessmetrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori also uses "Top down" approach, where maximal candidate item set and search item set used. The not frequent item set is remove from maximal candidate item set and at last final result into maximal candidate item set. The frequent item set generated from maximal candidate item set.

Hashing technique is used to improve the efficiency of the apriori algorithm. it work by creating a dictionary (hash table) that stores the candidate item sets as keys, and the number of appearances as the value. Initialization start with zero and Increment the counter for each item set that you see in the data.

Association rule

Association rule is important aspect of data mining. It is used to discover frequent pattern, Association, Connection or on the other hand casual structures among sets of products in value-based databases and other data stores. The volume of information is expanding significantly as the information produced by day by day exercises. Subsequently, mining association rules from bulky data in DB is best interested area for many industries. The strategies for finding association rules from the information have generally centered around recognizing connections between things, which demonstrate customer behavior[4]. For some applications, it is hard to discover strong relationship among information things at low or on the other hand crude levels of abstraction because of the deficiency of information at those levels. The Strong association can found at high level of abstraction represent sensible knowledge[1].

Suppose I={I1, I2, I3….. I_n} be the set of items. Let D is set of database transaction where each transaction T is set of item. Each transaction has identifier TID. Now A is the set of item in particular transaction.

There are two important measures of association rule that is support and confidence.

Support(S): It can be define as it is probability or percentage of transaction in D that contain $A \cup B$ or in other word we can say that it is ratio of occurrences of items and total num of transactions.

$$\text{Support } (A \rightarrow B) = P(A \cup B)$$

$$S\ (A \rightarrow B) = \frac{\text{Amount of transaction A \& B}}{\text{Total Transaction}}$$

Confidence(C): It can be define as it is conditional probability or percentage of the transaction containing A also contain B.

$$\text{Confidence}(A \rightarrow B) = P(B|A)$$

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

The rules that satisfy both min_sup(Minimum support threshold) min_conf(Minimum confidence threshold) that are strong association rule. Both values between 0% and 100%

Min_Sup : As we know that It is prespecified. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is known as the support count of that item set. so the item set whose support satisfy the min_sup value is known as frequent item set.

Min_conf : It is also prespecified and by satisfying Min_conf threshold generated rules are become strong.

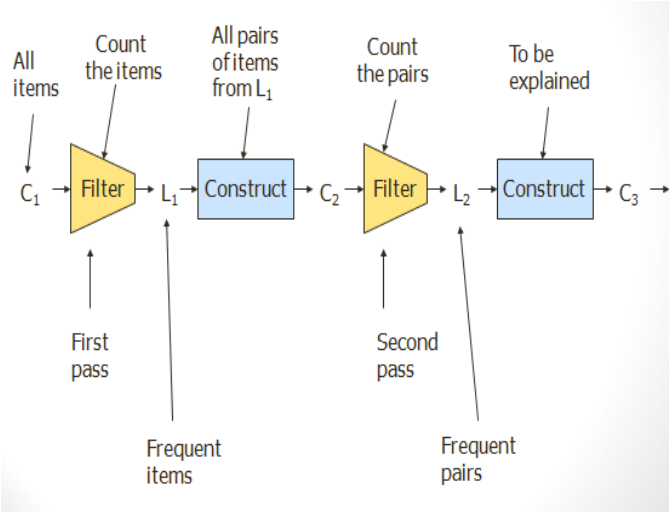
Ariori Algorithm

Market container examination is finished by many organizations with a specific end goal to recover item sets that are visit and together buy by client and furthermore discover client buying habits. Apriori algorithm is a widely used technique. it is used to find those combinations of item sets.apriori work on bottom-up manner. it work on 2 steps scan and prune steps. In scan step it generate candidate item set and in prune step it reduce size of candidate item set by reducing itemsets whose support below the min_sup threshold.

Apriori is known as bottom-up breadth first search method. apriori algorithm is iterative approach level-wise search, where k-itemset is used to generated (k+1) item set. Apriori algorithm uses Prior knowledge of frequent item set property Apriori algorithm based on 2 Pass process.

Pass1:- In Pass-1 system read the basket and count the number of occurrences of each items. It requires memory proportional to the no of items.T will be the transaction in the database. Each item support count is match with min_sup value if it satisfy min_sup then item will become frequent item set. L_1 table created with generated frequent item set.

Pass-2 :- In Pass-2 system read basket again and count pair of items which is generated from L_1 table. Again support of pair of item set match with min_sup count if it satisfy min_sup value then consider to be frequent. L_2 table create with this frequent item set. It require memory proportional to square of frequent item only plus list of frequent items.



Support count = of itemset is the number of transactions that contain the itemset.

Frequent itemset = If the support of an itemset satisfies a pre-specified minimum support threshold, then it is a frequent itemset.

There is some limitation of apriori. As we realize that the apriori calculation is work on bottom-up direction so it begin from the smallest arrangement of frequent item set and move upward way until the point when it achieves huge incessant item set. In the event that the span of frequent item set is vast at that point number of times that calculation goes through database is equivalent with estimate.At the point when item set is too expansive then calculation work slower and set aside greater opportunity to give the outcome so result is into the performance hit.

There are many methods to improve the efficiency of apriori algorithm. There is Hash-based technique (hashing itemsets into corresponding buckets.

Apriori with hashing Algorithm

As we know that apriori algorithm has some weakness so to reduce the span of the hopeful k-item sets, Ck hashing technique is used. Our hash based Apriori execution, utilizes the data structure that specifically speaks to a hash table. Specifically the 2-itemsets, since that is the way to enhancing execution. This calculation utilizes a hash based procedure to minimize the quantity of applicant item sets created in the 1st pass. It is guaranteed that the quantity of item sets in C2 produced utilizing hashing can be smalled, so that the output required to decide L2 is more efficient.

For instance, while scanning every transaction in the database to create the Frequent1-itemsets,L1, from the candidate 1-itemsets in C1, we can produce the majority of the 2-itemsets for every transaction, hash(i.e) map into the diverse bucket of a hash table structure, and increment the complementary bucket count. A 2-itemset whose

complementary bucket count in the hash table is below the min_sup threshold cannot be frequent and thus should be reduce from the candidate set. So hash based apriori reduce the no of candidate k-item set.

Steps:

1. Scan database transaction.generate frequent-1 item set. Then after generate frequent-2 item set.

2. Let take hash table of size 7.

3. For each bucket appoint a candidate sets utilizing the ASCII estimations of the itemsets.

4. Each bucket in the hash table has a count, which is expanded by 1 every item an item set is hashed to that bucket.

5. If the bucket count is satisfy the min_sup threshold value than bit vector is set to 1, otherwise is set to 0.

6. The candidate pairs that bit vector bit is not set that are removed.

TID	ITEMS
1	A,B,C
2	B,D
3	B,C
4	A,B,D
5	A,C
6	B,C
7	A,C
8	A,B,C,E
9	A,B,C

Table: Transaction Data

Now, Hash table is created using hash function that is as below.

H(x, y)=((Order of X)* 10+(Order of Y))mod 7

Bucket address	0	1	2	3	4	5	6
Bucket count	2	2	4	2	2	4	4
Bucket contents	{A,C} {C,E}	{A,E} {A,E}	{B,C} {B,C} {B,C} {B,C}	{B,D} {B,D}	{B,E} {B,E}	{A,B} {A,B} {A,B} {A,B}	{A,C} {A,C} {A,C} {A,C}

The hash table H_2 is created for C_2 by scanning transaction from above transaction table. Suppose the min_sup is 3 then the item set that are stored in bucket 0,1,3,4 are not frequent and remove from the list. They should not include in C_2.

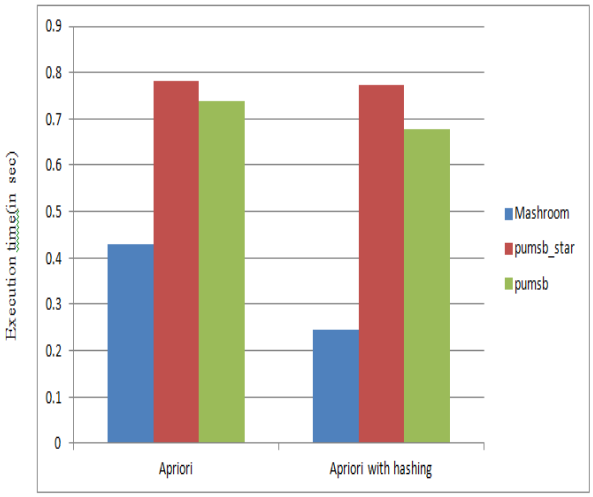
Comparison of algorithms

Apriori	Apriori with hashing
Prons: 1) Algo uses info from previous steps to produce the frequent itemsets 2) Easy to implement	**Prons:** 1) Reduce the number of scans 2) Remove the large candidates that cause high Input/output cost
Cons: 1) Database need to scan at every level 2) Uses more space and memory and time 3) In case of large database it is not efficient	**Cons:** 1) As DB size increase the size of bucket also increase. 2) For Large DB it is difficult to handle Hash table and candidate set. 3) execution time is less for small db.

Comparative result:

Here we have used 3 large dataset named are Mashroom dataset, pumsb_star dataset, pumsb dataset. The below table shows that 3 different datset and their size. The dataset size is in the term of transaction. Below tble shows the result of apriori and apriori with hashing algo. The result proves that apriori with hashing technique provide a better result with minimum time.

Dataset (size)	Apriori	Apriori with hashing
Mashroom (8124)	0.431	0.244
pumsb_star (49046)	0.783	0.773
Pumsb (49046)	0.739	0.677



Above graph represent the time vs algorithm which clearly shows that the hashing with apriori provide better result on large data set.

Conclusion:

Above result proves that the apriori with hashing provide better results in less amount of time as compare to apriori. We try to reduce more time with accurate result using map reduce concept.

References:

[1] Han J, Kamber M. Data Mining : Concepts and Techniques. Higher Education Press,2001..

[2] Warnia Nengsih "A Comparative Study On Market Basket Analysis And Apriori Association Technique" Politeknik Caltex Riau –Indonesia 2015 IEEE

[3] Sudhanshu Shekhar Bisoyi; Pragnyaban Mishra; S. N. Mishra"Weighted frequent multi partitioned itemset mining of market-basket data using MapReduce on YARN framework"-2016 IEEE

[4] Surbhi K. Solanki, Jalpa T. Patel"Survey on association rule",(2015),212 – 216

[5] Zhang Chunsheng ,Li Yan "The Visual Mining Method of Apriori Association Rule Based on Natural Language"-2016 IEEE.

[6] Thanmayee, H R Manjunath Prasad "Revamped Market-Basket Analysis Using In-Memory Computation Framework"-2017 IEEE.

[7] Ashish Shah "Association Rule Mining with Modified Apriori Algorithm using Top down Approach"-2016 IEEE.

[8] O. Yahya, O. Hegazy, Ehab Ezat. "An Efficient Implementation of Apriori Algorithm Based on Hadoop-Mapreduce Model.", Proc. of the International Journal of Reviews in Computing. (2012). Vol. 12: 59–67.

[9] T. Karthikeyan and N. Ravikumar, "A Survey on Association Rule Mining," International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), pp. 5223-5227, 2014.

[10] Abhijit Sarkar, Apurba Paul, Sainik Kumar Mahata, DeepakKumar, Modified Apriori Algorithm to find out Association Rules using Tree based Approach.

[11] Kaushal vyas,Shipa sherasiya " Modified apriori algorithm using hash based technique".IJARIIE-ISSN(O)-2395-4396.

[12] Rupali, Gaurav gupta"Apriori Based Algorithms And Their Comparisons". ISSN: 2278-0181

[13] http://fimi.ua.ac.be/data/