IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

A Hash based Mining Algorithm for Maximal Frequent Item Sets using Hashing

Vaishali Galav[1], Deepak Jain[2]

School of VIT University, Chennai[1]

Rajasthan technical University[2]

Abstract: Data mining plays important role in many applications like market-basket analysis, cross marketing field etc. In data mining, Association Rule Mining (ARM) finds the interesting relationship between across of various items in a given database. In this paper, we propose a new association rule mining algorithm called Hash Based Frequent item sets-Quadratic probing (HBFI-QP) in which hashing technology is used to resolve primary collisions in vertical data format of the data base. But Quadratic probing also suffer from secondary clustering. This secondary clustering problem solve by using double hashing technique (HBFI-DH). The proposed technique generates the exact set of maximal frequent item sets directly by removing all non-maximal item sets. The proposed technique access the data fastly and efficiently compare with other hashing technique.

Keywords: Frequent Item Sets, data mining, Association Rule Mining (ARM), double hashing technique (HBFI-DH).

I. INTRODUCTION

Now a days Data Mining is growing very quickly and is play key role in our lives. Data mining is a systematic way to discover the knowledge, extracting the information of data and their relationship, For example, E–commerce, web mining, Market basket analysis, catalog-design etc. The goal of data mining is Classification and Prediction. In Classification, data is sorted in various groups. For example, in Sales market they sort the customer information using customer-id or any kind of customer card. In Prediction, accuracy of a continuous variable is predicted. As pointed out in Market Analysis and Management, data mining is useful to collect and store the huge amount of sales data in the same order as they receive order and identify the best product for different customers. It uses prediction to find out the way to attract new customer and thus improve the quality of business decisions. One of the most important data mining problems is mining association rules.

Association rule is finding frequent patterns, correlation of database, association, set of items or objects in transaction database. For example given a set of transaction's , the challenging task in data mining is forming a rule which will be predict the occurrence of item set based on the occurrence of other item set in the transaction.

An Example of market Basket Transactions show in table1

TID	items
1	{Bread,milk}
2	{Bread,Tea,Jam,Eggs}
3	{Milk,Tea,Jam,Cola}
4	{Bread,Milk,Tea,Jam}
5	{Bread,Milk,Tea,Cola}

Association rule is useful to discover the relationship hidden large data item sets. This relationship can be represented in the form of association rule or set of frequent items.

In example, Extracted the data that show in above table.

{Tea} $\longrightarrow$ {Bread}

From the table it is clear that, thereis a strong relationship between Tea and Bread, because most of customer who buy Tea they also buy bread. Mining association rules can be divided into 2sub problems. First all sets of items have to be found that are consist the sufficient number of transactions above the minimum (support) requirement .These item sets are known as large item sets. Now all large item setsare obtained and apply the association rules that can be generated in a straightforward manner. There are various technique have been proposed to discover the large item sets, Generally , first construct a candidate set of large item sets based on some heuristics, and then find out the subset that contains large item sets. This process can be done iteratively means that the large item sets find out in one iteration will be used as the basis to generate the candidate set for the next iteration. For example, in the k^{th} iteration, all large item sets containing k items, large k item sets are generated. In the next iteration, to construct a candidate set of large (k +1) item sets, a heuristic is used to expand some large k item sets into a (k+1)item sets.

II. RELATED WORK

R. Agrawaletal proposed Method for finding sets of item in large database An Effective Hash Based Algorithm for Mining Association rule[5] which work an algorithm

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

DHP(Direct hashing and Pouring), The proposed of DHP has 3 major factor, one is efficient large itemset generation , other is reduce the transaction database size and third is database scan the data.

R. Agrawaletal propose Another technique is Apriori Algorithm [2]for finding sets of item in large database .Apriori is a basic algorithm to understand association rules and generate candidate item sets. Using hashing and pruning technique maximum forward reference found. Hashing is next method to find out the item sets from large databases.

Ming-syanchenetal proposed Efficient Data Mining for Path Traversal Patterns**[4]**In this paper we consists two step First step is to derive an algorithm to convert original sequence of log data into a set of maximal forward references. Second step is to derive an algorithm to determine frequent traversal patterns from Maximum Forward (MF) references obtained. For determining large reference sequences, hashing and pruning techniques are used. Another way is applying large references in batch, so as to reduce the required number of database scans.
Jong Sooetal proposed a Hash-Based Method with Transaction Trimming for Mining Association Rules [7] In this paper; we work on mining association rules that find the items in a large database item. Association rule is finding frequent patterns, correlation of database, association, set of items or objects in transaction database. For example given a set of transaction's ,find rule that will be predict the occurrence of an item set based on the occurrence of other item set in the transaction. Association rule is useful to discover the relationship hidden large data item sets. This relationship can be represented in the form of association rule or set of frequent items.

A.M.J et al proposed linear probing [6], this paper, we propose an algorithm, HBMFI-LP used hashing technology to store the database in vertical data format. To avoid hash collisions, linear probing technique is used. The proposed of the algorithm to generates the set of maximal frequent item sets directly by removing all non-maximal item sets A linear probing, we find the hash table sequentially starting from the original hash table location. If a location is occupied, next location we check. We wrap around from the last table location to the first table location if necessary..

III. PROPOSED WORK

3.1 HASHING

A **hash function** is any **function** that can be used to map data of arbitrary size to data of fixed size. The values returned by a hash function are called hash values, hash codes, hash sums, or simply hashes.

When two keys map to the same location in the hash tableis called collision .When two items are placed in same block, we have to find a systematic way for placing the second item in the hash tabe. This method is known as **collision resolution**. There two method for resolving collisions are chaining (close addressing) and open addressing. In open addressing consists three main techniques are linear probing, quadratic probing and double hashing.

Separate Chaining:
In separate chaining, each bucket is independent, and has some sort of list of entries with the same index. Collision resolution by chaining combines linked representation with hash table .when two or more records hast to the same location, these records are constituted into a singly linked list called a chain.

Open addressing:
In open addressing, all item records are stored in the hash table itself. When a new item has to be inserted, to found the place that item has to be inserted, starting with the hashed-to slot and proceeding in some sequence, until an unoccupied slot is found. When searching for an entry, the hash table are scanned in the same sequence, until either the target record is found. This approach is also know as closed hashing.

Quadratic Probing: In quadratic probing, for example, the original hash location is i. If a location is occupied, then check the location $i+1^2, i+2^2, i+3^2$……wrap around from the last table location to the first table location if necessary.
Hash Function H (key)=key mod n…….(1)
Quadratic Probing=(hf (key)+$c1.i + c2.i^2$)mod n…..(2)
Here hf is hash function and c1 and c2 are constant that value is 1……..n.

HBMFI-QP Algorithm

Input:
D, a database of transactions
Output:
Method:

1. Generate the vertical format of the transactional database.
2. N-itemsets are to be hashed
3. Linked list of transactions for each itemset is created with count in its header node.
4. generatefrequentitem
5. iffrequentitem is NULL then go to 9.
6. Remove any subsets that are included in another itemset from FI to generate MFI.
7. Find all combinations of the MFI
8. go to 1
9. return MFI

Generally the structure of the transactional database are in two different ways – horizontal data format and vertical data format. In this work, we used vertical format is used to store the transactions of database. In vertical data format, the data is represented in the form of item-tidset. The sample transactional database is shown in table 2 . In

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

table2 we have two columns first column is itemset and second column tid set where itemset contains number of items and tidset contains transaction identifiers. Quadratic probing technique is used to resolve collisions,

Table 2: The transactional database in 1st level

Itemset	Tidset
T1	I1, I3, I4, I5, I7
T2	I1, I2, I4, I6, I8
T3	I1, I4, I6, I7, I1
T4	I1, I2, I5, I8, I1
T5	I2, I3

In this table the transactions item are based on the hash function as
h(k) = (order of item Key) mod n. Here, we define that n is 7. So, the item T1 is stored in 1st
Location of hash table. In the Transaction in which Item T1 is present that are connected to the linked list. Each list contains a header in which the count of the transactions linked to the particular item is maintained. Thus quadratic probing technique is resolved the collision problem. Similarly, this method is usedfor every items in the transactional database. It can be shown in Fig.1 that T2 is located at 2, T3 at 3, T4 at 4 and T5 at location 5. The hash table shown in the Fig. 1, we know that the items T1, T2, T3, and T4 are the frequent items that occur with minimum requirement is 3 .The m-itemsets can be generated by taking all the combinations of frequent itemsets from the previous table and intersection of the tidsets of the corresponding itemsets. The 2-itemsets generated that shown in table 3.

Fig. 1 Table for transactional database in the 1st level connected by link list

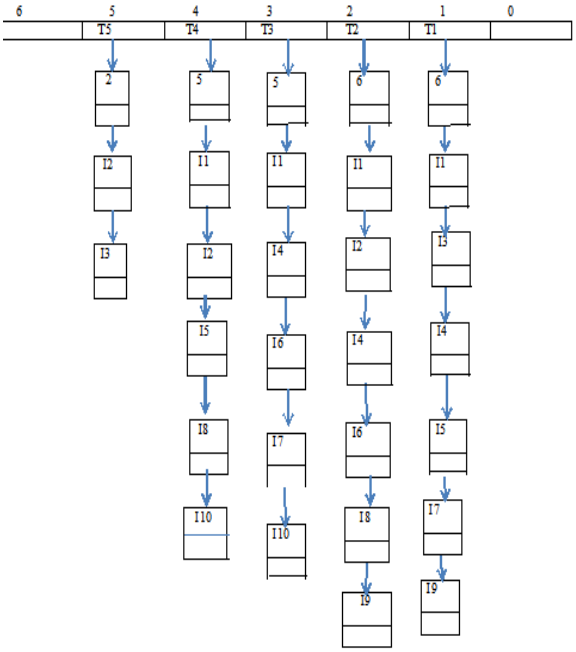
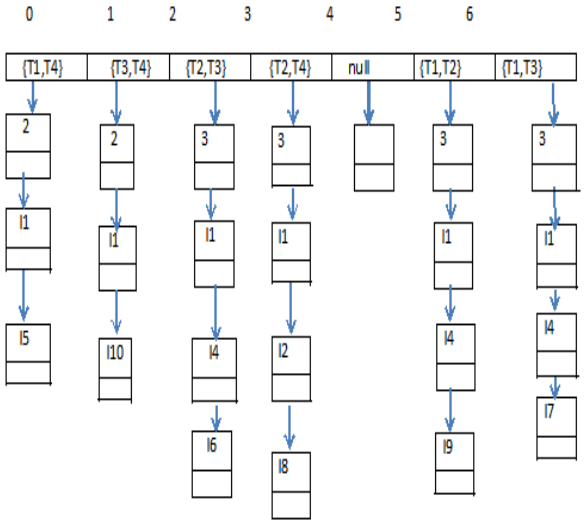


Table 3: Transactional database in second level

Itemset	Tidset
{T1, T2}	I1, I4, I9
{T1, T3}	I1, I4, I7
{T1, T4}	I1, I5
{T2, T3}	I1, I4, I6
{T2, T4}	I1, I2, I8
{T3, T4}	I1, I10

The itemsets in the 2nd level are on the hash function, h(k) = ((order of A)*10 + order of B) mod n. Here we consider n is 7. Using hash function, itemsets {T1, T2}, {T1, T3}, {T1, T4}, {T2, T3}, {T2, T4}, {T3, T4} are placed in locations 5, 6, 0, 2, 3, 1 respectively. Here collision occur for {I1, I3} and {I3, I4}. Both are trying toplaced in location 6. Then to solve this collision quadratic probing technique is used , {I3, I4} is placed in location 1. The hash table for the second level is shown in Fig. 2.

Fig. 2 Table for transactional database in the 2nd level connected by link list



In the second level, the itemsets {T1, T2}, {T1, T3}, {T2,T3}, {T2, T4} are frequent itemsets which in table 2 . Now, we combine all the possible frequent item sets are {T1, T2}, {T1, T3}, {T2, I3}, {T2, T4}}. From these, find maximally frequent itemsets are {T1, T2}, {T1, T3}, {T2, T3}, {T2, T4}. Now generated the 3-itemsets from maximal frequent itemsetsof 2nd level. Show in Table 4.

Table 4: Transactional database in 3rd level

Itemset	Tidset
{T1,T2,T3}	I1, I4
{T1,T2,T4}	I1
{T2,T3,T4}	I1

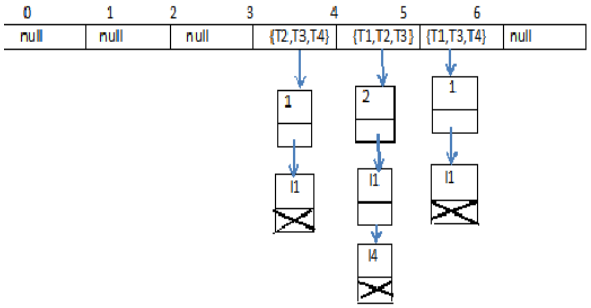
The itemsets in the 3rd level are based on the hash function, h(k) = (((order of A)*100 +(order of B)*10 +

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

order of C) mod n). Here we consider that n is 7. Using this hash function, itemsets {T1, T2, T3}, {T1, T2, T4}, {T2, T3, T4} are placed in locations 4, 5, 3 respectively. The hash table for the third level is shown in Fig. 3.



So the frequent maximal itemsets is {T2,T3,T4}, {T1,T2,T3},{T1,T3,T4}.

Double hashing:

Double hashing is used to resolve hash collision. Interval between probes is computed by another hash function. Double hashing reduces clustering in a better way to compute the Linear probing and quadratic probing. The increments for the probing sequence are computed by using a second function.

Hash funchionh1(k)=kmodN…….(1)

Second Hash function h2(k)= 1+(kmod(N-2))…….(2)

Double hashing =(h1(k)+ih2(k))mod n……(3)

Here N is table size and h is a hash function.

Existing algorithm:

Step-1: Start

Step-2: Simply scan the given transactional database tomake table of items with item count and theircorresponding transactions.

Step-3: Generate table L1

Step-4: This is similar to apriori join step. For ith levelcombine items to generate all possible ithleveltransaction in table Ck by using table Lk-1.Then frequency count I discovered for each combinationsand generate linked list structure and allocate items instructure.

Step-5: Apply hash function for each item in Ck

Step-6: If collision not occurs at point t

Go to step 4

Step-7: If collision occurs

Apply second hash function at position t to moveexisting value on empty slot in hash table.

Step-8: Check the new position in hash table is empty?

If empty:

Put the value in that place

Else if it a collision so go to step 6

Step 10: End the process when frequent itemset is found.

We used vertical format is used to store the transactions of database. In vertical data format, the data is represented in the form of item-tidset. The sample transactional database is shown in table. In table2 we have two columns first column is itemset and second column tidsetwhere itemset contains number of items and tidset contains transaction identifiers.

Itemset	Tidset
T1	I1, I3, I4, I5, I7
T2	I1, I2, I4, I6, I8
T3	I1, I4, I6, I7, I1
T4	I1, I2, I5, I8, I1
T5	I2, I3

In this table the transactions item are based on the hash function as

h(k) = (order of item Key) mod n. Here, we define that n is 7. So, the item T1 is stored in 1st Location of hash table.T2 is located at 2, T3 at 3, T4 at 4 and T5 at location 5. The hash table shown in the Fig. 1, we know that the items T1, T2, T3, and T4 are the frequent items that occur with minimum requirement is 3.Transactional database in second level. For second level hash function is

h(k)=[order of x*10+order of y]mod n

For combination of three itemset the hash function will be,

h(k)=[(order of x*100)+(order of y*10)+order of z]mod n

Similarly for other combinations do according linear increment in multiplication.

The second hash function for removal of hash collision is

h(k,i)=h1(k)+i*h2(k)]mod m

CONCLUSION

The proposed algorithm represented the data in vertical format and resolved the collision and find the frequent maximal item set. Quadratic probing resolved the collision problem in very simple manner and reduced the effect of clustering to get a good result.

But it is also suffering from second clustering problem in this paper presented HBFI, an algorithm for finding frequent item sets. Our experimental results demonstrate that HBFI-DH is better than other hash based methods because it efficiently map the item sets in the hash table and it also avoids the primary clustering problem and secondary clustering. The vertical data format representation of the database leads to the easy manipulations on hash data structure.

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

ACKNOWLEGEMENT

We express sincere thanks to our project guide **Prof. Sivagami,** CSE with Specialization in Big Data Analytics, VIT University Chennai Campus for their guidance and support.

REFERENCE

[1] K. Gouda and M.J.Zaki, "Efficiently Mining Maximal Frequent Itemsets", in Proc. of the IEE Int. Conference on Data Mining, San Jose, 2001.
[2] R. Agrawal, T. Imielienski and A. Swami, "Mining association rules between sets of items in large databases. In P. Bunemann and S. Jajodia, editors, Proceedings of the 1993 ACM SIGMOD Conference on Management of Data, Pages 207- 216, Newyork, 1993, ACM Press
[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", in Proceedings of the 20th International Conference on Very Large Databases (VLDB'94), Santiago de Chile, September 12-15, pages 487-499, MorganKaufmann, 1994.
[4] "Efficient Data Mining for Path Traversal Patterns",Ming-syanchen, Jong soo park ,Philips s.yu.
[5] Aggarwal, C.C. and P.S. Yu, "Mining large itemsets for association rules", in Bulletin of the
[6] IEEE Computer Society Technical Committee
[7] **"A** Hash based Mining Algorithm for Maximal Frequent Item Sets usingLinearProbing**",**A.M.J. Md. ZubairRahman, P. Balasubramanie and P. Venkata Krihsna1 Kongu Engineering College, Perundurai, Tamilnadu, India School of Computing Sciences, VIT University, Vellore
[8] "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules ",JongSoo
[9] Park, Member, IEEE, Ming-Syan Chen, Senior Member, IEEE, and Philip S. Yu, Fellow, IEEE
[10] A Hash Based Frequent Itemset Mining usingRehashing"
[11] Sirisha Aguru1 Department of Computer Scince and Engineering1, Sri Vasavi Engineering College Pedatadepalli, India.
[12] A Hash based Mining Algorithm for Maximal Frequent Item Sets using "Linear Probing A.M.J. Md. ZubairRahman, P. Balasubramanie and P. VenkataKrihsnaKongu Engineering College, Perundurai, Tamilnadu, IndiaSchool of Computing Sciences, VIT University, Vellore"