TO IDENTIFY THE LOCATION THROUGH DATA MINING ALGORITHM BY USING VERTICAL DATA FORMAT

V.K. Singh
Bansal College of Engineering
Mandideep Raisn(MP), India

ABSTRACT

To deal with very large database in supply – chain management is very crucial issue. The important factor of supply chain management is to find the particular location where a group of items are frequently required or consumed. As we know that the emerging technology gives the facility to collect tremendous amount of data. The data about customers , distributors , wholesalers , suppliers to the different cities are available . so it becomes necessary to manage the data. The selection of different locations is a big problem from the data. There are many items supplied to different locations and a single location may required many items. In this paper I am going to discuss the problem of selection of those locations which require group of items frequently and the solution of the above is proposed with the help of data mining technique. For this I assumed the data about items and locations.The proposed method includes the Association Rule mining. Vertical Data Format has been used to find out the frequent item sets required or consumed by the location. That item set is used to find primary and secondary locations.

Keywords: *Supply – chain management , Location selection , Data Mining , Association Rule , Apriori Algorithm.*

I. INTRODUCTION

Changing market condition and technological innovation are forcing the business to adaptive and competitive with these changes. At present most of the companies are trying to identify and optimize their supply chain to different locations.

Now a days the main and basic challenges in supply chain are to plan a strategy to manage the resource and meet the demand , to select the desired location where the company will supply the group of items and services whenever and wherever require.

To attain a good performance for remaining competitive the selection of the location can be important. Briefly , data mining is referred as knowledge discovery in database or mining of knowledge from a large amount of data. Mining is an important backend process for deriving business intelligence. by applying data mining approach we can identify a subset of items from thousand of database transactions. Association rule is a widely used data mining technique that searches through an entire data set for rules relating the nature and frequencies of relationship or association between entities , in this paper, association rules are used to identify a subset of key locations.

(a) Location selection

In general , most companies especially manufacturer , enterprises , wholesalers suppliers are consisted of supplier network. The selection of particular location is multi attribute decision making problem.

How to choose a particular location of group of items is an imperative issue in the management of modern business organization. Understanding a group of items require is important to ensuring a well functioning supply network. The selection of particular location problem is a multi – criterion problem which includes both quantitative and qualitative. Quantitative in terms of location , means the number of items required and qualitative means quality of items required.

In this paper I have assumed all the location posses qualitative factors and emphasis on development of potential subset of items supplied which can efficiently smooth. The supply chain process of the organization.

(b) Association Rule

If we think of the universe as the set of items available at the store , then each item has a Boolean variable representing the presence or absence of the item. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated. These patterns can be represented form of Association Rule. For example , the information that customers who purchase computers also tend to buy printers at the same time is represented in association rule

Computer => printer [support = 2% , confidence = 60%]

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for association rule means that 2 % of all the transactions under analysis show that computer and printers are purchased together. A confidence of 60% means that 60 % of the customers who purchased a computer also bought the printers. Typically , association rules are considered interesting if they satisfy both minimum support threshold and minimum confidence threshold. Such thresholds can be set by users or domain experts.

(c) Frequent Item sets

Let I = { I_1 , I_2, ….. I_m] be a set of items. Let D , the task relevant data be a set of database transaction where each transaction T is set of items such that T subset of I each transaction is associated with an identifier , called TID. Let A be a set of items. A transaction T is said to contain A if and only if A is subset of T. An association rule is an implication of the form A => , where A is subset of I and A intersection B = ø. The rule A => B holds in the transaction set D with support s , where s is the percentage of transactions in D that contain A union B. This is taken to be the probability , P(A union B) the rule A => B has confidence c in the transaction set D , where c is the percentage of transaction in D containing A that also contain B. This is taken to be the conditional probability , P(B|A)

Support (A=>B) = P(A union B)

$$Confidence(A=>B) = P(B|A)$$

Two major steps of the Apriori algorithm are the join and prune steps. The join step is used to construct new candidates sets. Higher level candidate item sets (C_i) are generated by joining previous level frequent item sets are L_{i-1} with itself. the prune step helps in filtering out candidate item sets whose subsets are frequent. This is based on the anti – monotonic property as a result of which every subset of a frequent item set is also frequent. Item set not supporting to the minimum threshold must be discarded before further joining.

The main notation for association rule mining that is used in Apriori algorithm is following (1) A k- item set is a set of K items (2) the set C_k is a set of candidate k item set that are potentially frequent (3) the set L_k is subset of C_k and is the set of k item set that are frequent.

(d) Algorithm

1. L_1 = frequent items of length 1.
2. For (k = 1 ; L_k! = 0 ; K++) do.
3. C_{k+1} = candidates generated from L_k
4. For each transaction t in database D do.
5. Increment the count of all candidates in C_{k+1} that are contained in t.
6. L_{k+1} =, candidates in C_{k+1} with minimum support
7. End do
8. Return the set L_k as the set of all possible frequent item set.

II. PROPOSED METHOD

The market trends are based on the concept of demand and supply. If the chain of supply and demand get disturbed deadlock like condition occurs. Such as a particular location is suffering from irregular demand. If we find out the particular location which require particular items or a group of items are frequent consumed in the location , then we can avoid the deadlock condition. The designed approach helps to identify clusters of location.

I have divided the approach in to two phases includes identification of item set which are very frequently used and then with the technique of association rule , classifies location. In the second phase I have identified the primary and secondary locations. The primary location set has a cluster of items which used frequently and secondary locations used remaining items.

(a) Vertical Data Format

Mining can be performed on the data set intersecting the transaction data set of every pair of frequent single item. Here we can understand the process of mining frequent item set by explaining the vertical data format.

First , we transform the horizontally formatted data to the vertical format by scanning the data set once. The support count of an item set is simply the length of the transaction set of item. Starting with k =1 , the frequent k-item set can be used to construct the candidate (k+1) item set based on the Apriori property. The computation is done by intersection of the transaction set of frequent k – item set to compute the transaction set of the corresponding (

k+1) item set. This process repeats , with k incremented by 1 each time , until no frequent item set

(b) Phase – 1 : Identify frequent item set

(i) Apply the Apriori algorithm using vertical data format on Location database in order to identify k – frequent item set

(ii) Identify the location of these k frequent item set and then take the intersection first and then union among location of the frequent k item set in order to find out locations.

(iii) Calculate the probability factor :

Pf = count(A) "C_j" / count(A) where A includes all items and C_j refer to supply items in location set.

(c)Phase -2 : identify primary and secondary Location set

(i) Identify all the distinct items which are supply to the different locations

(ii) Identify the missing items.

(iii) Find out the primary locations where a group of items supplied.

(iv) Take the rest locations set as secondary locations.

III. CASE STUDY

A case is shown here to describe the proposed method.

The table 1 consist of location id and items supplied to the locations. There are 12 locations and 8 items. And second table has the details of locations in the city Bhopal.

L_id	Items supplied to the Location
L_1	I_1 , I_4 , I_5 , I_3
L_2	I_2 , I_7 , I_6
L_3	I_1 , I_3 , I_5 , I_4
L_4	I_2 , I_5 , I_6
L_5	I_3 , I_4 , I_8 , I_2
L_6	I_1 , I_5 , I_{57}
L_7	I_2 , I_6 , I_8
L_8	I_5 , I_7 , I_2 , I_4
L_9	I_2 , I_1 , I_3
L_{10}	I_4 , I_6 , I_7 , I_2
L_{11}	I_1 , I_4 , I_8
L_{12}	I_1 , I_2 , I_3 , I_5

Table – 1

Location_id	Location name in the City
L_1	Jahangirabad
L_2	Bittan Market
L_3	New Market
L_4	Shaapura
L_5	Anand Nagar
L_6	Barkheda Market
L_7	Saket Nagar
L_8	Shakti Nagar
L_9	Nishatpura
L_{10}	Bag Sewaniya
L_{11}	Misrod
L_{12}	Koh-e-Fiza

Table – 2

Now I am going to identify the maximum frequent item set. Here I assumed the threshold min_sup = 3. In the table below I have use "X" to show pruning of the item set which does not follow the threshold of min_sup. The process of finding frequent k – item set is as follows :

Items	Locations	Support
I_1	L_1,L_3,L_6,L_9,L_{11},L_{12}	6
I_2	L_2,L_4,L_5,L_7,L_8,L_{10},L_{12}	7
I_3	L_1,L_3,L_5,L_9,L_{12}	5

I_4	$L_1,L_3,L_5,L_8,L_{10},L_{11}$	6
I_5	L_1,L_4,L_6,L_8,L_{12}	5
I_6	L_2,L_4,L_7,L_{10}	4
I_7	L_2,L_6,L_8,L_{10}	4
I_8	L_5,L_7,L_{11}	3

Table - 3 Frequent 1 – item set

Items	Locations	Support
I_1	$L_1,L_3,L_6,L_9,L_{11},L_{12}$	6
I_2	$L_2,L_4,L_5,L_7,L_8,L_{10},L_{12}$	7
I_3	L_1,L_3,L_5,L_9,L_{12}	5
I_4	$L_1,L_3,L_5,L_8,L_{10},L_{11}$	6
I_5	L_1,L_4,L_6,L_8,L_{12}	5
I_6	L_2,L_4,L_7,L_{10}	4
I_7	L_2,L_6,L_8,L_{10}	4
I_8	L_5,L_7,L_{11}	3

Table - 4 After Pruning

Items	Locations	Support	
I_1,I_2	L_{12}	1	x
I_1,I_3	L_1,L_3,L_9,L_{12}	4	
I_1,I_4	$L_1,L_3,L11$	3	
I_1,I_5	L_1,L_6,L_{12}	3	
I_1,I_6	Nil	0	x
I_1,I_7	L_6	1	x
I_1,I_8	L_{11}	1	x
I_2,I_3	L_5,L_{12}	2	x
I_2,I_4	L_5,L_8,L_{10}	3	
I_2,I_5	L_8,L_{12}	2	x
I_2,I_6	L_2,L_4,L_7,L_{10}	4	
I_2,I_7	L_2,L_8,L_{10}	3	
I_2,I_8	L_5,L_7	2	x
I_3,I_4	L_1,L_3,L_5	3	
I_3,I_5	L_1,L_{12}	2	x
I_3,I_6	Nil	0	x
I_3,I_7	Nil	0	x
I_3,I_8	L_5	1	x
I_4,I_5	L_1,L_8	2	x
I_4,I_6	L_{10}	1	x
I_4,I_7	L_8,L_{10}	2	x
I_4,I_8	L_5,L_{11}	2	x
I_5,I_6	L_4	1	x
I_5,I_7	L_6,L_8	2	x
I_5,I_8	Nil	0	x
I_6,I_7	L_2,L_{10}	2	x
I_6,I_8	L_7	1	x
I_7,I_8	Nil	0	x

Table - 5 Frequent 2 – item set

Items	Locations	Support
I_1,I_3	L_1,L_3,L_9,L_{12}	4
I_1,I_4	$L_1,L_3,L11$	3
I_1,I_5	L_1,L_6,L_{12}	3
I_2,I_4	L_5,L_8,L_{10}	3
I_2,I_6	L_2,L_4,L_7,L_{10}	4
I_2,I_7	L_2,L_8,L_{10}	3
I_3,I_4	L_1,L_3,L_5	3

Table - 6 After Pruning

Items	Locations	Support	
I_1,I_4,I_5	L_1,L_3	2	x
I_1,I_3,I_4	L_1	1	x
I_2,I_4,I_7	L_8,L_{10}	2	x
I_2,I_4,I_6	L_{10}	1	x

Table - 7 Frequent -3 item set

After prusing frequent item set in nil, Thus , the following method of vertical data format , we try to find out the maximum k-item set i.e. 3 – item set but it does not satisfy min_sup = 3 thus maximum k- item set that satisfy min_sup = 3 is 2 – item set.

So from above table we got six item set which are as follows :

(I_1,I_3) , (I_1,I_4) , (I_1,I_5) , (I_2,I_4) , (I_2,I_7) , (I_3,I_4)

Now we transform the above maximum 2 item set into a location , so that it can help to identify our primary location. To spot out the primary location from above item set. Probability factor (Pf value) is to be calculated for each location set.

Location probability factor.

Maximum frequent Item set **Location set j** **Pfj**

1. I_1 , I_3 L_{12} — $L_1 , L_3 , L_9 ,$ — $4 / 12 = 1/3$
2. I_1 , I_4 — L_1 , L_3 , L_{11} — $3 / 12 = ¼$
3. I_1 , I_5 L_{12} — $L_1 , L_3 , L_6 ,$ — $4 / 12 = 1/3$
4. I_2 , I_4 — L_5 , L_8 , L_{10} — $3 / 12 = ¼$
5. I_2 , I_7 — L_2 , L_8 , L_{10} — $3 / 12 = ¼$
6. I_3 , I_4 — L_1 , L_3 , L_5 — $3 / 12 = ¼$

As we see from above table , we identify two frequent item set { I_1 , I_3 } and { I_3 , I_5} having highest probability factor among others. Thus selecting those items which are having high probability factor consist of the following location :

$I_1 , I_3 \longrightarrow L_1 , L_3 , L_9 , L_{12}$

$I_1 , I_5 \longrightarrow L_1 , L_3 , L_6 , L_{12}$

From the above if we sort common location , we find five location i.e.

$L_1 , L_3 , L_6 , L_9 , L_{12}$ i.e. { Jahangirabad , New Market , Barkheda BHEL , Nishatpura , Koh – a – Fiza }

This is the key locations for our frequent item sets i.e. I_1 , I_3 , I_5 it means that out of 12 items there are only three items which were frequently required by five consecutive locations. The manufacturer or supplier have to concentrate for supplying items I_1 , I_3 and I_5 to locations L_1 , L_3 , L_6 , L_9 and L_{12}.

If we want to identify secondary locations for other frequent items. Then according to the above table :

$I_1 , I_4 \longrightarrow L_1 , L_3 , L_{11}$

$I_2 , I_4 \longrightarrow L_5 , L_8 , L_{10}$

$I_2 , I_7 \longrightarrow L_2 , L_8 , L_{10}$

$I_3 , I_4 \longrightarrow L_1 , L_3 , L_5$

By selecting the items we have five items { $I_1 , I_2 , I_3 , I_4 , I_7$ } and six locations are {$L_1 , L_2 , L_3 , L_5 , L_8 , L_{10}$ } i.e. {Jahangirabad , Bittan market , New Market , Anand Nagar , Shakti Nagar , Bag Sewaniya }

As we observed that locations L_1 and L_3 lie in both primary and secondary locations while the items I_1 and I_3 are also lie in primary and secondary frequent item sets. Moreover from the above , we have seen that the item sets { $I_6 , I_8 , I_9 , I_{10} , I_{11} , I_{12}$ } are neither lie in primary and secondary frequent item sets . it means that those items can not be consider is frequent as there is negligible demand , and also the locations { L_4 , L_7 , L_{11} } i.e. { Shapura , Saket Nagar , Misrod } did not lie in primary and secondary locations , it means that those locations are not good for supplying the items , therefore manufacturer or suppliers can avoid these locations.

IV. CONCLUSION

As data is growing with an exponential rate . it is very important to arrange it properly. The market is changing with huge number of products and their

suppliers. An organization must have a cluster of key locations so that it can monitor relationship with them for supplying the required items.

It is a challenging task to develop an approach for the selection of potential locations. This paper proposed a model using the concept of association rule which can be used in sorting of potential locations In the city. The vertical data format has its limitations and disadvantages during the association process I have assumed 10 items and 12 locations , which is quite easy and feasible task. But the actual database may contain thousand of transactions and may be some hundreds of locations within the city. It may be possible to handle but it is quite time consuming task.

REFERENCES

[1] R. agarwal and R. Srikant , Fast algorithm for mining association rules in Proc 1994 int. conf very large database (VLDB '94) , sep 1994.

[2] Simon Fong , serena chan "Mining online user's access records for web bussines intelligence" in proceedings of ICDM ' 2002 , PP 759 – 762.

[3] Data mining concept and Techniques , 2nd edition , Jiawei Han and Micheline Kamber.

[4] Data mining Introductory and advanced topics , Margret H. Dunham eight edition , Pearson Publications.