International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Identifying frequent patterns in E-Commerce using Partition Algorithm

P. Regina[1], Prof. M. V. Ramana Murthy[2]

[1]Research Scholar, Department. of CSE, Dravidian University, Kuppam, Aurora's PG college, Uppal, Hyderabad, Telangana, India

[2]Professor & Head of Mathematics & IT, MGIT, Hyderabad, Telangana, India, Former: Department of Mathematics & Computer Science, Osmania University

ABSTRACT

Internet emerged as one of the important tools of communication in the recent years and E-commerce has gradually grown with it and has given rise to a new world of doing business. It has drawn attention to apply data mining techniques to identify frequent patterns for improving business strategies. The main aim of Frequent pattern discovery is to find frequently occurring itemsets in large databases. Frequent pattern mining is the most important step in mining association rules to show items that have same patterns in the database, appear together. In E-commerce, frequently occurring product purchase combinations are essential to model user preference. In this paper we look at the existing algorithms which are used to identify the association rules and discusses how they are extend to find frequent patterns in E-commerce.

Keywords : Frequent pattern discovery, association rules, E-commerce.

I. INTRODUCTION

Electronic commerce is playing a vital role in business today. E-commerce provides good customer management, better strategies for marketing and sales, wide range of products and more efficient operations. Data mining techniques are used for pattern discovery in E-commerce. In discovery model the system automatically discovers hidden information in the data. An example of discovery model is supermarket database, which is mined to discover particular group of customers to target by analyzing the purchasing patterns. [1]. Two fundamental goals of data mining techniques are prediction and description.

The prevalent descriptive data mining techniques is Association rule mining [2], which is extensively used in marketing and retail communities. Mining association rules are particularly useful for discovering relationships among items from large databases [3]. The "market-basket analysis" which performs a study on the habits of customers is the source of motivation behind Association rule mining.

II. ASSOCIATION RULES

Association rule is expressed as $X \Rightarrow Y$ where X and Y are the non empty sets of items. The meaning is the transaction set which contains X tends to contain Y. There are two types of frequent patterns- frequent sets which are unordered collection of items and frequent sequences which are ordered collection of item sets. In this paper we will focus on frequent item sets[5].

Two measures called support and confidence are introduced by Agarwal to quantify the significance of association rule. The itemsets having support and confidence greater than the specified minimum confidenceor support are considered to be frequent item sets. Support is how often X & Y occur together as a % of the total transactions(S%). Confidence(C) measures how much a particular item is dependent on the other. A transaction t is said to support an item I_i if I_i is present in t. Let D be the database of transactions and I be the set of items. X ⇒ Y holds with confidence c if c% of transactions in D that support X also support Y and X ⇒ Y holds support s in the transaction set D if s% of transaction support X ∪ Y .

III. BASIC PROBLEM DEFINITION

The input data for frequent pattern mining consists of records each of which contains a set of items.

Eg. Customers buy groceries from an online store. The task is to find all the subsets that occur more frequently than the others, whose count is greater than some user defined threshold , which lets say is minsup.

Let I = { Bread, Butter,Milk,Dal,Rice,cheese}

Transaction ID	Items sold
t1	Bread , Butter, Milk
t2	Bread , Butter, Dal, Rice
t3	Bread , Butter, Milk, Oil
t4	Wheat, Cheese, Oil
t5	Bread , Butter, Milk, Cheese

Here items { Bread , Butter, Milk} has the frequency of 60% since it is present in 3 of the 5 records. Items having the frequency < minsup are not frequent item sets.

Rule R is Bread, Butter-->Milk, if this is given as an association rule then

Support of R = count(X) / |D| = 3 / 5 = 0.60 = 60%

where X is Bread, Butter, Y={Milk} and D={t1,t2,t3,t4,t5}

Confidence of R = support(X ∪ Y) / support(X) = 3 / 4 = 0.75 = 75%.

The rule says that 75% of records that contain Bread & Butter also contain Milk.

IV. MINING ASSOCIATION RULES

Association rules are mined by following two phase strategy:

Phase I: Frequent item set algorithms are applied on the database D to retrieve all the frequent item sets along with their support.

Phase 2: The output of phase 1 is used to form interesting association rules

Transaction ID	Items sold
t1	Bread , Butter, Milk
t2	Bread , Butter, Dal, Rice
t3	Bread , Butter, Milk, Oil
t4	Wheat, Cheese, Oil
t5	Bread , Butter, Milk, Cheese

Phase 1: Mining Frequent item sets with count.

Let's assume minsup = 0.5

{Bread}: 4 {Butter}: 4 { Milk}:3 {Bread,Butter}:4 {Bread,Milk}:3

{Butter, Milk}:3 {Bread,Butter,Milk}:3

Phase 2: Form Association Rules from phase 1

Table 1. Bread→Butter (0.5, 1) ⇒ In a transaction whenever Bread appears Butter also appears which is indicated by 1

Butter→Bread (0.5, 1)

Bread→Milk (0.5,0.75) ⇒ In a transaction whenever Bread appears ¾ times Milk appears.

Milk→Bread (0.5, 1

Several rules can be formed from table 1. For example

R1: {Bread, Butter, Milk} = confidence= 3/5=0.60
R2: {Bread→Butter, Milk} = confidence= 3/4=0.75
R3: {Butter→Bread, Milk} = confidence= 3/4=0.75
R4: {Milk→Bread, Butter,} = confidence= 3/3=1 etc

V. WHY FREQUENT ITEM SETS ARE IMPORTANT IN ASSOCIATION RULES

Many business applications, banking sectors, Government service portals, scientific and engineering fields make use of databases. So efficient management of the data is an important task. The newly extracted information or knowledge may be applied to data management, query processing and decision making in all the above said fields. With the explosive growth of data, which today is called as big data , mining information and knowledge from the very large volume of databases has become one of the major challenges for data management and mining sector.

The frequent itemset mining is motivated by problems such as market basket analysis. Association rules can be applied to mine market basket database to state that if some items are purchased in transaction, then it is likely that some other items are purchased as well. Set of items purchased by each customer in an individual transaction is treated as a tuple in market-basket dataset. Finding all association rules is valuable for guiding future sales promotions , management of customers and stock inventory management etc..

Frequent itemsets are mined especially to discover all association rules from a given transaction database D that have support and confidence greater than minsup.

VI. THE PARTITIONING ALGORITHM

The portioning algorithm published by Navathe et al in VLDB-1995 is a very elegant algorithm to handle bottleneck of Apriori. It requires two disk scans of the data. It first partitions the data into different partitions which are only logical.

Disk Scan 1: During the first disk scan of the dataset Arpiori type mining is applied to pool locally frequent item sets based on some user defined minsup.

Disk Scan 2: During the second scan globally frequent item sets are pooled based on some user defined minsup.

Let's consider the following data set: let the minsup be 60%.

Partition 1

Transaction ID	Items sold
1	Bread , Butter, Milk,tomato
2	Bread , Butter,Rice, Potato
3	Bread , Butter, Milk, Oil
4	Cheese, Oil, Tomato, Potato
5	Bread , Butter, Milk, Cheese

Partition 2

6	Bread , Tomato, Potato,onion
7	Tomato, Potato, onion,Bread
8	Potato, Chillies, cheese,Tomato
9	Tomato, Potato,Rice, onion
10	Onion,Lemon,brinjal

After Scan 1: Locally Frequent items which are >=3 are given as output since minsup is 60%:

Partition 1

Items sold	count
Bread	4
Butter	4
Milk	3
Bread, Butter	4
Bread, Milk	3
Butter, Milk	3
Bread , Butter, Milk	4

Partition 2

Items sold	count
Potato	4
Tomato	4
onion	3
Tomato , Potato	3
Tomato , Onion	3
Potato, Onion	3
Tomato, Potato, onion	3

After scan 2: Globally frequent items and itemset are generated.

Items from Partition 1 and Partition 2 which have global frequency >=6 since minsup is 60% are selected as frequent ones.

Items sold	count
Bread	6
Tomato	6
Potato	6
Tomato, Potato	6

The analysis shows that whenever customers buy Tomatoes they also buy potatoes most of the time. Most frequently bought items are Bread, Tomatoes and Potatoes.

VII. CONCLUSION

The main aim of this paper was to apply data mining techniques on e- commerce database to find good customer purchase patterns to provide better service to the customer. There are various algorithms used in classification technique but I have discussed only a simple Partitioning algorithm.

VIII. REFERENCES

1. Data Mining –Vikram Pudi R. Radha Krishna 2012
2. F Bodon, 2003. “A Fast Apriori Implementation”, In B. Goethals and M. J. Zaki, editors, Proceedings of the IEEE
3. ICDM Workshop on Frequent Itemset Mining Implementations, Vol.90 of CEUR Workshop Proceedings.
4. Yu-Chiang Li, Jieh-Shan Yeh, Chin-Chen Chang, 2005. “Efficient Algorithms for Mining Share-Frequent
5. Itemsets”, In Proceedings of the 11th World Congress of Intl. Fuzzy Systems Association.