An Implementation of Apriori Algorithm on a Super Bazaar Database for Association Rule Mining

Dr. S. D. Mundhe

Director, Sinhgad Institute of Management and Computer Application (SIMCA), Narhe
director_mca_simca@sinhgad.edu

Mr. D.R. Vidhate

Assistant Professor, College of Computer Application for Women, Satara,
vidhatedhananjay@rediffmail.com

***Abstract*—Super bazaar is one of the formats of retail business which is self service shop. Knowledge mining is an approach which refers to the extraction of unknown information from large super bazaar databases. In knowledge mining, Apriori is an algorithm for finding association rules. This paper takes transactional data and applies Apriori algorithm to find hidden patterns. This paper will help super bazaar owners in understanding the customer behaviour more easily and giving maximum profit to the super bazaar business.**

***Keywords*—Super Bazaar, Knowledge Mining, Consumer, Manager, Buying Behaviour.**

I. Introduction

Consumer is a king in super bazaar business. The study of buying behaviour of consumer is vital in decision making process of super bazaar business. Association rules mining is an important branch of knowledge mining research. Frequent pattern analysis is a technique of association rule mining which allows a researcher to systematically identify buying patterns from the database. Market Basket Analysis is a knowledge mining technique that is widely used to identify consumer patterns such that if customer buys certain group of items then customers are likely to buy another group of items. Apriori algorithm is widely used algorithm to generate strong buying patterns from consumer purchase data.

II. ASSOCIATION RULE MINING

Association rule mining is used for finding frequent patterns and associations among sets of items in transactional databases, relational databases, and other information repositories. [2] An association rule is the relationship between two disjoint itemsets, X and Y.

An association rule is of the form:- X => Y

X => Y: - When X occurs, Y also occurs.

Given a set of items $I = \{I_1, I_2, \ldots, I_m\}$ and a database of transactions $D = \{t_1, t_2, \ldots\ldots, t_n\}$ where $t_i = \{I_{i1}, I_{i2 \ldots} I_{ik}\}$ and $I_{ij} \in I$, an association rule is an implication of the form X=>Y where $X, Y \subseteq I$ are sets of items called itemsets and $X \cap Y = \emptyset$.

III. FREQUENT ITEMSETS

Finding frequent itemsets are those with frequency larger than or equal to a user specified minimum support. The identification of sets of items, products and characteristics which often occur together in the given database can be seen as one of the most basic tasks in frequent itemset mining. The association rule mining can be reduced to mining frequent itemset. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence [5].

IV. MARKET BASKET ANALYSIS

It is a very useful technique for finding out co-occurrence of items in consumer shopping baskets. Such information can be used to provide the super bazaars with information to understand purchasing behavior of consumer in super bazaar. Market basket analysis is an important component of analytical system in retail organizations to determine the placement of goods, designing sales promotions for different segments of customers to improve customer satisfaction and hence the profit of the supermarket. [3]

V. SUPPORT

It is the measure of how often the collections of items in an association occur together as percentage of all transactions. Support(s) for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contains $X \cup Y$. A low support rule is not profitable to promote items that customers seldom buy together. So, support is often used to eliminate uninteresting rules. Association rule find all set of items that has support greater than minimum support. Support could be absolute or relative.

VI. CONFIDENCE

Confidence for an association rule X=>Y is the ratio of the number of transaction that contain both antecedent and consequent to the number of transaction that contain only antecedent. A rule with low confidence is not meaningful. Confidence (α) for an association rule x=>Y is the ratio of number of transactions that contains X U Y to the number of transactions that contains X.

VII. MINIMUM THRESHOLD VALUES

The strength of an association rule can be measured in terms of its support and confidence. The rules derived from itemsets with high support and high confidence. The number of association rules discovered is affected by a user's decision concerning the minimum support threshold and minimum confidence threshold. It may be decided on the basis of number of transactions in database. support and confidence values occur between 0% and 100%.

VIII. OBJECTIVES

It has the following main objectives:

1. To identify hidden patterns from transactional database.
2. To study the usefulness of found patterns to maximum profit of the super bazaar business.

IX. APRIORI ALGORITHM

Apriori algorithm is very effective algorithm of association rule which finds data associations. The two basic steps can be summarized as:

a) **Joining**: In this step candidate itemsets are joined.

b) **Pruning**: In this step frequent itemsets are discovered and used whereas non-frequent itemsets are discarded.[2]

Steps in Apriori Algorithm:

1) Take transactional data as input from super bazaar database.

2) Find frequent itemsets from transactional data.

3) Generate strong association rules from frequent itemsets.

4) Take decisions based on the rules.

X. APRIORI IMPLEMENTATION

In super bazaar business consumers are mostly purchasing the items cloths, personal, stationary, toys and food.

Purchase by consumers:

Following is a transactional data by consumers which contains a list of 10 different transactions in a super bazaar. For simplicity, we have below given table:

A-Cloths, B-Personal, C- Stationary, D- Toys, E- Food

Transaction IDs	List of Item IDs
T1	A,B
T2	A,D
T3	A,D,E
T4	C,D,E
T5	B,D,E
T6	A,C,D,E
T7	B,D,E
T8	D,C,E,B
T9	A,D,E,B
T10	D,E

1) Find all Frequent Itemsets:

Step1: Scan all the transactions in the database to get candidate 1-itemsets, C1.

Item ID	Items	Support
A	Cloths	5
B	Personal	5
C	Stationary	3
D	Toys	9
E	Food	8

Step2: Use minimum support count=3 to get frequent 1-itemsets, L1

. Item ID	Items	Support
A	Cloths	5
B	Personal	5
C	Stationary	3
D	Toys	9
E	Food	8

Step3: Generate candidate 2-itemsets, C2 from frequent 1-itemsets, L1

Item ID	Items
A,B	Cloths, Personal
A,C	Cloths, Stationary
A,D	Cloths, Toys
A,E	Cloths, Food
B,C	Personal, Stationary
B,D	Personal, Toys
B,E	Personal, Food
C,D	Stationary, Toys
C,E	Stationary, Food
D,E	Toys, Food

Step4: Scan all the transactions in the database to get candidate 2-itemsets, C2.

Item ID	Items	Support
A,B	Cloths, Personal	2
A,C	Cloths, Stationary	0
A,D	Cloths, Toys	4
A,E	Cloths, Food	3
B,C	Personal, Stationary	1
B,D	Personal, Toys	4
B,E	Personal, Food	4
C,D	Stationary, Toys	3
C,E	Stationary, Food	3
D,E	Toys, Food	8

Step5: Use minimum support count=3 to get Frequent 2-itemsets, L2.

Item ID	Items	Support
A,D	Cloths, Toys	4
A,E	Cloths, Food	3

B,D	Personal, Toys	4
B,E	Personal, Food	4
C,D	Stationary, Toys	3
C,E	Stationary, Food	3
D,E	Toys, Food	8

Step6: Generate candidate 3-itemsets, C3 from Frequent 2-itemsets, L2 and pruning using the apriori property.

Step7: Scan all the transactions in the database to get candidate 3-itemsets, C3.

Item ID	Items
A,D,E	Cloths, Toys, Food
B,D,E	Personal, Toys, Food
C,D,E	Stationary, Toys, Food

Step8: Use minimum support count=3 to get frequent 3-itemsets, L3.

Item ID	Items	Support
A,D,E	Cloths, Toys, Food	3
B,D,E	Personal, Toys, Food	3
C,D,E	Stationary, Toys, Food	3

Step9: Generate candidate 4-itemsets, C4 from frequent 3-itemsets, L3 and pruning using the apriori property

Item ID	Items
A,B,D,E	Cloths, Personal, Toys, Food
A,C,D,E	Cloths, Stationary, Toys, Food
B,C,D,E	Personal, Stationary, Toys, Food

Pruning the above itemsets
{A,B,D,E}, {A,C,D,E} &{ B,C,D,E} we get
After pruning {A, B, D, E} it is found that the itemset
{A, B, D} is not frequent thus violating apriori property.
After pruning {A, C, D, E} it is found that the itemset
{A, C, D} is not frequent thus violating apriori property.
After pruning {B, C, D, E} it is found that the itemset
{B, C, D} is not frequent thus violating apriori property.

Thus C4=∅ and algorithm terminates. It indicates that it has found all frequent itemsets. This completes apriori algorithm.

2. Generating Association Rules from Frequent Itemsets:

The rules are generated using following method:
For every nonempty subset s of I, output the rule
"s → (I-s)"
if support_count(I)/support_count(s)>= minimum confidence threshold.

We have the list of frequent itemsets

Item ID	Items	Support
A,D,E	Cloths, Toys, Food	3
B,D,E	Personal, Toys, Food	3
C,D,E	Stationary, Toys, Food	3

Generating all non empty subsets foe each frequent itemsets I
For I = {A,D,E}
all non empty subsets are {A},{D},{E},{A,D},{A,E},{D,E}
For I = {B,D,E}
all non empty subsets are {B},{D},{E},{B,D},{B,E},{D,E}
For I = {C,D,E}
all non empty subsets are {C},{D},{E},{C,D},{C,E},{D,E}

Consider minimum confidence threshold=60%

For **I= {A, D, E}** association rules generated are as below:
Rule1:- {A, D} →E
Confidence=Support_Count(A, D, E)/Support_Count (A, D)
= 3/4 = 75%.
Rule2:- {A, E} →D
Confidence=Support_Count(A, E, D)/Support_Count (A, E)
= 3/3 = 100%
Rule3:- {D, E} →A
Confidence=Support_Count(D, E, A)/Support_Count (D, E)
= 3/8 = 37%
Rule4:- A→ {D, E}
Confidence=Support_Count(A, D, E)/Support_Count (A)
= 3/5 = 60%.
Rule5:- D→ {A, E}
Confidence=Support_Count(D, A, E)/Support_Count (D)
= 3/9 =33.33%
Rule6:- E→ {A, D}
Confidence=Support_Count(E, A, D)/Support_Count (E)
= 3/8 =37%

For **I= {B, D, E}** association rules generated are as below:
Rule7:- {B, D} →E
Confidence=Support_Count(B, D, E)/Support_Count (B, D)
= 4/4 = 100%
Rule8:- {B, E} →D
Confidence=Support_Count(B, E, D)/Support_Count (B, E)
= 4/4 = 100%
Rule9:- {D, E} →B
Confidence=Support_Count(D, E, B)/Support_Count (D, E)
= 4/8 = 50%
Rule10:-B → {D, E}
Confidence=Support_Count(B, D, E)/Support_Count (B)
= 4/5 = 80%

Item ID	Items	Support
A,D,E	Cloths, Toys, Food	3
B,D,E	Personal, Toys, Food	3
C,D,E	Stationary, Toys, Food	3

Rule11:-D →

{B, E}
Confidence=Support_Count(D, B, E)/Support_Count (D)
= 4/9 = 44%
Rule12:-E → {B, D}
Confidence=Support_Count(E, B, D)/Support_Count (E)
= 4/8 = 50%

For **I= {C, D, E}** association rules generated are as below:
Rule13 :-{ C, D} → E
Confidence=Support_Count(C, D, E)/Support_Count (C, D)
= 3/3 = 100%
Rule14:- {D, E} → C
Confidence=Support_Count(D, E, C)/Support_Count (D, E)
= 3/8 = 37%
Rule15:- {C, E} → D
Confidence=Support_Count(C, E, D)/Support_Count (C, E)
= 3/3 = 100%
Rule16:-C → {D, E}
Confidence=Support_Count(C, D, E)/Support_Count (C)
= 3/3 = 100%
Rule17:- D →{C, E}
Confidence = Support_Count (D, C, E)/Support_Count (D)
= 3/9 = 33.33%
Rule18:-E → {C, D}
Confidence = Support_Count (E, C, D)/Support_Count (E)
= 3/8 = 37.50%

Selected association rules generated from frequent itemsets {A, D, E}, {B, D, E}, {C, D, E} are:

Rule1	= {A, D} →E	= {Cloths, Toys} →Food
Rule2	= {A, E} →D	= {Cloths, Food} →Toys
Rule3	= A→ {D, E}	= Cloths→{Toys, Food}
Rule4	= {B, D} →E	= {Personal, Toys} →Food
Rule5	= {B, E} →D	= {Personal, Food} →Toys
Rule6	= B → {D, E}	= Personal → {Toys, Food}
Rule7	= {C, D} → E	= {Stationary, Toys} → Food
Rule8	= {C, E} → D	= {Stationary, Food} → Toys
Rule9	= C → {D, E}	= Stationary → {Toys, Food}

XI. Conclusion

The selected rules are considered as strong association rules because they satisfies minimum confidence threshold. From above rules it is to be concluded that purchase preference of consumers is clothes, food, toys and it is found that their association is with items like toys and food. It is observed that a consumer mostly purchases items personal, toys, food and their association is with items food and toys. It is also observed that consumer purchases items stationary, toys, food and their association is with items food and toys. Super bazaar industry has a bright prospect in our country. This paper puts forward a knowledge mining methodology for super bazaar industry based on Apriori algorithm. This algorithm of knowledge mining was used to find frequent itemsets from transaction dataset and derive association rules. Apriori is most suitable for transactional databases. The results shows that the Apriori algorithm is effective, precise and it can extract valuable information from databases.

References

[1] Association Rule Mining using Apriori algorithm for work-related beliefs of Generation X and Generation Y.
[2] Improvement in Apriori Algorithm with New Parameters.
[3] Association Rule Mining using Apriori Algorithm: A Survey.
[4] Market Basket Analysis for a Supermarket based on Frequent Itemset Mining.
[5] Margaret H. Dunham, S. Sridhar, “Data Mining: Introductory and Advanced Topics” Pearson Education, Inc., 2006.
[6] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20thVLDB conference, pp 487–499.
[7] The Apriori algorithm: Data Mining Approaches Is to Find Frequent Item Sets from a Transaction Dataset.
[8] Association Rules Mining: A Recent Overview.
[9] Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms.
[10] Review paper on finding Association rule using Apriori Algorithm in Data mining for finding frequent pattern.
[11] Survey on various improved Apriori Algorithms.