Association Rule Mining case study

- **Objective of the Project**
- •The objective of this project is to understandassociation rule mining and its applications in marketbasket analysis.•We have used two real-world datasets and carried outassociation rule mining and market basket analysis.•In one case, we have used the R programminglanguage, and in the other, we have used Pythonprogramming language to solve the problems.•We studied various rules and also investigated therelationship between support, confidence, and lift

Association Rule Mining case study

- **Market Basket Analysis (MBA)**
- MBA in retail setting
- Find out what are bought together
- Cross-selling
- Optimize shelf layout
- Product bundling
- Timing promotions
- Discount planning (avoid double-discounts)
- Product selection under limited space
- Targeted advertisement, personalized coupons, itemrecommendations
- Usage beyond Market Basket
- Medical (one symptom after another)
- Financial (customers with mortgage account also have asaving account)

Association Rule Mining case study

- **Transaction Information**

Transaction No.	Item 1	Item 2	Item 3	Item 4	…
100	Beer	Diaper	Chocolate	Cheese	
101	Milk	Chocolate	Shampoo		
102	Beer	Wine	Vodka		
103	Beer	Cheese	Diaper	Chocolate	
104	Ice Cream	Diaper	Beer		
…					

Customer No.	Age	Income	Saving_acct	Children	Mortgage
100	>50	High	Yes	Yes	Yes
101	35-50	Mid	No	No	No
102	<35	High	Yes	No	Yes
103	>50	Mid	Yes	No	Yes
104	<35	Low	No	Yes	No
…					

Association Rule Mining case study

- Actionable Rules
- Wal-Mart customers who purchase Barbie dolls have a60% likelihood of also purchasing one of three types ofcandy bars
- Trivial Rules
- Customers who purchase large appliances are verylikely to purchase maintenance agreements
- Inexplicable Rules
- When a new hardware store opens, one of the mostcommonly sold items is toilet bowl cleaners

Association Rule Mining case study

- **Learning Frequent Itemsets**
- A descriptive approach for discoveringrelevant and valid associations amongitems in the data.
- **If buydiapersBuy beerThen**
- The itemset corresponding to this rule is {Diaper, Beer}
- Itemset: A collection of items.
- Frequent Itemset: An itemset that occurs often in data.
- Often, finding frequent itemsets is enough.

Association Rule Mining case study

My friend, Bill, an 85 years old man, told me a joke in a partylast Friday:

An old man is celebrating his 103rd birthday.

“I will hold my 104 th

birthday party next year. You are all welcome to join me,” he announces to his guests proudly.

“How do you know you will still be alive then?” one of his guests asks.

“Because very few people died between the age of 103 and 104,” hereplies.

Explain the logic of the old man and provide your comments

Association Rule Mining case study

- The old man's logic: P{103+ & died} is low; so 1 -P{103+ & died} is high

- Common knowledge: P{103+ & died} = P{103+} *P{died|103+}, where P{103+} is low.

- So the low value of P{103+ & died} is due to P{103+},while P{died|103+} is still high.

Association Rule Mining case study

Conf(X

Y)= Prob(Y | X)= P(X and Y) / P(X)= Support(X

Y) / Support(X)

While support is used to prune the search space and only leavepotentially interesting rules, confidence is used in a second step tofilter rules that exceed a minimum confidence threshold.

A problem with confidence is that it is sensitive to the frequency of theconsequent (Y) in the dataset.

Caused by the way confidence is calculated, if Y has higher support, itwill automatically produce higher confidence values even if there is noassociation between the items.

Association Rule Mining case study

Leverage (X

Y) = Prob(X and Y) -(P(X) P(Y))

-

Leverage measures the difference between X and Y appearing togetherin the data set and what would be expected if X and Y were statisticallydependent.

-

The rationale in a sales setting is to find out how many more units(items X and Y together) are sold than expected from the independentsales.

-

Using minimum leverage thresholds at the same time incorporates animplicit frequency constraint. For example, for setting a minimumleverage threshold to 0.01% (corresponds to 10 occurrences in a dataset with 100,000 transactions) one can first use an algorithm to find allitemsets with minimum support of 0.01%, and then filter the founditem sets using the leverage constraint.

-

Because of this property, leverage also can suffer from the rare itemproblem.

Association Rule Mining case study

Conviction (X

Y)= P(X). P(not Y)/P(X and not Y)= (1-Support(Y)) / (1 –Confidence(X Y))

Conviction compares the probability that X appears without Y if theywere dependent on the actual frequency of the appearance of Xwithout Y.

In that respect, it is similar to lift. However, unlike lift, it is a directedmeasure.

Association Rule Mining case study

- Two case studies were carried out, one using R and the other usingPython Programming.
- The first case study deals with a dataset with approximately 9,800transaction records of a shop for one month. We applied the apriorialgorithm to find out the most frequently purchased items, extractedrules by specifying threshold minimum values for support andconfidence. We also created subsets of the rules by specifyingspecific items from the transaction data.
- In the second case, we used a large dataset of the sales records ofan online store. The dataset was extracted from the UCI repository.
- We extracted rules based on specified support and confidencevalues and based on specific countries.
- We plotted support, lift, and confidence values to understand theirrelative interactions.

Association Rule Mining case study

Case Study I

This case deals with a medium-sized grocery shop in the USA.

The dataset consists of transactions over a period of one month. Theraw data was available in the form of a

comma separated variable

(CSV) file.

The package “arules” in R has been used for carrying out theassociation rule mining and market basket analysis.

Initial analysis revealed that there were 9,835 transactions in thedataset with 169 unique items being purchased by the customers.

The dataset was very sparse with a density of 0.02609146.

Top 5 “most frequent items” and their corresponding number oftransactions were found to be: Whole Milk: 2513, Other Vegetables:1903, Rolls/Buns: 1809, Soda: 1715, and Yogurt: 1372.

The minimum and the maximum size of transactions were found tobe 1 and 32 respectively, the mean and median values being 4.409and 3.000 respectively.

Only 8 items were found to have their support values higher than 0.1.These items were: bottled water, other vegetables, rolls/buns, rootvegetables, soda, tropical fruits, whole milk, and yogurt.

Association Rule Mining case study

Case Study I (contd…)

We also identified the most frequently purchased items by tuning the valueof the parameter “topN” in the R function “itemFrequencyPlot”.

In order to extract the association rules from the transaction dataset, weused the function “apriori” in R. The default values of “support” and“confidence” in the “apriori” function are 0.1 and 0.8 respectively.

The default values turned out to be too high and we did not get any rulesextracted from the transaction dataset.

In order to extract rules, we changed the “support” and “confidence” valuesto 0.006 and 0.25 respectively. With this setting, we could extract 463 rulesfrom the dataset.

Among the 463 rules, the length of the rules was found to be varyingbetween 2 to 4. The number of rules corresponding to sizes 2, 3, and 4 was150, 297, and 16 respectively.

The minimum and maximum values of “support” for the rules were 0.006101,and 0.074835 respectively. The corresponding values for “confidence” were0.2500 and 0.6600 respectively. The minimum and the maximum values forthe “lift” of the rules were found to be 0.9932 and 3.9565.

We also extracted subsets of rules including only some specific items ofinterest and saved the rules in a “CSV” file

Association Rule Mining case study

Case Study I (contd…)

We also identified the most frequently purchased items by tuning the valueof the parameter “topN” in the R function “itemFrequencyPlot”.

In order to extract the association rules from the transaction dataset, weused the function “apriori” in R. The default values of “support” and“confidence” in the “apriori” function are 0.1 and 0.8 respectively.

The default values turned out to be too high and we did not get any rulesextracted from the transaction dataset.

In order to extract rules, we changed the “support” and “confidence” valuesto 0.006 and 0.25 respectively. With this setting, we could extract 463 rulesfrom the dataset.

Among the 463 rules, the length of the rules was found to be varyingbetween 2 to 4. The number of rules corresponding to sizes 2, 3, and 4 was150, 297, and 16 respectively.

The minimum and maximum values of “support” for the rules were 0.006101,and 0.074835 respectively. The corresponding values for “confidence” were0.2500 and 0.6600 respectively. The minimum and the maximum values forthe “lift” of the rules were found to be 0.9932 and 3.9565.

We also extracted subsets of rules including only some specific items ofinterest and saved the rules in a “CSV” file