Association Rules in Retail Business and New Product Development

CSE5230 Assignment 3 Group paper

10/6/2008

Zihong Chen(21311242)
Wei Tian(20599986)
Yong Yang(20494793)
Faculty of Information Technology Clayton, 3168

Contents

Abstract

Due to the development the world wild web and the information process of enterprise, the growth of information and data are rapider then before. To collect the useful information and discard the useless is interesting topic in data mining. Association rules is a technology that mines the underlay relationship from such huge information and database. The purpose of this paper is going to discuss the association rules in data mining and how apply this technology in different domain.

1.0 Introduction

This paper is divided into three main parts. In part one, our group explains what association rule is. Firstly, we brief discuss the background of association rules. Then we explain how association rules work and the main algorithm in association rules, which is Apriori algorithm. We use electric store database to explain the basic idea of Apriori algorithm. In the second and third part, our group discusses two case which uses association rules technology in different domain. The first case is focus on using temporal association rules on timestamped transactions which exist in retailer business and daily life frequently**.** The main topic of the second case study is about the apriori algorithm applied in the new product development.

1.1 Association rules background

In 1993, an Association rule was firstly introduced by Rakesh Agrawal in “Mining Association Rules between Sets of Items in Large Database”. This paper describes how to solve basket data type problem via association rules technology [7]. Assume we have a database of transaction, and each transaction is a set of items. The task of association rules is to find out all association rules which satisfy the given minimum support and minimum confidence. In other words, an association rules is a data mining technology to discovery all rules in a given data set rather than conform weather a rule holds. In most case, association rule is applied in retail business. The purpose of association rule in retail business case is to discovery the behavior mode of customer consume from the large retail transaction database, such as what products are bought together by customer in usual, how one product affects other product in process of promotion. This kind of information will help the manager to make a

decision in marketing advertising, store shelf design and classify products. Association rules technology also applied to other domain for other purpose as well, for example, association rules are used in mining business along with calendar schemas. A success story of applying association rules is Australia Health Insurance Commission [5]. After applying the association rules in HIC, the researchers find that an overpayment problem in HIC system. After solving this overpayment, the Australia HIC saved $500,000 per state per year. In 1993, Rakesh Agrawal provides the Apriori algorithm to implement the association rules [7]. Until recently, many association rules algorithms are developed based on Apriori algorithm, for example, sampling algorithm[2], partition algorithm[1] and CDA and DDA (data parallel algorithm)[6]. This paper will describe the Apriori algorithm.

1.2 Problem Statement

The purpose of association rules is to discover all rules by a specified minimum support and minimum confidence from database. To explain the idea of association rule, this paper will go though with an electric store transaction database as an example. The transaction record the items are bought every day in the electric store. Assume the following table is a part of transaction from an electric store database.

Transaction	Items
t1	Mouse, Monitor, Keyboard
t2	Mouse, Keyboard
t3	Mouse, Mp3, Keyboard
t4	CD, Mouse
t5	CD, Mp3

Table 1: Example of Transaction

Table 2 is the notation we use in this paper, we also give the description of each term, and give an example associate with electric store database.

Term	Description	Example
$I = \{I_1, I_2, \ldots, I_m\}$	Itemsets	I = {Mouse, Monitor, Keyboard, Mp3, CD}
$D = \{t_1, t_2, \ldots, t_n\}$	A set of transaction in Database	$D = \{t_1, t_2, t_3, t_4, t_5\}$
t_n	The n-th transaction in database	t_1 = {Mouse, Monitor, Keyboard}
s	Support	Supp(Mouse, Keyboard) = 0.6
α	Confidence	Conf(Mouse, Keyboard) = 1

X, Y	Items	Mouse, Keyboard
X => Y	Association rule	Mouse => Keyboard
L	Set of large itemsets	Depend on given minimal support and confidence.
l	Large itemset, l $\in$ L	
C	Set of candidate itemsets	It is also called potential large itemset.
c	Candidate itemset, c $\in$ C	

Table 2: Notation

A formal mathematic definition of association rules is given by Rakesh Agrawal. Rakesh Agrawal claim that: "Given a set of items I = {I1, I2, …, Im} and a database of transactions D = {t_1, t_2, …, t_n} where t_i = {I_{i1}, I_{i2}, …, I_{ik}} and $I_{ij} \in$ I, an association rules is an implication of the form X => Y where X, Y $\subset$ I are sets of items called itemsets and $X \cap Y = \phi$ [93]". The association rule problem is to find out all association rules X => Y whose support and confidence are greater than or equal than the specify minimal support and confidence. Hence there are two factor will affect the association rules, support and confidence. "The support(s) for an association rules X => Y is the percentage of transaction in the database that contains X $\cup$ Y. the confidence or strength (α) for an association rule X => Y is the ratio of the number of transactions that contain X $\cup$ Y to the number of transactions that contain X[4]". The confidence indicate how strength the rule is, and the support indicates the frequently of the association rule occurs in the whole database. The formula of support is:

$$\text{Support}(X) = |\{t \mid X \in t, t \in D\}| / |D|$$

The formula of confidence in association X => Y is:

$$\text{Confidence}(X, Y) = \text{Support}(X, Y) / \text{Support}(X)$$

For example, Support (Mouse, Keyboard) = 3 / 5 = 0.6. Support (Mouse) = 4 / 5 = 0.8. Confidence (Mouse, Keyboard) = Support (Mouse, Keyboard) / Support (Mouse) = 0.6 / 0.8 = 0.75. The table 3 shows a part of support and confidence for association rules in the electric store database.

X=>Y	s	α
Mouse=>Keyboard	60%	75%
Keyboard=>Mouse	60%	100%
CD=>Mouse	20%	50%
Keyboard=>Monitor	20%	33.3%
Monitor=>Keyboard	20%	100%
Monitor=>Mp3	0%	0%

Table 3: Support and Confidence

The statistics in table 3 shows when a keyboard is purchased, a mouse is purchased as well. In addition, the keyboard and mouse transaction occurs six times per ten transactions. On the other hand, the data indicates the customer have never bought monitor and mp3 together. Such this kind of information is useful. The manager would like to arrange mouse and keyboard on the same shelf in the electric store. Due to the confidence of association rules monitor => mp3 is 0%, it tell the manager to avoid advertising and promoting mp3 with monitor.

1.3 Na ïve Algorithm

The association rules problem can be decomposed into two steps.

1. Find large itemsets.
2. Generate all rules base on the large itemset.

In the following section, our paper explains the algorithm to generate the association rule firstly, and then explain use Apriori algorithm to find the large itemset.

Large item set also called frequent item set, whose support over the given threshold (minimal support). The na ïve approach to find the large item set is straightforward but expensive. Suppose the size of a given set of item is *m*, then the potential number of large itemset is 2^m. because of regardless of empty set, the potential number of large itemset is $2^m - 1$. When value of *m* is 5, the potential number is 31. When value of *m* is 10, the potential number is 1048575. Due to the explosive growth of potential number, the task of association rule algorithm is to reduce the number of potential large itemset. The potential large itemset are also called candidates, and the set of candidates is called candidates itemset. Until recently, many association rules forces on reducing the size of candidates itemsets, such as Apriori, Sampling.

Once candidate itemsets are generated, the second step is straightforward. The algorithm 1 shows how to generate association rules from the large itemsets [4].

Input:
D //Database of Transactions
I //Items
L //Large Itemsets
s //Support
α //Confidence
Output:
R //Association Rules satisfying s and α
AssociationRuleGen Algorithm:

R = ϕ;

for each l $\in$ L do

for each x $\subset$ l such that x $\neq$ ϕ do

if support(l) / support(x) $\geq$ α then
R = R {X => (1 – X)};

Algorithm 1: Association Rule Generate Algorithm

To demonstrate this algorithm, we assume the input minimal support is 30% and minimal confidence is 50%. According to value of minimal support is 30%, we get the large itemset:

L = {{Mouse}, {Keyboard}, {Mp3}, {CD}, {Mouse, Keyboard}}

After scanning the subset in large itemset, we get l = {Mouse, Keyboard}. Then the second for loop scan the item in l, there are two subset in item, {Mouse} and {Keyboard}. The formula support(l) / support(x) $\geq$ α in this algorithm are applied to generate the valid association rules. The confidence of Mouse => Keyboard is:

$$\frac{Support(Mouse, Keyboard)}{Support(Mouse)} = \frac{60}{80} = 0.75$$

0.75 is greater than minimal confidence, hence, the algorithm add association rule, Mouse=>Keyboard, into association rule set R. Sequence, this algorithm compute the confidence of Keyboard=>Mouse:

$$\frac{Support(Mouse, Keyboard)}{Support(Keyboard)} = \frac{60}{60} = 1$$

1 satisfies the condition of association, then add association rule Keyboard=>Mouse into R. this algorithm keep generate the association rules until no further subset in large itemset. The next section of this paper will discuss the Apriori algorithm to generate the large itemset.

1.4 Apriori Algorithm

The follow lemma is used in Apriori algorithm to generate the large itemset[93].

Lemma: Any subset of a large itemset must be large

In other words, this lemma means if a large itemset is greater then specify minimal support, all its subset greater then this minimal support as well. Margaret H.Dunham claim large itemset is downward closure [data mining]. The figure 1 illustrates the relationship of large itemset and its subsets. In figure 1, there are four items {A, B, C, D}. The lines indicate the subset relationship between two items. For example {A} is a subset of {A, B}. So the large itemset lemma says in the path of subset relation, if an item is large, then other items above it are large as well. For example, if the item {ACD} in figure 2 is large, then {AC}, {AD}, {CD} is large, subsequence, {A}, {C}, {D} is large. On the other hand, if any subset is small, then, {ACD} is small.

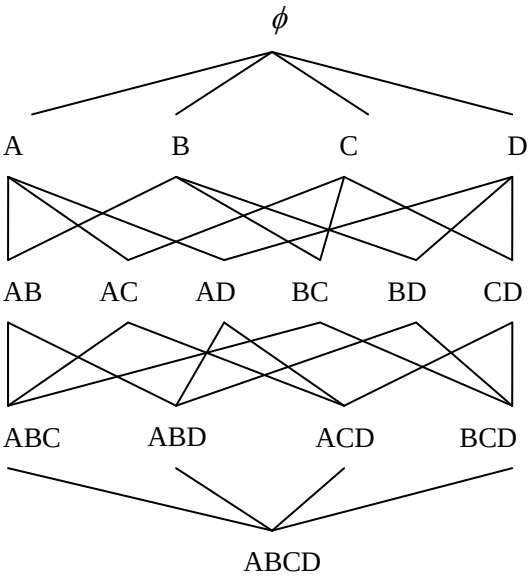


Figure 1

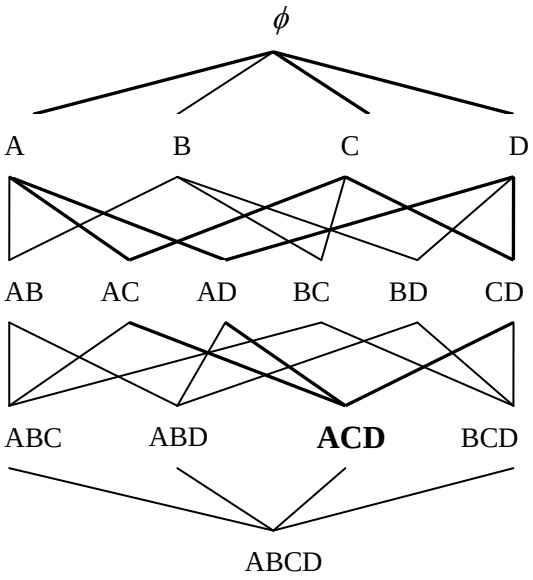


Figure 2

The idea of Apriori algorithm is to generate the candidate itemset from last large itemsets. Once the candidate itemset is generated, scan the whole candidate itemset, then generate the next large itemsets. During the process of Apriori algorithm, the size of candidate itemset is increased by one, denote it as Ci, and i is the size of candidate or i-th pass in the Apriori algorithm. Ci will generate the Li large itemset. Then Li joins with other items in it and generates the Ci+1 candidate itemset in the next pass. The following Apriori-gen Algorithm represents generate candidate itemset Ci from large itemset Li-1. All single items are treated as candidates in the first pass of Apriori-gen Algorithm.

```
Input:
    L_{i-1}      // Large itemsets of size i -1
Output:
    C_i      // Candidates of size i
Apriori-gen algorithm

    C_i = φ

    for each I ∈ L_{i-1} do
      for each J ≠ I ∈ L_{i-1} do
            If i – 2 of the elements in I and J are equal then
                C_k = C_k ∪ (I ∪ J);
```

Algorithm 2: Apriori-gen Algorithm

The figure 3 shows the whole process of Apriori algorithm. The Apriori algorithm is given in algorithm 3.

```
Input:
    I // Itemsets
    D       // Database of transactions
    S // Support
Output:
    L//Large itemsets
Apriori algorithm:
    K = 0; // k is used as the scan number.
    L = φ ;
    C_1 = I;
    Repeat
      k = k + 1;
      L = φ;
      for each I_i ∈ C_k do
            c_i = 0;      // Initial counts for each itemset are 0;
      for each t_j ∈ D do
            for each I_i ∈ C_k do
                  if I_i ∈ t_j then
                        c_i = c_i + 1;
      for each I_i ∈ C_k do
            if c_i ≥ (s X |D|) do
                  L_k = L_k ∪ I_i;
      L = L ∪ L_k;
      C_{k+1} = Apriori-Gen(L_k)

    Until C_{k+1} = φ ;
```

Algorithm 3: Apriori Algorithm

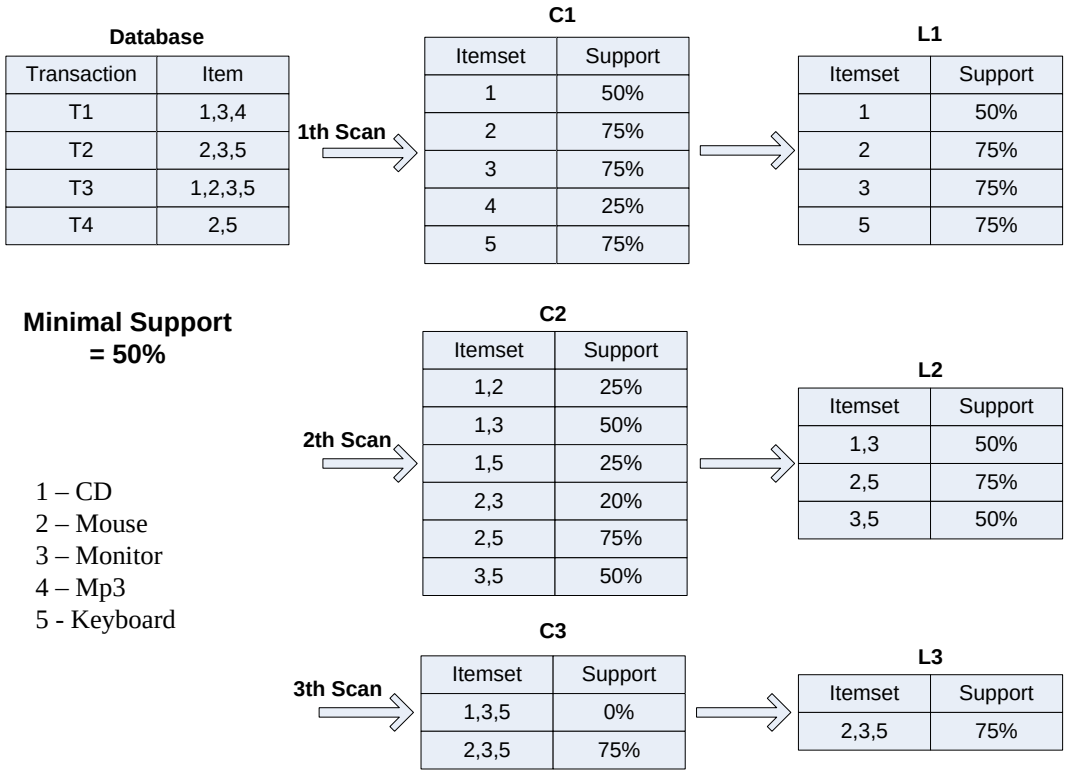


Figure 3: Apriori Algorithm

In figure 3, use number to stand for the items in database base, such as 1 stand for CD. Apply the Apriori algorithm to the electric store data, we get four candidate item in the first scan {1}, {2}, {3}, and {5}. In the second scan, Aproid-Gen algorithm (algorithm 2) was applier to L1 to generate the candidate item. Hence you have 3 + 2 + 1 = 6 candidate in second scan. In the third level scan, we join the item in large itemset. The rule to generate the candidate in this level is to combine the itemset with all others which has 1 (i – 2) same item. i indicates the i-th scan. According to this rule, {1, 3} and {3, 5} will combine, and generate the new candidate {1, 3, 5}. {1, 3} and {2, 5} do not satisfy the requirement, because there is not exactly one common item in these two itemset. After third scan, we have large itemset, L3. There are only one itemset in L3. Hence not new candidate will generate, the algorithm is completed.

2.0 Case study 1

2.1 Introduction

In the real world, data in the databases always is stored with the time sequence. An item always has an associationship which based on time with other items. In other words, compared with the normal time, the frequency of an item may increase in a particular time period, such as the seasonal selling of food. Traditional mining association rules focus on finding the unordered association between items. In this article, the temporal association rule which is an association rule will be attached importance to. Compared with the classic association rule, the temporal association rule is hold during the specific time intervals. In this paper, the temporal association rule which is hold during the specific time will be explored by user-given calendar schemas. The well-known Apriori algorithm will be used to support the research.

In this article, firstly, the background about temporal association rules will be indicated. Secondly, temporal association rules will be defined in terms of calendar patterns. The next section, Apriori algorithm which is a basic algorithm applied in association rule will be extended to support to discovering the temporal association rules. For example, Apriori is used to find the large itemsets for temporal association rules in temporal database. Also, the techniques which are proposed to optimize the results will be indicated. In the fourth section, the evaluation of extended algorithms will be presented. Finally, a conclusion will be given.

2.2Background and challenges

Association rule was proposed to use in transactions for capturing the concurrence of items. In other words, the usual objective is to find the regularities of appearance of certain events and the temporal relationships which are hold between different events. Temporal association rules are suitable to be used in transactions of business and daily life. Take example for a transaction of supermarket, we may discover that the relationship of turkey and pumpkin pie is not close in daily life. They are seldom sold. However, when Thanksgiving coming, the frequency that they sell together increase. In other word, there is a temporal association rule of the form turkey -> pumpkin pie in Thanksgiving. There is an example that a temporal association rule between chocolates and flowers on Saint Valentine's Day in every year.

Temporal association rules are more useful in mining business timestamped transactions. The reason is that time varying features exist in most of the real world. For example, the company can change their strategy along with varying time to maximize their benefits.

The challenges of applying temporal association rules is to discovery all temporal association rules from a set of timestamped transactions, and generating large itemsets for supporting discovering temporal association rules.

Calendar-based temporal association rules

2.2.1 Calendar pattern

Calendar schemas are proposed to support the discovering of temporal association rules, such as (year, month, day). For example, a calendar pattern which likes every 14th day of February of every year can be achieved. Each calendar pattern will be designed to define a group of time intervals.

The form is R= (Gn: Dn, Gn-1: Dn-1,….,G1:D1) with a valid constraint. The attribute Gi is names such as year, month, week and hour. Di is the value. The constraint valid is achieved by computing Dn *Dn-1 *…*D1. A simple calendar-based pattern on R(calendar schema) is (dn, dn-1,…,d1). This calendar pattern means that a set of time intervals can be described as “the d1th G1 of the d2th G2,…, of dnth Gn.” For example, the calendar schema is (year, month, week), and the calendar pattern is (1, 1, 3). Then, the means can be described as “the third week of every month in every year”.

2.2.2 Temporal association rules

In this article, we assume that every transaction is ordered. When a transaction happened in a period of specific time interval, the transaction will be timestamped. Thence, a set of data which bond transactions and a calendar schema will be obtained. A form such like(r, e) is proposed. This form means a pair of association rule and calendar pattern. R represents an association rule which satisfies the minimum support and confidence constraint, and the association rule should occur during each time interval of the calendar pattern e which is consistent with R.

According to the results of analysis, there are two kind of temporal association rules. The first one is called temporal association rules w.r.t precise match. Another one is called temporal association rules w.r.t. fuzzy match. For example, if an association rule of the form Christmas pudding -> turkey along with the calendar pattern every day in every December, the association rules can be called temporal association rules

w.r.t. precise match. However, for some transactions, the precision of matching in the temporal association rules w.r.t. precise match is required too high to achieve. Instead of precise match, another one is temporal association rules w.r.t. fuzzy match. For example, the association rule of form Christmas pudding -> turkey may not occur on every day of every December, but the rules can satisfy the precision which more than 80% of every December. In this situation, association rules do not have the high precision, but they can be hold during enough number of time intervals which belong to the relevant calendar pattern.

2.4 Algorithm for finding large itemsets

2.4.1 Overview

Two steps should be considered while mining temporal association rules. The first step is that all large itemsets are obtained by using algorithms. Also the all itemsets satisfy all calendar patterns which on the proposed calendar schema. The second one is how to generate the temporal association rules by the achieved large itemsets and the corresponding calendar patterns.

Apriori algorithm is used to find the large itemsets w.r.t. precise and fuzzy match. Three terms which are calendar schema, a group of timestamped transactions and a minimum support, need to be considered when get large itemsets with precise match. There is a additional term which need to be considered on fuzzy match. The additional term is a match ratio.
The processing of each basic time interval consists of three phases. In the first phase, candidate large itemsets are producted for the basic time intervals. in the second phase, all large itemsets will be found for the basic time interval by exploring the timestamped transactions and updating the supports of candidate large itemsets. Finally, in the last phase, the large itemsets for each calendar pattern which can generalize the basic intervals will be updated by using the exist large itemsets.

2.4.2 Generating candidate large itemsets

It is not useful to generating candidate large itemsets when using Apriori algorithm directly. In this case study, the important point is not discovering association rules which only exist during basic time intervals. Two optimization techniques are proposed. the two techniques are temporal aprioriGen and horizontal pruning which work together to generate a algorithm named temporal-Apriori. There are two Theorems. The first theorem is that “temporal- apriori for precise match is equivalent to direct-apriori for precise match”. the second theorem is that “temporal-apriori for

fuzzy match is equivalent to direct-apriori for fuzzy match".

2.4.3 Evaluation

Synthetic data sets will be used in the experiments which are for evaluating the performance of the temporal-Apriori algorithm. The following items are parameters for data generation.
|D| --number of transactions per basic time interval. Default value is 10,000
|T| -- average size of the transactions. Default value: 10
|I| -- average size of the maximal large itemsets. Default value: 4.
|L| -- number of per-interval itemsets. default value: 1,000
N – number of items. default value: 1000
Pr – pattern-ratio. Default value: 0.4
Np—number of calendar patterns per pattern itemset. Default value: 40.
Pattern itemsets are decided by considered several calendar patterns into each pattern. Large pattern itemsets are achieved from pattern itemsets by repeatedly considering itemsets. If a chosen pattern itemset has a calendar pattern which can covers the basic time interval, the pattern itemset will be used as a per-intercal itemset. by contraries, the pattern itemset will be ignored. In this experiment, the size of data sets varies from 739MB to 5.41GB. As the result, temporal-apriori is faster than direct-apriori when generating temporal association rules w.r.t precise match and fuzzy match.

2.4.4 Conclusion

In this article, temporal association rules are explained. Temporal association rules are useful when are applied in the timestamped transactions. The background of temporal association rules is outlined; the challenges of discovering the temporal association rules also are indicated. In this article, temporal-apriori algorithm is proposed to solve the challenge.

Firstly, temporal association rules are association rules with corresponding temporal patterns based on calendar schemas. The knowledge of calendar schema and calendar pattern is described in detail. Then two temporal association rules w.r.t precise match and temporal association rules w.r.t. fuzzy match are proposed. Moreover, Apriori algorithm which is the basic algorithm of association rules is modified to satisfy discovering the temporal association rules w.r.t both precise match and fuzzy match. Temporal-apriori algorithm is proposed by exploring the relationships between calendar patterns. Lastly, the performance of algorithm is tested by using large date sets. The results prove that the temporal- apriori algorithm which proposed in this article is faster than direct-apriori algorithm when discovering temporal association rules.

3.0 Case Study 2

3.1 Introduction

The new product development (NPD) is a crucial factor during the competition among different enterprises. The purpose of the NPD is to fulfill the diverse customers' needs and wants. Accordingly, the enterprises can get more reputation and profits. The coming of information technology leads the business going into a new age. Especially the database helps the business developed a lot. The data mining technology which can extract useful information from the huge database is an effective tool during the decision making. In this case study, the product of cosmetics has taken as a research object. The data is collected and then form a rational database. To develop new cosmetic product, the data mining technology is applied to extract the relation between varies cosmetics products. The Apriori algorithm has selected to do this association rule discovery.

3.2 The challenge of the NPD

The new product development is an important factor which can enhance the enterprises' competitive. However, it is difficult to decide the preference of the customers. Furthermore, the preference of the customers is always changing. A new product need time to introduce. As the result, the new product should satisfy the needs of the customers when it is put in the market. Otherwise, the product is fail which will force the supplier into a reactionary mode. [3]

Secondly, it is difficult to explore the whole database to get useful information. Inefficient collecting of data will be a waste of time and money, making the database becomes "data dumps". By the data mining technologies, the knowledge can be extracted and combined with the researchers' knowledge. The final information will be efficient evidence to help the enterprises to find the customer-oriented goods to develop new products to meet the customers' needs and wants. [3]

3.3 The challenge of the research

This case study takes the cosmetics as an example. Different from some other products, the cosmetics is a kind of personal product. It is required both the male and female from different age. Thus, the cosmetics product should satisfy different

customer segments. The wants and needs of the customers is the main resource during the process of NPD. [3]

Secondly, cosmetic is a kind of product which includes different collection of cosmetics to fulfil the different demands. For instance, different skin type needs different skin care products. Different functions of the cosmetics increase the NPD's difficulty. [3]

In the case study, the data of the product, customers and the transactions will be focused to do the association rule discovery.

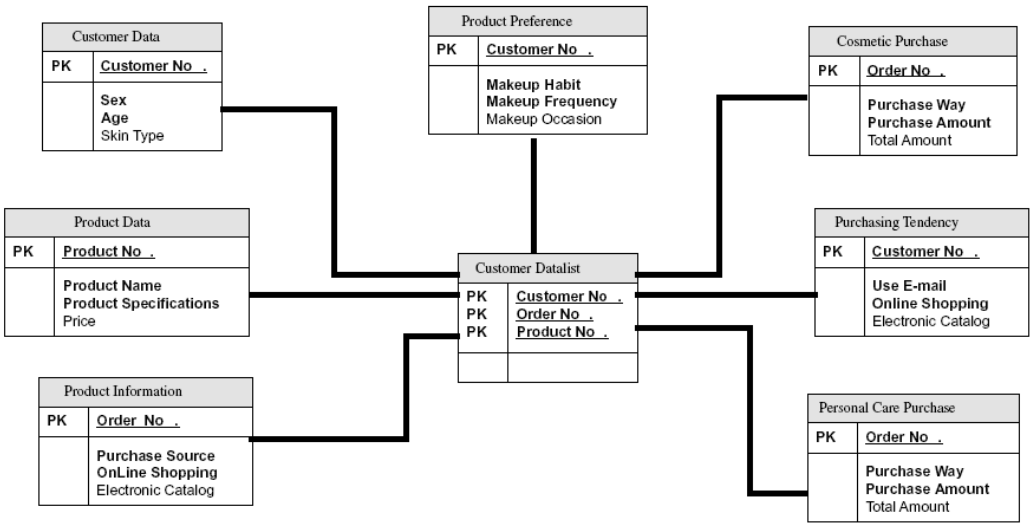
3.4 The Background

The data are collected from Taipei, Taichung, Kaoshiung and Hisinchu cities between May and October 2004. They are four large cities in Taiwan. The way to collect these data is to interview the customer who has consumed or use the cosmetics from the cosmetics' store or channel in these cities by questionnaire. These questionnaires including 7 tables: customer basic data table; customer product preference data table; customer skin attitudes data table; product function data table; customer suggestion data table; product data table and transaction data table. 1400 questionnaires has sent while 1009 valid returned.

3.5 The rational database

A rational database is the basic preparation for this case study. The rational database can give the clear relationship of each data. Consequently, the rational database is a useful way for mining data in the database. In the study, the database includes 8 entities, 7 relationships and 47 attributes. [3]

The rational database used in this case is as follow:



3.6 Review of the Association Rule and Apriori Algorithm

In this case, $Conf(X \rightarrow Y)$ represents if the transaction involves item set X, it will contains the item set Y in a high opportunity. To get the association rules among these item sets, there are two measure standard: the minimum support which is named as Minsup and the minimum confidence which is named as Minconf. Therefore, the basic pricinple of Apriori algorithm can be divided into two steps: the first step is to generate a large item set which is larger than the Minsup. The second step is abstract the association rules from large item set. The rules should obey these two conditions:

$$Sup(X \cup Y, D) \geq Min\sup$$

$$Conf(X \rightarrow Y) \geq Minconf$$

The detail of how to realize the Apriori algorithm has been described in the front of this paper. [3]

3.7 Data mining process

The purchase behavior of the customer is crucial in the product development and it can be detected from the database. In this case, the association rule can not only focus on the single customer but also the collection of the customers such as the family. Thus, the information should be selected from different rage of customer including different age, sex. The information can be categorized into product knowledge and marketing knowledge as well. And the product knowledge will help to do the NPD.

Abstracting the information about the product knowledge is needed in the NPD process. The Customers' opinions and experience should be considered as well. Then this information will be combined with the current information to do the NPD. As the result, the new product can meet the customers' requirements. [3]

The process of mining implementing the Apriori algorithm is to use different combination of decision variables such as customer sex+age+residential area+skin type+product sytle etc. [3]

Personal care products

Complicated skin type

Table 1
Association rules of basic personal care and complicated skin type (min sup = 12%, min conf = 60%)

Rule	Sup	Conf	Lift	Consequent	Antecedent				
R_{A1}	13.4	65.2	1.621	Complicated skin type	Cleansing powder	Age = 25–29	Makeup remover	Sex = F	Facial cleanser
R_{A2}	13.4	63.7	1.592	Pure white moisturizer	Cleansing powder	Age = 25–29	Makeup remover	Sex = F	Facial cleanser
R_{A3}	13.3	60.4	1.503	Complicated skin type	Moisture rich toner	Age = 25–29	Makeup remover	Sex = F	Facial cleanser
R_{A4}	13.6	68.6	1.476	Age = 25–29	Cleansing powder	Pure white moisturizer	Moisture rich toner	Sex = F	Facial cleanser
R_{A5}	13.5	64.7	1.338	Makeup remover	Cleansing powder	Moisture rich toner	Age = 25–29	Sex = F	Facial cleanser

This is the result of the first step of data mining. In this situation, the minimum

support is 12%. From this table we implement the Apriori algorithm, then the customer who buys the cleansing powder, makeup remover and facial cleanser would have the complicated skin type. [3]

Oily skin type

Association rules for basic personal care and oily skin type (min sup = 8%, min conf = 55%)

Rule	Sup	Conf	Lift	Consequent	Antecedent			
R_{B1}	8.70	68.20	2.100	Cleansing powder	Delicate care toner	Pure white moisturizer	Moisture rich toner	Sex = F Facial cleanser
R_{B2}	10.00	60.40	1.860	Cleansing powder	Cleansing cream	Pure white moisturizer	Moisture rich toner	Sex = F Facial cleanser
R_{B3}	9.70	59.20	1.823	Cleansing powder	Area = Kaoshiung	Moisture rich toner	Makeup remover	Sex = F Facial cleanser
R_{B4}	8.80	60.70	1.813	Cleansing cream	Nourishing cream	Pure white moisturizer	Moisture rich toner	Sex = F Facial cleanser
R_{B5}	8.60	58.60	1.805	Cleansing powder	Moisture rich cream	Pure white moisturizer	Moisture rich toner	Sex = F Facial cleanser
R_{B6}	9.50	58.30	1.796	Cleansing powder	Refreshing moisturizer	Area = Kaoshiung	Makeup remover	Sex = F Facial cleanser
R_{B7}	9.80	69.70	1.742	Pure white moisturizer	Cleansing powder	Area = Kaoshiung	Moisture rich toner	Sex = F Facial cleanser
R_{B8}	9.00	68.10	1.702	Pure white moisturizer	Eye and lip makeup remover	Cleansing powder	Moisture rich toner	Sex = F Facial cleanser
R_{B9}	11.30	66.70	1.666	Pure white moisturizer	Foaming cleanser	Cleansing wipes	Cleansing powder	Sex = F Facial cleanser
R_{B10}	9.30	57.40	1.435	Pure white moisturizer	Cleansing powder	Area = Kaoshiung	Makeup remover	Sex = F Facial cleanser
R_{B11}	11.30	65.80	1.419	Moisture rich toner	Foaming cleanser	Cleansing wipes	Cleansing powder	Sex = F Facial cleanser
R_{B12}	8.30	66.70	1.379	Makeup remover	Pure white moisturizer	Area = Taipei	Moisture rich toner	Sex = F Facial cleanser
R_{B13}	8.90	62.20	1.342	Moisture rich toner	Pure white moisturizer	Area = Taipei	Makeup remover	Sex = F Facial cleanser
R_{B14}	13.60	63.50	1.313	Makeup remover	Cleansing powder	Pure white moisturizer	Moisture rich toner	Sex = F Facial cleanser
R_{B15}	9.00	61.50	1.272	Makeup remover	Eye and lip makeup remover	Cleansing powder	Moisture rich toner	Sex = F Facial cleanser
R_{B16}	10.90	58.20	1.255	Moisture rich toner	Cleansing powder	Refreshing moisturizer	Makeup remover	Sex = F Facial cleanser
R_{B17}	8.90	56.70	1.222	Moisture rich toner	Area = Kaoshiung	Pure white moisturizer	Makeup remover	Sex = F Facial cleanser
R_{B18}	9.50	55.20	1.190	Moisture rich toner	Refreshing moisturizer	Area = Kaoshiung	Makeup remover	Sex = F Facial cleanser
R_{B19}	9.30	57.40	1.188	Makeup remover	Cleansing powder	Area = Kaoshiung	Pure white moisturizer	Sex = F Facial cleanser

This is the result of the first step. After the Apriori algorithm, the conclusion comes as the women who have the oily skin type like to us the delicate care toner, pure white moisturizer, moisture rich toner and facial cleanser products. [3]

Normal skin type

Association rules of basic personal care and normal skin type (min sup = 11%, min conf = 50%)

Rule	Sup	Conf	Lift	Consequent	Antecedent		
R_{C1}	11.30	57.90	1.783	Cleansing powder	Refreshing moisturizer	Age = 25–29	Makeup remover
R_{C2}	11.60	54.70	1.685	Cleansing powder	Refreshing moisturizer	Moisture rich toner	Facial cleanser
R_{C3}	11.30	66.70	1.666	Pure white moisturizer	Foaming cleanser	Cleansing wipes	Cleansing powder
R_{C4}	13.10	65.90	1.647	Pure white moisturizer	Cleansing powder	Moisture rich toner	Makeup remover
R_{C5}	13.40	63.70	1.592	Pure white moisturizer	Cleansing powder	Age = 25–29	Makeup remover
R_{C6}	11.30	71.90	1.548	Age = 25–29	Foaming cleanser	Cleansing wipes	Cleansing powder
R_{C7}	13.60	68.60	1.476	Age = 25–29	Cleansing powder	Pure white moisturizer	Moisture rich toner
R_{C8}	11.30	65.80	1.419	Moisture rich toner	Foaming cleanser	Cleansing wipes	Cleansing powder
R_{C9}	13.40	65.20	1.406	Moisture rich toner	Cleansing powder	Age = 25–29	Makeup remover
R_{C10}	13.30	50.70	1.392	Refreshing moisturizer	Moisture rich toner	Age = 25–29	Makeup remover
R_{C11}	13.30	54.50	1.361	Pure white moisturizer	Moisture rich toner	Age = 25–29	Makeup remover
R_{C12}	13.60	50.40	1.342	Sex = F	Cleansing powder	Pure white moisturizer	Moisture rich toner
R_{C13}	13.60	63.50	1.313	Makeup remover	Cleansing powder	Pure white moisturizer	Moisture rich toner
R_{C14}	11.30	62.30	1.288	Makeup remover	Foaming cleanser	Cleansing wipes	Cleansing powder
R_{C15}	11.60	58.10	1.251	Age = 25–29	Refreshing moisturizer	Moisture rich toner	Makeup remover

The result of the table reveals that the people who has the normal skin type tend to consume the cleansing powder and refreshing moisturizer and makeup remover. [3]

Special personal care

Association rules of special personal care products (min sup = 5%, min conf = 50%)

Rule	Sup	Conf	Lift	Consequent	Antecedent		
R_{D1}	5.40	51.90	1.832	Black-head treatment	Acne mask	Mask	Area = Taipei
R_{D2}	6.60	59.10	1.831	Eye gel	Acne mask	Mask	Sun protection lotion
R_{D3}	5.70	57.90	1.794	Sun protection lotion	Acne mask	Black-head treatment	Mask
R_{D4}	5.10	60.80	1.565	Sun protection lotion	Eye cream	Area = Kaoshiung	Age = 25–29
R_{D5}	6.10	72.10	1.552	Age = 25–29	After shave lotion	After shave balm	Shaving foam
R_{D6}	6.60	72.70	1.465	Sun protection lotion	Acne mask	Mask	Age = 25–29
R_{D7}	7.10	67.60	1.455	Age = 25–29	Eye cream	Mask	Sun protection lotion
R_{D8}	5.80	53.40	1.424	Area = Kaoshiung	Eye cream	Sun protection lotion	Age = 25–29
R_{D9}	6.80	66.20	1.424	Age = 25–29	Concentrated treatment	Sun protection lotion	Eye cream
R_{D10}	5.80	50.00	1.414	Mask	Eye cream	Sun protection lotion	Age = 25–29
R_{D11}	6.00	50.00	1.414	Mask	Age = 30–34	Area = Kaoshiung	Eye cream
R_{D12}	7.20	63.00	1.356	Age = 25–29	Lengthening mascara	Sun protection lotion	Eye cream
R_{D13}	5.70	52.60	1.355	Sun protection lotion	Eye gel	Eye cream	Age = 25–29
R_{D14}	6.20	50.00	1.332	Sun protection lotion	Acne mask	Eye gel	Mask
R_{D15}	6.20	56.50	1.137	Eye cream	Acne mask	Eye gel	Mask

The table tells that the women in Taipei always consider the black-head treatment, acne mask and mask. [3]

Cosmetics products

Association rules of cosmetics products (min sup = 10%, min conf = 55%)

Rule	Sup	Conf	Lift	Consequent	Antecedent		
R_{E1}	13.60	55.50	2.429	LipColor	Thickening type mascara	Lipstick	Nail enamel
R_{E2}	10.00	62.40	2.318	Age = 20–24	Lipglide	Area = Taipei	Nail enamel
R_{E3}	10.80	60.60	1.859	Loose powder	Thickening type mascara	LipStick	Sun protection foundation
R_{E4}	12.80	57.40	1.761	Loose powder	Calming make-up base	Sun protection foundation	Age = 25–29
R_{E5}	15.70	67.10	1.723	Powder foundation	Nail top coat	Nail base coat	Nail enamel
R_{E6}	12.00	66.10	1.698	Powder foundation	Lipstick	Lip color	Age = 25–29
R_{E7}	10.10	63.70	1.637	Powder foundation	Lip gloss	Moisture rich lipstick	Ravishing type mascara
R_{E8}	10.90	56.40	1.626	Sun protection foundation	Ravishing type mascara	Thickening type mascara	Lipstick
R_{E9}	10.00	55.40	1.600	Sun protection foundation	Moisture rich lipstick	Long-bristle type mascara	Lip color
R_{E10}	12.00	55.40	1.598	Sun protection foundation	Long-bristle type mascara	Lip color	Age = 25–29
R_{E11}	10.10	59.80	1.536	Powder foundation	Waterproof mascara	Lip color	Lipstick
R_{E12}	10.30	59.60	1.531	Powder foundation	Ravishing type mascara	Lip color	Lipstick
R_{E13}	12.80	56.60	1.454	Powder foundation	Calming make-up base	Sun protection foundation	Age = 25–29
R_{E14}	10.20	64.10	1.376	Age = 25–29	Lip gloss	Ravishing type mascara	Thickening type mascara
R_{E15}	12.80	71.30	1.350	Nail enamel	Calming make-up base	Sun protection foundation	age = 25–29
R_{E16}	10.00	60.40	1.297	Age = 25–29	Long-bristle type mascara	Lip color	Lipstick

Similar as described above, this table says that the women always buy the lipcolor, thickening type, lipstick and nail enamel together. There are potential relationship among these item sets. [3]

After several steps of algorithm, the information can be abstracted as follows:

Basic personal care product collection:

Basic personal care product collection

Skin type	Product collection	Collection ID	
Complicated skin type	Makeup remover + cleansing powder + moisture rich toner + pure white moisturizer	A1	a1 + a4 + a7 + a8
Oily skin type	Cleansing powder + moisture rich toner + pure white moisturizer + delicate care toner	A2	a4 + a7 + a8 + a6
	Makeup remover + cleansing powder + moisture rich toner + refreshing moisturizer	A3	a1 + a4 + a7 + a9
	Cleansing cream + cleansing powder + moisture rich toner + pure white moisturizer	A4	a2 + a4 + a7 + a8
	Eye and lip makeup remover + cleansing powder + moisture rich toner + pure white moisturizer	A5	a3 + a4 + a7 + a8
Normal skin type	makeup remover + foaming cleanser + refreshing moisturizer + moisture rich toner	A6	a1 + a5 + a9 + a7

Special personal care product collection:

Special personal care product collection

Product collection	Collection ID	
Black-head treatment + acne mask + mask + sun protection lotion	B1	b1 + b2 + b3 + b6
Eye gel + acne mask + mask + sun protection lotion	B2	b4 + b2 + b3 + b6
Eye gel + sun protection lotion + mask	B3	b4 + b6 + b3
Shaving foam + after shave balm + after shave lotion	B4	b7 + b8 + b9

Cosmetics product collection:

Cosmetic product collection

Product collection	Collection ID	
Calming make-up base + sun protection foundation + loose powder + nail enamel	C1	c1 + c2 + c3 + c9
Sun protection foundation + loose powder + lipstick + thickening type mascara	C2	c2 + c3 + c6 + c4
Sun protection foundation + calming make-up base + thickening type mascara + lipgloss	C3	c2 + c1 + c4 + c8
Thickening type mascara + lipcolor + lipstick + nail enamel	C4	c4 + c5 + c6 + c9

These tables reveal the relationship among different product and the characterizes of the customers. This information is useful to help the companies to do the decision making and products management. The researcher can find concealed needs and wants of the customers in these kinds of data mining. Therefore, the new product can be developed based on this useful information. Then, the new product can adapt the demands from the market. [3]

3.6 Analysis of the result

The wants and needs of the customers are complicated and changing all the time. If the companies can always keep this concept which is to fulfill the customers' demand in their mind and provide better service, they will win in the fury competition today.

It is important to collect information including the data about different customers, their preferences, and their experience. These data is used to do the data mining. The association rule is an important evidence to help the companies to do the decision making. To abstract useful relationship among different item set is a vital process to get information. Actually, there are varies algorithm in association rules discovery. In this case, the Apriori algorithm has been implemented. Different algorithms can be applied in different domains, area and situation. [3]

In this case, the Apriori algorithm helps to mining data from large database. The decision variables are set to do segmentation data mining. As the result, there are lots of relationship has been abstract from the database. From these relationships, for example, for oily skin type people, they tend to use the cosmetics which can remove the oil from their skin, the discovery should be combined with the basic knowledge of the companies to find the actual needs and wants of the customers. As the result, if the

new product is specified for the oily skin type people, the basic condition is the new product can remove the oil from the skin. This is an important process during the NPD. [3]

From this case study, the data mining and the association rules discovery is important and useful for the NPD. The Apriori algorithm can be implemented for many other purposes as well. The data mining is an efficient tool in many other areas too. [3]

4. Conclusion

Reference

[1] Askoka Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. *Processing of the International Very Large Databases Conference*, pages 432-444, 1995;

[2] Hannu Toivonen. Sampling large database for associatin rules. *Proceedings of the International Very Large Databases Conference*, page 134-145, 1996;

[3] Liao S.H. & Hsien C.L. & Huang S.P., *Mining product maps for new product development, Expert Systems with Applications*, Volume 34, Issue1, P50-62, January 2008

[4] Margaret H. Dunham. *Data Mining Introductory and Advanced Topics*. Prentice Hall, New Jersey, 2002;

[5] Marisa S. Viveros, John P. Nearhos and Michael J. Rothman, Applying Data Mining Techniques to a Health Insurance Information System, In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan and Nandlal L. Sarda eds., *Proceedings of the 22nd International Conference on Very Large Data Bases (VLDB)*, Mumbai, India, pp. 286-293, September 1996;

[6] Rakesh Agrawal and John C. Shafer. Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):962-969, December 1996;

[7] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. *Proceedings of the ACM Internation Conference on Management of Data*, pages 207-216, 1993;