ISSN 2277-3061

# Forecasting Personal Shopping Behavior

YEH, Hsiaoping

Department of Marketing & Distribution Management, National Kaohsiung University of Science & Technology, Taiwan

hpyeh2000@gmail.com

## ABSTRACT

Data mining (DM) techniques make efforts to discovery knowledge from data. Aiming to finding patterns, association rule (AR) computing algorithms seem to be one to be adopted on variety applications. To be originally claimed for best analyzing customer shopping goods in baskets, Apriori, the first AR algorithm, has been discussed and modified the most by researchers. This study adopts Apriori algorithm to forecast individual customer shopping behavior. This study finds that customer shopping behaviors can be comprehended better in a long run. With Apriori mining and the examining principles proposed by this study, customer purchase behaviors of no matter constant purchase, stopping purchasing habitual goods, and starting to purchase goods that never bought before is able to be recognized. However, impulse purchase, including purchase for holidays, is unable to be discovered.

## Indexing terms/Keywords

Data mining, customer shopping behavior, association rule, Apriori, retailing

ISSN 2277-3061

## INTRODUCTION

Data systems nowadays have been widely adopted in businesses. The massive data collected from business transactions have made entrepreneurs realize to use the data for supporting their business decision makings (Clifford, 2008). Knowledge Discovery in Database (KDD) (Piatesky-Shapiro, 1991) therefore becomes a wave of essential concept to extract useful and valid knowledge from data even though the knowledge is intuitive or previously unknown. The processes and algorithms of DM aim to find patterns that describe underlying relationships in the data. Fields of data usually are dependent yet recessive. If such dependent and recessive patterns could be discovered with DM techniques, for example, AR and clustering, the results would be useful for businesses and industries to make important decisions. AR mining looks for frequently occurring patterns in the data and is often used for market basket analysis. The famous result is the diaper-beer rule in baskets. Benefits for retailers are better shelf management, goods supply, and market promotion.

Basket analysis discovery of a highly relation between beers and diapers already circulates with many versions. A chain store in the USA even forecasted the time of being pregnant of single high school girls from their shopping goods. Although the truth is dispended, the stories told from transaction data, based on AR mining, reveal meaningful results to contribute to retail planning and management.

Customization techniques, to accommodate differences among individuals, enable the dynamic insertion, personalization, or suggestion of contents in any format that is relevant to the individual user, based on the user's implicit/explicit behavior and preferences. There are various ways to set levels of interestingness for discovering so called strong (or interesting) ARs (Piatetsky-Shapiro, 1991) but none of the settings is recognized to be appropriate or "accurate" rule-mining for a business (e.g., organization, industry), a procedure (e.g., firm's marking strategy, business model), or people (e.g., customers). While customers are claimed to be beneficial from DM, the purposes of this study are twofold. In addition to suggest how to decide values of interestingnesses for generating ARs for each individual customer, i.e., the strong rules for each individual customer even s/he shops at different retailing stores, this study also provides principles to interpret changes on each customer's purchasing goods in her/his shopping baskets. Providing a reliable shopping list suggestion for a customer before s/he enters the store, designing a better customized promotion for each customer, and attracting customers visiting the store repeatedly would be on the corner for the retailers.

## RELATED WORK AND REVIEW

Piatetsky-Shapiro (1991) defined AR as that there exists one type of association among a set of events. Agrawal and his colleagues (Agrawal et al., 1993; Agrawal and Srikant, 1994) then defined and proposed an association between two different sets of events within a database. For two events **A** and **B**, an AR {**A**}→{**B**} interprets that event **B** occurs while **A** occurs where $\mathbf{A} \cap \mathbf{B} = \varnothing$. AR originates with the market basket analysis. It is used to describe the relationship among goods in a shopping basket. If customers buy goods in itemset **A** will then often buy goods in itemset **B**, rule {**A**}→{**B**} is worthy for retailers to take notice of items in **A** and **B**. Customer shopping behaviors are therefore able to be predicted based on this conditional relation of itemsets **A** and **B** appearing in a period of time of customer shopping baskets. Retailers can extract important features from ARs which are translated into knowledge for product or marketing designs.

Based on Agrawal et al. (1993), AR is to find valuable associations of items in itemset **I** where $\mathbf{I}=\{i_1, i_2, i_3, \ldots i_R\}$ is a set of all purchased items over a transaction set **T** where $\mathbf{T}=\{t_1, t_2, t_3, \ldots t_C\}$ represents a set of itemsets where each itemset contains items in shopping basket at each purchase. Hence, $t_n$, $\forall n$=1~C, is a subset of **I**. The frequency of items appearing in shopping baskets can therefore be defined as $\sum(i_m)$=Count($t_n \mid i_m \subseteq t_n, i_m \subseteq \mathbf{I}, \forall t_n \in \mathbf{T}$). Take the transaction data in Table 1 for example, cookie and beer are both purchased at transaction Nos 1, 3, and 5, i.e., $t_1$, $t_3$, and $t_5$. Hence, the frequency of purchasing both cookie and beer is 3, i.e., $i_m$ = {cookie, beer} and $\sum(i_m)$=3.

**Table 1. Receipt data with binary interpretations**

| Transaction sequence | Purchased items | Binary data | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | cookie | milk | beer | egg | tissue | soda | coffee |
| 1 | Cookie, milk, beer, egg, tissue, soda, coffee | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | Beer, egg, soda | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 3 | Cookie, beer, tissue, soda, coffee | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 4 | Milk, egg, tissue, soda | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 5 | Cookie, beer, egg, tissue, soda, coffee | 1 | 0 | 1 | 1 | 1 | 1 | 1 |

They further proposed an AR algorithm, Apriori (Agrawal and Srikant, 1994). With Arpiori, *support* and *confidence* are two measures (definitions are in equations (1) and (2)), named as interestingness afterwards, used to describe the degree of association of **A** (named antecedent) and **B** (named consequent) for the rule {**A**}→{**B**}. *Support* denotes the frequency of buying goods all in **A** and **B** for a period of transactions, i.e., the probability of buying all items in **A** and **B**. *Confidence* denotes the frequency of buying goods in **B** under the condition of buying goods in **A**, i.e., the conditional probability of buying all items in **B** while goods in **A** are bought.

$$Support(\{\mathbf{A}\}\rightarrow\{\mathbf{B}\}) = \frac{\sum(\mathbf{A}\sqcup\mathbf{B})}{N} = p(\mathbf{A}\cup\mathbf{B}) \quad (1)$$

$$Confidence(\{\mathbf{A}\}\rightarrow\{\mathbf{B}\}) = \frac{\sum(\mathbf{A}\sqcup\mathbf{B})}{\sum(\mathbf{A})} = p(\mathbf{B}|\mathbf{A}) \quad (2)$$

where $\sum$ means the total counts of purchasing goods under the total transactions, *N*, in a period of time.

Take data in Table 1 for example, considering an AR {beer}→{coffee}, the values of it's *support* and *confidence* are $\frac{3}{5}$ (=0.6) and $\frac{3}{4}$ (=0.75), respectively. It interprets that, for the period of transactions, the probability of buying both beer and coffee is 0.6 and the probability of buying coffee while beer is bought is 0.75.

Mining data by Apriori has been currently applied on financial services such as the important effects of customer paying debt loans (Theresia and Beta, 2012; Li and Wang, 2013), on biomedical researches such as finding causes of DNA to diseases (Liang et al., 2010), on tele-communication such as individual customer paid mobile service interests (Yao and Shu, 2009), and on web safety such as web monitoring to detect illegal or suspicious intrusions (Lee and Salvatore, 1998; Lee et al., 2002). Han et al. (2000) and Lawrence et al. (2001) have testified that AR mining shopping basket data is able effectively to predict personalized shopping behavior.

An AR with high *support* and *confidence* is called strong (or interesting) rule and is potentially useful for a system. To decide whether a rule being considered to be valuable, levels of *support* and *confidence* are first determined. That is, ARs are extracted for system usages when their values of *support* and *confidence* have to be greater than thresholds of minimum *support* and minimum *confidence*. Since numbers of different items in basket may be large, a set of frequent itemsets (i.e., items often purchased) is first to be derived by adopting minimum *support*. This is, so called "join", the first step of Apriori. Strong ARs are then discovered, by the second step "prune", with the rules' *confidences* being greater than minimum *confidence*.

To be noted, AR is not used to discover single items often appearing in shopper baskets but to identify two (or more) different items relatively appearing in the baskets together. *Support* indicates frequencies of the occurring patterns in a rule and *confidence* denotes the strength of this implication. Other than *support* and *confidence*, which are subjective measures, many subjective and objective ones are proposed to derive strong ARs (Geng and Hamilton (2006) summarized 38 measures) as well as finding valuable rules by visualization procedures (Klemettinen et al., 1994). Besides Apriori, similar logics to Apriori (generally grouped as Apriori-like algorithms and there is a survey in Han et al. (2000)) are an interesting topic in the research field of AR. Subjective interestingnesses are based on the background of the problem, the knowledge of the domain, and the expectation of the experts. They are not represented by strict mathematical formulas because of the variance of knowledge, requirement and environment. On the other hand, objective ones are designed to evaluate the generality and reliability of the ARs. It is generally accepted that there is no single interestingness that is perfect and applicable to all problems. Usually different ones are complementary and can be applied at different applications or phases for matching the properties of the problems (Tan et al., 2002; Geng and Hamilton, 2006; Zhang, 2009).

The DM methods are mostly applied in large datasets. Mining valuable ARs is very likely to generate numerous rules from which it is difficult to build a model or summarize useful information. It is often desirable to pay attention to only those rules which may have reasonably high *support* and *confidence*. However, some important information may be filtered out while the remaining rules may be obvious or already known by setting high *support* and *confidence* thresholds (i.e., minimum *support* and minimum *confidence*). Reducing rules is an effective way, yet it may cause problem of losing some important information for businesses. Hence, the interestingness measures play an important role on mining ARs and have been well studied. Yet the evaluation is even more an important phase of the process than analysis comparing to other DM methodologies (Zhang, 2009).

As the database size increases nowadays, Apriori becomes a design trade-off between accuracy and efficiency with these two control parameters. Chen et al. (1996) argue that several business applications require mining transaction data to capture customer behaviors in a very frequent basis. The efficiency of DM could be a more important issue than the requirement for a complete accuracy of the results. They quoted that "Missing some marginal cases with *support* and *confidence* levels at the borderline may have little effect on the quality of the solution to the original problem….Allowing imprecise results can in fact significantly improve the efficiency of the mining algorithms". They further suggest a simple approach to help mitigate this problem is to gradually increase the threshold values of *support* and *confidence* until a manageable size of rules is generated. A technique of relaxing the support factor based on the sampling size is devised in Park et al. (1995) to achieve the desired level of accuracy. To a business or organization view, this study argues, efficiency might be pursued over accuracy; comprehending individual customer accurate shopping behavior, however, is the most important manners.

Regardless of all the measures, surveys, and discussions on the interestingnesses for Apriori or Apriori-like algorithms, this study believes that *support* and *confidence* hold traits of comprehension and calculating simplicity and still are convenient and suitable for mining ARs yet the values should be case-based. Determinations for the values of *support* and *confidence* should be individualized and even seasonal on account of different shopping sites and different shoppers.

ISSN 2277-3061

## METHODOLOGY

When discovering, evaluating, and even predicting individual customer shopping patterns and behaviors, neither any expert's subjective determination nor any objective interestingness measures but the customer's shopping lists themselves would best define the measures and levels of thresholds. That is, no matter which mining algorithms applied and measures adopted within algorithms, the most important issue is to correctly predict customer shopping behaviors and purchasing goods. Omitting the complexities of system execution, this study believes that support and confidence are applicable to heuristically describe the basic but comprehensive manners of AR.

Since understanding customer shopping behaviors would be the best analyzing data in a long timeline, four friends and relatives of the author are asked to participate in the study. In order not to affect and manipulate the subjects' consuming behaviors, none the study details are revealed to them. The backgrounds of the subjects are listed in Table 2. The subjects provided their shopping receipts, with all the POS (point of sales) data from April 2012 to September 2013. They are all constantly shopping at Carrefour and Costco in their living cities. The basic shopping information of these subjects is in Table 3.

**Table 2. The backgrounds of the research subjects (in 2012).**

| Background | Subject A | Subject B | Subject C | Subject D |
|---|---|---|---|---|
| Gender | Male | Female | Male | Female |
| Age | 56 | 57 | 44 | 50 |
| Blood type | B | O | O | O |
| Academic degree | Graduate | Graduate | College | College |
| **Birth city** | Kaohsiung, Taiwan | Tainan, Taiwan | Pingtung, Taiwan | Taipei, Taiwan |
| **Living city** | Kaohsiung, Taiwan | Kaohsiung, Taiwan | Tainan, Taiwan | Taichung, Taiwan |
| **Profession** | Construction corporation | Kaohsiung Harbor Bureau | Police department | Police department |
| **Job title** | General manager | Section chief | Sergeant | Section chief |
| **Work place** | Kaohsiung | Kaohsiung | Tainan | Taichung |
| **Marriage** | Married | Married | Married | Married |
| **Annual personal income (NTD)** | $1,200,000 | $1,000,000 | $600,000 | $800,000 |
| **Annual household income (NTD)** | $2,000,000 | $1,800,000 | $900,000 | $1,500,000 |
| **Family members** | | | | |
| **Living together (age)** | Wife (57) | Husband (58) | Mother (82), Wife (39), Daughter (14), Sons (12 & 3) | Husband (53), Daughter (18) |
| **Not living together (age)** | Mother (84), Son (32), Daughter (31) | Mother (89), Son (31), Daughter (28) | -- | Father (86), Son (24) |
| **Children's school city** | Boston, USA (daughter) | Chicago, USA (daughter) | Tainan, Taiwan (all children) | Taipei, Taiwan (son), Taichung, Taiwan (daughter) |

**Table 3. Descriptions of the subjects' shopping habits and data sizes.**

| | Subject A | Subject B | Subject C | Subject D |
|---|---|---|---|---|
| **The most shopping grocery store** | Carrefour in Kaohsiung | Costco in Tainan | Carrefour in Pingtung<br>Costco in Kaohsiung | Costco in Taichung |
| **Monthly times in average** | 1.9 | 2.3 | 1<br>1 | 1.4 |
| **Amount spent each time in average (NTD)** | $1,800 | $4,100 | $1,200<br>$3,700 | $3,400 |
| **Products each time in average** | 6 | 11 | 6 | 8 |
| **Usual shopping hours the most** | Wed. 19:00~21:00<br>Thr. 19:00~21:00 | Fri. 18:00~20:30<br>Sun. 18:00~20:30 | Sat. 18:00~19:00<br>Sun. 14:00~16:00 | Mon. 18:00~19:30<br>Tue. 18:00~19:30 |
| **Data size collected** | 34 | 41 | 18<br>16 | 26 |
| **Total Aprior mining runs** | 21 | 21 | 10<br>12 | 15 |

According to "Taiwan Import/Export Goods Classification" posted in September 2013, goods possible sold in Taiwan are classified into 21 categories, 97 chapters, and 1374 codes. 1374 codes are adopted to classify purchasing items for Apriori mining in this study. Packages applied to Apriori mining for researchers are usually WEKA, SQL Server 2005, Statistica, SAS EM, and Clementine. This study uses a self-coded in C language due to limitations of the above (Chang, 2013). The hardware adopted is ordinary 79 desktops with CPU of Intel Pentium DC G860 LGA-1155, motherboard of ASUS P8H61-M LX3 PLUS R2.0, DRAM of Kingston DDR3-1333 4GB and hard disk of WD Caviar-Green SATA3 500G 64MB.

The study is not only to find individual customer shopping patterns of item combinations, but also to discover her/his shopping behavior changes as well as levels of interestingnesses for Apriori. The dataset for the first Apriori run is each subject's shopping POS data from the first six months. The ARs from the first run is used for this subject's shopping behavior basis. Noted, researchers would take any time period of dataset but avoid dataset across two or more seasons. In addition, data on special holidays such as Christmas should better be eliminated. A dataset includes data across seasons or on holidays may be reveal different shopping patterns. The second run for Apriori mining is the dataset with the time of the first six months plus one more transactions, denoted as +1. The third run is the data including the data of +1 and one more transaction, denoted as +2, and so forth. The total times of mining execution for each subject's shopping dataset are listed in the last row of Table 3.

For the four subjects with five shopping datasets, the minimum execution time of a dataset is about 10 hours and the maximum 113 hours. The execution time depends on the variety of a subject's purchasing items not the transaction time period. All results take 10.8GB. The major purpose of obtaining a frequent itemset with a minimum *support* at the first step of Apriori is to delete some less frequent items for increasing the mining efficiency. Avoiding prejudgment, in this study, all items in the POS data are taken into account to generate all possible rules. Besides, since the value of minimum *confidence* is left to be discussed in the next section, ARs generated in this study are collected at all levels of *confidence* from the minimum conditional probabilities of 0.1 to 1.0

## RESULTS

### Forecasting accuracy of personal shopping behaviors by Apriori rules

The forecasting accuracy rates are calculated by comparing items of one actual transaction with the ARs mined from all previous transaction data. For example, to predict the transaction time of +1, ARs are derived from the data of the first six months; and to predict the transaction time of +2, ARs are derived from the dataset of +1. Figure 1 depicts the forecasting accuracy rates for all subjects at two grocery stores. The definition of forecasting accuracy rate, in this study, is the matched numbers of ARs, based on the actual purchased items from POS data, to all derived ARs by Apriori for all the previous purchases, i.e., the proportion of being that the ARs actually happen. For example, at *confidence* 0.8, if there are 12 rules derived by Apriori from the first six months dataset and 3 rules matched at +1's shopping receipt, the forecasting accuracy rate is 25%. Since *confidence* is a decisive measure to generate strong ARs, in this study, different ARs at

different levels of *confidence*, from 0.1, 0.2 to 1.0, are all mined in order to evaluate which levels of *confidence* are best to predict individual subject's shopping goods.

For subject A shopping at Carrefour, setting minimum *confidence* around at 0.7~0.8 is able to predict 55%, at most, of subject A's shopping baskets. However, setting minimum *confidence* at 0.5 would predict about 10%~12% of subject B's shopping baskets at Costco. Obviously, subject A's shopping baskets are more steady than other subjects' because of the higher forecasting accuracy rates. Many trend lines in Figure 1 gradually shift upward. It indicates that customer shopping behaviors can be comprehended better in a long run. Furthermore, this study also finds that in order to more correctly predict a customer shopping basket, minimum *confidence* value has to be adjusted accordingly.

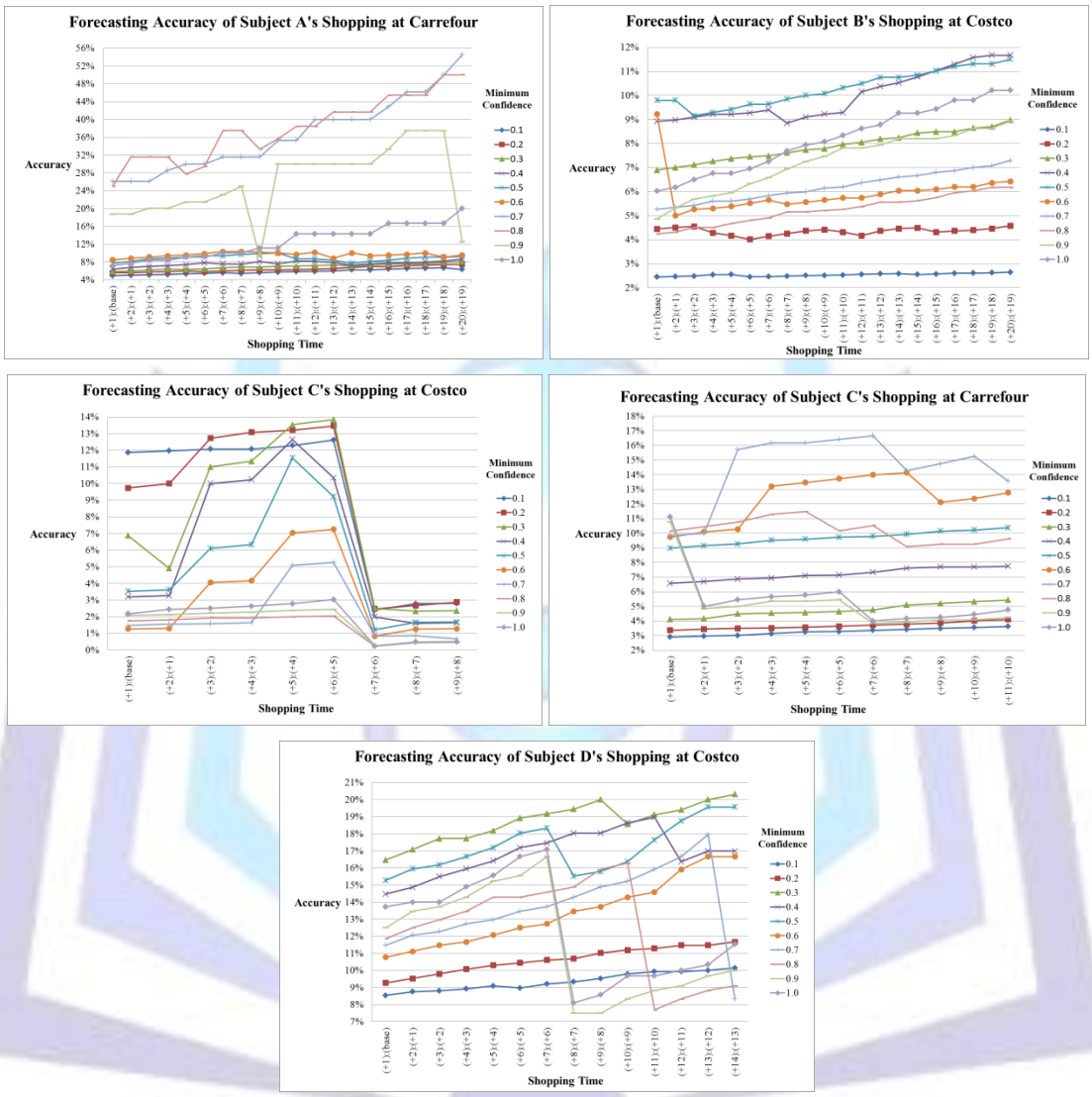


**Figure 1. Forecasting results for the subjects**

## Interpretations of personal shopping behaviors

From Figure 1, there are patterns of accuracy steeply decreasing for some levels of minimum *confidence*. This study finds that these patterns reveal certain subjects' shopping behavior changes based on the definition of *confidence*, items on transaction receipts, and items within the discovered ARs. Customer shopping behaviors would change under some circumstances such as seasonal needs, holiday needs, family members change, age grown, etc. Therefore, pursuing high forecasting accuracy rates with different levels of minimum *confidence* for different individual customers is not the only index to evaluate and predict customized shopping behaviors. Transaction time, rule's value of *confidence*, and accuracy rate should all be taken into account.

An accuracy rate steeply decreases only under one condition, in this study, of that the numbers of ARs (the denominator of the rate) generated by Apriori steeply increase. Since the numbers of the matched rules (the numerator of the rate) in this study are usually less than 10 and sometimes around 15 due to the numbers of items in baskets, they would not be possible to cause a sudden decrement on an accuracy rate. Further, basket items of the following transactions added for Apriori mining are the reason to affect the numbers of discovered rules. Table 4 concludes conditions of the changes of forecasting accuracy rates. These conditions are effective in the period of time of datasets.

ISSN 2277-3061

**Table 4. Conditions for shifts of forecasting accuracy**

| **Accuracy shifts** | | **Numerator** | | |
|---|---|---|---|---|
| | | **unchanged** | **increasing*** | **decreasing*** |
| **denominator** | **unchanged** | unchanged | increasing | decreasing |
| | **decreasing gently** | increasing | increasing | shifting uncertain |
| | **increasing gently** | decreasing | shifting uncertain | decreasing |
| | **increasing steeply** | decreasing | shifting uncertain | decreasing |
| **Conditions** | | | | |
| **denominator** | **for all cases** | 1. The matched rules stay unchanged. Previous purchasing goods stay and **no** new goods are added in the basket.<br>2. The matched rules changed. Some previous purchasing goods **may not** be and new goods, never bought before, are added in the basket. | 1. The matched rules include all old and new ones. Previous purchasing goods stay and new goods, never bought before, are added in the basket.<br>2. The old matched rules decrease plus more new ones. Some previous purchasing goods are **not** and new goods, never bought before, are added in the basket. | 1. The old matched rules decrease and no new ones added in. Some previous purchasing goods are not in the basket.<br>2. The old matched rules decrease plus less new ones. New goods, never bought before, are added and some previous purchasing goods are **not** in the basket. |
| | **increasing steeply** | 3. New goods, never bought before, are added in the basket. Yet, the goods are **not** shown in the matched rules.<br>4. New goods, never bought before are **only** bought in the basket **one time**, no matter if the goods are shown in the ARs afterwards. The new bought goods will never be shown in the matched rules.<br>5. New goods are continuously bought from now on. The goods will be shown in the ARs afterwards. The new bought goods will be shown in the matched rules on the time of +1. If the new goods are continuously bought more than three times, they will be shown in the matched rules on the time of +1 and +2, and so forth.<br>6. New goods are continuously bought twice. The goods will **not** be shown in the ARs afterwards. The new bought goods will **not** be shown in the matched rules on the time of +1 until they are shown in the ARs.<br>7. New goods are bought occasionally. It will depend on the actual goods combinations in baskets no matter if they are shown in the ARs. | | |

*****: The changes of numerators often small since the numbers of the matched rules of the actual purchasing goods with the discovered ARs usually would not be large.

Items in baskets would be no much variance for a customer while the forecast accuracy rate gradually increases. The shifts of accuracy rates reveal important information for system designers to follow: (1) an accuracy steeply decreases if items never bought before appearing on the transaction receipt; (2) an accuracy will then increase (gradually or steeply) if the items in situation (1) continuously bought in the following transactions; (3) an accuracy decreases (gradually or steeply) if items stop being purchased and then the accuracy will increase if some other items usually bought are still kept purchasing; (4) no indication from accuracy shifts for the behavior of impulse purchases and holiday needs. The first two shopping behaviors also imply the behavior of seasonal purchases.

Noted the degree of accuracy decreasing in situation (3) would be less than that in situation (1) for the same customer at the same grocery store except for the numbers of items being quite small. It is because when a new item never bought before once appears in the basket, a big amount of ARs suddenly discovered, for some levels of *confidence* in the latter. Yet it will take a period of time to have effect on the discovered ARs in the former.

POS system knows whether A and/or B are in baskets, same does DM system. The study further examines details how items **A** and **B** in rule {**A**}→{**B**} appear in baskets based on shifts of *confidence* values. For goods in itemsets **A** and **B**, rules {**A**}→{**B**} and {**B**}→{**A**} represent different conditional purchases. Evaluating shifts of *confidence* values of {**A**}→{**B**} and {**B**}→{**A**} together enables to figure out **A** and **B** actually being in the basket. The principles summarized in Tables 5 and 6.

ISSN 2277-3061

**Table 5. The relations of the shift of confidence values with goods purchase**

| shift of *confidence* | goods purchase |
|---|---|
| **increasing** | Times of buying {**A**}↑, times of buying {**A**} and {**B**}↑ (i.e, both {**A**} and {**B**} are in the basket) |
| **decreasing** | Times of buying {**A**}↑, times of buying {**A**} and {**B**} unchanged (i.e., only {**A**} in the basket) |
| **unchanged** | Times of buying {**A**} unchanged, times of buying {**A**} and {**B**} unchanged (i.e., {**A**} is not in the basket but {**B**} uncertain) |
| **fixed at 1** | Times of buying {**A**}=times of buying {**A**} and {**B**} (i.e., if {**A**} in the basket, {**B**} will be; if not, {**B**} uncertain) |

**Table 6. Judgments for goods purchase from shifts of rules' confidence values**

| **shift of *confidence* for rule {B}→{A}** | **shift of confidence value for rule {A}→{B}** | | | |
|---|---|---|---|---|
| | **increasing** | **decreasing** | **unchanged** | **fixed at 1** |
| **increasing** | both {**A**} and {**B**} are in the basket | N/A | N/A | both {**A**} and {**B**} are in the basket |
| **decreasing** | N/A | N/A | {**A**} is **not** in the basket but {**B**} is | {**A**} is **not** in the basket but {**B**} is |
| **unchanged** | N/A | {**A**} is in the basket but {**B**} is **not** | **neither** {**A**} **nor** {**B**} are in the basket | **neither** {**A**} **nor** {**B**} are in the basket |
| **fixed at 1** | both {**A**} and {**B**} are in the basket | {**A**} is **not** in the basket but {**B**} is | **neither** {**A**} **nor** {**B**} are in the basket | both {**A**} and {**B**} are in the basket or **neither** {**A**} **nor** {**B**} are in the basket |

Tables 7~11, listing actual purchases (with "v"s) of all single items in the matched rules with the best forecasts, coincide the principles in Tables 5~6. For goods within the matched rules at the minimum *confidence* of 0.7 and 0.8, subject A's shopping baskets do not show much variation, so do subject B's. However, the prediction accuracy for subject B, 11% by setting at 0.4~0.5, is lower than that for subject A. DM systems to predict this type of customer shopping behavior can set the minimum *confidence* with the highest forecasting accuracy rates before patterns change.

Subject C at Costco starts to buy adult diapers at shopping point +6. The accuracy rates therefore steeply decrease at this point. However, based on his shopping patterns at Carrefour, beers and cigarretes stop being purchased at +4 and accuracy rates steeply decrease at different shopping points with different minimum *confidence* levels. For subject D at Costco, the patterns are the same as subject C's at Carrefour but existing even more steep decrements. It shows that subject D stop buying cigarrets at +5.

**Table 7. Shopping details for subject A at Carrefour**

| **Items** | **Shopping Time** | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | +1 | +2 | +3 | +4 | +5 | +6 | +7 | +8 | +9 | +10 | +11 | +12 | +13 | +14 | +15 | +16 | +17 | +18 | +19 | +20 |
| meat paste | v | v | v | | v | v | v | v | v | | v | v | v | v | v | | v | v | v | v |
| sports drinks | v | v | v | v | v | v | v | | v | v | v | v | v | v | v | v | v | v | | v |
| soda | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v |

**Table 8. Shopping details for subject B at Costco**

| **Items** | **Shopping Time** | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | +1 | +2 | +3 | +4 | +5 | +6 | +7 | +8 | +9 | +10 | +11 | +12 | +13 | +14 | +15 | +16 | +17 | +18 | +19 | +20 |
| frozen foods | v | v | | v | v | v | v | v | | v | v | v | v | v | | v | v | v | v | v |
| fresh milk | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v | v |
| crystal sugar | v | v | | v | v | v | v | v | | v | v | v | v | v | | v | v | v | v | v |
| disposable heat pads | | | | | | | | | v | v | v | v | v | v | v | | | | | |

**Table 9. Shopping details for subject C at Costco**

| Items | Shopping Time | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | +1 | +2 | +3 | +4 | +5 | +6 | +7 | +8 | +9 |
| cooked chicken | v | v | | v | v | v | v | | v |
| baby diaper | v | v | | v | v | | v | v | |
| toys | | v | | v | | v | | v | |
| dog canned food | v | v | v | v | v | v | v | v | v |
| adult diaper | | | | | | v | v | v | v |
| canned dog food | | | | | | v | v | v | v |

**Table 10. Shopping details for subject C at Carrefour**

| Items | Shopping time | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | +1 | +2 | +3 | +4 | +5 | +6 | +7 | +8 | +9 | +10 | +11 |
| cravings | | v | v | v | | v | v | v | | v | v |
| canned beer | v | v | v | v | | | | | | | |
| imported cigarette | v | v | v | v | | | | | | | |
| gardening fertilizer | v | v | v | v | v | v | v | v | v | v | v |

**Table 11. Shopping details for subject D at Costco**

| Items | Shopping time | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | +1 | +2 | +3 | +4 | +5 | +6 | +7 | +8 | +9 | +10 | +11 | +12 | +13 | +14 |
| whole grain bakery | v | v | v | v | v | v | v | v | v | v | v | v | v | v |
| imported cigarette | v | v | v | v | | | | | | | | | | |
| contact lens drop | v | v | | v | v | | v | v | | v | v | | v | v |
| laundry detergent | v | | | v | | | v | | | v | | | v | |

## CONCLUSIONS AND DISCUSSIONS

DM techniques are still at the stage of chasm. Take AR for example, research in this area continues to propose incremental refinements in algorithms, but very little literature describes how the discovered ARs are used. While DM has been perceived to be a potentially powerful tool, the real benefit of DM for business intelligence (BI) has not been fully recognized (Wang and Wang, 2008). Further, even though the general task of KDD is the automatic extraction of novel, useful, and valid knowledge from large sets of data (Seung et al. 2010), most DM methods are bound to discover any knowledge that satisfies the chosen criterion of usefulness and validity. This includes typically very many rules that are already known to users.

In this study, the basic definition and heuristic evaluation of AR enable to efficiently and appropriately forecast individual shopping behavior from her/his shopping baskets. Purchase behaviors of no matter constant purchase, stopping purchasing habitual goods, and starting to purchase goods that never bought before is able to be recognized. However, impulse purchase, including purchase for holidays, is unable to discover by Apriori mining. The results are very important and valuable contribution for retailers to design forecasting systems of customer shopping behaviors.

In the real world, data and customer behaviors are changing on a continuous basis and thus the idea of building a static DM model that is subsequently used for a fixed period of time may no longer be appropriate. The mining model has to be amenable very quickly (Baesens et al., 2009). The best design is to let systems allow automatic updates in order to be adopted on the real-time decision making. Furthermore, a system, BI or expert system, enabling to make correct prediction of customer shopping behaviors, forecasting accuracy rate is not the sole indication to follow. Interestingness measures adoption and their thresholds determination for the quality of the extracted rules, either subjective or objective, more or fewer rules, neither are uniquely considered in a long run.

Finally, brand names of purchased items are not taken into account in this study; hence, mining combinations of purchasing goods with their brand names will be the first future study. This study posits that when a customer's shopping habit changes, the minimum *confidence* values to highly forecast her/his shopping basket might also change to other levels. There might be principles to adjust minimum *confidence*. Although the data in this study cannot verify this postulate probably due to the short time period of data collection, it would be another future study to be addressed.

## REFERENCES


[1] Agrawal, R., Imieliński, T. and Swami, A. (1993), "Mining association rules between sets of items in large databases", SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data, New York, Vol.22, pp.207-216.

[2] Agrawal, R. and Srikant, R. (1994), "Fast algorithms for mining association rules in large databases", VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco, pp.487-499.

[3] Baesens, C.M., Martens, D. and Vanthienen, J. (2009), "50 years of data mining and OR: upcoming trends and challenges", *Journal of the Operational Research Society*, Vol.60, pp.S16-S23.

[4] Chang, S.-J. (2013), "The application of Apriori on customized shopping behavior", Thesis of Marketing and Distribution Management, National Kaohsiung First University of Sci. & Tech., Taiwan.

[5] Chen, M.-S., Han, J.-W. and Yu, P.S. (1996), "Data mining: an overview from a database perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol.8, n.6, pp.866-883.

[6] Clifford, L. (2008), "Big data: How do your data grow?", *Nature*, Vol.455, pp. 28-29.

[7] Geng, L. and Hamilton, H. J. (2006), "Interestingness measures for data mining: a survey", *ACM Computing Surveys*, Vol. 38 No. 3.

[8] Han, J.-W., Pei, J. and Yin, Y. (2000), "Mining frequent patterns without candidate generation", Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD '00), ACM Press, pp.1-12.

[9] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. and Verkamo, A.I. (1994), "Finding interesting rules from large sets of discovered association rules", in: N. R. Adam, B. K. Bhargava & Y. Yesha (Eds), *Third International Conference on Information and Knowledge Management* (CIKM'94), ACM Press, pp. 401-407.

[10] Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S. and Duri, S.S. (2001), "Personalization of Supermarket Product Recommendations", *Data Mining and Knowledge Discovery*, Vol.5, pp.11-32.

[11] Lee, W. and Salvatore, J.S. (1998), "Data mining approaches for intrusion detection", USENIXSS '98 Proceedings of the 7th Conference on USENIX Security Symposium, San Antonio, Vol.7, pp.6.

[12] Lee, W., Salvatore, J.S. and Kui, W.M. (2002), "Algorithms for mining system audit data", *Data Mining, Rough Sets and Granular Computing*, Vol.1, pp.166-189.

[13] Li, W.-J. and Wang, S.-M. (2013), "Research on assessment method for credit risk in commercial banks of China based on data mining", *Applied Mechanics and Materials*, Vol.303-306, pp.1361-1364.

[14] Liang, X., Xue, C.-X. and Huang, M. (2010), "Improved Apriori algorithm for mining association rules of many diseases", *Communications in Computer and Information Science*, Vol.107, pp.272-279.

[15] Park, J.-S., Chen, M.-S. and Yu, P.S. (1995), "Mining association rules with adjustable accuracy", *IBM Research Report*.

[16] Piatetsky-Shapiro, G. (1991), "Discovery, analysis, and presentation of strong rules", *Knowledge Discovery in Databases*, pp.229-248.

[17] Seung, K.M., Timothy, W.S. and Soundar, R.T.K. (2010), "A methodology for knowledge discovery to support product family design", *Annual Operations Research*, Vol.174, pp. 201-218.

[18] Tan, P., Kumar, V. and Srivastava, J. (2002), "Selecting the right interestingness measure for association patterns", KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

[19] Theresia, W.A. and Noranita, B. (2012), "Apriori application to pattern profile creditor relationships with credit ceiling in rural bank", ICISBC '11 Proceedings of the 1st International Conference on Information Systems for Business Competitiveness, Semarang, pp.75-80.

[20] Wang, H. and Wang, S. (2008), "A knowledge management approach to data mining process for business intelligence", *Industrial Management & Data Systems*, Vol. 108, No. 5, pp. 622-634.

[21] Yao, X.-L. and Shu, H.-Y. (2009), "Study on value-added service in mobile telecom based on association rules", SNPD '09 Proceedings of the 10th International Conference on Software Engineering: Artificial Intelligences, Networking and Parallel/ Distributed Computing, Daegu, pp.116-119.

[22] Zhang, Y. (2009), “Association rule mining in cooperative research”, Thesis of Science and Industrial Engineering, University of Missouri-Columbia.

Hsiaoping Yeh is an Assistant Professor in the Department of Marketing & Logistics Management at National Kaohsiung First University of Science and Technology in Taiwan. She holds a doctorate in Industrial Engineering from the University of Wisconsin-Madison. Her major field of study is in operations research and decision sciences. She also holds a Master of Science in Industrial Engineering – Manufacturing Systems from the University of Wisconsin-Madison. Her current research interests focus on data mining.