Engineering Applications of Artificial Intelligence 63 (2017) 85–97

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Revenue prediction by mining frequent itemsets with customer analysis

Cheng-Hsiung Weng

Department of Management Information Systems, Central Taiwan University of Science and Technology, Taichung 406, Taiwan, Republic of China

ARTICLE INFO

Article history:
Received 1 June 2016
Received in revised form
7 April 2017
Accepted 27 April 2017
Available online 18 May 2017

Keywords:
Data mining
Frequent itemsets
Recency Frequency Monetary (RFM),
weighted transaction
Revenue prediction

ABSTRACT

Conventional frequent itemsets mining does not take into consideration the relative benefit or significance of transactions belonging to different customers. Therefore, frequent itemsets with high revenues cannot be discovered through the conventional approach. In this study, we extended the conventional association rule problem by associating the frequency–monetary (FM) weight with a transaction to reflect the interest or intensity of customer values and focusing on revenue. Furthermore, we proposed a new algorithm for discovering frequent itemsets with high revenues from FM-weighted transactions with customer analysis. The experimental results from the survey data revealed that the top k frequent itemsets with high revenues discovered using the proposed approach outperformed those discovered using the conventional approach in the prediction of revenues from customers in next-period transactions.

1. Introduction

In the knowledge discovery in data domain, association rule mining (ARM) is an important data mining approach that can enable the discovery of consumer purchasing behaviors from transaction databases. Agrawal et al. (1993) first introduced the problem of ARM defining it as identifying all rules from the transaction data that satisfy the minimum support and confidence constraints. The discovery of interesting associations or correlations is helpful in many business decision-making processes (Han and Kamber, 2006).

However, general ARM does not take into consideration the relative benefit or significance of transactions belonging to different customers, and instead assumes that the importance of each customer is identical. In other words, every customer is of equal weight during the mining process. However, numerous studies in customer relationship management (CRM) have revealed that the contributions of customers to businesses and profit maximization differ. Therefore, the evaluation of customer value is necessary before designing effective marketing strategies.

Businesses have started applying data mining technologies to marketing planning. Their objective is to gain customer loyalty and discover the contribution of customer value. Recency–frequency–monetary (RFM) analysis depends on recency (R), frequency (F), and monetary (M) measures and is one of the most popular database marketing metrics for quantifying customer transaction histories. RFM scoring is a method for determining the score of current customers on the basis of their R, F, and M values, and has been proven to be highly effective in marketing database applications (Blattberg et al., 2008). Moreover, RFM analysis is a well-known, behavior-based data mining method, which extracts customer profiles by using specific criteria. Recently, the RFM model has been used for CRM applications such as customer segmentation (Shim et al., 2012; Dursun and Caber, 2016).

Because RFM analysis and market basket analysis (i.e., frequent pattern mining) are the two most important tasks in database marketing, this study extended the conventional association rule problem by associating a customer value (i.e., frequency–monetary (FM) weight, which is determined by applying the FM scoring method) with a transaction to reflect the interest or intensity of customer values. This facilitates the association of an FM weight parameter with each transaction, enabling the discovery of valuable patterns. In addition, we propose a new frequent itemsets frequency–monetary (FIFM)-weighted algorithm for identifying frequent itemsets from FM-weighted transactions for the prediction of customer revenue.

We addressed the following questions related to discovering frequent itemsets from FM-weighted transactions: (1) Do the top k frequent itemsets discovered using the proposed FIFM algorithm outperform those discovered using the conventional Apriori algorithm in terms of predicting customers' purchasing itemsets? (2) Do the top k frequent itemsets discovered using the proposed FIFM algorithm outperform those discovered using the conventional Apriori algorithm in predicting customer revenue?

The remainder of this paper is organized as follows. A review of related work is presented in Section 2. The problem definitions are

E-mail address: chweng@mgt.ncu.edu.tw

http://dx.doi.org/10.1016/j.engappai.2017.04.020
0952-1976/

provided in Section 3. The proposed algorithm and an example are illustrated in Section 4. Section 5 uses survey data to demonstrate the usefulness of the proposed algorithm. Conclusions and future work are discussed in Section 6.

2. Related work

The main purpose of this study was to discover frequent itemsets from transaction data with customer values (FM weights). In this section, we mainly explore the problems and some techniques related to association rules and customer value (RFM value). Finally, we discuss the differences between the applications of the present study and a 2014 study by Hu and Yeh.

2.1. Association rule mining

Association rule mining (ARM) is an important data mining approach that enables the discovery of consumer purchasing behavior from transaction databases. Association rules are interesting and unexpected association relationships among attributes in a database that satisfy minimum support and confidence constraints (Han and Kamber, 2006). Agrawal et al. (1993) first introduced the problem, defining it as identifying all rules from the transaction data that satisfy the minimum support and confidence constraints. In brief, an ARM algorithm involves two steps: (1) generation of all frequent itemsets that satisfy the minimum support constraint; and (2) generation of all association rules that satisfy the minimum confidence constraints from the already discovered frequent itemsets.

We reviewed previous related studies that used association rule techniques. Table 1 presents a summary of previous studies that reported association rule analyses for various purposes in business applications.

2.2. Weighted association rules

Each item is treated uniformly by most ARM algorithms. However, in real applications, a user may be more interested in the rules that describe the frequently occurring “fashionable” items. In addition, the user may wish to mine the association rules but place more emphasis on some items. In other words, frequent itemsets are typically mined from binary databases, and each item in a transaction may have a different significance. Lu et al. (2001) proposed the mixed-weighted association rules algorithm to address the problem of mining mixed-weighted association rules. Since then, researchers have proposed weighted frequent itemset mining (WFIM) algorithms that reflect the importance of items. Wang et al. (2004) extended the conventional association rule problem by associating a weight with each item in a transaction to reflect the interest or intensity of each item within the transaction. Yun and Leggett (2005) proposed a WFIM algorithm to generate more concise and crucial weighted frequent itemsets in large databases. Vo et al. (2013) proposed several algorithms for the rapid mining of frequent weighted itemsets (FWI) from weighted item transaction databases.

2.3. Recency–frequency–monetary (RFM)

On the basis of the CRM theory (Kalakota and Robinson, 1999; Peppard, 2000), various strategies have been developed for enhancing shopping rates, increasing sales of high-profit or price products, and retaining customers as long-term customers. RFM was defined by Hughes (2006) as follows: R is defined as the “last purchasing time;” F is defined as the “purchasing frequency in a specific period”; and M is defined as the “average amount of purchase in a specific period.” The RFM model can be used to effectively perform the process of clustering based on customer values. Business plans can be devised to extend the customers’ life cycle by implementing marketing projects (Linoff and Berry, 2002). RFM scoring is a method of determining the score of current customers from their R, F, and M values, and has been proven to be highly effective in marketing database applications (Blattberg et al., 2008).

Numerous studies have discussed the use of RFM values in recommendation systems. Liu and Shih (2005) suggested combining customer lifetime value (CLV) and RFM to analyze customers’ consumption properties and to provide a recommendation on the basis of these properties. In their studies, clustering techniques were employed to group customers according to the weighted RFM (WRFM) value. However, Li et al. (2006) proposed the timely RFM (TRFM) method instead of the WRFM for considering product property and purchase periodicity.

RFM concepts have been applied in various areas. Kim et al. (2010) proposed the use of an RFM engine for anomaly detection to minimize false alarms in network attacks and reduce the time required to respond to hacking events. Chan (2008) combined RFM with a CLV model to evaluate the segmented customers and then used a genetic algorithm to select more appropriate customers for each campaign strategy. Hsieh (2004) used a self-organizing map (SOM) neural network to identify groups of bank customers on the basis of repayment behavior and RFM behavioral scoring predictors.

Lin and Tang (2006) combined the RFM model to analyze customers’ values and group similar values together. They combined users’ recent behavior with incremental mining according to weight to mine relations based on the weight rather than analyzing all the data, thus reducing the calculation cost and time. This method can also be applied to ARM by using the Apriori algorithm. In addition, Chiang (2011) proposed a new procedure and an improved Recency–Frequency–Monetary–Discount–Return Cost (RFMDR) model to mine the association rules of customer values.

Table 1
Previous association rule studies in business applications.

Works	Techniques	Task
Ahn (2012)	Association rule	Cross-selling: product assignment
Chiang (2011)	Association rule	Mining customer values
Cil (2012)	Association rule	Supermarket layout
Huang et al. (2011)	Association rule	Business process management
Kuo et al. (2015)	Association rule and artificial immune network	Supplier selection and order quantity allocation
Le et al. (2013)	Association rule	Risk management in retail supply chain collaboration
Lee et al. (2012)	Association rule and decision tree	New product development
Lee et al. (2013)	Association rule	Cross-selling: marketing solution
Shim et al. (2012)	Association rule and sequential pattern	CRM strategy
Weng (2016)	Association rule	Sale associations of specific later-marketed products

Table 2
Comparison of the present study and a previous study.

Characteristic	Hu and Yeh (2014)	This study
Patterns	RFM-pattern	FM-customer-pattern
Customer identified	No	Yes
Mechanism to determine weight	No	Frequency, Monetary
Weight normalization	No	Yes
Customer-patterns Set	Subset	Complete set
Thresholds	Min-Recency, Min-Frequency, Min-Monetary	Min-FMsup
Evaluation criteria	Precision, Revenue	Precision, Recall, Revenue

However, customers' purchasing behavior may not be clear from the association rules generated in this study.

Because many retailers record transactions without collecting customer information, RFM customer patterns cannot be discovered using the existing approaches. Hu and Yeh (2014) defined the RFM pattern and developed a novel algorithm to discover a complete set of RFM patterns that can approximate the set of RFM customer patterns without customer identification information. An RFM pattern is defined as a pattern that not only occurs frequently but also involves a recent purchase and a higher revenue percentage. Although this proposed approach is efficient and can be effectively applied to discover RFM patterns, not all RFM customer patterns could be discovered using this approach. A comparison of the differences and applications between the present study and the study by Hu and Yeh is presented in Table 2.

From the preceding analysis, we obtained the following information: (1) In conventional frequent itemsets mining, the relative benefit or significance of transactions of various customers is not considered. (2) The set of RFM patterns that can approximate the set of RFM customer patterns without customer identification information was discovered (Hu and Yeh, 2014); however, the relative benefit or significance of transactions of various customers was not reflected completely. (3) The integration of customer value into the transactions to reflect the interest or intensity of each customer and subsequently discover valuable patterns is noteworthy.

To address the aforementioned limitations, in this study, we extended the conventional association rule problem by allowing a customer value (FM weight) to be associated with a transaction to reflect the interest or intensity of customer values. Subsequently, we mined frequent itemsets with high revenues from FM-weighted transactions. Finally, we explored the differences in the revenue of patterns discovered using the proposed approach.

3. Problem definitions

In this section, we define the problem of the method for discovering frequent itemsets from FM-weighted transactions. Let $I = \{it_1, it_2, \ldots, it_m\}$ be a set of itemsets. Let D be a set of database transactions in which each transaction T is a set of items such that $T \subseteq I$. A transaction T is considered to contain X if and only if $X \subseteq T$.

It is important to determine weight values of the transactions of customers before discovering frequent itemsets from FM-weighted transactions. For preventing attributes (frequency and monetary) with initially ranges from outweighing, we normalized frequency and monetary of a customer by using *min-max* normalization method.

Table 3
Transaction data.

TID	*CID*	*Itemsets*	*Revenue* (it_i)				∑ *Revenue* (it_i)	*Trading date*
			a	*b*	*c*	*d*		
1001	1	*a*, *b*, *c*	30	10	20		60	2016/1/1
1002	2	*a*, *b*, *d*	30	10		30	70	2016/1/1
1003	1	*a*, *b*, *c*	30	10	20		60	2016/1/3
1004	2	*a*, *b*, *d*	30	10		30	70	2016/1/4
1005	3	*a*, *c*, *d*	30		20	30	80	2016/1/5

Definition 1. The transaction weight of the transaction T_i of customer c is defined as follows:

$$W(T_i) = \frac{F_c^w + M_c^w}{2},$$

where

$$F_c^w = \frac{F_c - F_{\min}}{F_{\max} - F_{\min}},$$
$$M_c^w = \frac{M_c - M_{\min}}{M_{\max} - M_{\min}}.$$

F_{max}: the maximum count of transactions among all customers;
F_{min}: the minimum count of transactions among all customers;
F_c: the count of transactions of customer c;
M_{max}: the maximum monetary sum among all customers;
M_{min}: the minimum monetary sum among all customers;
M_c: the monetary sum for customer c.

Example 1. Five transactions are presented in Table 3, and the customer FM value after calculation is presented in Table 4. Table 4 indicates that the FM weight of transactions {1001, 1003} is 0.83, that of transactions {1002, 1004} is 1.00, and that of transaction {1005} is 0.00. Finally, the transaction data with the FM weights after calculation are presented in Table 5.

Definition 2. The FM-weighted support of an itemset X is defined as follows:

$$FMsup(X) = \frac{\sum_{T_i(X \subseteq T) \wedge (T \in D)} W(T_i)}{\sum_{T_i(T \in D)} W(T_i)}.$$

Example 2. From Table 3, $FMsup(a) = (0.83 + 1.00 + 0.83 + 1.00)/3.66 = 1.00$, $FMsup(ab) = (0.83 + 1.00 + 0.83 + 1.00)/3.66 = 1.00$, and $FMsup(ac) = (0.83 + 0.83 + 0.00)/3.66 = 0.45$.

Definition 3. Given a user-specified *FMsup* threshold σ_{FMsup}, an itemset X is a high *FMsup* itemset if $FMsup(X) \geq \sigma_{FMsup}$.

Example 3. Let the *FMsup* threshold be set as $\sigma_{FMsup} = 0.50$. In

Table 4
Customers' FM values.

CID	*F*	*M*	*FM*
1	(2-1)/(2-1) = 1/1	(120-80)/(140-80) = 40/60	((1/1) + (40/60))/2 = 0.83
2	(2-1)/(2-1) = 1/1	(140-80)/(140-80) = 60/60	((1/1) + (60/60))/2 = 1.00
3	(1-1)/(2-1) = 0/1	(80-80)/(140-80) = 0/60	((0/1) + (0/60))/2 = 0.00

Table 5
Transaction data with FM weight.

TID	*CID*	*Itemsets*	*FM-weight*
1001	1	*a, b, c*	0.83
1002	2	*a, b, d*	1.00
1003	1	*a, b, c*	0.83
1004	2	*a, b, d*	1.00
1005	3	*a, c, d*	0.00
Sum			3.66

Example 2, itemset *a* with $FMsup(a) = 1.00$ is a high *FMsup* itemset; however, itemset *ac* with $FMsup(ac) = 0.45$ is not a high *FMsup* itemset.

Theorem 1. *If $X \subset Y$, then $FMsup(X) \geq FMsup(Y)$. The use of the aforementioned FM-weighted support metric satisfies the downward closure property.*

Proof.
If $X \subset Y$, then $\sum_{T_i(X \subseteq T) \wedge (T \in D)} W(T_i) \geq \sum_{T_i(Y \subseteq T) \wedge (T \in D)} W(T_i)$.

$$FMsup(X) = \frac{\sum_{T_i(X \subseteq T) \wedge (T \in D)} W(T_i)}{\sum_{T_i(T \in D)} W(T_i)} \geq \frac{\sum_{T_i(Y \subseteq T) \wedge (T \in D)} W(T_i)}{\sum_{T_i(T \in D)} W(T_i)} = FMsup(Y),$$

$FMsup(X) \geq FMsup(Y)$.

Definition 4. An association rule is an implication of the form $X \Rightarrow Y$, where $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in FM-weighed transaction set *D* with support $FMsup(X \cup Y)$. Rule $X \Rightarrow Y$ has a confidence $FMconf(X \Rightarrow Y)$ in *D*, where $FMconf(X \Rightarrow Y)$ is the percentage of transactions in *D* containing *X* that also contain *Y*. The following is the formal expression of $FMconf(X \Rightarrow Y)$:

$$\text{FMconf}(X \Rightarrow Y) = \frac{FM\text{sup}(X \cup Y)}{FMsup(X)}$$

Example 4. Let the minimum *FMsup* and minimum *FMconf* to 50% and 80%, respectively. In Example 2, $FMsup(a) = (0.83 + 1.00 + 0.83 + 1.00 + 0.00)/3.66 = 1.00$ and $FMsup(ab) = (0.83 + 1.00 + 0.83 + 1.00 + 0.00)/3.66 = 1.00$. In addition, $FMconf(a \Rightarrow b) = 1.00/1.00 = 1.00$. Thus, the following association rule is discovered:

$a \Rightarrow b$(*FMsup* = 100%, *FMconf* = 100%)

To discover frequent (high-*FMsup*) itemsets from FM-weighted transactions, we must discover all the itemsets that satisfy the minimum *FMsup* threshold. The discovery of frequent (high-*FMsup*) itemsets is the most computationally expensive procedure in ARM. In the next section, we introduce the proposed methodology for mining frequent (high-*FMsup*) itemsets and association rules from FM-weighted transactions for analyzing databases.

4. Algorithm for mining frequent itemsets from FM-weighted transactions

We now explain the proposed approach (FIFM) and provide an example to illustrate the method for discovering the frequent (high-*FMsup*) itemsets from FM-weighted transactions.

4.1. Proposed algorithm

The proposed FIFM algorithm is outlined in Fig. 1. The procedure can be divided into three phases: (1) calculation of the FM value for each customer and subsequent appending of the FM value to each transaction; (2) calculation of the *FMsup* of each itemset and generation of the high-*FMsup* itemsets; and (3) generation of association rules from high-*FMsup* itemsets. The step-by-step computing process is as follows:

Step 1: Preprocess data.

For discovering frequent (high-*FMsup*) itemsets from FM-weighted transactions, each customer's FM value must be calculated in advance. Subsequently, each customer's FM value must be appended to the transactions in which the customer's purchasing behaviors are recorded.

Step 2: Discover high-FMsup itemsets from FM-weighted transactions.

In this study, a level-wise approach is used in the first phase to iteratively generate candidate itemsets of *k* items (C_k), and to then discover high-*FMsup* itemsets of *k* items (L_k). To proceed to the next level, we generate a candidate set C_{k+1} from L_k and repeat. Unlike conventional ARM approaches, the itemset's *FMsup* is between 1 and 0, rather than 1 or 0. In addition, we should sum the total *FMsup* for all transactions.

Step 3: Generate association rules from all frequent (high-FMsup) itemsets.

In this step, we calculate the *FMconf* values of association rules.

//**Step 1 Call the** *FM_cal* Subroutine
(1). For each customer, calculate the *FM* value.
(2). For each transaction, set the *FM* value according to customer ID.

//**Step 2 Call the** *FreqItemsets_gen* Subroutine
(1). For each item it_i, calculate its *FMsup*.
(2). Check whether the *FMsup* of each item it_i is no less than the minimum *FMsup* (σ_{FMsup}). If it is, put it into the set of frequent one-itemsets (L_1).
(3). Generate candidate set C_{k+1} from L_k.
(4). Compute the *FMsup* values of all itemsets in C_k and determine L_k.
(5). If L_{k+1} is null, go to Step 3; otherwise, set $k = k + 1$ and repeat steps (3)–(5).

//**Step 3 Call the** *AR_gen* Subroutine.
(1). If the *FMconf* of rule $X \Rightarrow Y$ is no less than the minimum *FMconf* (σ_{FMconf}), then generate association rule ($X \Rightarrow Y$).

Fig. 1. Proposed algorithm.

Finally, we identify the association rules with *FMconf* values no less than the minimum *FMconf* (σ_{FMconf}).

4.2. Example

An example is provided to illustrate the proposed data mining algorithm. The data set used in this example is presented in Table 3.

Step 1: Preprocess data.

We first calculate each customer's frequency value (*F*). Table 3 indicates that the maximum count and minimum count of transactions for all customers are 2 and 1. In addition, the count of transactions for customers #1, #2, and #3 are 2, 2, and 1, respectively. Definition 1 indicates that each customer's *F* can be computed as 1/1, 1/1, and 0/1.

Second, we calculate each customer's monetary value (*M*). Table 1 indicates that the maximum monetary sum and minimum monetary sum of all customers are 140 (70 + 70) and 80, respectively. In addition, the monetary sums for customers #1, #2, and #3 are 120, 140, and 80, respectively. Definition 1 indicates that each customer's *M* can be computed as 40/60, 60/60, and 0/60.

Finally, each customer's *FM* value can be computed as 0.83, 1.00, and 0.00. Furthermore, we append each customer's FM value to the transactions in which the customer's purchasing behaviors are recorded.

Step 2: Discover frequent (high-FM) itemsets from FM-weighted transactions.

Assume that we set the minimum *FMsup* (σ_{FMsup}) to 0.5. For each itemset stored in the transactions, we calculate the itemset's *FMsup* and examine whether the *FMsup* of each itemset is greater than or equal to the minimum *FMsup* (σ_{FMsup}). The frequent itemsets discovered in this study and through using Apriori approaches are presented in Tables 6 and 7, respectively. The revenue of the frequent itemsets discovered in this study and through using Apriori approaches are presented in Tables 8 and 9, respectively.

STEP 3. Generate association rules from all frequent (high-*FMsup*) itemsets.

We can generate rules from all frequent (high-*FMsup*) itemsets. We set the minimum *FMsup* (σ_{FMsup}) to 50% and minimum *FMconfp* (σ_{FMconf}) to 80%. For brevity, we present only some generated association rules in Table 10.

Compared with the conventional Apriori approach, some high-*FMsup* itemsets such as {*b*, *d*} can be generated only by using the proposed FIFM approach. However, the low-*FMsup* itemsets such as {*c*} and {*a*, *c*} can no longer be generated. Moreover, business administrators are more interested in the prediction performance of the revenue for the next period than in frequent itemsets merely.

Assume that we have the same transaction data as shown in Table 3 for the next periods; we compare the prediction performance regarding revenue for the proposed approach and the conventional Apriori approach. The revenues of top k ($k = 7$) itemsets discovered using the FIFM and Apriori approaches are calculated (Tables 8 and 9). Notably, the total revenue of top k ($k = 7$) itemsets discovered using FIFM is higher than the total revenue of that discovered using the Apriori approach. In addition, by using the proposed FIFM approach, both high-frequency and high-revenue itemsets can be discovered, rather than only high-frequency itemsets. Therefore, the item (*d*) with higher revenue (90) was discovered and item (*c*) with lower revenue (60) was not discovered using proposed FIFM.

Table 6
Top k (k=7) frequent itemsets discovered using FIFM.

Itemsets	*FM support*
a	1.00
b	1.00
d	0.55
a, *b*	1.00
a, *d*	0.55
b, *d*	0.55
a, *b*, *d*	0.55

Table 7
Top k (k=7) frequent itemsets discovered using Apriori.

Itemsets	*Support*
a	1.00
b	0.80
c	0.60
d	0.60
a, *b*	0.80
a, *c*	0.60
a, *d*	0.60

Table 8
Revenue of Top k ($k = 7$) frequent itemsets discovered using FIFM.

No	*Itemsets*	*FM support*	*Revenue*
1	*a*	1	150
2	*b*	1	40
3	*a*, *b*	1	160
4	*d*	0.55	90
5	*a*, *d*	0.55	180
6	*b*, *d*	0.55	80
7	*a*, *b*, *d*	0.55	140
Sum			840

Table 9
Revenue of Top k ($k = 7$) frequent itemsets discovered using Apriori.

No	*Itemsets*	*Support*	*Revenue*
1	*a*	1	150
2	*b*	0.8	40
3	*a*, *b*	0.8	160
4	*c*	0.6	60
5	*d*	0.6	90
6	*a*, *c*	0.6	150
7	*a*, *d*	0.6	180
Sum			830

Table 10
Some association rules identified using FIFM.

No	Association rule	FMsup	FMconf
1	$a \Rightarrow b$	100%	100%
2	$b \Rightarrow a$	100%	100%
3	$d \Rightarrow a$	100%	100%
4	$d \Rightarrow b$	100%	100%

5. Experimental results

We applied three data sets (Foodmart, Grocery-POS, and Supermarket datasets) to evaluate the performance of the proposed FIFM algorithm. The three datasets possess dissimilar sale properties (such as products, customers, and purchasing transactions).

5.1. Evaluation criteria

We speculated whether the top k frequent itemsets with higher support discovered using the proposed FIFM algorithm are more suitable for predicting customers' purchasing itemsets and existing revenue in the next-period transactions.

Based on the training dataset (D_A) and test dataset (D_B), we explore the performance in prediction of customers' purchasing itemsets existing in the next-period transactions. Therefore, we define two metrics (*Precision* and *Recall*) for predicting customers' purchasing itemsets in the next-period transactions. Moreover, we define the metric (*Revenue*) to investigate the performance in predicting revenue of customers' purchasing itemsets in the next-period transactions.

Therefore, there are three metrics (*Precision*, *Recall*, and *Revenue*) defined to evaluate the performance of the frequent itemsets that existed in the transaction dataset in the following periods discovered using the proposed approach. Let A and B be the sets of patterns generated from training dataset (D_A) and test dataset (D_B), respectively. $Revenue(A \cap B)$ represents the revenue generated from the sets of patterns ($A \cap B$). The three metrics can be defined as follows:

$$Precision = \frac{|A \cap B|}{|D_A|},$$

$$Recall = \frac{|A \cap B|}{|D_B|},$$

$$Revenue = Revenue(A \cap B),$$

where

$|A \cap B|$ denotes the transaction numbers of itemset ($A \cap B$) in transaction set (D_B);
$|D_A|$ denotes the transaction numbers in transaction set (D_A);
$|D_B|$ denotes the transaction numbers in transaction set (D_B);
$Revenue(A \cap B)$ denotes the revenue value of itemset ($A \cap B$) in transaction set (D_B).

5.2. Experimental results of foodmart dataset

5.2.1. Dataset description

The food market dataset (sample dataset from Microsoft) comprised market baskets from 1560 products and 8842 users (http://recsyswiki.com/wiki/Grocery_shopping_datasets). We restored the database (foodmart_mysql) and extracted the sales transactions from the tables (sales_fact_1997, sales_fact_1998, and sales_fact_dec_1998). All the selected data were cleaned and processed before further analysis. There were 323 and 320 transactions in 1997 and 1998, respectively. The length of the itemsets ranged from 13 to 873 items, and the average itemset length was 241 items. The price of the items ranged from US$0.5 to US$23.64.

The transaction sizes in the first, second, third, and fourth quarters were 161, 162, 160, and 160, respectively. We used the transactions in the first quarter as the training dataset. Subsequently, we selected the top k frequent itemsets with higher support to investigate the number of patterns in L_1, L_2, and L_3. The numbers of patterns in L_1, L_2, and L_3 for top k frequent itemsets are the same as those shown in Table 11. Therefore, investigating the performance difference between the two approaches (Apriori and FIFM) in predicting customers' purchasing itemsets and the revenue from customers in advance is noteworthy.

5.2.2. Comparison of precision and recall predictions

We set the minimum support to 0.0005 and varied the top k from 100 to 1000 to investigate the differences in measures (precision and recall) between the two approaches (Apriori and FIFM). Figs. 2, 4, and 6 present the precision differences between the two approaches in predicting customers' purchasing itemsets. Figs. 3, 5, and 7 present the recall differences between the two approaches in predicting customers' purchasing itemsets. The experimental results of the precision and recall values reveal that Apriori outperformed the proposed approach in measures (precision and recall). In other words, the patterns discovered using Apriori are more suitable in predicting itemsets from customers in next-period transactions.

5.2.3. Comparison of revenue predictions

We set the minimum support to 0.0005 and varied the top k from 100 to 1000 to investigate the revenue differences between the two approaches. Figs. 8–10 present the differences between the two approaches in predicting revenues generated from customers' purchasing itemsets. The experimental results of revenue value prediction reveal that the proposed approach (FIFM) outperformed Apriori in predicting revenues of itemsets from customers in next-period transactions.

5.2.4. Discovered association rules

We identified the association rules from top 5 L_2 frequent itemsets generated using the proposed FIFM approach and the Apriori approach. Table 12 shows that the support values (*FMsup*) of patterns discovered using FIFM are lower than the support values (*sup*) of patterns discovered using Apriori. This is because FIFM discovers frequent itemsets from FM-weighted transactions by considering customer value.

Furthermore, the association rules identified using FIFM and Apriori are different (Table 12). Because patterns discovered using FIFM outperformed those discovered using Apriori in predicting revenues of itemsets from customers in next-period transactions, the association rules (Table 12) generated using FIFM could be used in recommender systems to increase revenue, particularly for food retail businesses in promotion strategies.

Table 11
Top k frequent itemsets with higher support.

Frequent	*k = 300*		*k = 600*		*k = 900*		*k = 1200*		*k = 1500*	
itemset	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*
L1	300	300	600	600	900	900	1200	1200	1500	1500
L2, L3	0	0	0	0	0	0	0	0	0	0

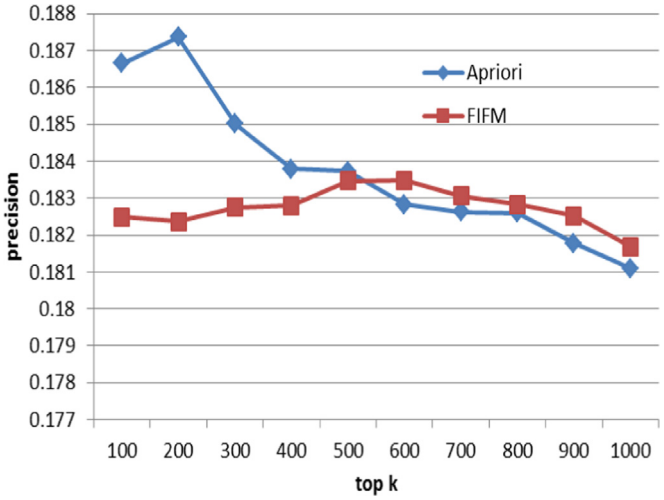

Fig. 2. Second quarter (precision).

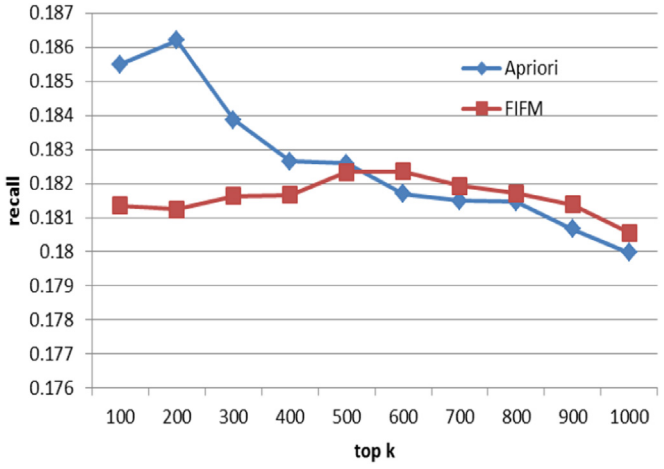

Fig. 3. Second quarter (recall).

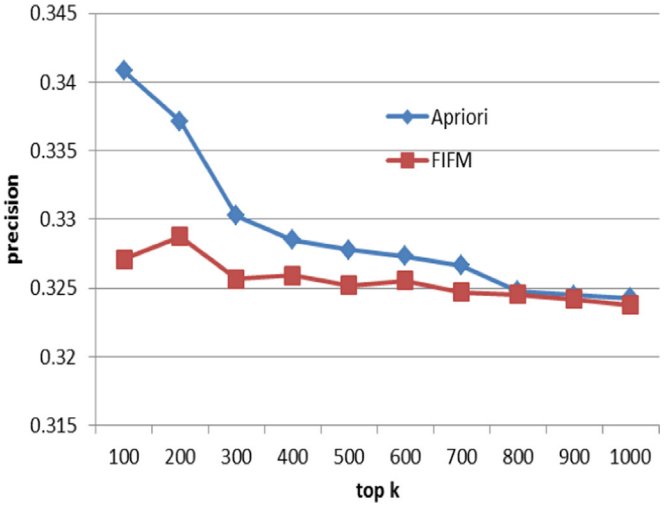

Fig. 4. Third quarter (precision).

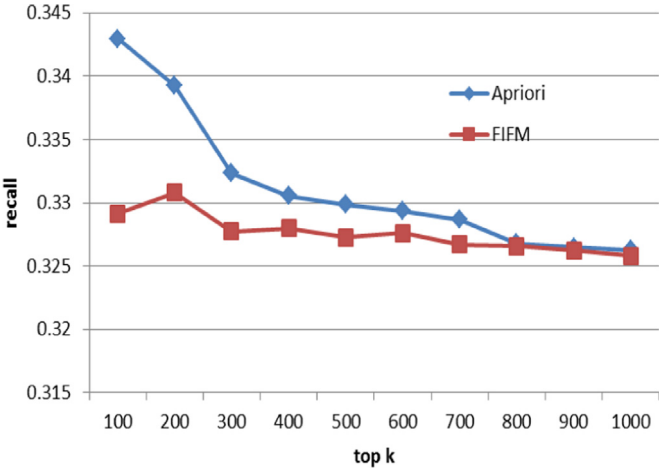

Fig. 5. Third quarter (recall).

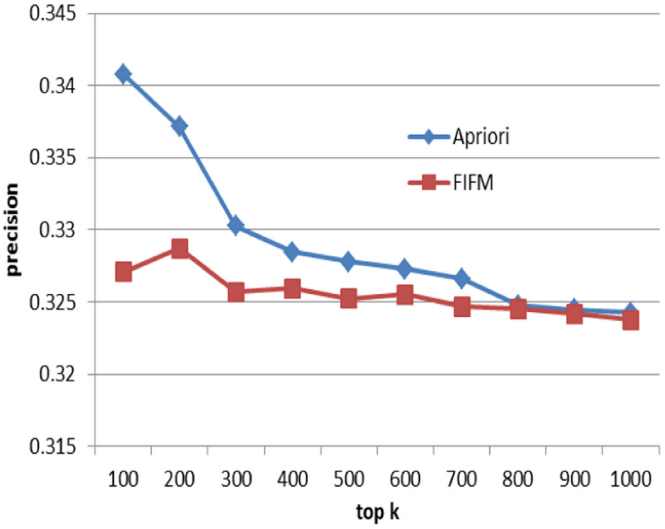

Fig. 6. Fourth quarter (precision).

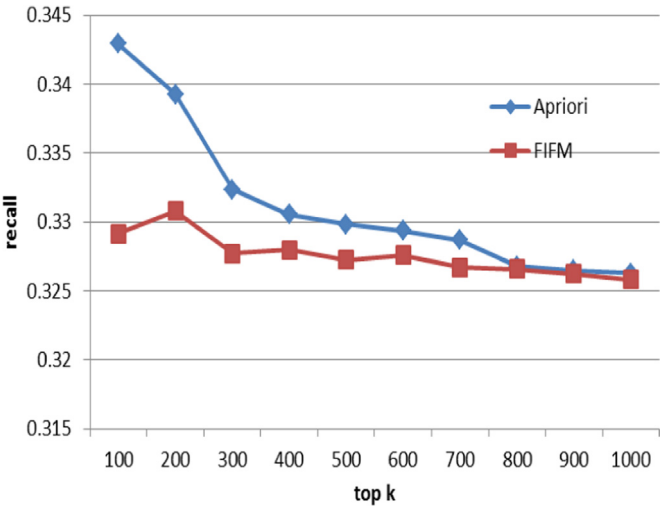

Fig. 7. Fourth quarter (recall).

5.3. Grocery-POS datasets

5.3.1. Dataset description

The Grocery-POS dataset comprises the sales records of a grocery store and an open dataset (Northwind) from Microsoft SQL Server 2008 R2. The Grocery-POS dataset recorded all transactions between 1996/7/4 and 1998/5/6. Each transaction contains information about the retail records in the grocery store. After we performed the necessary data preprocessing tasks, the dataset comprised 830 records of transactions related to 77 product types. The length of the itemsets ranged from 1 to 25 items, and the

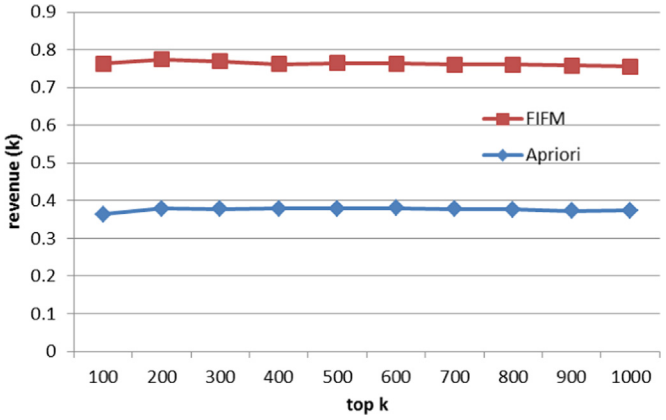

Fig. 8. Second quarter (revenue).

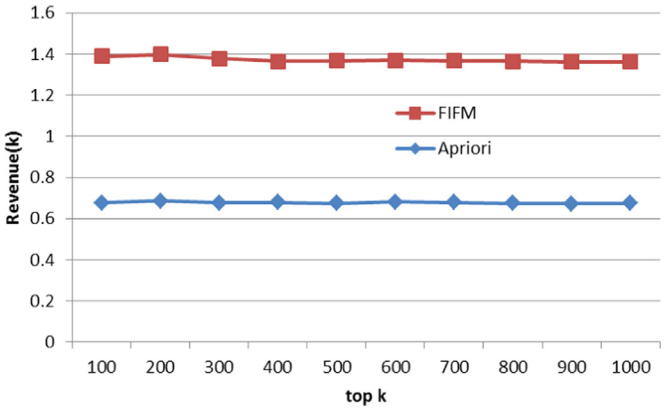

Fig. 9. Third quarter (revenue).

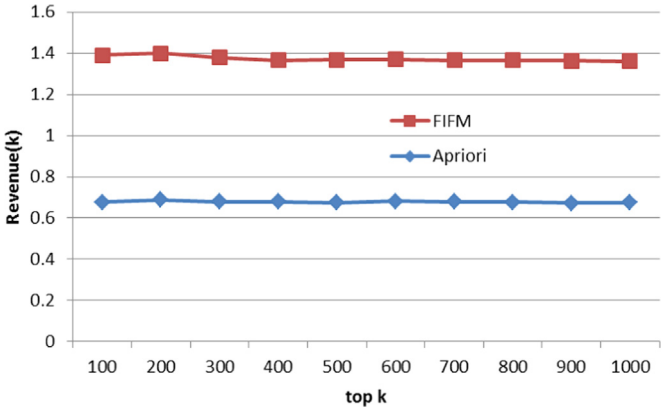

Fig. 10. Fourth quarter (revenue).

average itemset length was 3 items. The price of the items ranged from US$2 to US$263.5.

The transaction sizes in the first, second, third, and fourth quarters were 152, 185, 223, and 270, respectively. We used the transactions in the first quarter as the training dataset. Subsequently, we selected the top k frequent itemsets with higher support to investigate the number of patterns in L_1, L_2, and L_3 for the top k frequent itemsets with higher support. The number of patterns in L_1, L_2, and L_3 for top k frequent itemsets were different (Table 13). Therefore, investigating the performance difference between the two approaches (Apriori and FIFM) in predicting customers' purchasing itemsets and the revenue from customer in advance is noteworthy.

Table 12
Association rules (ARs) generated using top 5 L_2 frequent itemsets.

No	*FIFM*			*Apriori*		
	AR	*FMsup*	*FMconf*	AR	*sup*	*conf*
1	P1014⇒P1629	1.22%	36.15%	P1317⇒P1400	3.11%	14.29%
2	P1629⇒P1014	1.22%	27.86%	P1400⇒P1317	3.11%	15.15%
3	P1186⇒P2268	1.25%	36.81%	P1491⇒P1724	3.11%	19.23%
4	P2268⇒P1186	1.25%	24.75%	P1724⇒P1491	3.11%	16.67%
5	P1186⇒P2511	1.26%	37.18%	P1575⇒P2224	3.11%	19.23%
6	P2511⇒P1186	1.26%	25.99%	P2224⇒P1575	3.11%	13.51%
7	P1375⇒P1891	1.21%	20.95%	P1928⇒P2344	4.35%	20.59%
8	P1891⇒P1375	1.21%	41.40%	P2344⇒P1928	4.35%	21.21%
9	P1923⇒P2433	1.16%	20.44%	P1928⇒P2494	3.73%	17.65%
10	P2433⇒P1923	1.16%	18.62%	P2494⇒P1928	3.73%	18.75%

5.3.2. Comparison of precision and recall predictions

We set the minimum support to 0.001 and varied the top k from 10 to 100 to investigate the differences in measures (precision and recall) between the two approaches. Figs. 11, 13, and 15 present the precision differences between the two approaches in predicting customers' purchasing itemsets. Figs. 12, 14, and 16 present the recall differences of the two approaches for predicting customers' purchasing itemsets. The experimental results of the precision and recall values reveal that Apriori outperformed the proposed approach in measures (precision and recall). In other words, the patterns discovered using Apriori are more suitable in predicting itemsets from customers in next-period transactions.

5.3.3. Comparison of revenue predictions

We set the minimum support to 0.001 and varied the top k from 10 to 100 to investigate the revenue differences between the two approaches. Figs. 17–19 present the differences between the two approaches in predicting revenues generated from customers' purchasing itemsets. The experimental results of revenue value predictions reveal that the proposed approach outperformed Apriori in predicting revenues of itemsets from customers in the top 30 patterns in next-period transactions.

5.3.4. Discovered association rules

We identified the association rules from top 5 L_2 frequent itemsets generated using the proposed FIFM approach and the Apriori approach. Table 14 shows that the support values (*FMsup*) of patterns discovered using FIFM are lower than the support values (*sup*) of patterns discovered using Apriori. This is because FIFM discovered frequent itemsets from FM-weighted transactions by assessing customer value.

Furthermore, some association rules identified using FIFM and Apriori are different (Table 14). Because the patterns discovered using FIFM outperformed those discovered using Apriori in predicting revenues of itemsets from customers in next-period transactions, we suggest that marketing administrators of Grocery-POS businesses could use the association rules generated only using FIFM in recommender systems to design more efficient promotion strategies.

5.4. Supermarket dataset

5.4.1. Dataset description

The supermarket dataset comprised the sales data of a supermarket chain in Taiwan. The sales data, called SC-POS, comprised all the transactions from 20 branches between September 1, 2002, and September 28, 2002. Each transaction in SC-POS was a customer's shopping list containing purchased items, prices, and quantities. After we performed the necessary data preprocessing tasks, the dataset contained 4998 items and 405,285 transactions.

Table 13
Top k frequent itemsets with higher support.

Frequent	*k = 30*		*k = 60*		*k = 90*		*k = 120*		*k = 150*	
itemset	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*
L1	30	30	51	57	57	64	61	66	61	67
L2	0	0	9	3	33	26	59	54	89	83
L3	0	0	0	0	0	0	0	0	0	0

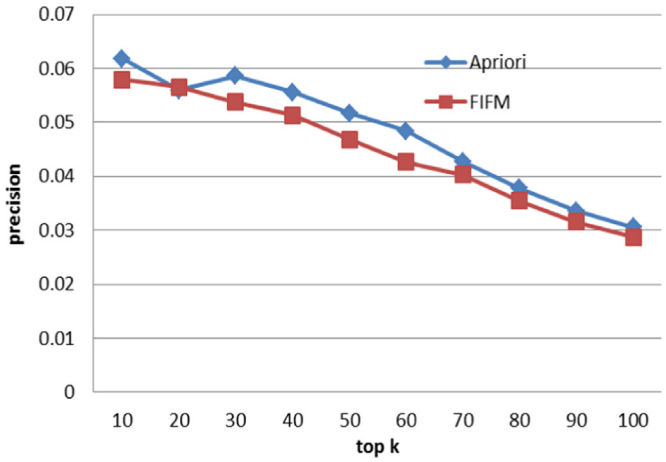

Fig. 11. Second quarter (precision).

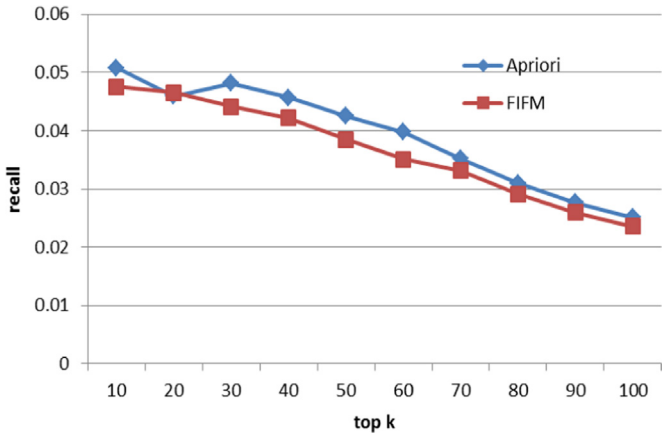

Fig. 12. Second quarter (recall).

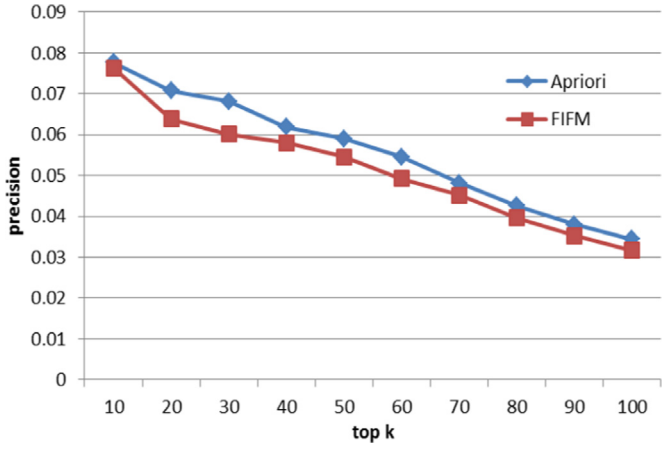

Fig. 13. Third quarter (precision).

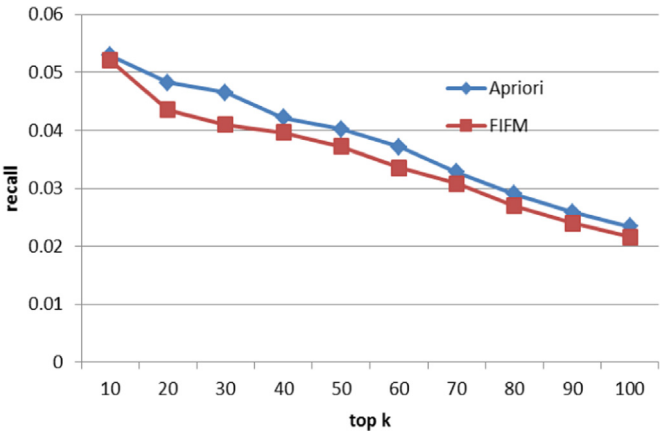

Fig. 14. Third quarter (recall).

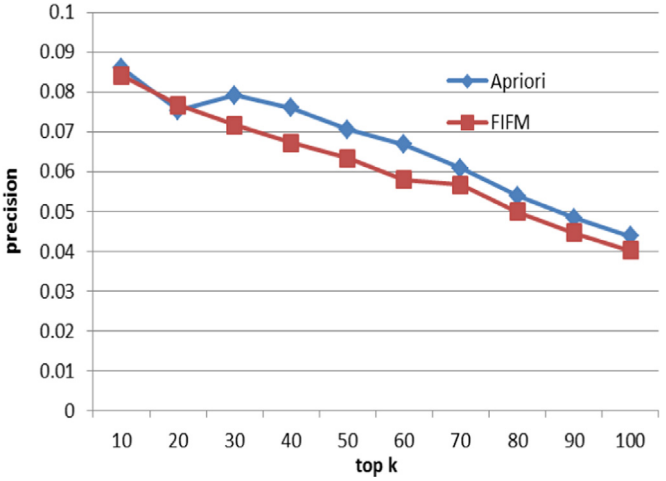

Fig. 15. Fourth quarter (precision).

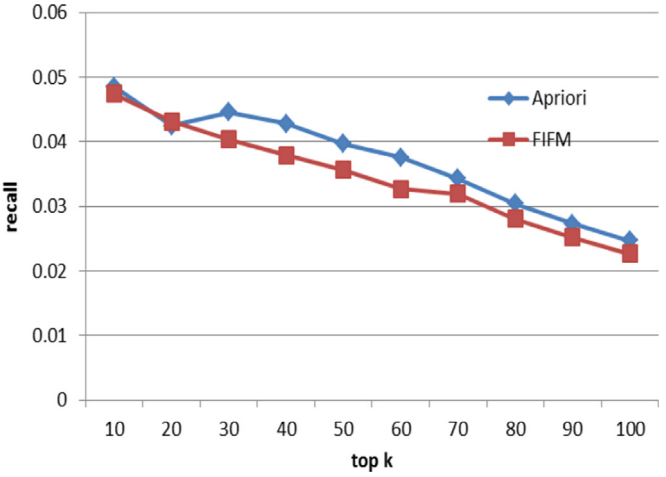

Fig. 16. Fourth quarter (recall).

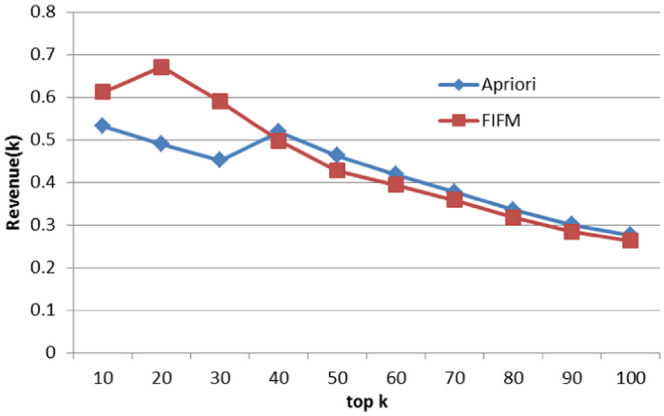

Fig. 17. Second quarter (revenue).

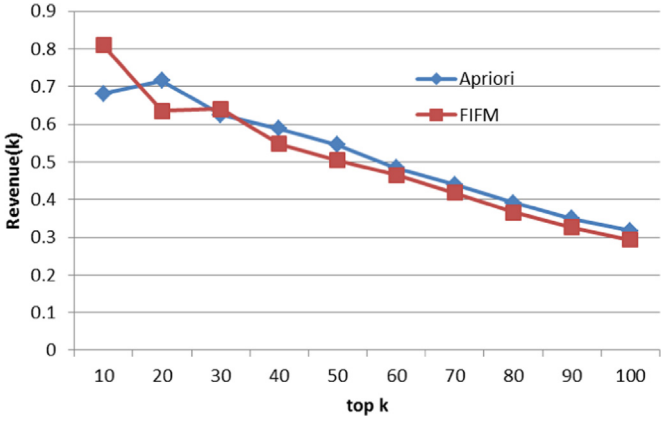

Fig. 18. Third quarter (revenue).

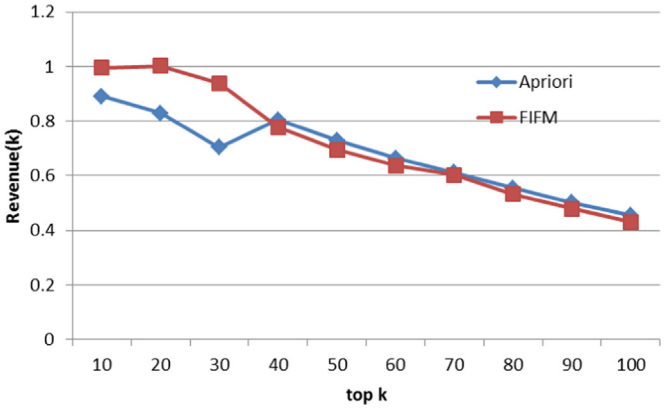

Fig. 19. Fourth quarter (revenue).

The length of the itemsets ranged from 1 to 67 items, and the average length of the itemsets was 5 items. The prices of the items ranged from NT$1 to NT$4800.

The transaction sizes in the first, second, third, and fourth weeks were 120,802, 89,665, 99,332, and 91,732, respectively. We used the transactions in the first week as the training dataset. Subsequently, we selected the top k frequent itemsets with higher support to investigate the numbers of patterns in L_1, L_2, and L_3 for top k frequent itemsets with higher support. We observed that the numbers of patterns in L_1, L_2, and L_3 for top k frequent itemsets were different (Table 15). Therefore, investigating the performance difference between the two approaches (Apriori and FIFM) in predicting customers' purchasing itemsets and the revenue from customer in advance is illuminating.

Table 14
Association rules (ARs) generated using top 5 L_2 frequent itemsets.

No	*FIFM*			*Apriori*		
	AR	*FMsup*	*FMconf*	AR	*sup*	*conf*
1	P1002⇒P1031	1.23%	45.38%	P1002⇒P1031	1.97%	37.50%
2	P1031⇒P1002	1.23%	32.36%	P1031⇒P1002	1.97%	23.08%
3	P1002⇒P1032	1.11%	40.85%	P1002⇒P1036	1.97%	37.50%
4	P1032⇒P1002	1.11%	70.47%	P1036⇒P1002	1.97%	37.50%
5	P1002⇒P1036	0.95%	35.01%	P1020⇒P1060	1.97%	60.00%
6	P1036⇒P1002	0.95%	43.94%	P1060⇒P1020	1.97%	27.27%
7	P1031⇒P1072	1.26%	33.02%	P1031⇒P1072	1.97%	23.08%
8	P1072⇒P1031	1.26%	35.64%	P1072⇒P1031	1.97%	23.08%
9	P1035⇒P1046	1.03%	46.95%	P1041⇒P1065	1.97%	37.50%
10	P1046⇒P1035	1.03%	82.51%	P1065⇒P1041	1.97%	37.50%

5.4.2. Comparison of precision and recall predictions

We set the minimum support to 0.0005 and varied the top k from 100 to 1000 to investigate the recall differences between the two approaches. Figs. 20, 22, and 24 present the precision differences between the two approaches in predicting customers' purchasing itemsets. Figs. 21, 23, and 25 present the differences in measures (precision and recall) between the two approaches in predicting customers' purchasing itemsets. The experimental results of the precision and recall values reveal that Apriori slightly outperformed the proposed approach in measures (precision and recall). In other words, the patterns discovered using Apriori are more suitable for prediction of itemsets from customers in next-period transactions.

5.4.3. Comparison of revenue predictions

We set the minimum support to 0.0005 and varied the top k from 100 to 1000 to investigate the revenue differences between the two approaches. Figs. 26–28 present the differences between the two approaches in predicting revenues generated from customers' purchasing itemsets. The experimental results of the revenue value predictions reveal that the proposed approach outperformed Apriori in predicting revenues of itemsets from customers in next-period transactions. In addition, the proposed approach slightly outperformed Apriori in predicting revenues in the top 200 patterns.

5.4.4. Discovered association rules

We identified the association rules from top 5 L_2 frequent itemsets generated using the FIFM approach and the Apriori approach. Table 16 shows that the support values (*FMsup*) of patterns discovered using FIFM are lower than the support values (*sup*) of patterns discovered using Apriori. This is because FIFM discovers frequent itemsets from FM-weighted transactions by assessing customer value.

The association rules of top 5 L_2 frequent itemsets identified using FIFM and Apriori were the same (Table 16). However, some association rules were only generated from FIFM in the top 200 patterns because FIFM outperformed Apriori in predicting revenues (Figs. 26–28).

Because patterns discovered using FIFM outperformed those discovered using Apriori in predicting revenues of itemsets from customers in next-period transactions, we suggest that marketing administrators of supermarket businesses could use the association rules generated by using FIFM in recommender systems to increase revenue.

Table 15
Top k frequent itemsets with higher support.

Frequent	*k = 300*		*k = 600*		*k = 900*		*k = 1200*		*k = 1500*	
itemset	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*	*FIFM*	*Apriori*
L1	283	286	531	537	778	791	915	1006	915	1229
L2	17	14	69	63	122	109	166	194	166	271
L3	0	0	0	0	0	0	0	0	0	0

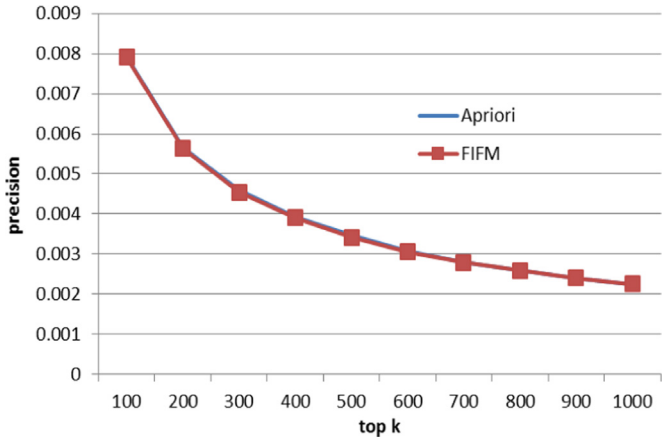

Fig. 20. Second week (precision).

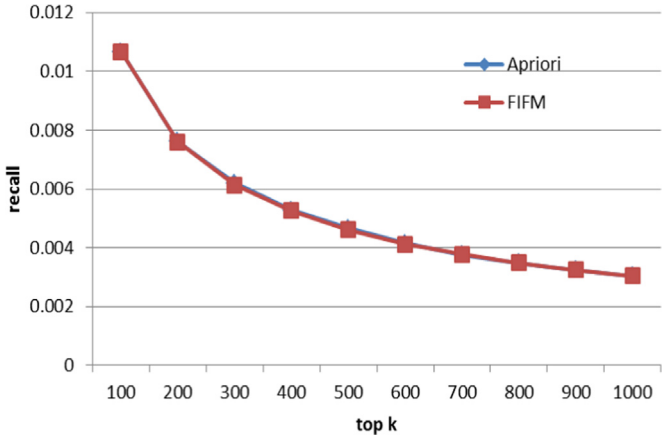

Fig. 21. Second week (recall).

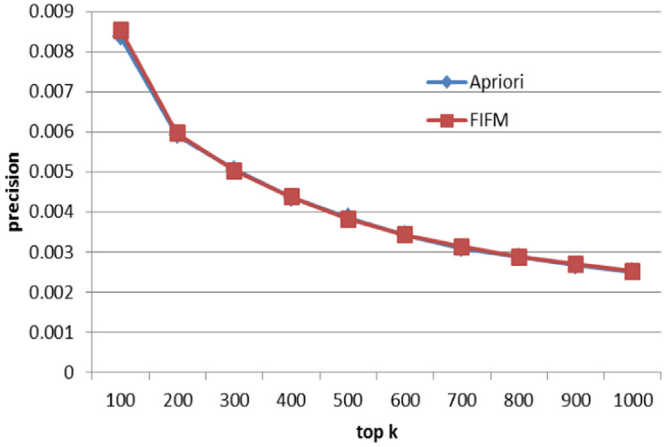

Fig. 22. Third week (precision).

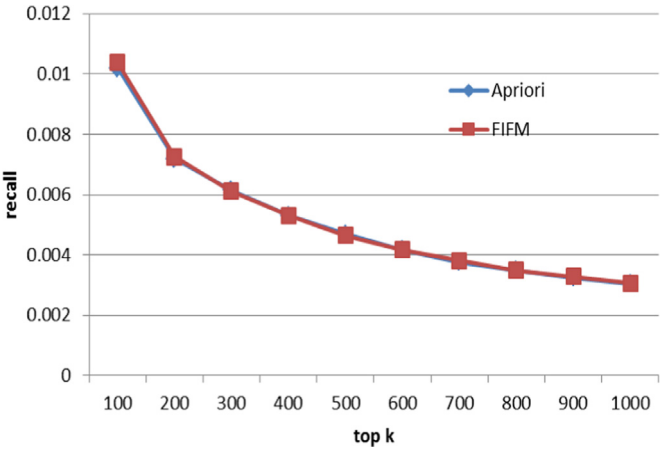

Fig. 23. Third week (recall).

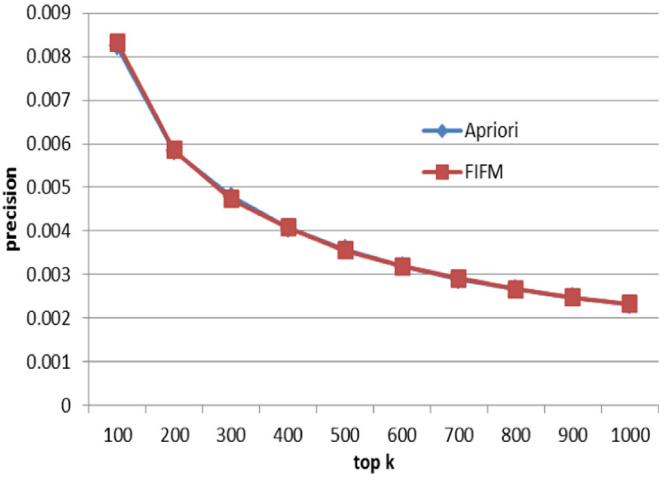

Fig. 24. Fourth week (precision).

5.5. Summary and discussion

In this study, we explored the differences in prediction measures (precision recall and revenue) among patterns (discovered using Apriori and the proposed FIFM algorithms) exhibited by customers in next-period transactions. The experimental results are summarized as follows.

According to Figs. 2–7, 11–16, and 20–25, the frequent itemsets generated using the Apriori algorithm outperformed the patterns generated using the FIFM algorithm in the measures (precision and recall). Therefore, the frequent itemsets generated using the Apriori algorithm are more suitable for predicting the itemsets of customer purchases in next-period transactions.

According to Figs. 8–10, 17–19, and 26–28, the patterns generated using the FIFM algorithm outperformed the frequent itemsets generated using the Apriori algorithm in the revenue measure. Therefore, the patterns generated using FIFM algorithm are more suitable

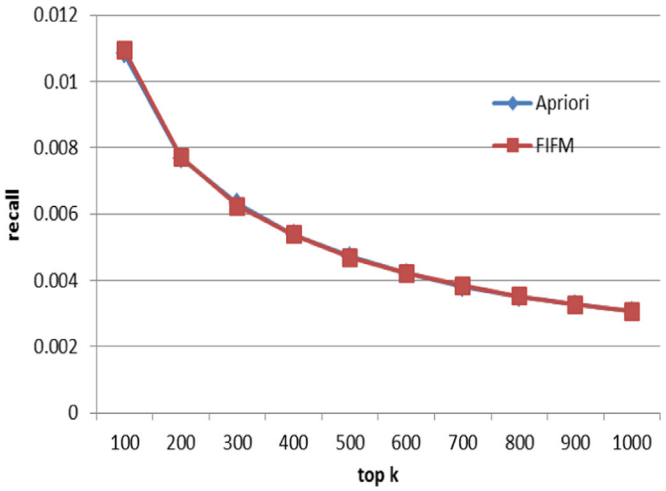

Fig. 25. Fourth week (recall).

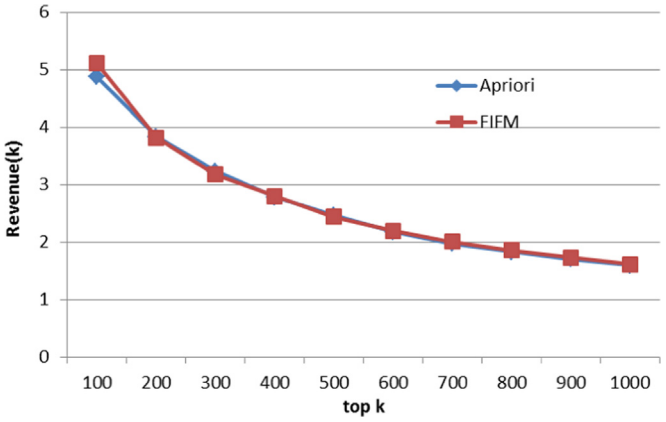

Fig. 26. Second week (revenue).

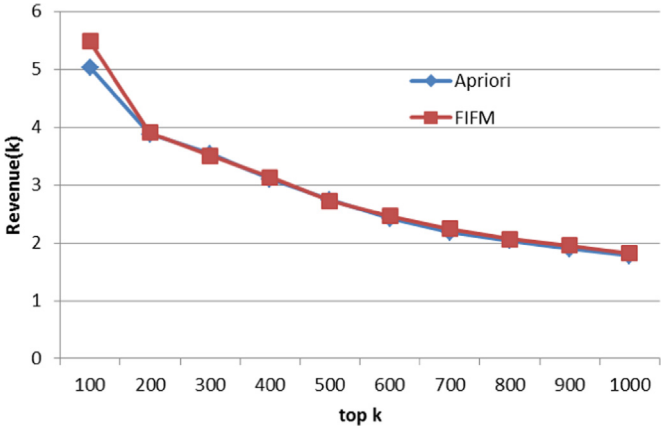

Fig. 27. Third week (revenue).

for predicting customers' revenue in next-period transactions.

6. Conclusion

Because of their practicality, ARM algorithms have been used in various applications and data sets. This study is the first to introduce the most efficient method for discovering frequent itemsets from transactions by considering the customer's FM value. Furthermore,

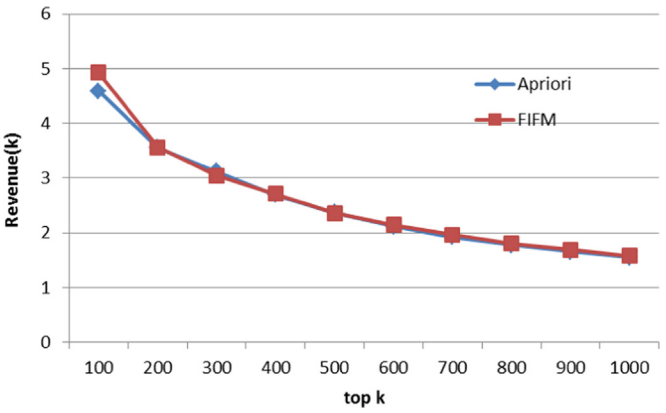

Fig. 28. Fourth week (revenue).

Table 16
Association rules (ARs) generated using top 5 L_2 frequent itemsets.

No	*FIFM*			*Apriori*		
	AR	*FMsup*	*FMconf*	AR	*sup*	*conf*
1	P1001⇒P1002	0.04%	12.82%	P1001⇒P1002	0.74%	12.72%
2	P1002⇒P1001	0.04%	17.12%	P1002⇒P1001	0.74%	15.75%
3	P1001⇒P1004	0.03%	10.76%	P1001⇒P1004	0.67%	11.59%
4	P1004⇒P1001	0.03%	17.83%	P1004⇒P1001	0.67%	17.80%
5	P1001⇒P1007	0.03%	9.74%	P1001⇒P1007	0.53%	9.18%
6	P1007⇒P1001	0.03%	18.25%	P1007⇒P1001	0.53%	18.14%
7	P1002⇒P1013	0.03%	13.32%	P1002⇒P1013	0.54%	11.42%
8	P1013⇒P1002	0.03%	24.47%	P1013⇒P1002	0.54%	23.41%
9	P1002⇒P1025	0.03%	12.86%	P1002⇒P1025	0.56%	11.84%
10	P1025⇒P1002	0.03%	31.78%	P1025⇒P1002	0.56%	32.50%

we proposed a new algorithm, FIFM, to discover frequent itemsets from FM-weighted transactions. Experimental results from the survey data reveal that the proposed approach can enable the discovery of interesting and valuable patterns that have not been discovered using conventional approaches. In addition, the top k frequent itemsets discovered using the proposed FIFM approach outperformed those derived using the conventional approach in the prediction of revenue from customers in next-period transactions.

Several issues remain to be addressed. First, we focused on discovering frequent itemsets from FM-weighted transactions rather than using all the transactions with equal weights. In some applications, business administrators may be interested in items with different weights. In addition, applying new additional measures to augment the support-confidence framework to discover interesting association rules could be productive. To address this issue, a more efficient algorithm should be designed. Finally, we suggest that the proposed approach be refined to discover frequent itemsets from FM-weighted transactions.

Acknowledgements

This research was supported by the Ministry of Science and Technology of the Republic of China under contract MOST 105-2410-H-166-002.

References

Agrawal, R., Imieliński, T., Swami, A. 1993. Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD. Washington, DC, USA. pp. 207–216.

Ahn, K.I., 2012. Effective product assignment based on association rule mining in retail. Expert Syst. Appl. 39 (16), 12551–12556.
Chan, C.C.H., 2008. Intelligent value-based customer segmentation method for campaign management: a case study of automobile retailer. Expert Syst. Appl. 34 (4), 2754–2762.
Chiang, W.Y., 2011. To mine association rules of customer values via a data mining procedure with improved model: an empirical case study. Expert Syst. Appl. 38 (3), 1716–1722.
Cil, I., 2012. Consumption universes based supermarket layout through association rule mining and multidimensional scaling. Expert Syst. Appl. 39 (10), 8611–8625.
Dursun, A., Caber, M., 2016. Using data mining techniques for profiling profitable hotel customers: an application of RFM analysis. Tour. Manag. Perspect. 18, 153–160.
Han, J.W., Kamber, M., 2006. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco.
Hsieh, N., 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Syst. Appl. 27 (4), 623–633.
Hu, Y.H., Yeh, T.W., 2014. Discovering valuable frequent patterns based on RFM analysis without customer identification information. Knowl. Based Syst. 61, 76–88.
Huang, Z., Lu, X., Duan, H., 2011. Mining association rules to support resource allocation in business process management. Expert Syst. Appl. 38 (8), 9483–9490.
Hughes, A.M., 2006. Strategic Database Marketing. McGraw-Hill.
Kalakota, R., Robinson, M., 1999. e-Business Roadmap For Success. Addison Wesley Longman Inc, New York, USA.
Kim, H.K., Im, K.H., Park, S.C., 2010. DSS for computer security incident response applying CBR and collaborative response. Expert Syst. Appl. 37 (1), 852–870.
Kuo, R.J., Pai, C.M., Lin, R.H., Chu, H.C., 2015. The integration of association rule mining and artificial immune network for supplier selection and order quantity allocation. Appl. Math. Comput. 250, 958–972.
Le, H.Q., Arch-Int, S., Nguyen, H.X., Arch-Int, N., 2013. Association rule hiding in risk management for retail supply chain collaboration. Comput. Ind. 64 (7), 776–784.
Lee, C., Song, B., & Park, Y. (2012). Design of convergent product concepts based on functionality: An association rule mining and decision tree approach. Expert Systems with Applications, 39(10), 9534–9542.
Lee, D., Park, S.H., Moon, S., 2013. Utility-based association rule mining: a marketing solution for cross-selling. Expert Syst. Appl. 40 (7), 2715–2725.
Li, L.H., Lee, F.M., Liu, W.J., 2006. The timely product recommendation based on RFM method. In: Proceedings of International Conference on Business and Information. Singapore.
Lin, C.S., Tang, Y.Q., 2006. Application of incremental mining and customer's value analysis to collaborative music recommendations. J. Inf. Technol. Soc. 6 (1), 1–26.
Linoff, G.S., Berry, M.J.A., 2002. Mining the Web: Transforming Customer Data into Customer Value. John Wiley and Sons, New York, NY.
Liu, D.R., Shih, Y.Y., 2005. Integrating AHP and data mining for product recommendation based on customer lifetime value. Inf. Manag. 42 (3), 387–400.
Lu, S., Hu, H., Li, F., 2001. Mining weighted association rules. Intell. Data Anal. 5 (3), 211–225.
Peppard, J., 2000. Customer relationship management (CRM) in financial services. Eur. Manag. J. 18 (3), 312–327.
R.C. Blattberg B.D. Kim S.A. Neslin. 2008. Database Marketing: Analyzing and Managing Customers (Chapter 12, Series eds. J. Eliashberg). Springer. New York, USA.
Shim, B., Choi, K., Suh, Y., 2012. CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. Expert Syst. Appl. 39 (9), 7736–7742.
Vo, B., Coenen, F., Le, B., 2013. A new method for mining frequent weighted itemsets based on WIT-trees. Expert Syst. Appl. 40 (4), 1256–1264.
Wang, W., Yang, J., Yu, P., 2004. WAR: weighted association rules for item intensities. Knowl. Inf. Syst. 6 (2), 203–229.
Weng, C.H., 2016. Identifying association rules of specific later-marketed products. Appl. Soft Comput. 38, 518–529.
Yun, U., Leggett, J.J., 2005. WFIM: weighted frequent itemset mining with a weight range and a minimum weight. In: Proceeding of the 2005 SIAM International Conference on Data Mining (SDM'05). Newport Beach, CA. pp. 636–640.