Received 2 June 2015
Revised 9 October 2015
Accepted 30 October 2015

Knowledge discovery of digital library subscription by RFC itemsets

Cheng-Hsiung Weng
Central Taiwan University of Science and Technology, Taichung, Taiwan

Abstract

Purpose – The paper aims to understand the book subscription characteristics of the students at each college and help the library administrators to conduct efficient library management plans for books in the library. Unlike the traditional association rule mining (ARM) techniques which mine patterns from a single data set, this paper proposes a model, recency-frequency-college (RFC) model, to analyse book subscription characteristics of library users and then discovers interesting association rules from equivalence-class RFC segments.

Design/methodology/approach – A framework which integrates the RFC model and ARM technique is proposed to analyse book subscription characteristics of library users. First, the author applies the RFC model to determine library users' RFC values. After that, the author clusters library users' transactions into several RFC segments by their RFC values. Finally, the author discovers RFC association rules and analyses book subscription characteristics of RFC segments (library users).

Findings – The paper provides experimental results from the survey data. It shows that the precision of the frequent itemsets discovered by the proposed RFC model outperforms the traditional approach in predicting library user subscription itemsets in the following time periods. Besides, the proposed approach can discover interesting and valuable patterns from library book circulation transactions.

Research limitations/implications – Because RFC thresholds were assigned based on expert opinion in this paper, it is an acquisition bottleneck. Therefore, researchers are encouraged to automatically infer the RFC thresholds from the library book circulation transactions.

Practical implications – The paper includes implications for the library administrators in conducting library book management plans for different library users.

Originality/value – This paper proposes a model, the RFC model, to analyse book subscription characteristics of library users.

Keywords Data processing, Knowledge management, Discovery tools, Data retrieval, Library administration

Paper type Research paper

Introduction

In the research literature, most studies that contain the terms "library" and "data mining" are not talking about traditional library data, but rather using the word *library* in the context of software libraries. Ladwig and Miller (2013) considered this distinction to be important for decisions regarding locating off-site storage, assessing collection development, maintaining departmental libraries and so forth. With a thorough understanding of book subscription characteristics needs of the students at each of their

This research was partially supported by the National Science Council of the Republic of China under contract NSC 102-2410-H-166-002.

The Electronic Library
Vol. 34 No. 5, 2016
pp. 772-788

0264-0473
DOI 10.1108/EL-06-2015-0086

colleges, library administrators can conduct efficient library management plans for books. Furthermore, this could be a useful reference to develop book circulation strategies. Therefore, analysing library users' preferences, such as for books in print or electronic, is still an important issue for libraries to explore.

This study proposes a new method for analysing characteristics of library users. This new method, the recency-frequency-college (RFC) scoring method, is designed to analyse book subscription characteristics of library users. In particular, the researchers are interested in formulating the RFC model to represent the patterns of usage of digital libraries associated with different types of users.

Compared to the traditional association rule mining (ARM) techniques, which mine patterns from a single data set, this study first applies the RFC model to analyse book subscription characteristics of library users and then segments library users into several groups. After this, the transactions (book subscriptions) are partitioned into several sub-RFC data sets according to the users' RFC scores. Finally, the RFC association rules from each sub-RFC data set are calculated, rather than the original data set.

In this study, the difference in precision of frequent itemsets discovered by proposed and traditional approaches are considered. The differences of library book subscription among clusters (sub-RFC data sets) by inter-cluster analysis are also investigated. Moreover, the library book subscription patterns of students in specific clusters by intra-cluster analysis are examined.

Literature review

This study proposes an RFC model to score book subscription characteristics of library users from their book-borrowing records and segment users into several groups. In this section, the issue and theoretical basis of the operation model as well as some techniques related to clustering customer values are explored. Related recency-frequency-monetary (RFM) and ARM studies are reviewed.

Recency-frequency-monetary

Based on the customer relationship management theory (Kalakota and Robinson, 1999; Peppard, 2000), library managers should implement a value analysis of library users and determine how to further provide better library services. RFM was defined by Hughes (1994) as follows:

- *recency* (R) is defined as "last purchasing time";
- *frequency* (F) is defined as "purchasing frequency in a specific period"; and
- *monetary* (M) is defined as "average amount of purchasing in a specific period".

The RFM model can effectively carry out the process of clustering based upon customer values. Business plans can be made to extend the customer's life cycle by implementing marketing projects (Linoff and Berry, 2002). RFM scoring is a way to determine the score of current customers from their recency, frequency and monetary values and has proven very effective when applied to marketing databases (Blattberg *et al.*, 2008). Various strategies are developed for the purpose of enhancing shopping rates, increasing high profit/price product sales and retaining customers to become long-term customers.

Numerous studies have discussed the usage of RFM values in recommender systems. Liu and Shih (2005) suggested combining customer lifetime value and RFM to analyse

customers' consumption properties and to give a recommendation based upon these properties. In their studies, clustering techniques were used to group customers according to the weighted RFM (WRFM) value. However, Li *et al.* (2006) proposed the timely RFM method, instead of the WRFM, to take product property and purchase periodicity into consideration.

RFM concepts have been applied in different areas. Kim *et al.* (2010) proposed utilizing an RFM engine for anomaly detection to minimize false alarms in network attacks and decrease the time needed to respond to hacking events. Chan (2008) combined RFM with a customer life time value model to evaluate segmented customers and then used a genetic algorithm to select more appropriate customers for each campaign strategy. Hsieh (2004) used a self-organizing map neural network to identify groups of bank customers based on repayment behaviour and RFM behavioural scoring predicators.

Several researchers considered RFM variables when developing clustering models. For example, Aggelis and Christodoulakis (2005) proposed using the *k*-means algorithm and two-step cluster analysis for clustering active e-banking users using a pyramid model. Wu *et al.* (2009) applied the RFM model and *k*-means method in the value analysis of the customer database of an outfitter in Taiwan to establish strong relationships and eventually consolidate customer loyalty for highly profitable long-term customers.

Association rules

ARM is an important data mining approach that can discover consumer purchasing behaviours from transaction databases (Han and Kamber, 2006). Agrawal *et al.* (1993) first introduced the problem, defining it as finding all rules from transaction data that satisfy the minimum support and the minimum confidence constraints. Because of its great success and widespread usage, many variants of ARM algorithms have been proposed. These algorithms can be roughly classified into three categories, according to the data types they can handle:

(1) nominal/Boolean data (Agrawal *et al.*, 1993; Agrawal and Srikant, 1994);

(2) ordinal data (Chen and Weng, 2008); and

(3) quantitative data (Chen *et al.*, 2010; Weng, 2011).

Mining frequent itemsets in ARM plays a crucial role (Agrawal *et al.*, 1996). At present, most of the frequent itemsets mining algorithms are improved or derivative algorithms based on Apriori (Agrawal and Srikant, 1994) or frequent pattern (FP)-growth (Han *et al.*, 2000). Moreover, other efficient methods for mining frequent itemsets are also proposed, such as H-mine (Pei *et al.*, 2001), Index-BitTableFI (Song *et al.*, 2008) and so forth.

To mine interesting rules, a correlation measure is used to augment the support – confidence framework of association rules. Aggarwal and Yu (1998) supplemented the support – confidence framework with additional measures of interesting rules based on statistical significance and correlation analysis. This leads to correlation rules of the form $X \Rightarrow Y$ [support, confidence and correlation]. There are various correlation measures including *lift*, χ^2, cosine and all-confidence (Han *et al.*, 2007). The problem of rule interestingness has been studied by many researchers. Brin *et al.* (1997) proposed using *lift* and χ^2 as correlation measures. Aggarwal and Yu (1998) studied the weakness of the support – confidence framework and then proposed the strongly collective itemset model for association rule generation.

The discovery of interesting association or correlation relationships is helpful in many business decision-making processes (Ahn, 2012; Kamsu-Foguem *et al.*, 2013). Huang *et al.* (2011) introduced an ARM-based approach to mine interesting resource allocation rules from event logs. Na and Sohn (2011) proposed the analysis of association rule for predicting changes based on the time series data of various interrelated world stock market indices. Chen *et al.* (2013) applied properties of propositional logic into consideration and proposed an algorithm for mining highly coherent rules. Le *et al.* (2013) proposed a new algorithm to remove sensitive knowledge from the released database based on the intersection lattice of frequent itemsets.

Association rules in recency-frequency-monetary

Lin and Tang (2006) combined the RFM model with the analysis of customer values to classify customers with similar characteristics into the same group. After this process, the similar vector matrix was used to calculate the degree of similarity relationship between users to assemble them conveniently. Finally, through the concept of collaborative filtering, the authors used the recommendation model as the method of individual recommendation that suggests music to users. Chiang (2011) proposed a new procedure and an improved the recency, frequency, monetary, discount, return cost (RFMDR) model to mine association rules of customer values.

However, for library users, searching for relevant literature is the most important thing, regardless of the literature's price in the library context. Determining the monetary (M) value of the library user is not necessary in the library context. Therefore, the RFM model is not suitable for library users. Library users at different colleges may have different book reading characteristics. Therefore, RF (recency-frequency) measures are retained from the RFM model, and the new measure named C (college) is appended to create a new RFC (recency-frequency-college) model to analyse library users and segment library users into different groups according to their RFC values in advance.

This study partitions a single data set into several sub-RFC data sets by using library users' RFC values to discover useful patterns from each RFC group of library users' transaction records. When discovering association rules from different sub-RFC data sets, some association rules may simultaneously exist in multiple sub-RFC data sets. That is, the same association rules (the same left-hand side (LHS) and right-hand side (RHS)) may have different support and confidence in different sub-RFC data sets. Therefore, it will be necessary to define a new form of association rule to identify which sub-RFC data set (also called *cluster of customers*) is more suitable to apply the association rules discovered.

Data mining in libraries

Nicholson (2006) proposed a framework for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Bibliomining* is concisely defined as the data mining techniques used to extract patterns of behaviour-based artefacts from library systems. The bibliomining process includes identifying a topic, creating a data warehouse, refining data, exploring data and evaluating results (Shieh, 2010). Hajek and Stejskal (2012) applied clustering (*k*-means) and extraction of attributes from real public library data to uncover similarities in the services provided by public libraries. ARM is also popularly used in the analysis of library circulation (Hwang and Lim, 2002; Lunfeng *et al.*, 2012; Song and Wei, 2011).

To the best of our knowledge, this study is the first to integrate clustering (RFC) and association rule techniques in analysing digital library subscriptions. To describe the differences between our approach and prior studies, the authors provide comparison data in Table I.

Proposed approach

The proposed RFC model is first introduced to determine library users' RFC values. Then, the problem of how to discover the RFC association rules from library user transactions is defined, and the proposed procedures for discovering useful RFC patterns are detailed.

Recency-frequency-college model

An RFC score model is constructed, modified from the RFM model, to determine library users' RFC values. According to the definition of the RFM model proposed by Hughes (1994), the three attributes (R, F and M) are equal in importance. That is, these three attributes for evaluating customer value should be equally weighted. However, the RFM model was originally developed to analyse products within each transaction for each customer at a specific time in the retail industry (Miglautsch, 2000). The library user who borrows books from a library only considers whether the content of the book matches his/her needs or not, rather than the price of the book. Therefore, the original RFM model is not suitable for analysing book subscription characteristics of the library users. Thus, a new model (RFC model) is proposed which uses the category (C) instead of the monetary value (M). The three key attributes (R, F and C) are defined as follows:

(1) *Recency (R)*: The R value represents the date of the user's last transaction. Because the R value contributes to the RFC scoring determination, a numeric value is necessary. The date threshold (R_T) is set to determine the R value of the user's last transaction. For example, if the date threshold (R_T) is set to 2012/2/1, then a user who has conducted his/her last transaction before 2012/2/1 is characterized by $R = 0$, whereas one who has conducted his/her last transaction after 2012/2/1 will have $R = 1$.

(2) *Frequency (F)*: The F value indicates the count of a user's transactions within the period of 2011/2/1 to 2012/2/1. Because the F value contributes to the RFC scoring determination, a numeric value is necessary. The count threshold (F_T) is set to determine the F value of the user's transactions. For example, set the count threshold (F_T) as 3. If the count of a user's transactions is more than 3, then that user's frequency value is characterized by $F = 1$, otherwise $F = 0$.

Table I. Comparison between this study and literature

Work	Techniques	Application (aim)
This study	Clustering (RFC) association rule	Analysis of digital library subscriptions
Hajek and Stejskal (2012)	Clustering (*k*-means)	Analysis of user behaviour in a public library
Hwang and Lim (2002)	Association rule	New library book recommendations
Lunfeng *et al.* (2012)	Association rule	Analysis of book-lending service
Song and Wei (2011)	Association rule	Analysis of library circulation

(3) College (C): The C value indicates the college of the library user (student). Table II shows the colleges used in this study. For example, for students of "College of Management", their college value are characterized by $C = 1$.

Problem definitions

In this section, the problem of how to discover the RFC association rules from library user transactions is defined.

Definition 1. Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of itemsets. Let D be a set of transactions where each transaction T is a set of items such that $T \subseteq I$. Recency (R), frequency (F) and college (C) values of each transaction T refer to the last trading date, total number of transactions and college of each library user, respectively. The RFC values of each transaction T are defined as $(RFC)_T = \{r_T, f_T, c_T\}$, where r_T, f_T and c_T are the recency, frequency and college values of transaction T in the data set D.

Example 1. There are five transactions indicated in Table III. When the last trading date of the transaction of the library user is after 2011/2/1, the *Recency* value (r_T) is set to 1; otherwise, to 0. When the total number of transactions of the user is greater than 3, the *Frequency* value (f_T) is set to 1; otherwise, to 0. The colleges used in this study are shown in Table II. If the student belongs to "College of Management", then the user's college value is $C = 1$.

After determining each library user's RFC value, each transaction's RFC value is determined as the same as the library user's RFC value. For example, the RFC value of library user UID No. 1 is (1, 0, 1) and, thus, the RFC value of TID No. 1001 is (1, 0, 1).

Definition 2. A transaction T is said to contain X if and only if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in data set D with *support* s, where s is the percentage of the transaction in D that contains $X \cup Y$. Rule $X \Rightarrow Y$ has a *confidence* c in D, where c is the percentage of transactions in D containing X that also contain Y. The following are the formal expressions for $Support(X \cup Y)$ and $Confidence(X \Rightarrow Y)$, respectively:

$$Support(X \cup Y) = \frac{|X \cup Y|}{|D|},$$

Code	College
0	College of Health Science
1	College of Management
2	College of Nursing

Table II. College codes

TID	UID	Itemsets	$(RFC)_T$
1001	1	*a, b, c*	(1, 0, 1)
1002	2	*a, b, d*	(1, 0, 1)
1003	5	*a, e, f*	(1, 0, 1)
1004	6	*a, b, d, e, g, h, i*	(1, 0, 2)
1005	7	*a, b, d, f, x, y, z*	(1, 0, 2)

Table III. RFC values of five transactions

where $|D|$ denotes the number of transactions in D and $|X \cup Y|$ denotes the number of transactions containing $X \cup Y$ in D, and

$$Confidence(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}.$$

Example 2. Table III shows a data set (D) containing five transactions and that $Support(a \cup b) = 4/5 = 0.80$ and $Confidence(a \Rightarrow b) = 0.80/1.00 = 0.80$. Thus, the following association rule is discovered:

$$a \Rightarrow b\,[support = 80\%, confidence = 80\%].$$

If the minimum support and minimum confidence are both set to 65 per cent, an association rule ($a \Rightarrow b$) can be generated. However, an association rule ($a \Rightarrow d$) cannot be generated, because the support of the itemsets $\{ad\}$ is 60 per cent, smaller than the minimum support (65 per cent).

Definition 3. Let $D = \cup_{i=1}^{k} D^{(RFC)_i}$. The RFC association rule, $X \Rightarrow Y$, generated from the RFC data set ($D^{(RFC)_i}$) is defined as follows:

$$X \Rightarrow Y[support, confidence, D^{(RFC)_i}].$$

Example 3. Table III shows two RFC data sets (D^{101} and D^{102}). From the RFC data set (D^{102}), $Support(a \cup d) = 2/2 = 1.00$ and $Confidence(a \Rightarrow d) = 1.00/1.00 = 1.00$. Thus, the following RFC association rule is discovered:

$$a \Rightarrow d\,[support = 100\%, confidence = 100\%, D^{102}].$$

The RFC data set (D^{101}) shows the other RFC association rule, $\{a \Rightarrow b\,[support = 66$ per cent, $confidence = 66$ per cent, $D^{101}]\}$. Therefore, from the two RFC data sets (D^{101} and D^{102}), the following RFC association rules are discovered:

$$a \Rightarrow b\,[support = 66\%, confidence = 66\%, D^{101}];$$

and

$$a \Rightarrow b\,[support = 100\%, confidence = 100\%, D^{102}].$$

With the help of the new RFC association rule form $\{X \Rightarrow Y[support, confidence, D^{(RFC)_i}]\}$, it can be seen that the RFC rule $\{a \Rightarrow b\}$ is more suitable to apply in the RFC data set (D^{102}), because the rule holds higher support and confidence. The RFC rule $\{a \Rightarrow b\}$ means that Nursing College students who borrowed books recently and infrequently like to borrow book (b) when borrowing book (a), with a confidence level of 100 per cent.

From the above discussion, with the form $\{X \Rightarrow Y\,[support, confidence, D^{(RFC)_i}]\}$, various RFC association rules from different RFC data sets (library user segments) can be discovered to identify library users' book subscription characteristics.

Proposed procedures

The proposed procedures for discovering useful RFC association rules are further explained. The proposed framework is illustrated in Figure 1. The procedures can be divided into five phases:

(1) select the data set and preprocess data;

(2) calculate a library user's RFC value by using the proposed RFC model;

(3) cluster library transaction segments by using the library users' RFC values;

(4) extract RFC association rules from library transaction segments; and

(5) evaluate the results of the experiment by comparing RFC association rules discovered from library transaction segments.

Step 1: preprocess data. For various reasons, real-world data sets are highly susceptible to noisy, missing and inconsistent data. Low-quality data lead to low-quality results. Therefore, it is necessary to preprocess the data set to make knowledge discovery. The first step is to delete records with missing values or inaccurate values, eliminating redundant attributes. After this, the data are transformed into a specific format for determining library users' RFC values.

Step 2: determine library users' RFC values. The RFC attributes are equal in weight (i.e. 1:1:1). Define the scaling of RFC attributes. The scaling of attribute *R* is (1, 0) and 1 which refers to the library user who has conducted his/her last transaction after 2012/2/1. The scaling of the *F* attribute is (1, 0) and 1 which refers to the count of the transactions a user who has conducted more than 30. The scaling of the *C* attribute is (2, 1, 0) and 2, which indicates College of Nursing.

Step 3: cluster library user transactions. Cluster transactions of library users into several clusters (segments) according to library users' RFC values. That is, if a library user belonged to cluster No. 1, all of his/her transactions (book-borrowing records) will be assigned to cluster No. 1.

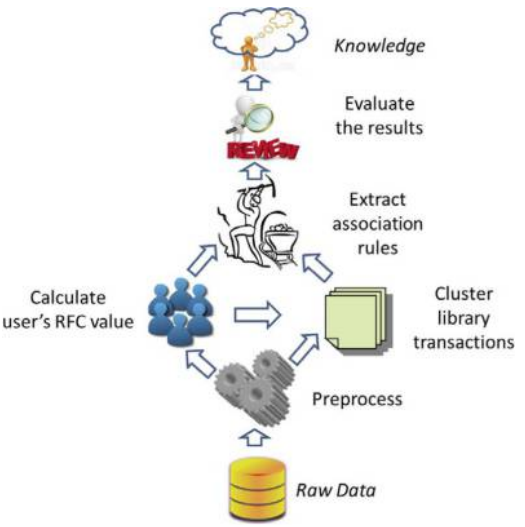

Figure 1. The proposed approach

Step 4: extract the association rules. After the preceding steps are completed, RFC association rules generated from different clusters (segments) will be inspected for diverse levels of support and confidence values. RFC association rules reveal what kinds of books are usually borrowed together by library users in different clusters.

Step 5: evaluate the results. Evaluate RFC association rules generated from all clusters using intra-cluster analysis and inter-cluster analysis. It is useful to help library managers to identify characteristics within the same cluster through intra-cluster analysis and to understand the difference between the clusters with inter-cluster analysis.

During the inter-cluster analysis phase, rules generated from different clusters reveal the different book subscription characteristics of library users in different clusters. Understanding the difference in book subscription characteristics of library users among different clusters can help library administrators provide different services to meet the specific requirements in different library user clusters.

The comparison of the proposed approach and literature

To describe the differences and application between the proposed algorithm and the Apriori algorithm, comparison data are provided in Table IV.

Empirical case study

In this section, the experimental results are discussed. The results are based on the association rules generated from the University CTUST library transactions. A total of 109,902 transactions from 8,762 students were collected. Each transaction gave information about the student's book-borrowing record. All selected transactions were cleaned and processed for further analysis in advance.

First, the library users are divided, as well as their transactions, into three clusters (College of Health Science, College of Management and College of Nursing) based on the *C* value in the RFC model. Then, the characteristics of three clusters are identified by comparing pair-wise rules. Second, each intra-cluster (such as the College of Nursing) is further explored to identify the book subscription characteristics of students at a specific college.

Prediction precision

It is interesting to investigate if the frequent itemsets discovered by the proposed approach are more suitable to predict library users' subscription itemsets that exist in the transactions in the next time periods. Therefore, the *Precision* measure is used to evaluate prediction precision of frequent itemsets discovered by the proposed approach.

Table IV.
Comparison of the proposed approach and Apriori algorithm

Characteristic	Apriori algorithm (Agrawal *et al.*, 1993)	Proposed algorithm
Mechanism to cluster users	None	RFC
Mechanism to cluster transactions	None	RFC
Data set	Single data set without clustering	Several RFC sub-data sets after clustering
Type of rule presentation	General association rules	RFC association rules
Application situation	Single market	Several market segments

Let A and B be the sets of patterns generated from the training and test data set, respectively. The *Precision* measure can be defined as follows:

$$precision = \frac{|A \cap B|}{|A|}.$$

Transactions are partitioned into two subsets (subset-A and subset-B). Subset-A is set to be a training data set and Subset-B to be a test data set. Subset-A recorded all transactions between 2007/1/2 and 2009/12/31. Subset-B recorded all transactions between 2010/1/2 and 2011/12/31.

First, the library users' RFC values are determined from subset-A. The date threshold (R_T) is set to "2009/1/2" and then a user who has conducted his/her last transaction after 2009/1/2 is characterized by $R = 1$; otherwise $R = 0$. The count threshold (F_T) is set as 2. If the count of a specific user's transactions is greater than 2, the user's frequency value is characterized by $F = 1$; otherwise $F = 0$. The C value indicates the college of the library users (students). The colleges' values in this study are shown in Table II. The number of students in RFC clusters in Subset-A is shown in Table V.

Table VI shows that precision of the proposed approach is always higher than that of the traditional approach (Apriori) across the three colleges. Therefore, the frequent itemsets discovered by the proposed approach outperform the traditional approach (Apriori) in predicting library users' subscription itemsets over the examined time periods.

Inter-cluster analysis

The differences of library users' book subscription characteristics of three clusters (health science, management and nursing) are investigated. Based on the RFC model, the number of frequent itemsets in the Health Science College is D^{-0} (3,150 students), in the Management College is D^{-1} (2,472 students) and in the Nursing College is D^{-2} (3,140 students).

R	F	C	Students	R	F	C	Students	R	F	C	Students
0	0	0	156	0	0	1	106	0	0	2	125
0	1	0	167	0	1	1	96	0	1	2	114
1	0	0	383	1	0	1	338	1	0	2	327
1	1	0	930	1	1	1	539	1	1	2	859
	Total		1,636		Total		1,079		Total		1,425

Table V. The number of students in the RFC clusters

	Management college		Nursing college		Health Science college	
Year	This study (%)	Apriori (%)	This study (%)	Apriori (%)	This study (%)	Apriori (%)
2010	30.8	0.0	39.6	0.2	25.4	0.0
2011	33.4	0.3	21.9	0.2	38.1	0.2
AVG	32.1	0.1	30.7	0.3	31.7	0.1

Table VI. Prediction precision values of the proposed approach and Apriori

Table VII shows that the number of frequent itemsets in the Health Science College is 750, more than the other two colleges (Management and Nursing). That reflects the diversity of books borrowed by the students at the Health Science College.

From the results in Table VIII, the number of the same frequent itemset (L_1) existed in the Management College and Health Science College and is 416. It means that students at the Management College and Health Science College like to borrow the same books more than the other college pairs.

With *support* threshold ($\sigma_{sup} = 0.005$), the number of the same frequent itemset (L_1) in the three colleges is determined to be 318. Compared to Table VII, the proportions of the same frequent itemsets over the frequent itemsets in three colleges are 53.8 per cent (318/591) in the Management College, 46.2 per cent (318/688) in the Nursing College and 42.4 per cent (318/750) in the Health Science College. Therefore, there are about 40 per cent of the same frequent itemsets (books) borrowed in the three colleges.

Furthermore, the books borrowed frequently (frequent itemsets) by students at each college were investigated to understand if the books subscribed are the same or not. Tables IX shows the top five books borrowed frequently by students at the colleges (Health Science, Management and Nursing). It is interesting to note that the top five books subscribed together by students at each college are totally different. With *support* threshold ($\sigma_{sup} = 0.005$) and *confidence* threshold ($\sigma_{conf} = 0.4$), the association rules discovered by the Apriori algorithm from the University CUTST library data set are shown in Table X. Rule No. 1 indicates that the students of university CUTST like to borrow two novels entitled "A person lives 5 years" and "150 cm life". With such a low *support* threshold ($\sigma_{sup} = 0.005$), only one association rule from the University CUTST library data set is discovered.

One may wonder if students from different colleges borrowed different books. The results in Table XI show that most of the students at the College of Health Science and the College of Management like to borrow novels. Comparison of the association rules discovered from three colleges (Health Science, Management and Nursing) shows that most of the Nursing College students like to borrow textbooks, rather than novels.

Table VII.
Number of frequent itemsets with different support thresholds

College	Support 0.005	0.01	0.02	0.03	0.04
Management	591	162	22	6	0
Nursing	688	137	13	2	2
Health Science	750	216	31	8	3

Table VIII.
Number of the same frequent itemsets (L_1) in different colleges

College	Management	Nursing	Health Science
Management		357	416
Nursing	357		375
Health Science	416	375	

College	Itemsets	Count	Sup (%)
Health Science	Oolong Police Station, Ganso Crazy Family	35	1.1
	GTO, Oolong Police Station	26	0.9
	Dramatic Baseball Comic, One Piece	25	0.8
	One Piece, Ganso Crazy Family	25	0.8
	Bleach, One Piece	25	0.8
Management	Building B, 11th Floor; We Do Not Get Married, Okay?	20	0.8
	Building B, 11th Floor; City, 11th Floor, Block B, Part II	18	0.7
	City, 11th Floor, Block B, Part II; We Do Not Get Married, Okay?	17	0.7
	Ardour; Seduction Poisoning	15	0.6
	A Person Lives 5 Years; 150cm Life	15	0.6
Nursing	Nursing Test Tips; Nursing Administration	24	0.8
	Medical and Surgical Nursing; Nursing Administration	23	0.7
	Nightingale and Modern Nursing; Nursing Administration	20	0.6
	Paper Sculpture Modeling; Creative Skillfully Decorated	20	0.6
	Gakuen Alice; Skip. Beat!	18	0.6

Table IX.
Top five frequent itemsets (L_2) in colleges discovered by the proposed approach

No.	LHS	RHS	Count	Sup (%)	Conf (%)
1	A Person Lives 5 Years	150 cm Life	44	0.5	61.1

Table X.
Association rule discovered by the Apriori approach

College	No.	LHS	RHS	Count	Sup (%)	Conf (%)
Health Science	1	Ha Bang, My Boss	Fate Hunter	19	0.6	63.3
	2	Hitman Reborn!	One Piece	18	0.6	58.1
	3	Ingot Crazy Family	Ganso Crazy Family	18	0.6	90.0
	4	A Person Lives 5 Years	150cm Life	17	0.5	85.0
	5	Dream Dreams	Howl	16	0.5	66.7
Management	1	Building B, 11th Floor	We Do Not Get Married, Okay?	20	0.8	54.1
	2	Building B, 11th Floor	City, 11th Floor, Block B, Part II	18	0.7	48.6
	3	City, 11th Floor, Block B, Part II	We Do Not Get Married, Okay?	17	0.7	40.5
	4	Ardour	Seduction Poisoning	15	0.6	48.4
	5	A Person Lives 5 Years	150 cm Life	15	0.6	57.7
Nursing	1	Nursing Test Tips	Nursing Administration	24	0.8	72.7
	2	Paper Sculpture Modeling	Original and Skillful Decorations	20	0.6	71.4
	3	Gakuen Alice	Skip. Beat!	18	0.6	48.6
	4	Gakuen Alice	Special A	16	0.5	43.2
	5	Practical Pediatric Physiology	Pediatric Physiology	16	0.5	59.3

Table XI.
Top five association rules discovered by the proposed approach

Intra-cluster analysis

Table XI, shows that most of the Nursing College students like to borrow textbooks, rather than novels. It would be interesting to analyse which books are borrowed with different recency and frequency thresholds (R_T and F_T).

The frequency threshold (F_T) is set to be 30 and then the Nursing College students are divided into two groups (frequent and infrequent). There are 342 Nursing College students who borrowed books frequently and 2,798 Nursing College students who borrowed books infrequently.

Intra-cluster analysis shows that the 342 Nursing College students who borrowed books frequently also borrowed novels frequently. However, the 2,798 Nursing College students who borrowed books infrequently only borrowed textbooks.

Moreover, the recency threshold (R_T) is set to be 2012/2/1 and then the 2,798 Nursing College students who borrowed books infrequently are divided into two groups (recent and non-recent). There are 1,768 Nursing College students who borrowed books recently and infrequently. There are 1,030 Nursing College students who borrowed books non-recently and infrequently.

The results from Table XII show that:

- Nursing College students who borrowed books non-recently and infrequently like to borrow textbooks (such as "Interpersonal Relationships and Communication" and "Nursing Administration"); and
- Nursing College students who borrowed books recently and infrequently like to borrow textbooks (such as "Nursing Test Tips" and "Nursing Administration").

Note that the textbook "Nursing Test Tips" is useful for nursing certification. In addition, these students also like to borrow reference books (such as "Paper Sculpture Modeling" and "Original and Skillful Decorations"), in addition to textbooks.

Summary and management implications

The goal of this experiment is to analyse book subscription characteristics of library users and to give some management information to library administrators. In preparation, comparison data are provided to describe the differences between our study and Apriori study, as shown in Table XIII.

On the basis of the RFC patterns mining results, some suggestions are proposed for the administrators of college libraries. This could be a useful reference for developing book circulation strategies. The results and implications based on our findings are discussed below.

Table XII.
Top three association rules discovered by concerning different recency values

Recency	No.	LHS	RHS	Count	Sup (%)	Conf (%)
No	1	Human Relations	Nursing Administration	6	0.6	42.9
	2	Introduction to Critical Care	Critical Care Nursing	6	0.6	54.5
	3	Interpersonal Relationships and Communication	Nursing Administration	6	0.6	85.7
Yes	1	Nursing Test Tips	Nursing Administration	11	0.6	73.3
	2	Pediatric Physiology	Pediatric Physiology	9	0.5	47.4
	3	Paper Sculpture Modeling	Original and Skillful Decorations	9	0.5	69.2

Inter-cluster analysis finds that students at the colleges borrowed different books from the library. One interesting finding is that most of the Nursing College students like to borrow textbooks, rather than novels. This is different from the students at the Health Science College and the Management College, who like to borrow novels. In addition, some of the Nursing College students also like to borrow reference books. Therefore, the library administrator should utilize different library management plans of books for different colleges.

Intra-cluster analysis finds that the recency value (recent or non-recent) helps to gain insight into the book subscription characteristics of students at the Nursing College. Their book subscription characteristics are further investigated with intra-cluster analysis. It is discovered that the students at the Nursing College who borrowed books recently and infrequently liked to borrow reference books, rather than textbooks. Therefore, the library administrator should conduct specific library management plans about reference books for students at the Nursing College.

Conclusion and future works

This study first proposes the RFC model to analyse the book subscription characteristics of library users and also discovers interesting association rules from equivalence-class RFC data sets in the library context. Experimental results from the survey data show that the precision of the frequent itemsets discovered by the proposed approach outperforms the traditional approach (Apriori) in predicting library users' subscription itemsets using sequential time periods. Furthermore, the proposed approach can help to discover interesting and valuable patterns for library book circulation in different library user segments. With the help of the RFC association rules discovered, library administrators can understand differences among library user segments and provide better services in the form of book circulation in the library.

There are several issues that remain to be addressed in future research. First, clustering library users correctly according to their RFC values is very important for library user segments. In the future, I hope to combine other clustering methods to cluster library users. In addition, it is currently assumed that the RFC thresholds were assigned based on expert opinion. In the future, I will attempt to automatically infer the RFC thresholds from the raw data, thereby avoiding the acquisition bottleneck. Finally, because learning environments in which the users of library services are changing radically, I consider that combining user profiles in bibliomining is an interesting research issue. Integrating change mining into bibliomining will be helpful in the analysis of library circulation.

Table XIII. Comparison of proposed approach and Apriori

Characteristic	Apriori algorithm (Agrawal *et al.*, 1993)	Proposed algorithm
Prediction precision	0.2%	31.5%
Inter-cluster analysis	No	Yes
Intra-cluster analysis	No	Yes

References

Aggarwal, C.C. and Yu, P.S. (1998), "A new framework for itemset generation", *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM, New York, NY*, pp. 18-24.

Aggelis, V. and Christodoulakis, D. (2005), "Customer clustering using RFM analysis", *Proceedings of the 9th WSEAS International Conference on Computers, World Scientific and Engineering Academy and Society (WSEAS), Wisconsin*, p. 2.

Agrawal, R., Imieliński, T. and Swami, A. (1993), "Mining association rules between sets of items in large databases", *ACM SIGMOD Record*, Vol. 22 No. 2, pp. 207-216.

Agrawal, R. and Srikant, R. (1994), "Fast algorithms for mining association rules", *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), San Jose, CA*, Vol. 1215, pp. 487-499.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. (1996), "Fast discovery of association rules", *Advances in Knowledge Discovery and Data Mining*, Vol. 12 No. 1, pp. 307-328.

Ahn, K.I. (2012), "Effective product assignment based on association rule mining in retail", *Expert Systems with Applications*, Vol. 39 No. 16, pp. 12551-12556.

Blattberg, R.C., Kim, B.D. and Neslin, S.A. (2008), "Chapter 12", in Eliashberg, J. (Ed.), *Database Marketing: Analyzing and Managing Customers*, Springer, New York, NY.

Brin, S., Motwani, R. and Silverstein, C. (1997), "Beyond market baskets: generalizing association rules to correlations", *ACM SIGMOD Record*, Vol. 26 No. 2, pp. 265-276.

Chan, C.C.H. (2008), "Intelligent value-based customer segmentation method for campaign management: a case study of automobile retailer", *Expert Systems with Applications*, Vol. 34 No. 4, pp. 2754-2762.

Chen, C.H., Lan, G.C., Hong, T.P. and Lin, Y.K. (2013), "Mining high coherent association rules with consideration of support measure", *Expert Systems with Applications*, Vol. 40 No. 16, pp. 6531-6537.

Chen, C.L., Tseng, F.S. and Liang, T. (2010), "Mining fuzzy frequent itemsets for hierarchical document clustering", *Information Processing and Management*, Vol. 46 No. 2, pp. 193-211.

Chen, Y.L. and Weng, C.H. (2008), "Mining association rules from imprecise ordinal data", *Fuzzy Sets and Systems*, Vol. 159 No. 4, pp. 460-474.

Chiang, W.Y. (2011), "To mine association rules of customer values via a data mining procedure with improved model: an empirical case study", *Expert Systems with Applications*, Vol. 38 No. 3, pp. 1716-1722.

Hajek, P. and Stejskal, J. (2012), "Analysis of user behavior in a public library using bibliomining", *Advances in Environment, Computational Chemistry and Bioscience*, pp. 339-344.

Han, J., Cheng, H., Xin, D. and Yan, X. (2007), "Frequent pattern mining: current status and future directions", *Data Mining and Knowledge Discovery*, Vol. 15 No. 1, pp. 55-86.

Han, J., Pei, J. and Yin, Y. (2000), "Mining frequent patterns without candidate generation", *ACM SIGMOD Record*, Vol. 29 No. 2, pp. 1-12.

Han, J.W. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA.

Hsieh, N.C. (2004), "An integrated data mining and behavioral scoring model for analyzing bank customers", *Expert Systems with Applications*, Vol. 27 No. 4, pp. 623-633.

Huang, Z., Lu, X. and Duan, H. (2011), "Mining association rules to support resource allocation in business process management", *Expert Systems with Applications*, Vol. 38 No. 8, pp. 9483-9490.

Hughes, A.M. (1994), *Strategic Database Marketing*, Probus Publishing, Chicago, IL.

Hwang, S.Y. and Lim, E.P. (2002), "A data mining approach to new library book recommendations", *Digital Libraries: People, Knowledge, and Technology*, Springer, Berlin Heidelberg, pp. 229-240.

Kalakota, R. and Robinson, M. (1999), *e-Business Roadmap for Success*, 1st ed., Addison Wesley Longman, New York, NY.

Kamsu-Foguem, B., Rigal, F. and Mauget, F. (2013), "Mining association rules for the quality improvement of the production process", *Expert Systems with Applications*, Vol. 40 No. 4, pp. 1034-1045.

Kim, H.K., Im, K.H. and Park, S.C. (2010), "DSS for computer security incident response applying CBR and collaborative response", *Expert Systems with Applications*, Vol. 37 No. 1, pp. 852-870.

Ladwig, J.P. and Miller, T.D. (2013), "Are first-circulation patterns for monographs in the humanities different from the sciences?", *Library Collections, Acquisitions, and Technical Services*, Vol. 37 Nos 3/4, pp. 77-84.

Le, H.Q., Arch-Int, S., Nguyen, H.X. and Arch-Int, N. (2013), "Association rule hiding in risk management for retail supply chain collaboration", *Computers in Industry*, Vol. 64 No. 7, pp. 776-784.

Li, L.H., Lee, F.M. and Liu, W.J. (2006), "The timely product recommendation based on RFM method", *Proceedings of International Conference on Business and Information, Singapore*.

Lin, C.S. and Tang, Y.Q. (2006), "Application of incremental mining and customer's value analysis to collaborative music recommendations", *Journal of Information, Technology and Society*, Vol. 6 No. 1, pp. 1-26.

Linoff, G.S. and Berry, M.J. (2002), *Mining the Web: Transforming Customer Data into Customer Value*, John Wiley and Sons, New York, NY.

Liu, D.R. and Shih, Y.Y. (2005), "Integrating AHP and data mining for product recommendation based on customer lifetime value", *Information and Management*, Vol. 42 No. 3, pp. 387-400.

Lunfeng, G., Huan, L. and Li, Z. (2012), "The application of association rules of data mining in book-lending service", *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, Sichuan*, pp. 761-764.

Miglautsch, J.R. (2000), "Thoughts on RFM scoring", *The Journal of Database Marketing*, Vol. 8 No. 1, pp. 67-72.

Na, S.H. and Sohn, S.Y. (2011), "Forecasting changes in Korea composite stock price index (KOSPI) using association rules", *Expert Systems with Applications*, Vol. 38 No. 7, pp. 9046-9049.

Nicholson, S. (2006), "The basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services", *Information Processing and Management*, Vol. 42 No. 3, pp. 785-804.

Pei, J., Han, J., Lu, H., Nishio, S., Tang, S. and Yang, D. (2001), "H-mine: hyper-structure mining of frequent patterns in large databases", *Proceedings IEEE International Conference on Data Mining/ICDM, IEEE, Washington, DC*, pp. 441-448.

Peppard, J. (2000), "Customer relationship management (CRM) in financial services", *European Management Journal*, Vol. 18 No. 3, pp. 312-327.

Shieh, J.C. (2010), "The integration system for librarians' bibliomining", *The Electronic Library*, Vol. 28 No. 5, pp. 709-721.

Song, W., Yang, B. and Xu, Z. (2008), "Index-BitTableFI: an improved algorithm for mining frequent itemsets", *Knowledge-Based Systems*, Vol. 21 No. 6, pp. 507-513.

Song, Y. and Wei, R. (2011), "Research on application of data mining based on FP-growth algorithm for digital library", *Second International Conference on Mechanic Automation and Control Engineering (MACE), IEEE, Weihai*, pp. 1525-1528.

Weng, C.H. (2011), "Mining fuzzy specific rare itemsets for education data", *Knowledge-Based Systems*, Vol. 24 No. 5, pp. 697-708.

Wu, H.H., Chang, E.C. and Lo, C.F. (2009), "Applying RFM model and k-means method in customer value analysis of an outfitter", *Global Perspective for Competitive Enterprise, Economy and Ecology*, Springer, London, pp. 665-672.

Corresponding author
Cheng-Hsiung Weng can be contacted at: chweng@mgt.ncu.edu.tw