

## Singapore Management University Institutional Knowledge at Singapore Management University

---

Research Collection School Of Information Systems

School of Information Systems

---

12-2002

# A data mining approach to library new book recommendations


San-Yih HWANG

Ee Peng LIM

Singapore Management University, [eplim@smu.edu.sg](mailto:eplim@smu.edu.sg)

**DOI:** [https://doi.org/10.1007/3-540-36227-4\\_23](https://doi.org/10.1007/3-540-36227-4_23)

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

---

### Citation

HWANG, San-Yih and LIM, Ee Peng. A data mining approach to library new book recommendations. (2002). *Digital Libraries: People, Knowledge, and Technology: 5th International Conference on Asian Digital Libraries, ICADL 2002 Singapore, December 11–14, 2002 Proceedings*. 2555, 229-240. Research Collection School Of Information Systems.

**Available at:** [https://ink.library.smu.edu.sg/sis\\_research/1036](https://ink.library.smu.edu.sg/sis_research/1036)

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [libIR@smu.edu.sg](mailto:libIR@smu.edu.sg).

# A Data Mining Approach to New Library Book Recommendations\*

San-Yih Hwang<sup>1</sup> and Ee-Peng Lim<sup>2</sup>

<sup>1</sup>Department of Information Management  
National Sun Yat-Sen University  
Kaohsiung 80424, Taiwan  
syhwang@mis.nsysu.edu.tw

<sup>2</sup>Centre for Advanced Information Systems  
School of Computer Engineering  
Nanyang Technological University  
Nanyang Avenue, Singapore 639798  
aseplim@ntu.edu.sg

**Abstract.** In this paper, we propose a data mining approach to recommending new library books that have never been rated or borrowed by users. In our problem context, users are characterized by their demographic attributes, and concept hierarchies can be defined for some of these demographic attributes. Books are assigned to the base categories of a taxonomy. Our goal is therefore to identify the type of users interested in some specific type of books. We call such knowledge *generalized profile association rules*. In this paper, we propose a new definition of rule interestingness to prune away rules that are redundant and not useful in book recommendation. We have developed a new algorithm for efficiently discovering generalized profile association rules from a circulation database. It is noted that generalized profile association rules can be applied to other kinds of applications, including e-commerce.

## 1 Introduction

Book recommendation in the digital library context is similar to product recommendation in electronic commerce. In the past few years, we have seen the emergence of many recommendation systems that provide personalized recommendation of various types of products and services, including news (GroupLens), web pages and research articles (citeseer.nj.nec.com), books (amazon.com), albums (CDNow.com), and movies (MovieFinder.com) [SKR01]. The basic idea behind recommendation techniques is to recommend products according to the users' preferences, which are either explicitly stated by the users or implicitly inferred from previous transaction records, web logs,

---

\* This research was supported by National Science Council, ROC, under grant NSC 90-2213-E-110-022.

or cookies data. Most recommendation techniques fall into two categories, namely *content-based filtering* and *collaborative filtering* [CACM92]. The content-based filtering technique characterizes product items by a set of content features and users' interest profiles by a similar feature set. A similarity function is then defined to measure the relevance of a product item and a user's interest profile. Product items having high degrees of similarity with a user's interest profile are then recommended to the user. This method assumes that common features exist between the product items and users in order to define a reasonable similarity function. Unfortunately, this assumption does not hold for many applications where the features of user interest profiles are incompatible with the products' features. Even when such common content features exist between product items and users, the fact that only product item with content features similar to that of a user will imply no surprising recommendation can ever be found by the content-based techniques. The *collaborative filtering* (also called *social filtering*) techniques address this problem by taking into account the given user's interest profile and the profiles of other users with similar interests [SM95]. Specifically, the collaborative filtering techniques look for similarities among users by observing the ratings they assign to products. Given a target user, the nearest-neighbor users are those who are most similar in terms of product rating assignments. These users then act as the "recommendation partners" for the target user, and a collaborative filtering technique will recommend product items that appear in the transactions of these recommendation partners but not in the target user's transactions. To realize collaborative filtering, many measures have been proposed to predict the rating a person will give to an un-rated product item, based on either simple calculation on the rating matrix, e.g., correlation, cosine and regression, or a more complicated probability model, e.g., Bayesian classifier and Bayesian network [BHK98], or a combination of both [PHLG00]. It has been shown that collaborative filtering techniques yield more effective recommendations [Pazz99, MR00].

However, while the idea of recommending to a given user those products in which his peers have shown interest has demonstrated its usefulness in many applications, it has its limitations. First, it provides no or limited interpretation for a recommendation. Second, the prediction is poor if the data is sparse. In other words, unless there is a sufficient number of common items rated by users, the prediction may be unreliable. Finally, collaborative techniques fail to recommend newly introduced products that have not yet been rated by users.

**Example.** Consider the new book recommendation system of the National Sun Yat-sen University (NSYSU) library. The NSYSU library currently houses over 600,000 volumes of books and bound periodicals, and this amount is increasing at the pace of 6% per year. In other words, each month there are about 3,000 new books, which is a long list and unlikely to be browsed by any individual. As there are only 6,000 students enrolled at NSYSU, statistics from the circulation system show that the check-out figures are very low. The average number of books ever borrowed by a patron is around 30, and 75% of the library's books have never been checked out. Also, the circulation system records a wide variety of patrons' demographic information, including address, gender, birthday, degree sought, program majored, work unit, and academic status. In the library domain, each book is well classified by experts ac-

ording to some classification scheme. The NSYSU library adopts a Chinese classification scheme and the Library of Congress classification scheme to classify oriental books and western books, respectively. It is an important task to recommend a small number of new books to patrons based on their interests derived from past circulation (check out) records.

The above example shows that it may not be appropriate to infer the interests of a patron from his or her check-out records and that it is important for recommendation systems in the context of new library books to incorporate demographic information of patrons and/or genres of books. In [Pazz99], Pazzani derived demographic information of students from their home pages and used classification techniques to identify the characteristics of users who liked a particular restaurant. In [Kim01], Kim et al. used decision tree techniques to infer the characteristics of customers who liked a particular product category, as opposed to an individual item as considered in [Pazz99]. However, aggregating demographic attribute values was not explored by either study.

Our work takes into account a wide variety of information in making new book (or product) recommendations, including customers' demographic information, book (product) attribute values, customers' borrowing (purchase) records, and the concept hierarchies on demographic and book (product) attributes. Specifically, our approach starts with the discovery of generalized association rules that determine the associations between types of customers and book types. Less interesting or redundant rules are then pruned to form a concise rule set. The above two steps are time consuming and conducted off-line. In step 3, the resulting rule set is then used for on-line promotion of new books (products). Due to space constraints, this paper will only focus on the first 2 steps of mining interesting generalized association rules. To give a more general discussion, we will use the terms 'book' and 'product' interchangeably.

This paper is structured as follows. In Section 2, we will formally define the problem of mining generalized profile association rules for new product recommendations. Section 3 describes the data mining algorithm we developed. Section 4 presents our interestingness measure and the approach we used to prune redundant rules. Finally, Section 5 concludes with a summary and discussion of future research directions.

## 2 The Problem

Our data mining approach to the new product recommendation is to identify a set of strong associations between types of customers and genres of products that frequently appear in a transaction database, followed by the recommendations by using these association rules. Suppose there are  $k$  demographic attributes with domains being  $D_1, \dots, D_k$  respectively. Let  $P = \{p_1, p_2, \dots, p_r\}$  be the set of product items. An aggregation hierarchy on the  $i$ 'th demographic attribute, denoted  $H(D_i)$ , is a tree with the set of leaves being equal to  $D_i$ , and an internal node represents a demographic type. A taxonomy on  $P$ , denoted  $H(P)$ , is a tree with the set of leaves being equal to  $P$  and internal nodes indicate product categories. A link in  $H$  represents an is-a relationship.

Each transaction in the transaction database may records the identifier of a customer, the products (books) s/he has purchased (checked out), and the time of the transaction. To facilitate mining generalized profile association rules, we group transactions of the same customer and include the customer's demographic information, resulting in a new type of transaction called *demographic-product transaction*. Specifically, the demographic-product transaction of the  $i$ 'th customer is represented as a tuple  $t_i = \langle d_{i,1}, d_{i,2}, \dots, d_{i,k}, p_{i,1}, p_{i,2}, \dots, p_{i,s} \rangle$  ( $1 \leq i \leq n, k \geq 1, s \geq 1$ ), where  $d_{i,j}$  is a leaf in  $H(D_j)$  that represents the  $j$ th demographic attribute value of the  $i$ th customer, and  $p_{i,t}$  is a leaf in  $H(P)$  that represents the  $t$ th product item that the  $i$ th customer has ever purchased. In the following discussion, unless otherwise stated, when we say a transaction we actually refers to a demographic-product transaction. Since our goal is to identify the associations between customer demographics types and product categories, the demographic values and product items presented in each transaction must be converted into demographic types and product categories respectively, resulting in a so called *extended transaction* [SA95]. Here we simply include all demographic types of each demographic value and all product categories of each product item appeared in the transaction. Therefore, the  $i$ 'th transaction can be translated to the extended transaction  $t_i' = \langle d_{i,1}', d_{i,2}', \dots, d_{i,u}', p_{i,1}', p_{i,2}', \dots, p_{i,m}' \rangle$  ( $1 \leq i \leq n, u \geq 1, m \geq 1$ ), where  $d_{i,j}'$ ,  $1 \leq j \leq u$ , and  $p_{i,j}'$ ,  $1 \leq j \leq m$ , are internal nodes in  $H(D_j)$  and  $H(P)$  respectively. We say that the transaction  $t_i$  supports a demographic type  $d' = (d_1, d_2, \dots, d_l)$  if  $\{d_1, d_2, \dots, d_l\} \subset t_i'$ , where  $t_i'$  is the extended transaction of  $t_i$ . Similarly, we say that  $t_i$  supports a product category  $c$  if  $c \in t_i'$ . A *generalized profile association rule* is an implication of the form  $X \rightarrow Y$ , where  $X$  is a demographic type and  $Y$  is a product category. The rule  $X \rightarrow Y$  holds in the transaction set  $T$  with a confidence  $c\%$  if  $c$  percent of the transactions in  $T$  that support  $X$  also support  $Y$ . The rule  $X \rightarrow Y$  has support  $s\%$  in the transaction set  $T$  if  $s$  percent of the transactions in  $T$  support both  $X$  and  $Y$ . Therefore, given a set of transactions  $T$  and several demographic aggregation hierarchies  $H(D_1), H(D_2), \dots, H(D_k)$  (each one representing the generalization of one demographic attribute), and one product taxonomy  $H(P)$ , the problem of mining generalized profile association rules from transaction data is to discover all rules that have support and confidence greater than the user-specified minimum support (called  $Min_{sup}$ ) and minimum confidence (called  $Min_{conf}$ ). These rules are named *strong* rules.

### 3 Identifying Generalized Profile Association Rules

Since the goal is to identify generalized profile association rules, the itemsets that will interest us are of the following form  $\langle d_{i_1}, d_{i_2}, \dots, d_{i_l}, p \rangle$ , where  $d_{i_j} \in$  is an internal node in  $H(D_{i_j})$  and  $p$  is an internal node in  $H(P)$ . Such itemsets are called *demographic-product itemsets*. By finding large (or frequent) demographic-product itemsets, one can easily derive the corresponding generalized profile association rules. In the fol-

lowing, we present our proposed **GP-Apriori** algorithm for generating frequent itemsets.

GP-Apriori is a slight modification to the original Apriori algorithm proposed in [SA95] for mining generalized association rules. Consider the classical problem of discovering generalized frequent itemsets from market basket databases, where all items in an itemset are product items and a taxonomy for all product items is given [SA95, HF95]. It is possible to directly employ the existing techniques to discover the generalized demographic-product itemsets. In other words, a (demographic-product) transaction can be visualized as a market basket transaction by treating both demographic attribute values and product items homogeneously as ordinary items. However, this straightforward approach is inefficient and may generate many useless rules with antecedent and consequent being of the same type (products or demographic attributes). This problem of unwanted rules can be easily addressed by modifying the way candidate itemsets are generated. Let  $L_k$  denote the frequent itemsets of the form  $\langle d_{i_1}, d_{i_2}, \dots, d_{i_k}, p \rangle$ . A candidate itemset  $C_{k+1}$  is generated by joining  $L_k$  and  $L_k$  in a way similar to the Apriori candidate generation algorithm [AS94], except that the  $k$  join attributes must include one product ( $p$ ) and the other  $k-1$  demographic attribute values (from  $d_{i_1}, d_{i_2}, \dots, d_{i_k}$ ).

Specifically, this modified approach works as follows. We first extend each transaction  $t_i = \langle d_{i,1}, d_{i,2}, \dots, d_{i,k}, p_{i,1}, p_{i,2}, \dots, p_{i,s} \rangle$  ( $1 \leq i \leq n, k \geq 1, s \geq 1$ ) in  $T$  as described above. The set of extended transactions is denoted  $ET$ . After scanning the data set  $ET$ , we obtain large demographic 1-itemsets  $L_1(D)$  and large product 1-itemsets  $L_1(P)$ . If an item is not a member of  $L_1(D)$  or  $L_1(P)$ , it will not appear in any large demographic-product itemset and is therefore useless. We delete all the useless items in every transaction of  $ET$  in order to reduce its size. The set  $C_1$  of candidate 1-itemsets is defined as  $L_1(D) \times L_1(P)$ . Data set  $ET$  is scanned again to find the set  $L_1$  of large demographic-product 1-itemsets from  $C_1$ . A subsequent pass, say pass  $k$ , is composed of two steps. First, we use the above-mentioned candidate generation function to generate the set  $C_k$  of candidate itemsets by joining two large  $(k-1)$ -itemsets in  $L_{k-1}$  on the basis of their common  $k-2$  demographic attribute values and the product attribute value. Next, data set  $ET$  is scanned and the support of candidates in  $C_k$  is counted. The set  $L_k$  of large  $k$ -itemsets are itemsets in  $C_k$  with minimum support. This algorithm is called “GP-Apriori” because it is an extension of Apriori algorithm for finding Generalized Profile association rules. The pseudo-code is eliminated for brevity.

#### 4 Pruning Uninteresting Rules

From the large demographic-product itemsets derived from the GP-Apriori algorithm described in the previous section, it is trivial to derive the generalized profile association rules that satisfy both  $Min_{sup}$  and  $Min_{conf}$ . However, some of the strong generalized

profile association rules could be related to each other in either the demographic itemset part (the antecedent) or the product itemset part (the consequent), and therefore the existence of one such rule could make some others not interesting. There has been a lot of work for measuring the interestingness of association rules on items [AL99, LHM99 PT00, SK98, JS02]. A rule  $A \rightarrow C$  is said to be a sub-rule of another rule  $B \rightarrow C$  if  $A \subseteq B$ . A rule that has confidence close to one of its sub-rules is considered not interesting. Many approaches that try to identify such rules and prune them have been proposed in the literature. With respect to generalized association rules, Srikant and Agrawal defined an interestingness measure that is used to prune descendant rules given an ancestor rule<sup>1</sup> [SA95]. In their work, a rule  $R$  is interesting if and only if for every close ancestor rule  $R'$ , the support of  $R$  is at least  $\gamma$  times higher than the expected support derived from  $R'$ , or the confidence of  $R$  is at least  $\gamma$  times higher than the expected support derived from  $R'$ , where  $\gamma$  is a user-specified threshold. The intuition is that if the support and confidence of a rule can be derived from any of its ancestor rules, this rule is considered uninteresting and can be pruned.

All the previous work described above favors more general rules because they have wider scope of application, and the more specialized rules will not be picked unless they are much stronger in terms of support or confidence. Take the library circulation data mining, to which our approach has been applied, as an example. The existence of a rule such as  $R_1$ : “engineering students”  $\rightarrow$  “mathematics books” will make a more specialized rule  $R_2$ : “CS students”  $\rightarrow$  “mathematics books” not interesting unless the later is much stronger in terms of support and confidence. While this approach is useful in some cases, it falls short in identifying those ancestor rules that are strong simply because some of the descendant rules are strong. Furthermore, to recommend product items, specificity of rules should be considered. That is, if a descendant rule has adequate support and confidence, it will make a better rule for product recommendation than its ancestor rules that have slightly higher or similar support and confidence. Suppose that the following rule is strong:  $R_1$ : “CS students”  $\rightarrow$  “computer books”. Then the rule,  $R_2$ : “CS students”  $\rightarrow$  “mathematics books”, must also be strong because every computer book is classified as a mathematics book by the library classification scheme. However, although  $R_2$  is more general, this rule may not be interesting if most transactions that support  $R_2$  also support  $R_1$ . We see  $R_2$  as interesting only when many students who major in CS have also been issued non-computer mathematics books. In this case, it makes sense to recommend non-computer mathematics books to CS students. Consider another association rule  $R_3$ : “engineering students”  $\rightarrow$  “computer books”. If  $R_1$  is strong, then  $R_3$  must satisfying minimum support. Again,  $R_3$  is not interesting if most transactions that support  $R_3$  come from those supporting  $R_1$ . In contrast, we will consider  $R_3$  as interesting if a sufficient number of engineering students who are not CS majors have also been issued with computer books.

---

<sup>1</sup> As defined in [SA95], a rule  $X \rightarrow Y$  is an ancestor of another  $X' \rightarrow Y'$  if  $X' \subseteq X$  and  $Y' \subseteq Y$ . Given a set of rules, a rule  $R$  is called a close ancestor of  $R'$  if there does not exist a distinct rule  $R''$  in the set such that  $R$  is an ancestor of  $R''$  and  $R''$  is an ancestor of  $R'$ .  $R'$  is said to be a descendant of  $R$  if  $R$  is an ancestor of  $R'$ .

Based on the above observation, we develop our “interestingness” measure as follows. Let  $\Pi$  be the set of all demographic attribute types, i.e.,  $\Pi = \text{internalnodes}(H(D_1)) \cup \text{internalnodes}(H(D_2)) \cup \dots \cup \text{internalnodes}(H(D_k))$ . For a given constant  $\gamma$ ,  $0 \leq \gamma \leq 1$ , rule  $D \rightarrow p$  is called  $\gamma$ -confident if its confidence is no less than  $\gamma \cdot \text{Min}_{\text{conf}}$ . We call a rule  $R_1: D' \rightarrow p_1$ , where  $D' \subseteq \Pi$  and  $p_1 \in P$ , a D-ancestor of another rule  $R_2: D'' \rightarrow p_2$  where  $D'' \subseteq \Pi$  and  $p_2 \in P$ , if  $p_1 = p_2$  and  $\forall d_1 \in D', \exists d_2 \in D''$ , such that  $d_1$  is equal to or an ancestor of  $d_2$  in the associated demographic concept hierarchy (i.e.,  $D''$  is more specific than  $D'$ ). Similarly, we call a rule  $R_1: D' \rightarrow p_1$  a P-ancestor of another rule  $R_2: D'' \rightarrow p_2$  if  $D' = D''$  and  $p_1$  is equal to or an ancestor of  $p_2$  in the product taxonomy. Also,  $R_2$  is called a D-descendant (P-descendant) of  $R_1$  if  $R_1$  is a D-ancestor (P-ancestor) of  $R_2$ . For example, both (CS students)  $\rightarrow$  (mathematics books) and (male, engineering students)  $\rightarrow$  (mathematics books) are D-descendants of (engineering students)  $\rightarrow$  (mathematics books), and (engineering students)  $\rightarrow$  (computer books) is a P-descendant of (engineering students)  $\rightarrow$  (mathematics books).

Given a set of strong rules and a given constant  $\gamma_1$ ,  $0 \leq \gamma_1 \leq 1$ , a generalized profile association rule  $R: D \rightarrow p$  is downward-interesting if

- $R$  does not have any D-descendant or for all close D-descendants of  $R$ ,  $R_1: D' \rightarrow p, R_2: D'' \rightarrow p, \dots, R_l: D^{(l)} \rightarrow p$ ,  $D - (D' \cup D'' \cup \dots \cup D^{(l)}) \rightarrow p$  (called *D-deductive rule*) is  $\gamma_1$ -confident, and
- $R$  does not have any P-descendant or for all close P-descendants of  $R$ ,  $R_1: D \rightarrow p_1, R_2: D \rightarrow p_2, \dots, R_l: D \rightarrow p_l$ ,  $D \rightarrow p - (p_1 \cup p_2 \cup \dots \cup p_l)$  (called *P-deductive rule*) is  $\gamma_1$ -confident.

Note that in the above definition, to determine whether a rule  $R: D \rightarrow p$  is downward-interesting, we do not consider the more specialized rule  $R': D' \rightarrow p'$ , where  $D' \subset D$  and  $p' \subset p$ . This is because if  $R': D' \rightarrow p'$  is strong, so is  $D' \rightarrow p$ , a D-descendant of  $R$ . Therefore, it suffices to consider only the D-descendants and P-descendants when it comes to determining downward-interestingness. The intuition behind the downward-interestingness measure is that a more general rule will interest us only when it cannot be represented collectively by some less general rules (i.e., D-descendants or P-descendants). In other words, if the deduction of a rule and its D-descendants (P-descendants) still present sufficiently high confidence ( $\gamma_1$ -confident), then this rule should be preserved. The downward-interestingness measure favors more specialized rules, and a more general rule is selected only if it can be generally applied to the specializations of its antecedents. It is important to prune out rules that are not downward-interesting because it is misleading to apply these rules for making recommendations. For example, if the rule  $R: D \rightarrow p$  is not downward-interesting because its P-deductive rule  $D \rightarrow p - (p_1 \cup p_2 \cup \dots \cup p_l)$  is not  $\gamma_1$ -confident, it does not make sense to recommend a product of category  $p - (p_1 \cup p_2 \cup \dots \cup p_l)$  to a customer characterized by  $D$ . However, when the set of descendant rules can be indeed fully represented by a more general, downward-interesting rule, the rule and its descendant rules will be preserved. Although the existence of such descendant rules



will not affect the effectiveness of the product recommendation, the large number of rules may impact performance. Therefore we propose to combine both downward-interestingness and the measure proposed in [SA95] (we call it upward-interestingness) and define a hybrid new interestingness measure as follows:

Given a set of strong rules and a given constant  $\gamma_2, \gamma_2 \geq 1$ , a generalized profile association rule  $R: D \rightarrow p$  is interesting if

- $R$  is downward interesting.
- For each close ancestor  $R'$  of  $R$  that are downward interesting, the confidence of  $R$  is at least  $\gamma_2$  times the expected confidence based on  $R'$ .

In this definition, in addition to being downward interesting, a rule must present sufficiently high confidence with respect to each of its ancestor in order to be considered interesting. Note that the expected confidence of a rule  $D \rightarrow p$  based on an ancestor rule  $D' \rightarrow p'$  is represented as  $conf(D \rightarrow p) \cdot \frac{\sup(p)}{\sup(p')}$ , where  $\sup(p)$  is

the support of  $p$ . Also, unlike the work [SA95] of Srikant and Agrawal which considers a rule as interesting if it has higher value in either support and confidence, we focus only on confidence as the goal is to identify the association between demographics and products.

The set  $\mathcal{R}$  of all strong rules can be seen as a partial order set (POSET)  $(\mathcal{R}, <)$ , where  $r_1 < r_2, r_1, r_2 \in \mathcal{R}$  if  $r_1$  is an ancestor of  $r_2$ . The constructive definition of our interestingness measure suggests a bottom-up traversal (for identifying downward interesting rules), followed by a top-down traversal (for identifying upward interesting rules). However, the difficulties of identifying downward interesting rules lie in the computation of the confidences of D-deductive and P-deductive rules. We approximate the confidence of a D-deductive rule by using the following theoretic results:

**Lemma 1<sup>2</sup>.** Let  $D', D'', \dots, D^{(l)}$  be mutually disjoint, the confidence of  $D - (D' \cup D'' \cup \dots \cup D^{(l)}) \rightarrow p$  is  $\frac{\sup(D, p) - \sup(D', p) - \sup(D'', p) - \dots - \sup(D^{(l)}, p)}{\sup(D) - \sup(D') - \sup(D'') - \dots - \sup(D^{(l)})}$ .

**Theorem 1.** Without loss of generality, let  $D', D'', \dots, D^{(l)}, 1 \leq i \leq l$ , be mutually disjoint. Assume that  $Conf((D^{(i+1)} \cup \dots \cup D^{(l)}) - (D' \cup \dots \cup D^{(i)}) \rightarrow p) \geq \gamma_1 \cdot Min_{conf}$ . If  $D - (D' \cup D'' \cup \dots \cup D^{(l)}) \rightarrow p$  is  $\gamma_1$ -confident,

$$\frac{\sup(D, p) - \sup(D', p) - \sup(D'', p) - \dots - \sup(D^{(l)}, p)}{\sup(D) - \sup(D') - \sup(D'') - \dots - \sup(D^{(l)})} \geq \gamma_1 \cdot Min_{conf}.$$

Note that to apply Theorem 1, the equation  $Conf((D^{(i+1)} \cup \dots \cup D^{(l)}) - (D' \cup \dots \cup D^{(i)}) \rightarrow p) \geq \gamma_1 \cdot Min_{conf}$  must hold. Refer to Figure 1, since  $D_4 \rightarrow p$  and  $D_5 \rightarrow p$  both have confidences higher than  $Min_{sup}$ , it is very likely that  $Conf((D_4 \cup D_5) - (D_1 \cup D_2 \cup D_3) \rightarrow p) \geq \gamma_1 \times Min_{conf}$ , where  $\gamma_1 < 1$ .  $((D_4 \cup D_5) - (D_1 \cup D_2 \cup D_3))$  is shown in shaded area in Figure 1.)

<sup>2</sup> We have skipped the proofs of all lemmas and theorems due to space constraints.

Therefore, this equation will hold in many cases. When computing the confidence of a D-deductive rule  $r: D - (D' \cup D'' \cup \dots \cup D^{(l)}) \rightarrow p$ , we first find a (maximum) set of mutually disjoint domains  $D', D'', \dots, D^{(i)}, 1 \leq i \leq l$ , and compute the confidence of  $D - (D' \cup D'' \cup \dots \cup D^{(i)}) \rightarrow p$ . If the confidence is less than  $\gamma_1 \cdot Min_{conf}$ , we drop the rule because it is likely that  $r$  is not  $\gamma_1$ -confident.

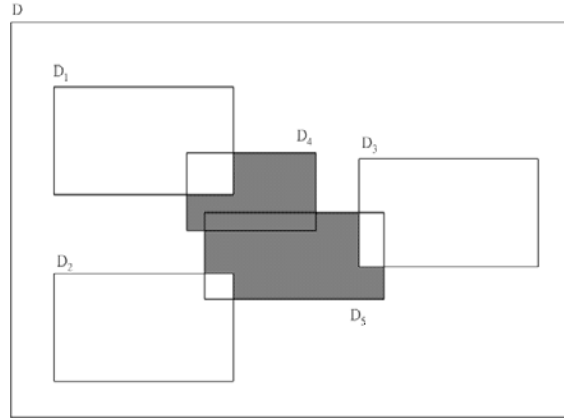


Fig. 1. Pictorial representation of  $D$  and its five sub-domains  $D_1, D_2, D_3, D_4$ , and  $D_5$ .

Now we discuss how to compute the confidence of a P-deductive rule  $r: D \rightarrow p - (p_1 \cup p_2 \cup \dots \cup p_l)$ . The transactions that support  $r$  must have included products that fall outside  $p_1 \cup p_2 \cup \dots \cup p_l$ . We say product categories  $p_j$  and  $p_i$  are siblings if they have a common parent in the respective concept hierarchy. Let  $NoSiblingTrans(D, p_i)$  denote the set of transactions that support  $(D, p_i)$  but does not support any sibling of  $p_i$ . Obviously, none of the transactions in  $NoSiblingTrans(D, p_i)$  supports  $r$ . Let  $NoSiblingSup(D, p_i)$  denote the ratio of the number of transactions in  $NoSiblingTrans(D, p_i)$  to the total number of transactions in the database. To calculate  $NoSiblingSup(D, p_i)$ , we associate a flag  $NoSibling$  on each product category of an extended transaction.  $NoSibling(p, et)$ , where  $p$  is a product category and  $et$  is an extended transaction, is equal to 1 if there exists no sibling of  $p$  in  $et$  and 0 otherwise.

Therefore,  $NoSiblingSup(D, p_i) = \frac{\sum_{et \text{ supports } (D, p_i)} NoSibling(p_i, et)}{n}$ , where  $n$  is the total number of transactions.

**Theorem 2.** If  $r: D \rightarrow p - (p_1 \cup p_2 \cup \dots \cup p_l)$  is  $\gamma_1$ -confident,

$$\frac{\sup(D, p) - \sum_{1 \leq i \leq l} NoSiblingSup(D, p_i)}{\sup(D)} \geq \gamma_1 \cdot Min_{conf}$$

$NoSiblingSum(D, p_i)$  for a demographic-product itemset  $(D, p_i)$  can be computed when counting the support for  $(D, p_i)$  by GP-Apriori described in Section 3, and such a

computation causes negligible overhead. Theorem 2 shows that  $\frac{\sup(D, p) - \sum_{1 \leq i \leq l} \text{NoSiblingSum}(D, p_i)}{\sup(D)}$  is an upper bound of the confidence of

$D \rightarrow p - (p_1 \cup p_2 \cup \dots \cup p_l)$ . Therefore, if the upper bound is less than  $\gamma_1 \cdot \text{Min}_{\text{conf}}$ , we drop the rule because it cannot be  $\gamma_1$ -confident.

For example, consider the four strong rules shown in Table 1. The bottom-up traversal starts with the rule “CS students  $\rightarrow$  computer books”, which is downward interesting because it does not have any D-descendant or P-descendant. Then we determine that the rule “Engineering students  $\rightarrow$  computer books” is not downward interesting because the D-deductive rule “Non CS-majored engineering students  $\rightarrow$  computer books” has low confidence 1/18 as shown below:

$$\begin{aligned} \frac{|E, \text{comp}| - |CS, \text{comp}|}{|E| - |CS|} &= \frac{\sup(E, \text{comp}) - \sup(CS, \text{comp})}{\sup(E) - \sup(CS)} = \frac{\sup(E, \text{comp}) - \sup(CS, \text{comp})}{\frac{\sup(E, \text{comp})}{\text{conf}(E \rightarrow \text{comp})} - \frac{\sup(CS, \text{comp})}{\text{conf}(CS \rightarrow \text{comp})}} \\ &= \frac{1/18 - 1/20}{1/6 - 1/15} = 1/18 < \text{Min}_{\text{conf}} \cdot \gamma = 20\% \end{aligned}$$

The rule “CS students  $\rightarrow$  math books” is not downward interesting either because the P-deductive rule “CS students  $\rightarrow$  (math – comp) books” has confidence no higher than 13/110 as shown below:

$$\frac{\sup(CS, \text{math}) - \text{NoSiblingSup}(CS, \text{comp})}{\sup(CS)} = \frac{4/75 - 1/22}{1/15} = 13/110 < \text{Min}_{\text{conf}} \cdot \gamma = 20\%$$

The rule “Engineering students  $\rightarrow$  math books”, however, is downward interesting because both the D-deductive rule “Non CS-majored engineering students  $\rightarrow$  math books” and the P-deductive rule “Engineering students  $\rightarrow$  (math-comp) books” have high confidences as shown below:

$$\begin{aligned} &\text{Conf}(\text{Non CS-majored engineering students} \rightarrow \text{math books}) \\ &= \frac{|E, \text{math}| - |CS, \text{math}|}{|E| - |CS|} = \frac{\sup(E, \text{math}) - \sup(CS, \text{math})}{\sup(E) - \sup(CS)} \\ &= \frac{\sup(E, \text{math}) - \sup(CS, \text{math})}{\frac{\sup(E, \text{math})}{\text{conf}(E \rightarrow \text{math})} - \frac{\sup(CS, \text{math})}{\text{conf}(CS \rightarrow \text{math})}} = \frac{1/8 - 4/75}{1/6 - 1/15} = 43/60 > \text{Min}_{\text{conf}} \cdot \gamma = 20\% \end{aligned}$$

$\text{Conf}(\text{Engineering students} \rightarrow (\text{math-comp}) \text{ books})$

$$= \frac{\sup(E, \text{math}) - \text{NoSiblingSup}(E, \text{comp})}{\sup(E)} = \frac{1/8 - 1/40}{1/6} = 3/5 > \text{Min}_{\text{conf}} \cdot \gamma = 20\%$$

In the subsequent top-down traversal (for testing upward-interestingness), no rules will be pruned. Therefore, at the end of traversal, only two rules remain: “Engineering

students→math books” and “CS students→computer books”. Note that if we simply adopt upward-interestingness, it is likely that all four rules will be preserved (because the confidences of “CS students→math books” and “Engineering students→computer books” could be higher than the estimated confidences derived from “Engineering students→math books”). As a result, ineffective recommendations, such as recommending pure math books to CS students or computer books to non-CS engineering students, will be subsequently made.

**Table 1.** Four example rules

Minconf = 25%  $\gamma$  = 0.8

Strong rules	confidence	support	NoSiblingSup
CS students→computer books	75%	1/20	1/22
CS students→math books	80%	4/75	Don't care
Engineering students→computer books	33.3%	1/18	1/40
Engineering students→math books	75%	1/8	Don't care

## 5 Conclusion

We have examined the problem of mining generalized profile association rules for recommending new books (products) that have no circulation (transaction) records and have proposed a novel approach to this problem. This approach starts with the identification of the associations between demographic types and product categories. We have developed an algorithm for this task. The obtained generalized profile associations rules are pruned by a new interestingness measure that favors special rules over general rules. We are in the process of evaluating the application of the discovered rules to recommend new books in the context of university libraries. Preliminary performance results will be shown during the conference presentation.

## References

- [AL99] Y. Aumann and Y. Lindell, “A statistical theory for quantitative association rules,” *Proc. of the 5<sup>th</sup> ACM SIGKDD Int’l. Conf. on Knowledge Discovery and Data Mining*, pp. 261-270, 1999.
- [AS94] R. Agrawal and R. Srikant, “Fast algorithm for mining association rules,” *Proc. of the 20<sup>th</sup> VLDB Conf.*, pp. 478–499, Sept. 1994.
- [BHK98] J. S. Breese, D. Heckerman and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” *Tech. Report, MSR-TR-98-12*, Microsoft Research, Oct. 1998.
- [CACM92] Special issue on information filtering, *Communications of the ACM*, 35(12), Dec. 1992.

- [HF95] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," *Proc. of the 21st VLDB Conf.*, pp. 420-431, 1995.
- [JS02] S. Jaroszewicz and D. A. Simovici, "Pruning redundant association rules using maximum entropy principle," *Proc. of 6'th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2002)*, Taipei, Taiwan, 2002.
- [Kim01] J. W. Kim, B. H. Lee, M. J. Shaw, H. L. Chang, and M. Nelson, "Application of Decision-tree Induction Techniques to Personalized Advertisements on Internet Storefronts," *International Journal of Electronic Commerce*, 5(3), pp. 45-62, 2001.
- [LHM99] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," *Proc. of the 5'th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, pp.125-134, N.Y. Aug., 1999.
- [MR00] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," *Proc. Of the 5'th ACM Conf. on Digital Libraries*, pp. 195-240, June 2000.
- [Pazz99] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, pp. 393-408, 1999.
- [PHLG00] D. Pennock, E. Horvitz, S. Lawrence and C. L. Giles, "Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach," *Proc. of the 6'th Conf. on Uncertainty in Artificial Intelligence (UAI-2000)*, pp. 473-480, 2000.
- [PT00] B. Padmanabhan and A. Tuzhilin, "Small is beautiful: discovering the minimal set of unexpected patterns," *Proc. of the 6'th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, pp.54-63, Aug. 2000.
- [SA95] R. Srikant and R. Agrawal, "Mining generalized association rules," *Proc. of the 21st VLDB Conf.*, pp. 409-419, 1995.
- [SK98] E. Suzuki and Y. Kodratoff, "Discovery of surprising exception rules based on intensity of implication," *Proc. of PKDD-98*, France, p.10-18, 1998.
- [SKR01] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-Commerce Recommendation Applications," *Data Mining and Knowledge Discovery*, 5(1), pp. 115-153, 2001.
- [SM95] U. Shardanand and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth,'" *Proc. Of the Conference on Human Factors in Computing Systems (CHI'95)*, pp. 210-217, 1995.