國立中山大學資訊管理研究所

碩士論文

運用 GP 關聯法則來做目標行銷

Targeted Advertising Based on GP-association Rules

研究生：蔡嘉文 撰

指導教授：黃三益 博士

中華民國九十三年七月

論文提要

學年度： 92

學期： 2

校院： 國立中山大學

系所： 資訊管理學系研究所

論文名稱(中)： 運用 GP 關聯法則來做目標行銷

論文名稱(英)： Targeted Advertising Based on GP-association rules

學位類別： 碩士

語文別： 英文

學號： 9142639

提要開放使用： 是

頁數： 44

研究生(中)姓： 蔡

研究生(中)名： 嘉文

研究生(英)姓： Tsai

研究生(英)名： Chai-Wen

指導教授(中)姓名： 黃三益

指導教授(英)姓名： San-Yih Hwang

關鍵字(中)： 目標行銷，GP關聯法則，產品推薦，階層概念

關鍵字(英)： Targeted Advertising, GP-association rules, Product Recommendation, Concept Hierarchies

致謝

兩年研究生涯一轉眼就過了，即將結束美好的學生生活。首先，我要感謝的就是我的指導教授-黃三益老師，國外跟台灣的時差大，老師常常都需要早起跟我們討論論文，並且提供很多想法和意見，跟著老師磨練兩年讓我成長了不少，感謝老師辛苦指導。另外還要感謝魏志平與林福林老師給予的幫助和建議，使論文能順利完成。

接著，我該感謝的是同研究室的夥伴們，錦銘、康迪、應翰、高彬，五個人聚在一起講話都沒什麼內容可言=.=，還有就是學長姊，俊凱、士民、豐文、雨廷，學弟妹，志吉、鎔碩、祐平、鈴雅，3024 的猴子家族，感謝一路上有你們的陪伴，為苦悶的研究生生活注入源源不絕的活力。

最後，最感謝的是我敬愛的父母以及家人，常常給予我鼓勵與加油，是我研究學習時的精神支柱，也是最強而有力的後盾，沒有你們就沒有今天的我。

嘉文

高雄西灣

2004.7

Abstract

Targeting a small portion of customers for advertising has long been recognized by businesses. In this thesis we proposed a novel approach to promoting products with no prior transaction records. This approach starts with discovering the GP-association rules between customer types and product genres that had occurred frequently in transaction records. Customers are characterized by demographic attributes, some of these attributes have concept hierarchies and products can be generalized through some product taxonomy. Based on GP-association rules set, we developed a comprehensive algorithm to locating a short list of prospective customers for a given promotion product. The new approach was evaluated using the patron's circulation data from OPAC system of our university library. We measured the accuracy of estimated method and the effectiveness of targeted advertising in different parameters. The result shows that our approach achieved higher accuracy and effectiveness than other methods.

中文摘要

針對特定群體之需求擬定產品及行銷策略長久以來為企業所重視，為了達到更有效率的市場區隔，本研究提出一個新的方法來推薦產品。剛開始先從交易紀錄資料中找出顧客屬性與產品類別經常一起時出現的關聯法則，其中顧客屬性與產品類別都包含了階層的概念。然後根據之前找到的關聯法則，在給定推薦產品情況下，利用演算法來找出目標顧客。方法評估上，先從校內圖書館的 OPAC System 取得讀者的借閱紀錄資料，藉由調整不同參數分別衡量預測的正確性與目標行銷的有效性，結果顯示跟其他方法相比，可以達到較高的預測的正確性與推薦有效性。

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Research Background

The effectiveness of targeting a small portion of customers for advertising has long been recognized by businesses. Traditional approach to targeted advertising is to (manually) analyze a historical database of previous transactions and the features associated with the (potential) customers, possibly with the help of some statistical tools and identify a list of customers most likely to respond to the advertisement of the product. With the advent of new technologies, it is advocated that automatic tools being used for identifying potential customers [WSJ94]. Research in this area has recently attracted considerable attention, for two main reasons. First, the amount of product/service information available to customers is ever-increasing, and hence it is desirable to help customers wade through the information to find the product/service they want. Second, understanding the needs of current and potential customers is an essential part of customer relationship management. The ability to accurately and efficiently identify the needs of customers and subsequently advertise products/services that they will find desirable opens up new possibilities to increase the customer retention, growth, and profitability of a business.

With the emergence of the Internet as a low latency, low cost channel, customer solicitation has reached unprecedented levels. Both web page banners and emails have been widely used for providing advertisements. Web page banners are the most common way of Internet advertising and the major revenue source of many on-line companies. Advertisements through emails are also very popular due to the high availability of long email lists compiled by many non-profit organizations or

individual corporations. Although advertisement on the Internet incurs low cost, its efficacy is constantly questioned. Many researches suggested that customers tend to overlook Web page banners, a phenomenon called banner blindness [BL98]. Unsolicited commercial emails, or called spam, cause legal problems as well as incurs substantial costs for both providers and email users. In fact, due to the rapid growth of spam in recent years, anti-spam laws have recently been (close to be) installed in many countries [SPAM03]. To this end, the old wisdom—targeting a few good customers while providing high incentives—has still been shown to be effective. However, the fundamental problem is senders find it difficult or expensive to target their email messages.

After identifying the characteristics of prospective customers, subsequent tasks can be performed, including designing attractive banners or email contents, locating the Web sites these customers often visit, and deciding where and when to post the banners or emails. Given a set of customer groups, a set of ads, the fondness probabilities between customer groups and ads, and other constraints, one can formulate an integer program for scheduling ads display so as to optimize the overall click-throughs or the revenue. For example, Langheinrich et al. [LNAK99] described the unintrusive targeted advertising problem, in which customer groups are characterized by the keywords entered to prevent privacy intrusion and the aim is to decide the display probability of an ad given a customer (or a keyword). This problem was expressed as a linear program and an efficient algorithm was proposed. It was then extended [Toml00, Naka02] to remedy some related problems (e.g., over-targeting problem [Toml00], multi-impressions and over-selling [Naka02]). In [AGM02], Adler et al. introduced the ad placement problem, in which each ad is associated with an access rate and a size and the goal is to choose a subset of ads for displaying in each time slot

so as to maximize the total revenue, which is proportional to the sum of the multiplications of display size and the frequency of each ad. More efficient algorithm was then given by Amiri and Manon [AM03].

1.2 Research Motivations and Objectives

However, it is non-trial to determine the fondness probabilities between customer groups and a given (to-be-promoted) product. Previous research on association rules [AIS93] [AS94] can be applied to identify the fondness relationships between customer demographics and products. However, since there are typically tens of thousands of brands in a given store, the support of association rules involving a given product is often small. These rules thus may not be identified simply because their support values are small.

The situation is even worse when the product is brand new or no prior (sales) records are available for the product. In this case, it is reasonable to look at the transaction records on other similar products, which shed the light on how customers like products of the same kind. Hwang et al. [HL02] extend the association rule discovery algorithm to identify generalized profile association rules (GP-association rules) between demographics and products, given a hierarchy for each demographic attribute and a hierarchical product classification scheme. The confidence of a GP-association rule specifies how a type of customers, identified by their demographics, likes a genre of products.

Given a set of GP-association rules, this thesis addresses the problem of identifying a set of customers who may like a given product. This problem is not as simple as it may look like. Consider the partial product taxonomy in Figure 1-1:

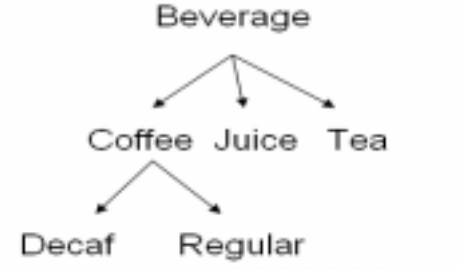

Figure 1-1: Partial product taxonomy

Suppose we want to promote a brand of regular coffee to targeted customers. Consider the confidence of an association rule: age20-50 $\rightarrow$ beverage, from which, the confidence of a more specific rule age20-50 $\rightarrow$ coffee can be estimated, which in turns serves as the basis for estimating the confidence of age20-50 $\rightarrow$ coffee.

This paper investigates product advertisement in an environment with the following features:

1. There exists a standardized, usually industry recognized, product hierarchy. This product hierarchy in some way determines the similarity between products.
2. Customer demographics exist that include an extensive set of attributes such as the name, address, gender, highest level of education, occupation, and family socioeconomic status. Some of these attributes have concept hierarchies that allow different levels of aggregation. This extensive demographic information allows the recommendation of products to certain types of customers, thereby enabling targeted marketing.
3. A transaction database that contains past customers' transaction records is available.

Such features exist in many application domains, such as membership stores that need to conduct targeted marketing for promoted products, online literature databases that

seek to recommend articles to their members, and libraries or bookstore that seek to promote fine books to their patrons.

With the above specializations in mind, we develop a data mining approach for advertising promoted products. Our approach starts with the identification of the associations between the types of customers and the product types frequently appeared in the transaction database, called GP-association rules. Based on the set of GP-association rules, we develop a mechanism for identifying the characteristics of prospective customers and the fondness probabilities given a product. This information is crutial for online promotion of products.

The discovered association rules serve as the guidelines for mapping between customers and new products. Our approach to this problem comprises three steps: (1) Develop algorithms to filter the GP association rules with respect to a product type *P*, and compute the estimated confidences and the applicable demographic domain (ADD) of the remaining rules with respect to *P*. (2) Partition the overlapping demographic attribute values into disjoint segments and estimate how each segment likes products of type *P*. (3) Promote the product of type *P* according to the information obtained in Step 2. The whole process is shown in figure 1-2:

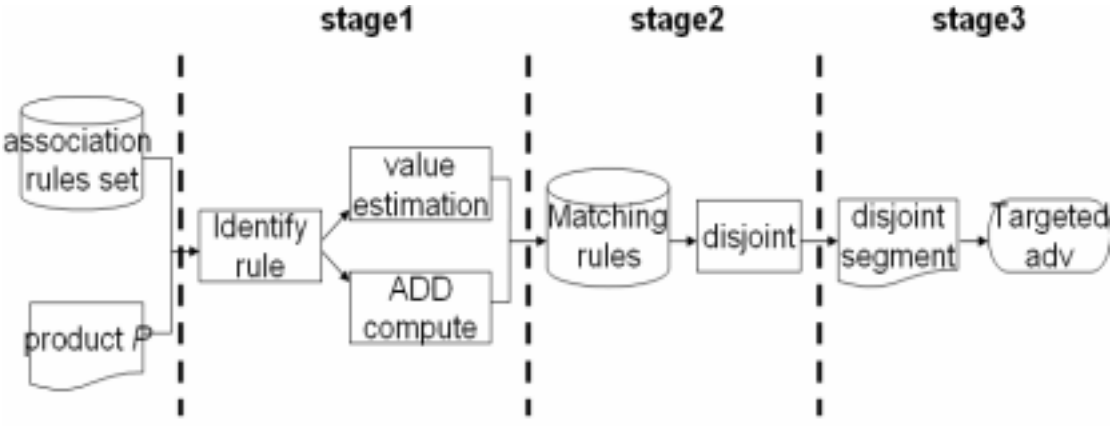

Figure 1-2: the process of target advertising

1.3 Thesis Organization

This thesis is structured as follows:

- Section 2, we review the problem of mining GP-association rules for product recommendations, the scheduling of internet banner advertisements and another approach for optimized association rules.
- Section 3 describes the algorithms of selecting a limited number of customers for promoting a given product based on the discovered GP-association rules.
- Section 4 demonstrates how to apply output of algorithm in targeted advertising.
- We evaluate the proposed method with previous research in section 5.
- Section 6 concludes with a summary and discussions of future research.

Chapter 2 Literature Review

2.1 GP-association Rules

This section describes how to locate associations between (generalized) products and (generalized) values of relevant demographic attributes. As pointed out by Agrawal and Srikant, the bottleneck of finding association rules lies in the enumeration of large item sets [AS94].

A demographic-product item set is of the form $< d_{i_1}, d_{i_2}, ..., d_{i_l}, c >$, where $d_{i_j} \in D_{i_j}$ is a type of a demographic attribute and $c \in C$ is a product genre. We say that a demographic-product item set is large if it is supported by no less than a user-specified number of transactions. The problem is how to identify all large demographic-product item sets, given a transaction database, several demographic concept hierarchies, and product taxonomy.

2.1.1 Extended transaction

Suppose there are k demographic attributes in domains $D_1, \ldots, D_k$, each of which contains demographic literals pertaining to a particular attribute. Let $P = \{p_1, p_2, ..., p_r\}$ be the set of product items and $C = \{c_1, c_2, ..., c_n\}$ be the set of product categories. Each product item in P must belong to one or more categories in C, and there exists a taxonomy on C, denoted $H(C)$, which is a directed acyclic graph with the set of vertices being C and the set of edges representing an is-a relationship. An aggregation hierarchy on the i'th demographic attribute, denoted $H(D_i)$, is a tree whose set of nodes is D_i. A link in H represents is-a relationship.

Each transaction in the transaction database may record the identifier of a customer, the products s/he has purchased, and the time of the transaction. To facilitate mining GP-association rules, we group transactions of the same customer and include the customer's demographic information, resulting in a new type of transaction called a *demographic-product transaction*. Let T be a set of demographic-product transactions, each of which is a tuple:

$t_i =< d_{i,1}, d_{i,2}, ..., d_{i,k}, p_{i,1}, p_{i,2}, ..., p_{i,s} > (1 \le i \le n, k \ge 1, s \ge 1)$, where $d_{i,j}$ is a leaf in $H(D_j)$ and represents the j'th demographic attribute value of the i'th customer, and $p_{i,j} \in P$ is the j'th product item that the i'th customer has purchased. Since the intention is to identify the associations between customer demographics and product genres, the product items presented in each transaction have to be converted into product categories. As a result, the i'th demographic-product transaction is converted as: $t_i =< d_{i,1}, d_{i,2}, ..., d_{i,k}, c_{i,1}, c_{i,2}, ..., c_{i,m} > (1 \le i \le n, k \ge 1, m \ge 1)$, 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 $c_{i,j}$ is a leaf node in $H(C)$ that represents the basic category of the j'th product item that the i'th customer has purchased.

In the following discussion, unless otherwise stated, when we refer to a transaction we actually mean a demographic-product transaction. In order to identify the associations between customer-demographic 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 appearing in the

transaction without duplication. Therefore, the *i*'th transaction can be translated to the extended transaction like this:

$t_i' = < d_{i,1}', d_{i,2}', ..., d_{i,u}', c_{i,1}', c_{i,2}', ..., c_{i,m}' > (1 \le i \le n, u \ge 1, m \ge 1)$, where $d_{i,j}'$, $1 \le j \le u$, and $c_{i,j}'$, $1 \le j \le m$, are nodes in $H(D_j)$ and $H(C)$ respectively. We say that the transaction t_i supports a demographic type $d' = (d_1, d_2, \ldots, d_l)$ if $\{d_1, d_2, \ldots, d_l\} \subset t_i'$, where t_i' is the extended transaction of t_i. Similarly, we say that t_i supports a product category c if $c \in t_i'$. GP-association rules are an implication of the form $X \rightarrow Y$, where $X \subset D_1 \cup D_2 \cup \ldots \cup D_k$ and $Y \in C$. 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 $H(D_1)$, $H(D_2)$, …, $H(D_k)$ (each one representing the generalization of one demographic attribute), and one product taxonomy $H(C)$, the problem of mining GP-association rules from transaction data involves discovering all rules that have support and confidence greater than the user-specified minimum support (called Min_{sup}) and minimum confidence (called Min_{conf}).

2.1.2 GP-Apriori algorithm

This approach is a slight modification to the algorithms proposed in [SA95] for mining generalized association rules. Consider the classical problem of discovering generalized large item sets from market-basket databases, where taxonomy on all items is assumed to exist [SA95]. We can employ the existing techniques to discover the generalized demographic-product item sets. To do so, a transaction can be visualized as a market-basket transaction by treating both demographic attribute

values and products homogeneously as ordinary items, so that associations between items can be obtained. 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 redundancy problem can be easily alleviated by modifying the way candidate item sets are generated. Let L_k denote the large item sets of the form $< d_{i_1}, d_{i_2}, ..., d_{i_k}, c >$. A candidate item set 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 (c) and the other k–1 demographic attribute values (from $d_{i_1}, d_{i_2}, ..., d_{i_k}$).

Specifically, this modified approach works as follows. We extend each transaction $t_i = < d_{i,1}, d_{i,1}, ..., d_{i,k}, c_{i,1}, c_{i,2}, ..., c_{i,m} > (1 \le i \le n, k \ge 1, m \ge 1)$ in T by adding all ancestors of $d_{i,j}$ from the concept hierarchy of the j'th demographic attribute, $1 \le j \le k$, and all ancestors of $c_{i,j}$, $1 \le j \le m$, from the product taxonomy. The set of extended transactions is denoted ET. After scanning the data set ET, we obtain large demographic one-item sets $L_1(D)$ and large product one-item sets $L_1(P)$. If an item is not a member of $L_1(D)$ or $L_1(P)$, it will not appear in any large demographic-product item set 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 one-item sets 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 one-item sets from C_1. A subsequent pass, say pass k, is composed of two steps. First, the above-mentioned candidate-generation function is used to generate the set C_k of candidate item sets by joining two large (k-1)-item sets 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-item sets are item sets in C_k with minimum support. This algorithm is called "GP-apriori" because it is an extension of the Apriori algorithm for finding GP-association rules. Its pseudo code is listed in Figure 2.

GP-apriori(T: a set of transactions, Min_{sup}: the user-specified support threshold): a set of GP-association rules

{

FOR (each transaction $t_i = < d_{i,1}, d_{i,1}, ..., d_{i,k}, c_{i,1}, c_{i,2}, ..., c_{i,m} > (1 \le i \le n, k \ge 1, m \ge 1)$)

{

Extend t_i by adding all ancestors of $d_{i,j}$ and $c_{i,j'}$, $1 \le j \le k$, $1 \le j' \le m$,

from the corresponding concept hierarchies;

Increment the count of every item in t_i;

}

$L_1(D)$= {c | c.count $\ge Min_{sup}$, c is a node in a demographic hierarchy};

$L_1(P)$= {c | c.count $\ge Min_{sup}$, c is a node in the product hierarchy};

$C_1 = L_1(D) \times L_1(P)$;

FOR (each extended transaction $t \in T$)

{

Delete any item in t that does not appear in either $L_1(D)$ or $L_1(P)$;

} // end of FOR loop

L_1 = All candidates in C_1 with minimum support;

k=2;

WHILE ($L_{k-1} \neq \varnothing$)

{

Join L_{k-1} based on the product attribute and k–2 common demographic attributes;

Put the result in C_k;

Delete all item sets $l \in C_k$ such that some k–1 demographic subset of l is not in L_{k-1};

FOR (each extended transaction $t \in T$)

{

Increment the count of all candidates in C_k that are contained in t;

} // end of FOR loop

L_k = {c | c $\in C_k$ and c.count $\ge Min_{sup}$};

$k = k + 1$;

} // end of WHILE loop

RETURN $L_1 \cup L_2 \cup \ldots \cup L_k$

}

Figure 2-1: Pseudo code for algorithm GP-apriori

2.2 Estimating Specialized Rules from Generalized Rules

Discovery of interesting association or sequential patterns from large amounts of transaction records will help marketing, decision making, and business management. Since mining association rules or sequential patterns has been a hot topic in recent research into knowledge discovery in databases [AS94] [AS95] [KMRTV94] [PCY95]. However, previous research has been focused on mining association rules at a single concept level, not consider the taxonomy hierarchy of product items.

Then some researchers brought out the idea of mining generalized association rules, in order to solve the problem of finding rules that contain product types at arbitrary levels of the concept hierarchy. [HF95] modified the single level Apriori to handle multiple level association rules by remapping the transaction database and performing the mining by a progressive deepening of the levels. Srikant and Agrawal [SA95] search for associations in given taxonomies, using support and confidence thresholds to guide the choice of level of abstraction.

To extend the work of mining single-level association rules to multiple-level ones, concept taxonomy should be provided and be used for generalizing primitive level concepts to high level ones. In general this should not be a problem because we can easily list the concept hierarchy. What deserves to be mentioned is the problem of results extracted from large amount of transactions, generalized as below:

1. Strong support is more likely to exist at high concept levels rather than at low concept levels.
2. It is unlikely to find many strong association rules at a primitive concept level.

Mining association rules at high concept level may often lead to the rules corresponding to common sense, or lead to some uninteresting attribute combinations. In order to remove uninteresting rules generated in mining process, the previous approach [PM94] quantify the usefulness or interest of a rule focused on how much the support of a rule was more than the excepted support. [SA95] also defined the interesting rules and claimed that their method can prune probably 40% to 60% redundant rules. The authors identify the interest rules by estimating the support and confidence of less general rule and see if the exist of rule can convey addition information. Consider the rule: $soft\ drink \rightarrow ice\ cream$ with 8% support and 60% confidence. If "soft drink" is a parent of "cola" and about a quarter of sales of "soft drink" are "cola", we could expect the rule $cola \rightarrow ice\ cream$ to have 2% support and 60% confidence. If the actual support and confidence for $cola \rightarrow ice\ cream$ are indeed around 2% and 60% respectively, the rule is considered redundant. The estimation of support and confidence in [SA95] is described next.

Let $\hat{X}$ and $\hat{Y}$ be the ancestors of X and Y respectively. The rules $\hat{X} \rightarrow Y$, $\hat{X} \rightarrow \hat{Y}$ or $X \rightarrow \hat{Y}$ are said to be ancestors of the rule $X \rightarrow Y$. Consider a rule $X \rightarrow Y$, and let Z = $X \cup Y$. The support of Z will be the same as the support of the rule $X \rightarrow Y$. Let $E_{\hat{Z}}[\Pr(Z)]$ denote the expected value of Pr (Z) given Pr ($\hat{Z}$), where $\hat{Z}$ is an ancestor of Z. Let Z = $\{z_1,\ldots,z_n\}$ and $\hat{Z}$ = $\{\hat{z}_1,\ldots,\hat{z}_j,z_{j+1},\ldots,z_n\}$, $1 \le j \le n$, where $\hat{z}_i$

is an ancestor of z_i. Then value of $E_{\hat{Z}}[\Pr(Z)]$ is defined as:

$$E_{\hat{Z}}[\Pr(Z)] = \frac{\Pr(z_1)}{\Pr(\hat{z}_1)} \times \ldots \times \frac{\Pr(z_j)}{\Pr(\hat{z}_j)} \times \Pr(\hat{Z}).$$

Similarly, let $E_{\hat{X} \to \hat{Y}}[\Pr(Y \mid X)]$ denote the expected confidence of the rule $X \to Y$ given the rule $\hat{X} \to \hat{Y}$. Let Y = $\{y_1,\ldots,y_n\}$ and $\hat{Y}$ = $\{\hat{y}_1,\ldots,\hat{y}_j, y_{j+1},\ldots,y_n\}$, $1 \le j \le n$, where $\hat{y}_i$ is an ancestor of y_i. Then value of $E_{\hat{X} \to \hat{Y}}[\Pr(Y \mid X)]$ is defined as:

$$E_{\hat{X} \to \hat{Y}}[\Pr(Y \mid X)] = E_{\hat{Z}}[\Pr(Z)] = \frac{\Pr(y_1)}{\Pr(\hat{y}_1)} \times \ldots \times \frac{\Pr(y_j)}{\Pr(\hat{y}_j)} \times \Pr(\hat{Y} \mid \hat{X})$$

The method mentioned above estimate the expected support and confidence from its parent rule directly. While this method can be easily used to infer support or confidence of a child rule, its accuracy remained questioned. Consider the following example: Coffee has three child products: decaf coffee, espresso, and regular coffee. If we have two strong rules R_1: age 30-50→coffee with confidence=70% and R_2: Age 30-50→decaf coffee with confidence=50%. The expected value of "age 30-50→regular coffee" can be inferred by considering both R_1 and R_2, rather than simply consulting R_1.

2.3 Scheduling of Web Advertisements

After the World-Wide Web was invented by Tim Berners-Lee in 1991, it grows at an amazing rate. Up to now, advertisement remains the single major source of revenue for most companies on web. In order for web advertisement to be effective, the research on this problem rose and conferred extensively.

The idea of optimal scheduling of web advertisements so as to maximize the number of click-through was proposed in [LNAK99]. The problem of maximize the number of click-through reduced to a transportation problem, a kind of linear programming (LP) problem, which is known to be one that can be solved efficiently by making use of special features of the problem. Now we review the LP model and see how this model can be used in targeted advertising.

We consider the attribute-driven advertisement problem and form of formalization proposed in [Naka02]. There are a fixed set of attributes such as the set of ‘time of day’ and ‘page category’. Suppose now we have the accurately estimated numbers of page views for each combination of attribute values and click-through rates of these advertisements such as shown in Table 2-1.

Time of day	Page category	Number of page views	Click through rate (%)		
			Ad 1	Ad 2	Ad 3
Evening	Travels	5000	2.0	1.5	1.0
Evening	Sports	5000	2.0	1.8	1.0
Afternoon	Travels	2500	2.0	1.8	1.5
Afternoon	Sports	2500	2.0	1.8	1.5

Table 2-1: Case of advertisement scheduling

The strategy is to select an advertisement with the highest click-through rate among the ads. Assume that each ad demands 5000 displays. The optimal strategy would be to select ad 1 for the first 5000 page views, followed by ad 2 for the next 5000 page views, and ended by ad 3 for the last 5000 page views, because (the click-through rate of ad 1) > (the click-through rate of ad 2) > (the click-through rate of ad 3) holds for all combinations of attribute values.

The problem is formalized as follows. Let i (=1, …,n) denote combinations of attribute values and j (=1, …,m) denote advertisements. The variable k_i, h_j and $c_{i,j}$ are defined as follows:

k_i: Estimated rate of page views for combination i of attribute values

h_j: Desired display rate for advertisement j

$c_{i,j}$: Estimated click-through probability for advertisement j and combination i of attribute values

The estimation of k_i is calculated based on the number of recent page views for combination i of attribute values. Desired display rate h_j is calculated from the contracted number of impressions by subtracting the number of actual impressions so far and considering the remaining contracted period. The estimation of $c_{i,j}$ is calculated based on the previous click-through rate for combination i of attribute values and advertisement j.

Let $d_{i,j}$ denote the display probability of advertisement j for combination i. The problem of optimal scheduling of web ads can be formalized as following:

Maximize $\sum_{i=1}^{n}\sum_{j=1}^{m} c_{i,j} k_i d_{i,j}$

Subject to

$$\sum_{i=1}^{n} k_i d_{i,j} = h_j,\ j = 1,...,m$$

$$\sum_{j=1}^{m} d_{i,j} = 1,\ i = 1,...,n$$

$$d_{i,j} \geq 0,\ i = 1,...,n,\ j = 1,...,m$$

The question of web advertisements scheduling mentioned above is reduced to a named transportation problem. Using the known algorithm it can be easily handle.

Similar to the LP model applied to web ads scheduling, the problem of targeted advertising can be formalize as transportation ones. When given a set of promotion product, the first job is to identify matching rules involves estimated confidence and disjoint demographic attribute with respect to each product. Then we can describe the LP model and solve it based on these data. Detailed illustration will be introduced in Section 4.

2.4 Optimized Association Rules

Association rules provide a useful mechanism for discovering co-occurrence items in large amounts of transactions. Optimized association rules are permitted to contain un-instantiated attributes and the problem is to determine instantiations such that either the support or confidence of the rule is maximized.

The optimized association problem, motivate by applications in marketing and advertising, was introduced in [FTTM96]. An association rule is of the form: $(D_1 \in [d_1, d_2]) \wedge C_1 \rightarrow C_2$, where D_1 is a numeric attribute, d_1 and d_2 are

un-instantiated variables, and C_1 and C_2 contain only instantiated variable. The authors proposed algorithms to determine the range of un-instantiated variable that maximize confidence or support of the rule.

For example, consider a table in a digital camera store database that contains selling information. The attributes of the table are gender, age, and products. Suppose the store is interested in offering a promotion to male customers who buy brand of Nikon. In this situation the age of male customer may be important, if the store can find male customer whose age between a_1 and a_2 are more likely to buy product Nikon through previous data, promotion activity will be effective.

The above case can be formulated as the following association rule $(age \in [a_1, a_2]) \wedge (gender = male) \rightarrow (product = Nikon)$, which results in two detailed problems. First, with a minimum support we try to solve the optimized confidence problem for the percentage of Nikon bought by male customers whose age within the interval is maximized. Second, with a minimum confidence we try to find which age interval that make the rule has maximum support.

The algorithm proposed in [FTTM96] has limitations in that it can only determine the single numeric attribute. A single attribute may not be adequate to describe the trend or property of data in complicated application. For example, suppose an airplane company is interested in doing promotion to customer in Taiwan flying to Australia. For this purpose the company needs to identify several period during which a sizable number are made. In this way we want to find the association rules like:

$(date \in [d_1, e_1] \vee date \in [d_2, e_2] \vee \mathrm{K} \vee date \in [d_k, e_k]) \wedge (start = Taiwan) \rightarrow (dest = NewYork)$

The power of expression can be strengthened by containing more than one un-instantiated attribute, and permitted the attribute to be either numeric or categorical form. The new algorithm proposed in [RS98] achieves the objective to handle such a problem and give us some idea.

[RS98] synthesize the related work in the past and generalize the optimized association rules problem in three ways: (1) association rules are permitted to contain disjunctions over un-instantiated attributes, (2) association rules are allowed to contain an arbitrary number of un-instantiated attributes, (3) un-instantiated attributes can be either categorical or numeric. Then the authors both present the algorithm to exploring the search space and use branch and bound techniques to prune the search space effectively.

The optimized association rule requires optimal instantiations to be computed for a un-instantiated association rule which has the form $U \wedge C_1 \rightarrow C_2$, where U is a conjunction of *m* un-instantiated atomic conditions over *m* distinct attributes, and C_1 and C_2 are arbitrary instantiated conditions. U_i denote an instantiation of U and is obtained by replacing variables in U with values. Under the above notation of association rules the problem of optimized association rules is how to determine the un-instantiated variables a_1 and a_2 for each of the following two cases:

- Optimized Confidence Problem: Given an un-instantiated rule R: $U \wedge C_1 \rightarrow C_2$, determine non-overlapping instantiations $U_1,\ldots,U_k$ of U such that $\sup(R) \geq$ *minsup* and conf(R) is maximized, where R is the rule $(U_1 \vee \mathrm{K} \vee U_k) \wedge C_1 \rightarrow C_2$.

- Optimized Support Problem: Given an un-instantiated rule R: $U \wedge C_1 \rightarrow C_2$,

determine non-overlapping instantiations $U_1,\ldots,U_k$ of U such that conf(R)$\geq$ *minconf* and sup(R) is maximized, where R is the rule $(U_1 \vee \mathrm{K} \vee U_k) \wedge C_1 \rightarrow C_2$.

The concept of mining optimized association rules is similar to the purpose of ours. Given a promotion product we want to identify a set of strong rules that maximize the total confidence in the hypothesis of a small group of target customers. When solving the optimized confidence or support problem, the product type must be determined in advanced and then finding the combination of un-instantiated attribute. The two problems are similar in that they both fix the product to be promoted. But in our problem, the demographic and product attribute both contain concept hierarchies, the method mentioned above cannot directly handle this problem. The appropriate algorithm will be proposed in the next section.

Chapter 3 Targeted Advertising

3.1 Problem Statement

Given a product, our first task is to identify a small set of customers for effective promotion by consulting the set of discovered GP-association rules. Targeted advertising is really a tradeoff between expense and effectiveness. The goal is to achieve a highly effective promotion by targeting only on a small number of customers. However, for a given product *p*, there could be a large number of rules with consequents matching *p*, and the antecedents of some rules could be very broad. Thus, it is inappropriate to simply choose rules with higher confidences and subsequently target customers whose demographics match these rules' antecedents.

For example consider the rule: R_1: Age 30-50→Coffee with confidence=70% and R_2: Age 30-50→Decaf Coffee with confidence=50%. When we need to promote a brand of decaffeinated coffee, it makes perfect sense to make use of R_2 and to assume that half of the customers of ages between 30 and 50 like decaf coffee. However, when it comes to the target identification of a regular coffee (with caffeine) product, the importance of R_1 should not be overemphasized, because many transactions that support R_1 could come from those supporting R_2. A more useful rule in such a case is R_1': "Age 30~50→Non-decaf Coffee". Unfortunately, the confidence of R_1' could be unknown and can only be estimated from the set of related strong rules. In our example, we can conclude that the confidence of R_1' is at least 20%[1]. Techniques for more accurately approximating the confidence of R_1' have to be developed.

[1] Note that some transactions could support both R_1 and R_2 (i.e., some customers like all kinds of coffee).

3.2 Value Estimation

We call a rule R_1: $D' \to P_1$ a P-ancestor of another rule R_2: $D'' \to P_2$ if $D' = D''$ and P_1 is equal to or an ancestor of P_2 in the product hierarchy. Conversely, R_2 is called a P-descendant of R_1. Based on the P-descendant relation (which is a partial order), we can form a lattice on the set of strong rules. Let the children of P in the product hierarchy be $P_1, P_2, \ldots, P_{m_P}$. Suppose a rule $R: D \to P$ has k immediate P-descendants $R_1: D \to P_1$, $R_2: D \to P_2$, …, $R_k: D \to P_k$ in the lattice. We would like to derive the way of estimating the confidence of $R_j: D \to P_j, k < j \le m_P$ so that it can be consulted when we need to recommend a product $p \in P_j$. For a given product type P, assume that in average each transaction that supports P also supports β_P of its children, where β_P, $\beta_P \ge 1$, is called the overlapping factor of P. Note that β_P can be computed using the support of P ($Sup(P)$) and those of its children ($Sup(P_i)$) by the following equation:

$$\beta_p = \frac{\sum_{i=1}^{m_p} Sup(P_i)}{Sup(P)}$$

Therefore, we have the following estimation:

$$Conf(R) \cdot \beta_P = \sum_{i=1..m_P} Conf(R_i) = \sum_{i=1}^{k} Conf(R_i) + \sum_{j=k+1}^{m_P} Conf(R_j)$$

Equivalently,

$$\sum_{j=k+1}^{m_P} Conf(R_j) = Conf(R) \cdot \beta_P - \sum_{i=1}^{k} Conf(R_i)$$

Assuming that the confidence of a rule $R_j: D \to P_j, k < j \le m_P$ is proportional to the

support of P_j, $Sup(P_j)$, we can estimate the confidence of R_j as follows:

$$Conf(R_j) \approx (Conf(R) \cdot \beta_P - \sum_{i=1}^{k} Conf(R_i)) \cdot \frac{Sup(P_j)}{\sum_{i=k+1}^{m_P} Sup(P_i)}, k < j \le m_P .$$

We define a value function of R: $_{D \to P}$ with respect to a child P_i of P, $1<i\le m_P$, denoted $Value(R, P_i)$, as follows.

$$Value(R, P_i) = \begin{cases} 0 \quad if\ 1 \le i \le k \\ (Conf(R) \cdot \beta_P - \sum_{i=1}^{k} Conf(R_i)) \cdot \frac{Sup(P_j)}{\sum_{i=k+1}^{m_P} Sup(P_i)}, if\ k < i \le m_P \end{cases},$$

Note that $Value(R, P_i)$ is defined as 0 for $1\le i\le k$ because the existence of a more specific strong rule $_{D \to P_i}$ makes the rule R: $_{D \to P}$ of no value when it comes to promoting a product $p\in P_i$. Here β_P can be empirically computed as the average number of children of P appeared in a transaction that supports P. Such a computation can be conducted when computing L_1 (P) in GP-Apriori, and the overhead is negligible.

The value function $Value(R, P_{ij})$ of R: $_{D \to P}$ with respect to a grandchild P_{ij} of P (i.e., P_{ij} is the j'th child of the i'th child P_i of P in the product hierarchy) can be similarly derived:

$$Value(R, P_{ij}) = \begin{cases} 0 \quad if\ 1 \le i \le k \\ Value(R, P_i) \cdot \beta_{P_i} \cdot \frac{Sup(P_{ij})}{\sum_{k=1}^{m_{P_i}} Sup(P_{ik})}, if\ k < i \le m_P \end{cases},$$

, where m_{P_i} and β_{P_i} are respectively the number of children and the overlapping factor of P_i, the parent of P_{ij}, in the product hierarchy.

The value of a rule with respect to a product type at lower level can be similarly induced and is not elaborated here. For completion, we also define the function *Value*(*R*, *P*) of *R*: $D \to P$ with respect to *P* as simply the confidence of *R*, *Conf*(*R*).

In our previous example, assume that Coffee has three children in the product hierarchy: Decaf Coffee, Espresso, and Regular and that the support of each child is the same. Besides, let the overlapping factor β_{Coffee} be 1.5. Thus, the value of R_1: Age 30-50 $\Rightarrow$ Coffee with respect to Regular Coffee or Espresso is $\frac{70\% \cdot 1.5 - 50\%}{2} = 27.5\%$, while the value of R_1: Age 30-50 $\Rightarrow$ Coffee with respect to Decaf Coffee is 0 (because of the existence of the more specific rule R_2: Age 30-50 $\Rightarrow$ Decaf Coffee).

The following lists the pseudo code of computing the value of a GP association rule *R* with respect to a product type *P*.

```
// Let Ancestors(P) denote the set of P and all ancestors of P in the product hierarchy,
// R.prod denote the consequent (product type) of R, and
// R.demo denote the antecedent (demographic type) of R
// ProdChild(R, P) returns a rule with antecedent being R.demo and consequent being a
   child and an ancestor of R.prod and P respectively.
// β(P) returns the β_P.

Value(R: a GP-association rule, P: a product type, GP: a set of GP-association rules):
a numeric value
01  {
02      IF R.prod ∉ Ancestors(P) RETURN NULL // error
03      IF R.prod=P RETURN R.conf;
04
```

```
05      R" = ProdChild(R, P);
06      IF R"∈GP RETURN 0;
07      PDecendants = {R'| R'.demo = R.demo, R'.prod ∈ Child(R.prod), R'∈GP};
08      ValueChild=R.conf *β(R.prod);
09
10      FOR each R' in PDecendants DO
11          ValueChild= ValueChild–R'.conf;
12      UncoveredChildren = {P'|P' ∈ Child(R.prod) and ∀R' ∈ PDecendants,
                                                          R'.prod≠P'}
13      TotalSupport = 0;
14
15      FOR each P' in UncoveredChildren DO
16          TotalSupport = TotalSupport+Support(P')
17      ValueChild= ValueChild*Support(R".prod)/TotalSupport;
18      R".conf = ValueChild;
19      RETURN Value(R", P, GP);
20  }
```

Figure 3-1: Pseudo code of value estimation

3.3 Identifying Applicable Demographics of a GP-association Rule

When we are given a product *p*, we first locate the leaf product category *Lp* in the product hierarchy such that $p \in Lp$. We then identify the set of rules whose consequents are equal to or ancestors of L_p in the product hierarchy. We call the set of rules *matching rule set* of *p*, denoted M_p. Now we can define a partial order on M_p based on the demographic parts (i.e., the antecedents) of the rules. Specifically, we call a rule R_1: $D' \rightarrow P_1$ a D-ancestor of another rule R_2: $D'' \rightarrow P_2$ if $P_1=P_2$ and $D' \supset D''$. Conversely, R_2 is called a D-descendant of R_1. Based on the P-descendant relation, we can define a lattice on the matching rule set. Consider a rule R: $D \rightarrow P \in M_p$ with *k*

immediate D-descendants: $D_1 \to P$, $D_2 \to P$,…, $D_k \to P$ in the lattice, we define the *applicable demographic domain* of *R* (with respect to *p*), denoted *ADD*(*R*, *p*), as $D - D_1 - D_2 - ... - D_k$. *ADD*(*R*, *p*) represents the set of customers that are chosen as target for advertising *p* when *R* is selected. Note that demographic domain D_i, $1 \le i \le k$, is excluded from *ADD*(*R*, *p*) because, when deciding the degree to which a product $p \in P$ should be promoted to customers with demographics D_i, one should consult the more specific rule $D_i \to P$, and *R* is consulted only when we need to decide whether to target a customer in $D - D_1 - D_2 - ... - D_k$ for promoting *p*. The following lists the pseudo code of computing the applicable demographic domain of *R*:

ADD(*R*: a GP-association rule, *GP*: a set of GP-association rules): a demographic type

```
01  {
02      DDecendants = {R'| R'. demo = Child(R.demo), R'.prod ∈ R.prod, ,R'∈GP};
03      RETURN R.demo− U_{R'∈DDecendants} R'.demo ;
04  }
```

Figure 3-2: Pseudo code of ADD computation

Example: Suppose there are four GP-association rules as listed in Table 3-1:

rule	Demographic	product	confidence
R_1	Age 20-50	Coffee	70%
R_2	Age 20-50	Decaf Coffee	50%
R_3	Age 20-30	Coffee	60%
R_4	Male, Age 20-50	Coffee	60%
R_5	Age 20-30	Regular Coffee	40%

Table 3-1: Example GP-association rules

Assume that coffee has three children in the product hierarchy: Regular Coffee, Espresso, Decaf Coffee, and their supports are listed in Table 3-2:

Product type *P*	*Sup*(*P*)
Regular Coffee	0.3
Espresso	0.1
Decaf Coffee	0.2

Table 3-2: Example supports of various brands of coffee

Note R_3 and R_4 are D-descendant of R_1, while R_2 is a P-descendant of R_1. Suppose we would like to promote a brand of regular coffee. The overlapping factor β_{Coffee} is 1.3. Obviously, the matching rule set is {R_1, R_3, R_4, R_5}. *Value* (R_1, Regular Coffee) = $(70\% \cdot 1.3 - 50\%)\frac{0.3}{0.3+0.1} \approx 30.3\%$ and *ADD*(R_1) = "female, Age 30-50" (i.e., {Age 20-50}−{Age 20-30}-{ Male, Age 20-50}). *Value*(R_3, Regular Coffee) =0 because a P-descendant R_5 exists. Further, *Value*(R_4, Regular Coffee) $= 60\% \cdot 1.3 \cdot \frac{0.3}{0.3+0.1+0.2} = 39\%$ and *Value*(R_5, Regular Coffee) = 40%. The *ADD* () and *Value* () of each rule in the matching rule set is summarized in Table 3-3:

Rule *R*	*ADD*(*R*)	*Value*(*R*, Regular Coffee)
R_1	Female, Age 30-50	30.3%
R_3	Age 20-30	0
R_4	Male, Age 20-50	39%
R_5	Age 20-30	40%

Table 3-3: Applicable demographic domain and estimated value

Of course, R_3 can be eliminated because it is of no value in regard to the promotion of *p*.

3.4 Identifying Disjoint Demographic Segments

Now we are given a product p and the matching rule set M_P, where each rule R in M_P is associated with a non-zero value $Value(R, L_P)$ and an applicable demographic domain $ADD(R)$. However, the applicable demographic domains in M_P may not be mutually disjoint. Consider our previous example, it is clear $ADD(R_4)$ and $ADD(R_5)$ are not disjoint, and it is not clear what value to be assigned to male customers of ages between 20 and 30. To solve this problem, we fragment the domain covered by M_P into a set of mutually disjoint segment. The value of each segment s, denoted $Value(s, p)$ is the average of $Value(R_i, L_P)$, where $s \subseteq ADD(R_i)$ and $R_i \in M_P$. In our above example, four segments are identified and their values are shown in Table 3-4:

segment	$ADD(s)$	$Value(s, p)$
s_1	Female, Age 30-50	30.3%
s_2	Male, Age 30-50	39%
s_3	Female, Age 20-30	40%
s_4	Male, Age 20-30	39.5%

Table 3-4: disjoint segment

In the following we show a general approach for identify a set of disjoint segments from a set of (possibly overlapping) ADDs:

// ***Graph*** (M_p) returns adjacency matrix that denote relation of rules in M_p
// ***EnumerateClique*** (G) returns a list of all cliques of graph G
// S_p denote set of disjoint segments

```
Disjoint_seg(M_p)
01  {
02      S_p = ∅
03      G = Graph (M_p)
04      Clique = EnumerateClique (G)
05      FOR each clique C in Clique DO {
06          Demo = ALL;
07          FOR each rule r in C DO {
08              Demo = Demo ∧ r.demo
09          }
10          S_p = S_p ∪ Demo;
11      }
```

Figure 3-3: Pseudo code of identifying disjoint demographic segments

Chapter 4 Identifying Targeted Customers

Given a product *p*, the next task is to identify a set of *N* targeted customers, where *N* is specified by the user according to the budget and other factors. Let S_p be the set of disjoint segments with respect to *p*. The problem of selecting the set of targeted customers can be formulated as the following linear program:

$$\text{Maximize} \quad \sum_{s \in S_P} Value(s, p) \cdot X_i \cdot |ADD(s)|$$

Subject to

$$\sum_{s \in S_P} X_i \cdot |ADD(s)| \leq N;$$
$$0 \leq X_i \leq 1.$$

The variable X_i denotes the ratio of customers in *ADD*(*s*) to be added to the target set for advertising *p*. Obvious, the above linear programming problem is a special case of the fractional 0-1 knapsack problem and can be efficiently solved by a greedy algorithm which incrementally selects the segment with the highest value until *N* customers are chosen. The pseudo code of this algorithm *Target-adv* is shown in Figure 4-1.

Value-based-targeted-adv(*p*: a product, *N*: number of advertising customers, *GP*: a set of GP-association rules): set of customers
{
S = ∅; //set of targeted customers
NT = 0; // number of accumulated targeted customers
M_p = {r: r is a rule in *GP* with consequent matching *p*};
S_p = the set of disjoint segments formed by the antecedents of rules in M_p;
FOR (each segment $s \in S_p$ listed in descending order of *Value*(*s p*)) **DO**

```
{
    S' ={c: c is a customer with demographics ADD(s)};
    IF (|S'|+|S| ≥ N) {
        Randomly choose N-|S| customers with demographics ADD(s) and add
    them to S;
        BREAK;
    }
     S = S'∪S;
}
RETURN S;
} // end of Target-adv
```

Figure 4-1: Pseudo code of selecting customers for advertising given a product p

The problem of selecting targets for advertising products can be further complicated by considering more constraints. For example, to advertise a set of products as web banners, we can view the set of products as a set of ads and the set of user segments as a set of user groups. More precisely, let the set of user segments be $\{s_1, s_2, \ldots, s_m\}$ and the set of products be $\{p_1, p_2, \ldots p_n\}$. v_{ij} denotes the value of s_i with respect to L_{p_j} as computed above, and X_{ij} denotes the display probability of p_j given a user segment s_i. Also, the desired display probability of p_j, determined based on its advertisement expense, is h_j, $1 \le j \le n$, and the probability that users in segment s_i show up is k_i, $1 \le i \le m$. The model can be formulated as follows:

Maximize $\sum_{i=1}^{m}\sum_{j=1}^{n} v_{ij} \cdot k_i \cdot X_{ij}$

Subject to

$$\sum_{j=1}^{n} X_{ij} = 1, 1 \le i \le m;$$

$$\sum_{i=1}^{m} k_i \cdot X_{ij} = h_j, 1 \le j \le n;$$

$$0 \le X_{ij} \le 1.$$

Again the above formula is a linear program, though in a much larger scale. Various techniques have been proposed for efficiently solving this problem [LNAK99, Toml00 Naka02]. Concise discussions of these techniques can be found in Chapter 2. Our main contribution here is on the identification of the way for computing v_{ij} from the set of discovered GP-association rules such that the existing (e.g., banner advertising) techniques can be applied, rather than proposing yet another efficient algorithm for solving the banner ads allocation problem.

Chapter 5 Performance Evaluations

In this section, we evaluated our targeted advertising algorithm using real world data. The data was collected from circulation system of the National Sun Yat-sen University (NSYSU) library. The circulation system records patron's demographic attribute and corresponding loan books. The demographic information of patron includes identification number, gender, address, grade, unit, and program majored. The circulation data was collected between May 1, 2003 and May 1, 2004. First, we measure the accuracy of estimated confidence with varied parameters in Section 5.1. Then we compare the recommendation accuracy of our method and the existing ones in Section 5.2 (e.g., minimum support / confidence for GP-Apriori and beta value for target advertising).

We use the circulation data of NSYSU library and apply our algorithm on advertising library books. First we accumulated the data from online public access catalog (OPAC) system from year 2003 May 1 to 2004 May 1. Over the one-year period, we accumulated 62568 book transactions, which belong to 5631 patrons. The OPAC records several demographic attributes of patrons and also contains a number of attributes about books, including call number, ISBN, subjects (keywords), authors, and location.

First we identified the patron's attributes that are highly related to books they checked out. We chose the identified number, unit, program majored, and degree of patron. Only degree and program majored used in the next step for finding GP-association rules as recommended by the experienced librarians at NSYSU library.

Generating GP-association rules

The concept hierarchy of degree and program is formulated as a framework of three levels. There are 8 nodes in the degree hierarchy and 38 nodes in the program hierarchy. Part of the concept hierarchies are shown in Figure 5-1:

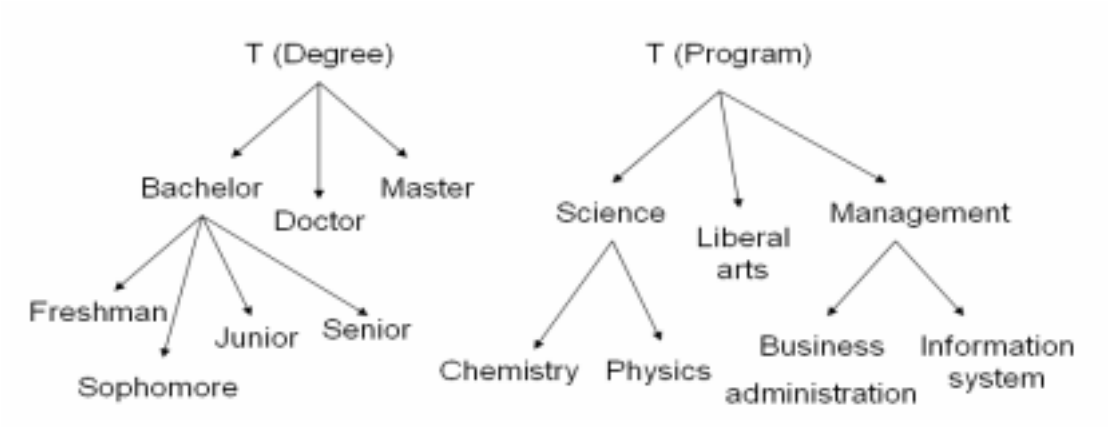

Figure 5-1: Partial concept hierarchies of (a) degree and (b) program

We adopted the first three levels of Chinese classification scheme for the book hierarchy of Chinese books and the first two levels of Congress classification scheme for the book hierarchy of Western books. The numbers of leaf nodes in Chinese book hierarchy and Western book hierarchy are 1,000 and 526 respectively. Then we discovered GP-association rules from the training data set using GP-Apriori. Finally we applied the approach described in Section 3 and 4 to choose a set of targeted patrons for advertising promoted book in the test data set.

5.1 Accuracy of Estimated Confidence

Comparing with previous work

The work proposed in [SA95] for estimating the confidence of a more specific rule from a generalized rule, as reviewed in Section 2.2, serves as the benchmark of our experiments.

In the process of generating GP-association rules, we set the minimum support ad minimum confidence at 0.0025[2] and 0.1 respectively. We first applied GP-Apriori and identified 718 GP-association rules. Then we randomly selected ten book types from the leaves of the experimental book classification scheme and identified the set of matching rule *Mp* for each book type *P*. For each matching rule, we calculated ADD (applicable domain demographic) and value. Table 5-1 shows the number of matching rules for each tested product type.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
50	53	52	30	50	54	12	44	51	18

Table 5-1: Number of matching rules for the ten randomly selected book

The confidence estimation method introduced in [SA95] is used to calculate the estimated confidence of each matching rule with respect to its ADD and *Lp*. Finally we scan the database to compute the real confidence for each matching rule. The average error of an estimation method (either ours or that reported in [SA95]) is defined as the absolute difference between the estimated confidence (or value) and the real confidence. The result of average errors of the ten products is shown in Figure 5-2. As can be seen, our method constantly incurs smaller errors compared to the

[2] The setting of 0.0025 for Min_{sup} mandates that each strong GP-association rule to be supported by no les than 10 transactions. The effect of Min_{conf} will be shown in Subsection.5.1.

method proposed in [SA95].

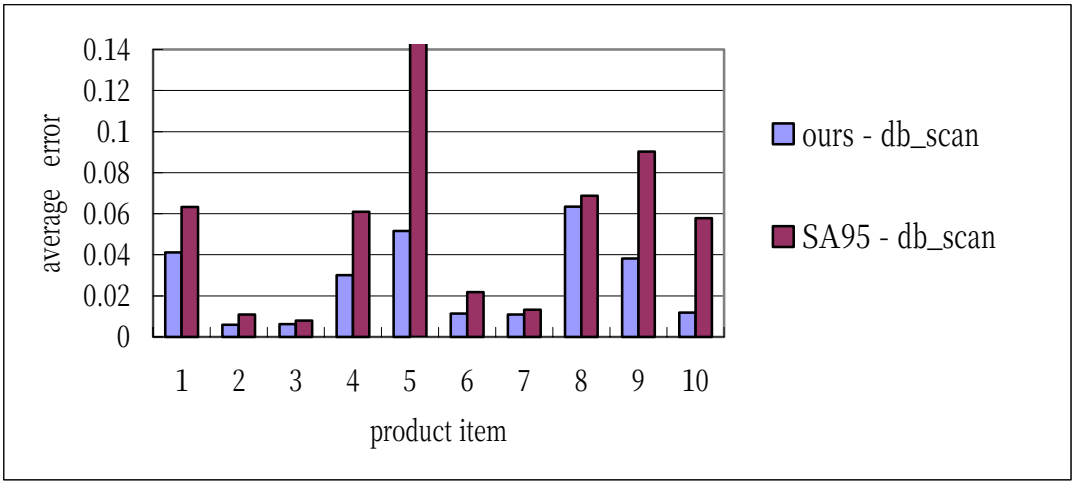

Figure 5-2: Average errors of our method and [SA95] under 10 randomly selected products

Effects of various parameters

In this experiment we evaluated the accuracy of the method proposed in Chapter 3 under various parameter settings. Recall that our method, when given a set of GP-association rules and a promotion product type *Lp*, identifies a set of matching rules and for each matching rule computes its value and applicable demographic domain (ADD), where the value is nothing but the estimate confidence of the matching rule ADD→*Lp*. To evaluate our method, we calculate the confidence of matching rule by scanning the entire transaction records and the method proposed in [SA95]. We will look at the differences between the real and estimated confidences with our approach under various values of minimum confidence.

Figure 5-3 shows the average numbers of GP-association rules at various Min_{conf} values while fixing Min_{sup} at 0.25%. As expected, the total number of GP-association rules decreases with the increase of Min_{conf}.

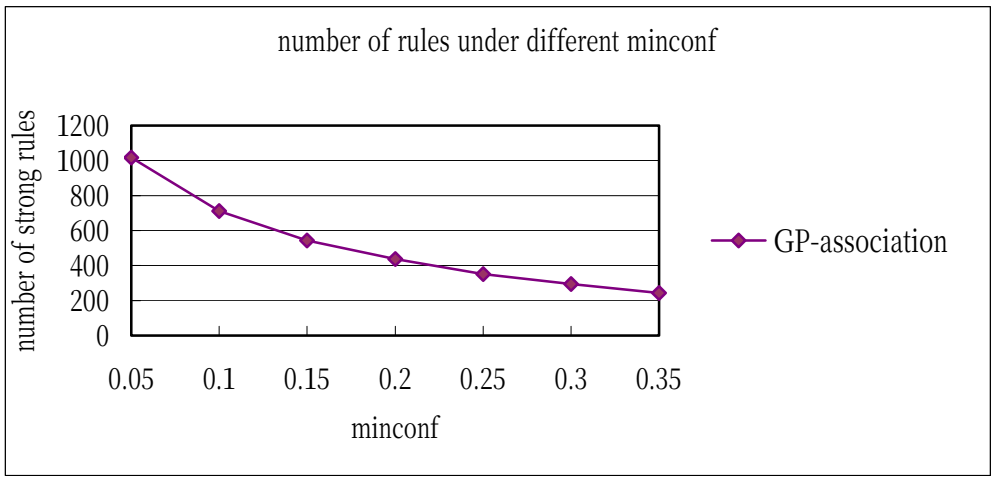

Figure 5-3: Number of strong rules (0.05 ~ 0.35)

We then compute the average errors of our targeted advertising approach and method in [SA95] under different minimum confidence (from 0.05 to 0.35). For a given Min_{conf}, we first selected ten book categories randomly and computed the their errors. The average errors are shown in Figure 5-4. As can be seen, in either method, the average error gradually increased with the increase of Min_{conf}. This is because higher Min_{conf} results in fewer GP-assocation rules, which make the estimation less accurate. Besides, the value estimated by our targeted advertising method is constantly more accurate under different Min_{conf}.

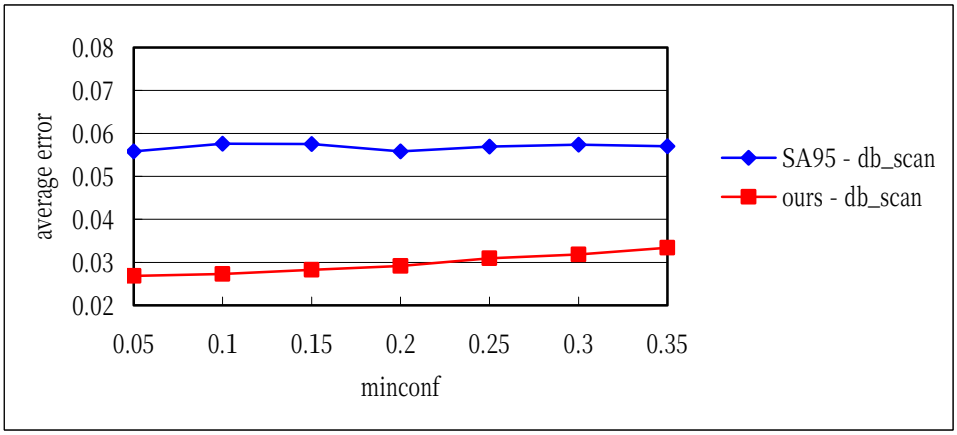

Figure 5-4: Average errors of the two value estimation methods under various Min_{conf}

Computation time of various methods

Now this experiment calculated the execution time of above three methods. We also selected ten products from book category randomly and measured the execution time of each approach, then average the record of each experiment. Figure 5-5 list the computation times for our method, the method reported in [SA95], and the database scan method. As can be seen the method of scanning database spent considerable time and the SA95 cost least. Our method cost a little more than method in SA95, but noticeable fast than scanning database.

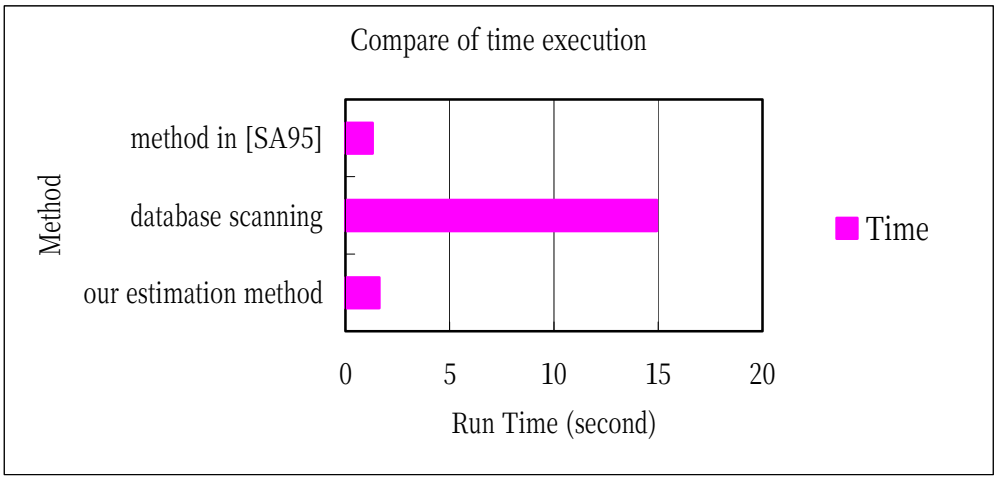

Figure 5-5: The computation time of our method, the method reported in [SA95] and the database scan method

5.2 Effectiveness for Targeted Advertising

This experiment aims to evaluate the effectiveness of applying our algorithm for targeted advertising. In addition to value-based target advertising approach, as described in Chapter 4, we also implemented similar target advertising algorithms by employing two other measures:

1. *confidence* of the matching rule, and
2. *confidence estimation* as reported in [SA95]

The algorithms using confidence and confidence estimation in [SA95] are named *Conf-based-targeted-adv*() and *EstConf-based-targeted-adv*() respectively. Their pseudo code, are listed in Figure 5-6 and 5-7.

Conf-based-targeted-adv(*p*: a product, *N*: number of advertising customers, *GP*: a set of GP-association rules): set of customers
{
    $S = \varnothing$; //set of targeted customers
    $NT = 0$; // number of accumulated targeted customers
    M_p = {*r*: *r* is a rule in *GP* with consequent matching *p*};
    FOR (each rule $r \in M_p$ listed in descending order of *r.Conf*) **DO**
    {
        S' ={c: c is a customer with demographics *r*.demo};
        IF ($|S' \cup S| \geq N$) {
            Randomly choose $N-|S|$ customers with demographics *r*.demo and add them to *S*;
            BREAK;
        }
        $S = S' \cup S$;
    }
    RETURN *S*;
} // end of ***Conf-based-targeted-adv***

Figure 5-6: Pseudo code of *Conf-based-targeted-adv*()

// Let r=D→P and Lp be the leaf in the product taxonomy that matches p
// *Est_Conf*(r, p) is the estimated confidence of D→Lp using the method in [SA95]

EstConf-based-targeted-adv(*p*: a product, *N*: number of advertising customers, *GP*: a set of GP-association rules): set of customers
{
    $S = \varnothing$; //set of targeted customers
    $NT = 0$; // number of accumulated targeted customers
    M_p = {*r*: *r* is a rule in *GP* with consequent matching *p*};
    FOR (each rule $r \in M_p$ listed in descending order of *Est_Conf*(*r p*)) **DO**
    {

```
        S' ={c: c is a customer with demographics r.demo};
        IF (|S'∪S | ≥ N) {
            Randomly choose N-|S| customers with demographics r.demo and add
        them to S;
            BREAK;
        }
         S = S'∪S;
    }
    RETURN S;
} // end of Conf-est-target-adv
```

Figure 5-7: Pseudo code of *EstConf-based-targeted-adv*

Finally, a baseline approach that randomly chooses patrons for recommendation is also implemented.

We adopted 5-fold cross validation and divided the books in our experimental data set into five sets. One set is assigned as the test data set while the other four sets serve as the training data set. Given a book in the test data set, we applied each of the above approach and identified a fixed number of patrons (called target set).

The measure of hit rate is done as follows. For each book p in the test data set, we compute a target set T_p of patrons. The hit ratio for p is the percentage of the patrons who had checked out p that are in T_p. Here we plot the cumulative gains chart [SD02] to show the performance of our method and the others. The cumulative gains chart consists of lift curve and baseline curve. The *value-based* approach and the other two methods are regarded as lift curves and the random approach serves as the baseline curve. As we know, the greater the area between the lift curve and the baseline, the better the model. The x-axis shows the maximum number of targeted customers (N). The y-axis shows the average hit ratio of ten randomly selected books.

The average hit ratio across 5 trials is reported. The average hit ratio of targeted advertising at various values of *N* (from 1000 to 5000 patrons), the number of target set, is reported. The result is shown in Figure 5-8.

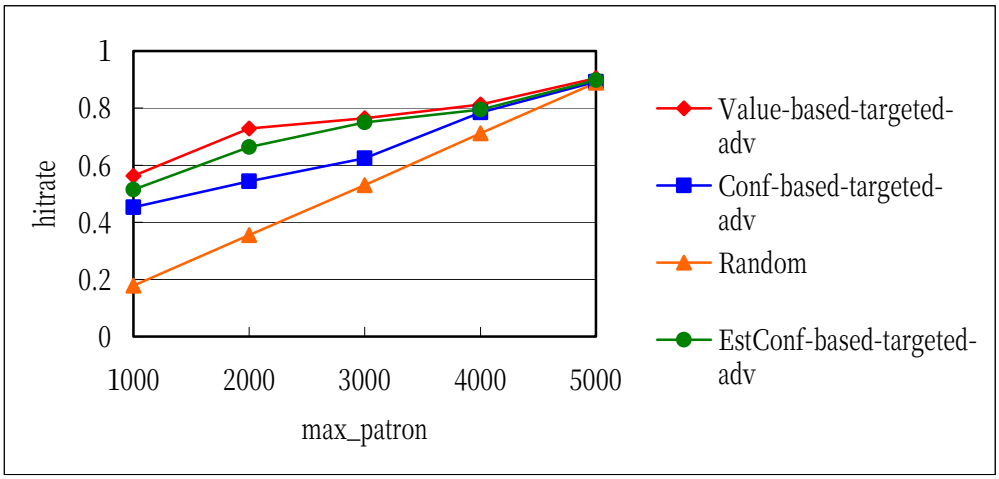

Figure 5-8: The cumulative gain charts of various methods

As can be seen, all four approaches incurs higher hit ratio as the increase of the size of the target set, and the increasing rates of *Value-based*, *Conf-est-based*, and *Conf-based* methods are much smaller than *random* method. Also the value-based method exhibits the best hit ratio, and the confidence-estimation based method performs better than the confidence-based method. This demonstrates the usefulness of applying GP-association rules to targeted advertising. As expected, when *N* is close to the total number of patrons (=5631), all four approaches exhibited similar high hit ratio.

Chapter 6 Conclusions

We have proposed a novel approach to promoting products with no prior transaction records. This approach starts with the identification of strong GP-association rules, the identified GP-association rules are used for locating a short list of prospective customers for a given promoted product, for which we have developed a comprehensive algorithm. We have evaluated the proposed approach using the library-circulation data obtained from our university library. We have shown that our proposed method for estimating confidence of a specialized rule from other related rules achieves the highest accuracy compared to the existing approach proposed in [SA95]. The target advertising approach that uses the estimated confidence, or called *value-based* approach, was also shown to achieve the most effective target advertising than the other methods.

This thesis deals with the target advertising of only one product at a time. Target advertising multiple products at the same time requires further research.

References

[AGM02] M. Adler, P. B. Gibbsons, Y. Martia, "Scheduling space sharing for Internet advertising," *Journal of Scheduling*, 5(2), pp. 103-119, 2002.

[AIS93] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington DC, pp. 207–216, 1993.

[AM03] A. Amiri and S. Menon, "Efficient Scheduling of internet banner advertisements," *ACM Transactions on Internet Technology*, 3(4), November, pp. 334-346, 2003.

[AS94] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proceedings of the 20th VLDB Conference*, pp. 478–499, Sept. 1994.

[AS95] R. AgrawaI and R. Srikant, "Mining sequential patterns," *Proceedings of the 11th International Conference on Data Engineering*, Taipei, Taiwan, March 1995.

[ASY98] C. Aggarwal, Z. Sun and P. S. Yu, "Online algorithms for finding profile association rules," *Proceedings of Fifth International Conference on Information and Knowledge Management* (CIKM98), pp. 86–95, 1998.

[AT01] G. Adomavicius and A. Tuzhilin, "Multidimensional recommender systems: A data warehousing approach," *Proceedings of Second International Workshop on Electronic Commerce* (WELCOM01), Heidelberg, Germany, 2001.

[BL98] J. P. Benway and D. M Lane, "Banner blindness: Web searchers often miss obvious links," Internetworking, 1(3), 1998, also available at http://www.internettg.org/newsletter/dec98/banner_blindness.html.

[FTTM96] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, "Mining optimized association rules from numeric attributes," *Proceeding of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, June 1996.

[HC80] R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics*, 4'th ed., MacMillan Publishing, New York, 1980.

[HF95] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," *Proceedings of the 21st VLDB Conference*, pp. 420–431, 1995.

[HL02] S. Hwang and E. Lim, "A data mining approach to new library book recommendations," *International Conference on Asian Digital Libraries*, 2002.

[HT97] R. V. Hogg and E. A. Tanis, *Probability and Statistical Inference*, 5'th ed., Prentice-Hall, N.J., 1997.

[KMRTV94] M. Klemettinen, H. ManniIa, P. Ronkainen, H. Toivonen, and A. I. Verkamo, "Finding interesting rules from large sets of discovered association rules," *Processing 3rd Int'l Conf. on Information and Knowledge Management*, pp. 401-408, Gaithersburg, Maryland, Nov. 1994.

[LNAK99] M. Langheinrich, A. Nakamura, N. Abe, T. Kamba, and Y. Koseki, "Unintrusive customization techniques for web advertising," *Computer Networks*, 31, pp. 1259-1272, 1999, also available on *Proceedings of the 8'th International World Wide Web Conference* (WWW1999)

[Naka02] A. Nakamura, "Improvements in practical aspects of optimally scheduling web advertising," *Proceedings of the* 11*'th International World Wide Web Conference* (WWW2002), pp. 536-541, 2002.

[PCY95] J.S. Park, M.S. Chen, and P.S. Yu, "An effective hash based algorithm for mining association rules," *Proceeding 1995 ACM-SIGMOD Int. Conf. Management of Data*, San Jose, CA, May 1995

[PM94] G. Piatesky-Shapiro and C. J. Matheus. "The interestingness of deviations," *In AAAI'94 Workshop on Knowledge Discovery in Databases*, Seattle, pp. 25-36, WA, July 1994.

[RS98] R. Rastogi and K. Shim, "Mining Optimized Association Rules with Categorical and Numeric Attributes," *Proceedings of the International Conference on Data Engineering*, Orlando, Florida, February 1998.

[SA95] R. Srikant and R. Agrawal, "Mining generalized association rules," *Proceedings of the 21st VLDB Conference*, Zurich, Switzerland, pp. 409–419, Sept. 1995.

[SA96] R. Srikant and R. Agrawal, "Mining quantitative association rules in large tables," *Proceedings of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, pp. 1–12, June 1996.

[SD02] T. Soukup and I. Davidson, *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, (*to be published by*) Wiley, 2002.

[SPAM03] Spam laws, see http://www.spamlaws.com/.

[Toml00] J. A. Tomlin, "An entropy approach to unintrusive targeted advertising on the Web," *Computer Networks* 33, pp. 767-774, 2000, also available on the *Proceedings of the* 11*'th International World Wide Web Conference* (WWW2000).

[WSJ94] "Using Computers to Divine Who Might Buy a Gas Grill," *The Wall Street Journal*, August 16, 1994.