Data Mining and Knowledge Discovery, 8, 7–23, 2004

Building an Association Rules Framework to Improve Product Assortment Decisions

TOM BRIJS tom.brijs@luc.ac.be
GILBERT SWINNEN gilbert.swinnen@luc.ac.be
KOEN VANHOOF koen.vanhoof@luc.ac.be
GEERT WETS geert.wets@luc.ac.be
Department of Applied Economic Sciences, Limburgs Universitair Centrum, Universitaire Campus, Gebouw D, B-3590 Diepenbeek, Belgium

Editor: Heikki Mannila

Received September 27, 1999; Revised June 19, 2000

Abstract. It has been claimed that the discovery of association rules is well suited for applications of market basket analysis to reveal regularities in the purchase behaviour of customers. However today, one disadvantage of associations discovery is that there is no provision for taking into account the business value of an association. Therefore, recent work indicates that the discovery of *interesting* rules can in fact best be addressed within a microeconomic framework. This study integrates the discovery of frequent itemsets with a (microeconomic) model for product selection (PROFSET). The model enables the integration of both quantitative and qualitative (domain knowledge) criteria. Sales transaction data from a fully automated convenience store are used to demonstrate the effectiveness of the model against a heuristic for product selection based on product-specific profitability. We show that with the use of frequent itemsets we are able to identify the cross-sales potential of product items and use this information for better product selection. Furthermore, we demonstrate that the impact of product assortment decisions on overall assortment profitability can easily be evaluated by means of sensitivity analysis.

Keywords: association rules, frequent itemset, product assortment decisions

1. Introduction

In the past, retailers saw their job as one of buying products and putting them out for sale to the public. If the products were sold, more were ordered. If they did not sell, they were disposed of. Blischok (1995) describes retailing in this model as a *product-oriented* business, where talented merchants could tell by the look and feel of an item whether or not it was a winner. In order to be successful, retailing today can no longer be just a product-oriented business. According to Blischok, it must be a *customer-oriented* business and superior customer service comes from superior knowledge of the customer. It is defined as the understanding of all customers' purchasing behavior as revealed through his or her sales transactions, i.e. *market basket analysis*.

Currently, the gradual availability of cheaper and better information technology has in many retail organizations resulted in an abundance of sales data. Hedberg (1995) mentions the American supermarket chain 'Wal-Mart' which stores about 20 million sales transactions

per day. This explosive growth of data leads to a situation in which retailers today find it increasingly difficult to obtain the right information, since traditional methods of data analysis cannot deal effectively with such huge volumes of data. This is where knowledge discovery in databases (KDD) comes into play.

Today, among the most popular techniques in KDD, is the extraction of association rules from large databases. While many researchers have significantly contributed to the development of efficient association rule algorithms (Agrawal et al., 1993, 1994, 1996; Brin et al., 1997; Park et al., 1995; Zaki et al., 1997), literature on the use of this technique in concrete real-world applications remains rather limited (Ali et al., 1997; Anand et al., 1997; Viveros et al., 1996). Nevertheless, the widespread acceptance of association rules as a valuable technique to solve real business problems will largely depend on the successful application of this technique on real-world data. Moreover, it has been claimed recently (Kleinberg et al., 1998) that the utility of extracted patterns (such as association rules) in decision-making can only be addressed within the microeconomic framework of the enterprise. This means that a pattern in the data is interesting only to the extent in which it can be used in the decision-making process of the enterprise to increase utility. In this perspective, currently one major disadvantage of associations discovery is that there is no provision for taking into account the business value of an association (Cabena et al., 1998). For instance, in terms of the interestingness of the associations discovered, the sale of an expensive bottle of wine together with a few oysters accounts for as much as the sale of a can of coke together with a packet of crisps. Therefore, we claim that the current output of association rule discovery methods is inadequate to support commercial decision-making in retailing.

In this paper, we tackle the problem of product assortment analysis and we introduce a concrete microeconomic integer-programming model for product selection (PROFSET[1]) based on the use of frequent itemsets. Furthermore, we demonstrate its effectiveness on real-world sales transaction data obtained from a fully automated convenience store.

The remaining part of this paper is structured as follows. Section 2 introduces the product selection problem in a retailing environment and presents an historical overview of the different techniques to measure product interdependencies. In Section 3, we introduce a product selection model based on the use of frequent itemsets. Section 4 presents the results of the empirical study. Finally, Section 5 summarizes our work and presents directions for future research.

2. The product selection problem

2.1. Problem situation

Determining the *ideal* product assortment has been (and still is) the dream of every retailer. From the marketing literature (Van der Ster and van Wissen, 1993) it is known that the optimal product assortment should meet two important criteria.

Firstly, the assortment should be *qualitatively* consistent with the store's image. A store's image distinguishes the retailer from its competition and is projected by means of its design, layout, services and of course its products. Therefore, retailers often distinguish between

basic products and *added* products. Basic products are products that should not be deleted from the assortment because they are the core materialization of the retailer's store formula. For example, for a typical convenience store, customers expect at least beverages, cigarettes, food and candy products in the assortment. Therefore, such products should not be removed. Otherwise, the assortment will not meet the basic expectations of customers who visit the store. In contrast, *added* products are chosen by the retailer to reinforce the store image and should be selected as to maximize cross-sales potential with *basic* products. Indeed, retailers are interested in adding items whose sales will not be made at the expense of currently stocked items but may help increase the sales of other items (sales complements) (Pessemier, 1980). For the convenience store, examples may include cigarette lighters, coffee whitener or tea warmers. This means that *added* products should be selected by the model based on their purchase affinity with *basic* products.

Secondly, because retailing organizations are profit-seeking companies, the product assortment should be *quantitatively* appealing in terms of the profitability that it generates for the retailer (i.e. the microeconomic framework). In Section 3.2, this quantitative element will be further defined.

From the above two criteria, it must be clear that the issue of 'product interdependencies' is critical to evaluate the position and contribution of a product within the assortment. Indeed, we believe that it is important to include *cross-selling effects* when selecting products for the optimal product assortment. This implies that one does not only have to look at the contribution of individual products, but one must also investigate the extent to which a product exhibits a significant positive *radiation effect* on other products in the assortment.

2.2. An historical overview of measuring product interdependencies

Since the idea of product interdependencies is critical for the product selection problem, we believe that it is useful to provide a short literature overview of this topic. Moreover, drawbacks of past techniques to measure product interdependencies will justify the use of frequent itemsets as an alternative in this paper. In general, previous measures can be classified into two major categories: association coefficients and interaction parameters.

2.2.1. Association coefficients. Already in the mid 70's and early 80's, in the marketing literature, Böcker (1978) and Merkle (1981) introduced a number of measures to investigate product interdependencies. Basically, coefficients were developed as follows. A matrix was built containing the frequencies of simultaneous purchases for all product pairs. Then, for each pair, an association coefficient was calculated to reflect the similarity in the sales of the two products. However, the matrix was built on the assumptions that symmetric and transitive relations exist between product sales. Similarity implies that purchase relations from product A to product B equal those from B to A. The assumption of transitivity was introduced to process the data coming from more than two concurrent purchases, i.e. when a relation exists between $A \Rightarrow B$ and between $B \Rightarrow C$, then it is assumed that there also exists a relation between $A \Rightarrow C$.

However, practical observations show that these assumptions are highly questionable. Furthermore, data storage problems are enormous since calculating all association

Table 1. An illustration of 7 multiple purchases.

TID	*A*	*B*	*C*	*D*	*E*	*F*	Number of items bought
1	1	0	1	1	0	0	3
2	0	0	0	0	1	1	2
3	0	0	0	0	1	1	2
4	0	1	1	1	1	0	4
5	1	0	1	1	1	0	4
6	1	1	1	1	1	0	5
7	1	0	1	1	0	0	3
Total item sales	4	2	5	5	5	2	23

coefficients for some 5000 items in a small supermarket requires the construction of a (5000×5000)-matrix! A similar idea as the one expressed by association coefficients is the Yule's Q-coefficient (Kendall and Stuart, 1979).

Since multiple purchases of products (for instance A, B, C and D are purchased together) are divided into two-way relations (AB, AC, AD, BC, BD and CD), one can show that the number of two-way relations will increase in proportion to the number of products (m) with a factor $m * (m - 1)/2$. As a consequence, products with an equal purchasing frequency will be treated unequally if they arise from purchases that differ with respect to the number of products purchased. This is illustrated by Tables 1 and 2, which show that products B and F are included in two purchases that differ in volume. The number of two-way relations adds to 7 for B and 2 for F (last row in Table 2). To correct for such unequal treatment, Böcker and Merkle suggest weighting all two-way relations with a factor $1/(m - 1)$. The resulting matrix of association frequencies is depicted in Table 3. Frequency data are normalized in order to take into consideration the unequal total amount purchased for each product.

Therefore, Merkle suggests using the following association coefficient (A_{ij}) with the respective results shown in Table 4:

$$A_{ij} = a/\min\{b, c\}$$

where

$a =$ the frequency of joint purchases of i and j
$b =$ the frequency of purchases of i
$c =$ the frequency of purchases of j

2.2.2. Interaction parameters. A second family of measures for interdependence is the so-called interaction parameters that are frequently used in log linear models in order to

Table 2. Matrix of association frequencies.

Item	*A*	*B*	*C*	*D*	*E*	*F*	Total
A	0	1	4	4	2	0	11
B	1	0	2	2	2	0	7
C	4	2	0	5	3	0	14
D	4	2	5	0	3	0	14
E	2	2	3	3	0	2	12
F	0	0	0	0	2	0	2
Total	11	7	14	14	12	2	60

Table 3. Matrix of association frequencies using weighting factor $1/(m-1)$.

Item	*A*	*B*	*C*	*D*	*E*	*F*	Total
A	0	1/4	$1+7/12$	$1+7/12$	7/12	0	4
B	1/4	0	7/12	7/12	7/12	0	2
C	$1+7/12$	7/12	0	$1+11/12$	$1+11/12$	0	5
D	$1+7/12$	7/12	$1+11/12$	0	11/12	0	5
E	7/12	7/12	11/12	11/12	0	2	5
F	0	0	0	0	2	0	2
Total	4	2	5	5	5	2	23

Table 4. Association coefficients A_{ij} calculated from Table 3.

Item	*A*	*B*	*C*	*D*	*E*	*F*
A	–					
B	0.125	–				
C	0.396	0.292	–			
D	0.396	0.292	0.383	–		
E	0.146	0.292	0.383	0.183	–	
F	0	0	0	0	1	–

calculate joint purchase probabilities (Hruschka et al., 1991). Although these models have a profound statistical background, they are limited in the number of products or categories they can handle. Mostly, they only include interactions between pairs of products or categories (first-order interactions) since computational problems for higher-order interactions become too cumbersome. Furthermore, these models typically use category interdependencies instead of product interdependencies because in the latter case, statistical significance of the interaction parameters between products becomes too low. For instance, let $Y_i (i = 1, \ldots, I)$ be a binary variable representing a purchase in category i and let X_i be a

binary variable indicating a sales promotion in category i, then a log linear model for joint purchase probabilities $P(Y_1, \ldots, Y_I)$ may be represented as:

$$\mathrm{Ln}P(Y_1, \ldots, Y_I) = a_0 + \sum_{i=1}^{I}(a_i + b_i X_i)Y_i + \sum_{i=1}^{I-1}\sum_{j=i+1}^{I}(a_{ij} + b_{iji}X_i + b_{iji}X_i)Y_i Y_j$$

a_i is the main effect of category i (the change of the log expected joint probabilities by a purchase of category i), a_{ij} the first-order interaction between two categories i and j. Interactions measure the deviation of the log observed joint probabilities from the log expected joint probabilities if only main effects are considered.

2.2.3. Frequent itemsets: a viable alternative. Given the shortcomings of the interdependency measures discussed above, we argue that frequent itemsets (Mannila, 1997) provide a viable alternative to the measurement of product interdependencies. In a retailing environment, a frequent itemset is a set of products that frequently occurs together in a set of shopping baskets. More formally, if D is a database of shopping baskets and X is a set of products (i.e. an itemset), then the frequency of this itemset X can be expressed as in Definition 1.

Definition 1. $s(X, D)$ represents the frequency of itemset X in D, i.e. the fraction of shopping baskets in D that contain X.

Consequently, if the frequency of the itemset X exceeds a user-defined frequency threshold σ, then this itemset X is called *frequent*.

Definition 2. An itemset X is called frequent in D, if $s(X, D) \geq \sigma$ with σ the minsup.

The concept of a frequent itemset offers several advantages compared to the other measures presented above. First of all, the measurement of interdependencies between products on the SKU^2-level seems to be empirically tractable. Secondly, the frequent itemsets approach enables the discovery of higher-order interactions (interactions between more than two products). Finally, problems with transitivity and symmetry are solved with the discovery of association rules (Agrawal et al., 1994). Indeed, association rules enable to distinguish between the confidence of the relationship $A \Rightarrow B$ and $B \Rightarrow A$, i.e. symmetry is not assumed, and if $A \Rightarrow B$ and $B \Rightarrow C$ are supported, the association rules algorithm may still conclude that $B \Rightarrow C$ does not support the user-defined support and confidence thresholds, i.e. transitivity is not assumed.

2.3. The search for interesting product combinations

The discussion how to measure product interdependencies can be viewed within a broader framework of exploring the *interestingness* of product associations. So far, we believe that three different approaches of interestingness can be distinguished.

Table 5. Economical interpretation of *interest*.

Outcome	Interpretation
Interest > 1	Complementarity effects between X and Y
Interest $= 1$	Conditional independence between X and Y
Interest < 1	Substitutability[3] effects between X and Y

First, a number of *objective* measures of interestingness have been developed in order to filter out non-interesting association rules based on a number of statistical properties of the rules, such as support and confidence (Agrawal and Srikant, 1994), intensity of implication (Guillaume et al., 1998), J-measure (Wang et al., 1998). Other measures are based on the syntactical properties of the rules (Liu and Hsu, 1996), or they are used to discover the least-redundant set of rules (Brijs et al., 2000; Toivonen et al., 1995). In terms of measuring purchase complementarities in a retail environment, two objective measures are very relevant, i.e. interest (Silverstein et al., 1998) and correlation (Liu et al., 1999 and recently Ahmed et al., 2000):

Definition 3. Interest (Silverstein et al., 1998).

$$s(A \Rightarrow B)/s(A) * s(B)$$

The nominator $s(A \Rightarrow B)$ measures the observed frequency of the co-occurrence of the items in the antecedent (A) and the consequent (B) of the rule. The denominator $s(A) * s(B)$ measures the expected frequency of the co-occurrence of the items in the antecedent and the consequent of the rule if both itemsets were conditionally independent. Table 5 illustrates the three possible outcomes for the interest measure and their associated economic interpretation for the interdependence between the items in the antecedent and consequent of the rule.

Definition 4. Positive correlation (Ahmed et al., 2000).

$$P(A \wedge B)/P(A) - P(B) \geq 0$$

Although both measures are suited to discover complementarity effects between product items, both of them fail to incorporate the monetary value of product associations. As a result, all of the existing objective interestingness measures fail to incorporate the monetary value of product associations and as such they do not really solve the issue of interestingness for the retailer.

Second, it was recognized that domain knowledge may also play an important role in determining the interestingness of association rules, a number of *subjective* measures of interestingness have been put forward, such as unexpectedness (Silberschatz and Tuzhilin, 1996; Padmanabhan and Tuzhilin, 1999), actionability (Adomavicius and Tuzhilin, 1997) and rule templates (Klemettinen et al., 1994). Again, the monetary value of product associations does not come into play and therefore, practice demonstrates that the usability of these approaches for purposes of product selection is rather low.

Finally, the most recent stream of research advocates the evaluation of the interestingness of associations in the light of the *micro-economic framework* of the retailer (Kleinberg et al., 1998). More specifically, a pattern in the data is considered interesting only to the extent in which it can be used in the decision-making process of the enterprise to increase its utility. Indeed, in the case of market basket analysis, eventually it is important for the retailer to adopt the discovered knowledge for improving his marketing decision-making and to increase profits. It is in the light of this idea that the PROFSET model is constructed since it takes into account the monetary value of product associations.

3. The PROFSET model for product selection

According to the problem situation described above (Section 2.1), a framework must be developed which is able to select a *hit list* of products, i.e. a selection of a user-defined number of products from the assortment that yields the maximum overall profit, taking into account background knowledge of the retailer. A simple solution to this problem, which is often used in practice, is to calculate the total profit contribution generated *per product* and then select those products, in addition to the *basic* products that have already been selected by the retailer, that contribute the most to the overall profitability. We call this the product-specific profitability heuristic. Although easy to calculate, it does not take cross-selling effects of products into account. In contrast, the PROFSET model, which we will introduce in this paper, implicitly takes into account cross-selling effects by using frequent itemsets (Mannila, 1997) from association rule mining (Agrawal et al., 1994). Before specifying the microeconomic optimization model, we will first introduce the parameters and components of the PROFSET model.

3.1. Model parameters

Gross margin:
Let: $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items in a retail store (i.e. the product assortment)

D be a set of transactions, where each transaction $T \subseteq I$
$SP(i)$ be the selling price at which product i is sold to the consumer
$PP(i)$ be the purchase price at which product i is purchased by the retailer
$f(i)$ be the number of times product i was purchased in a particular shopping basket T

Definition 5. $m(T)$ is the gross sales margin generated by sales transaction T.

$$m(T) = \sum_{i \in T} (SP(i) - PP(i)) * f(i)$$

Definition 6. $M(X)$ is the gross sales margin generated by frequent itemset X.

$$M(X) = \sum_{T \in D} m'(T) \quad \text{with} \quad \begin{cases} m'(T) = m(T) & \text{if } X = T \\ m'(T) = 0 & \text{otherwise} \end{cases}$$

Definition 6 is very crucial. It demonstrates that the gross sales margin $M(X)$ for a frequent set X is calculated as the sum of all gross sales margins of transactions T (i.e. $m(T)$) for which the items contained in that frequent itemset equal exactly those contained in the transaction T. Therefore, if $X = T$ (i.e. shopping basket T contains exactly the same items as frequent itemset X) then $m(T)$ adds up to $M(X)$ by setting $m'(T) = m(T)$. Otherwise, if $X \neq T$, then $m(T)$ is not added to $M(X)$ by setting $m'(T) = 0$. The reason for doing this is that we will use the sum of all $M(X)$ to approximate the total profitability of the assortment. Now, suppose that $m'(T) \neq 0$ when $X \subseteq T$ instead of $X = T$ with $[i_1, i_2]$ a frequent itemset[4] X and $\{i_1, i_2, i_4\}$ a sales transaction T. Clearly, $[i_1, i_2] \subseteq \{i_1, i_2, i_4\}$ but, because $[i_1, i_2]$ is frequent, it is known (Agrawal et al., 1996) that $[i_1]$ and $[i_2]$ will also be frequent. Consequently, $[i_1] \subseteq \{i_1, i_2, i_4\}$ and $[i_2] \subseteq \{i_1, i_2, i_4\}$ and thus the gross sales margin generated by sales transaction $\{i_1, i_2, i_4\}$ will add to $M([i_1, i_2])$, $M([i_1])$ and $M([i_2])$ even if i_4 is not selected for inclusion in the hit list. Thus, if $m'(T) \neq 0$ when $X \subseteq T$, then a single sales transaction increases the $M(X)$ parameter of *all* the frequent itemsets that are contained in that transaction. To summarize, a single sales transaction is allowed to contribute to the total profitability only once through the $M(X)$ parameter of the *frequent itemset* that contains the same items as those included in that transaction. Thus, X must be equal to T to prevent double counting.

Cost of products. Also product handling and inventory costs should be included in the model. Product handling costs refer to costs associated with the physical handling of the goods. Inventory costs include financial costs of stocking the items and costs of re-stocking, which are a function of replenishment frequency and the lead-time of the orders. In practice, however, these costs are often difficult to obtain, especially product handling costs. For reasons of simplicity, we assume that a total cost figure $Cost_i$ per product i can be obtained for each product.

3.2. Model components

The PROFSET optimization problem is operationalized by means of an integer-programming model containing two important components:

Objective function. The objective function represents the goal of the optimization problem and therefore reflects the microeconomic framework of the retail decision maker. It is constructed in order to maximize the overall profitability of the hit list. The gross margins $M(X)$ associated with the frequent sets X contribute in a positive sense to the objective function. Of course, this will only occur when a frequent set X is selected which is represented in the objective function by the Boolean variable P_X. In contrast, the $Cost_i$ associated with each individual product i contributes in a negative sense, but only if the product i is selected which is represented by a second Boolean variable Q_i.

Constraints:

(1) Because the final decisions need to be taken at the product level instead of at the *frequent itemset* level, we must specify which products i are included in each frequent itemset

X. This information can be obtained from the frequent sets during association rule mining.
(2) *Basic* products can be specified by forcing the model to select certain products.
(3) The *size* of the hit list is specified by the *ItemMax* constraint.

3.3. Model specification

Let L be the set of frequent itemsets X, and let P_X, $Q_i \in \{0, 1\}$ be the decision variables for which the optimization routine must find the optimal values. Q_i equals 1 as soon as any frequent itemset X in which it is included is set to 1 ($P_X = 1$) by the optimization routine.

$$\begin{aligned} \max \quad & \left(\sum_{X \in L} M(X) * P_X - \sum_{i \in L} Cost_i * Q_i \right) \\ \text{subject to} \quad & \forall X \in L, \forall i \in X: Q_i \geq P_X \qquad (1) \\ & \forall i \text{ is a } \textit{basic} \text{ product: } Q_i = 1 \qquad (2) \\ & \sum_{i \in L} Q_i = ItemMax \qquad (3) \end{aligned}$$

with P_X and Q_i Booleans.

By using frequent itemsets the objective function will give a lower bound, i.e. the *observed* amount of profit will be higher than indicated by the value of the objective function. The reason is that we consider frequent itemsets and thus *in*frequent itemsets will not add to the total profit amount in the objective function. This is justified because it is highly probable that infrequent itemsets exist because of random purchase behavior. Consequently, we claim that the objective function only measures the profit from structural, underlying purchase behavior.

3.4. Model calculation

The calculation of the PROFSET model is carried out by a Mixed Integer Programming (MIP) solver called CPLEX 6.5. CPLEX 6.5 (Bixby et al., 1999) is a commercial operations research software (www.cplex.com) and uses a branch-and-bound (with cuts) algorithm, which solves a series of Linear Programming (LP) sub problems to solve large MIPs. But, since a single MIP generates many LP sub problems, MIPs can be very computer intensive and require significant amounts of physical memory. The reason is that the branch-and-bound tree may be as large as 2^n nodes, where n equals the number of binary variables, such that a problem containing only 30 binary variables (i.e. the number of frequent itemsets in our study) could produce a tree having over one billion nodes! Therefore, typically a stopping criterion is being set, e.g. a time limit or a relative optimality criterion, the latter specifying that the search for the optimal solution is aborted if the current best solution is $x\%$ below the best possible integer solution.

The model contains as many binary variables as there are frequent itemsets discovered during association rules mining. Furthermore, the number of constraints of type (1) equals the number of frequent itemsets multiplied by the number of items contained in each frequent set since for each frequent set there are as many constraints as items contained in that frequent set. The number of type (2) constraints is dependent on the number of basic products that the retailer wants to specify. Finally, there is only one type (3) constraint. Details about the concrete number of variables, constraints and the execution time of the PROFSET model on our data are given in the next section on empirical results.

4. Empirical study

The empirical study is based on a data set of 27148 sales transactions acquired from a fully-automated convenience (FAC) store over a period of 5.5 months in 1998. The concept of the fully automated convenience store is closely related to that of the vending machine. However, as opposed to the product assortment of the typical vending machine, this new retail store offers a wider variety of products. Typically, a selection of about 200 products is included ranging from the typical product categories such as beverages, food, candy and cigarettes, to products like healthcare, pet food, fruit, batteries, film supplies (camera, roll of film), which are presented to the customer by means of a 8 m^2 window display. The product assortment of the store under study consists of 206 different items. However, the average sales transaction contains only 1.4 different items because, in this type of convenience store, customers typically do not purchase many items during a single shopping visit. With regard to the costs of each individual product in the assortment, detailed information about handling and inventory costs could not be obtained, so these will be considered equal for all products and therefore these costs are not included in the model.

Basically, the empirical study involves two important phases. In the first phase, frequent sets of products are discovered to represent structural purchase behavior (Section 4.1). Then, in the second phase, the PROFSET method is used to select a hit list of products from the assortment (see Section 4.2).

4.1. Mining for frequent itemsets

Because the objective function in the PROFSET method requires frequent itemsets as input, frequent itemsets were discovered from the database. An *absolute* support of 10 was chosen. This means that no item or set of items will be considered frequent if it does not appear in at least 10 sales transactions. As a consequence, we consider all itemsets X as non-frequent, i.e. describing random purchase behavior, if the itemset appears in less than 10 rows in the sales-transaction database. It could be argued that the choice for this support parameter is rather subjective. This is partially true, however, domain knowledge from the retailer can often indicate what level of support may be considered as relevant. Furthermore, within relatively small intervals, the model will be insensitive to alterations of the minimum support threshold. The reason is that when gross margins of products are within a relatively small range, frequent itemsets with relatively low support will not be able to influence the objective function. From the analysis, 523 frequent itemsets were obtained of size 1 or 2 with absolute

support ranging from 10 to 2833. The size of the *frequent* itemsets is rather small; this can however be explained by the small size of the average sales transaction. Although the PROFSET model does not use association rules as input, i.e. it uses only frequent itemsets, the discovery of association rules will be helpful for interpreting the output of PROFSET (as will be explained in the next section), and therefore they were also generated during this phase.

4.2. Product selection (PROFSET)

In order to make the comparison between PROFSET and the product specific profitability heuristic straightforward, we chose not to specify *basic* products in the model. Consequently, the model will be able to fully exploit cross-sales potential between items in the assortment without any restrictions. Furthermore, any *ItemMax* value can be selected to constrain the number of items for inclusion in the hit list. The PROFSET model contains 523 binary variables (one for each frequent itemset), 840 constraints of type (2) and one constraint of type (3). Calculation of the PROFSET model with CPLEX 6.5 (see Section 3.4 for details about CPLEX 6.5) took 0.080 seconds on a Windows NT server Pentium II 333 Mhz machine with 256 MB internal memory. Simulations with the PROFSET method indicated that, for this data set, important results can be identified:

1. Using PROFSET, some products with relatively low product-specific profitability but considerably high cross-selling effects are selected for the hit list.
2. The PROFSET method enables to assess the sensitivity of product assortment decisions and, as a result, allows to identify the importance of the impact of such decisions on the total profitability of the hit list.

4.2.1. Observation 1. We demonstrate observation 1 by a concrete example obtained from the empirical results with both product selection methods. Consider the products *tobacco brand* 'x' and *cigarette paper brand* 'y'. Table 6 illustrates the total margin[5] of each product and its according position (with regard to the total margin) within the entire product assortment. Table 6 shows that from the product-specific point of view, *tobacco* is the 17th most profitable product in the assortment and *cigarette paper* ends up 66th. So, when the maximum number of products allowed in the hit list is less than 66, according to the product-specific profitability heuristic, *cigarette paper* will not be included (see column 4). In contrast, simulations showed that for *ItemMax* equal to 35, the PROFSET method selects both products for inclusion in the hit list (see last column). This indicates that *cigarette paper* must have considerable cross-selling opportunities with one or more products that are included in the hit list.

In fact, this information can easily be obtained from examination of the association rules (Section 4.1)

$$\textit{Cigarette paper} \Rightarrow \textit{tobacco}\ [\textit{absolute sup} = 291,\ \textit{conf} = 1.00]$$
$$\textit{Tobacco} \Rightarrow \textit{cigarette paper}\ [\textit{absolute sup} = 291,\ \textit{conf} = 0.82]$$

Table 6. Total margin, position and selection for tobacco and cigarette paper.

	Total margin	Position	Prod profit	PROFSET
Tobacco brand 'x'	10353 BEF	17	•	•
Cigarette paper brand 'y'	3258 BEF	66	○	•

The above rules demonstrate that whenever a customer buys *cigarette paper*, he also buys *tobacco* (confidence = 100%) and that when a customer buys *tobacco* he will often also buy *cigarette paper* with it (confidence 82%), i.e. asymmetry. Concerning the products *tobacco* and *cigarette paper*, interest (see Definition 3) is equal to $76.7 \gg 1$ which indicates very strong complementarity effects between the two products. As a consequence, when treated together, these two products may represent a high total profit contribution indicating that it may be advised to select both products for the hit list instead of selecting only *tobacco*. Indeed, the total profit contribution of the frequent itemset $\{tobacco, cigarette\ paper\}$ makes this sales combination the 10th most profitable frequent itemset, and therefore PROFSET selected both products for inclusion in the hit list.

However, not all product combinations with high cross-selling potential are necessarily included in the hit list. The profit contribution of the sales combination must be sufficiently high for the items to be included in the hit list. For instance, the itemset $\{$*toothpaste*, *toothbrush*$\}$ has an interest of $2468 \gg 1$ (extremely high) and, according to the association rules generated in Section 4.1, they are always bought together. However, the support count of the itemset is equal to 11 (slightly above the minimum absolute support threshold). As a consequence, the total profit contribution of this itemset is insufficient to influence the product selection process.

This again illustrates that the microeconomic framework of the retailer directly determines the *interestingness* of the associations. Some associations (see example *toothpaste* and *toothbrush*) are very interesting from the statistical point of view (i.e. strong dependency between both items) but it is the microeconomic framework of the retailer that ultimately determines the profitability and thus the *real* interestingness of the association.

4.2.2. Observation 2. Concerning observation 2, the impact on total profitability caused by product assortment decisions can easily be assessed by means of sensitivity analysis. When for instance product i is deleted from the optimal set, and it is replaced by the best product i' outside the hit list, its impact on profitability can easily be observed by simulations in the optimization model.

Figure 1 illustrates the profit impact of the replacement of each product in the hit list. While most product replacements have only moderate profit implications (-2%), some products (6, 9, 20, 22) represent major profit drivers that should not be deleted from the hit list. This insight can help retailers to quantitatively evaluate product assortment changes. Furthermore, the replacement of items in the PROFSET model is based on dynamic reselection of products whereas for the product-specific profitability heuristic the product that replaces the exiting product will always be the one with the highest product-specific margin outside of the list. If this entrant happens to have no or small cross-selling effects with the

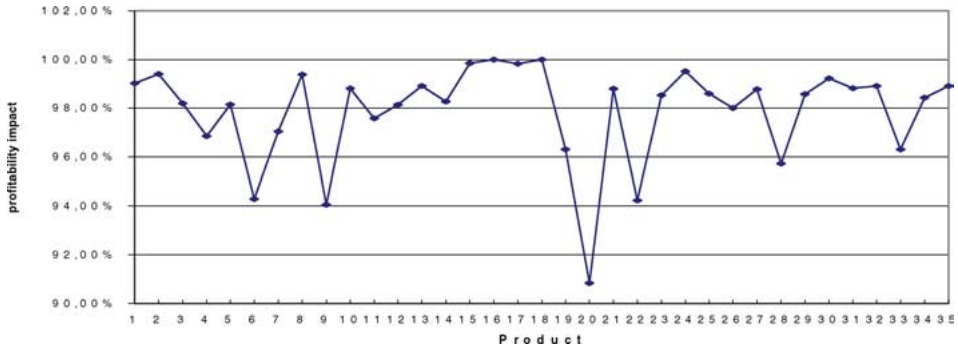

Figure 1. Profitability impact of replacement decisions.

items inside the hit list, selecting a product with a lower product-specific profitability but higher cross-selling effects with items inside the hit list could be more appropriate.

In fact, simulations revealed that for almost half of the items, dynamic reselection with the PROFSET model resulted in better overall profit compared to replacement with the product-specific profit heuristic. Obviously, the more cross-selling effects exist in the product assortment, the more impressive the profit improvement of the dynamic reselection will be, when compared to the product-specific profitability heuristic.

5. Conclusions and future research

5.1. Conclusions

In this paper, we proposed a microeconomic model for product selection based on the use of frequent itemsets obtained from association rule mining. More specifically, we integrated the notion of frequent itemsets into an integer-programming model taking into account some important microeconomic parameters that are often used by retailers to support their product selection decision-making process. The motivation for using frequent itemsets was partially supported by drawbacks of past measures to calculate product interdependencies. To empirically validate our model we used sales transaction data from a fully automated convenience store and compared the results with a frequently used method for product selection based on product-specific profitability. This comparison resulted in two major observations. Firstly, we showed that our model PROFSET select products that are truly interesting for the retailer, both in terms of qualitative and quantitative criteria, taking into account cross-selling effects between products. Secondly, we also showed that with our model, sensitivity analysis could easily be carried out, enabling the retailer to quantitatively assess the profitability impact of product assortment decisions.

Yet, the retailer should also consider the following limitation. The presented model is deterministic in nature. This means that the model assumes that when for itemset $\{X, Y\}$ the model does not select one of the items X or Y, consequently all profits related to this itemset will be lost. This is of course too simplistic, as customers do not always purchase

certain product combinations intentionally. Therefore, it may well be that a fraction of the sales related to that itemset may still be recovered. In fact, the impact of the assumption will depend on the availability of substitute products, either in the same or in other stores in the surrounding area of the current store. E.g. removing tobacco from the itemset {*tobacco*, *cigarette paper*} would cause the sales from cigarette paper to be lost almost entirely since the confidence of the rule *tobacco* $\Rightarrow$ *cigarette paper* is very high (82%). Therefore, the confidence of an association rule can provide an indication of the strength of the purchase relationship and thus the potential profits that are lost as a result of a product deletion. For products where there are sufficient substitute products available, or the confidence of the purchase relationship is low, the assumption given above causes an overestimated loss of profits. However, on the other hand, the loss of profits as calculated under the current assumption, provides an upper bound (worst case scenario) on the expected profit loss: information that retailers currently do not possess!

5.2. Future research

Three main topics will be issues for further research.

Firstly, we want to assess our model on supermarket data. It is expected that cross-selling effects are more manifestly present in supermarket data because consumers typically visit supermarkets to do one-stop-shopping. Given the size of a typical supermarket assortment, however, there is a possibility that we will not be able to carry out the analysis at the level of individual items but, instead, have to confine ourselves to an analysis within or between categories.

Secondly, when sales transaction data from multiple stores with different product assortments but more or less the same underlying purchase behavior can be obtained, it is possible to use the PROFSET method to construct an *ideal composite* product assortment. Indeed, when certain product combinations demonstrate to be very successful, the best product combinations obtained from multiple stores could be integrated in one *ideal* product assortment.

Finally, instead of including only gross margins from transactions for which the items contained in that transaction equally match the items in the frequent set (i.e. $X = T$), an alternate model would be to split the gross margin among all frequent itemsets that are contained in the transaction. While this may not influence the results for the current case study (since the average transaction length was only 1.4), the alternate model may be able to capture a higher percentage of transactions in sales data with higher transaction length (since the model will cover a higher percentage of transactions). However, the crucial point then is how much of the gross margin of a transaction should be allocated to each of the frequent sets that are contained in that transaction. Especially, the problem of frequent sets that are overlapping each other in the same transaction poses significant problems.

Acknowledgments

Tom Brijs is supported by the Fund for Scientific Research, Flanders. We acknowledge the reviewers for their valuable comments.

Notes

1. PROFSET uses the PROFitability per frequent SET to determine the optimal selection of products in terms of maximal total profit.
2. SKU = Stock Keeping Unit (an individual product identification).
3. Recall that substitutability indicates less than the expected level of mutual support.
4. Note that we use [..] to symbolize a frequent set and {...} to symbolize a sales transaction.
5. Total profit margin = number of items sold x unit profit margin of the product in Belgian Francs (BEF).

References

Adomavicius, G. and Tuzhilin, A. 1997. Discovery of actionable patterns in databases: The action hierarchy approach. In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97), pp. 111–114.

Agrawal, R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large databases. In Proceedings of ACM SIGMOD Conference on Management of Data, pp. 207–216.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. 1996. Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining, AAAI Press, pp. 307–328.

Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases, pp. 487–499.

Ahmed, K.H., El-Makky, N.M., and Taha, Y. 2000. A note on "Beyond market baskets: Generalizing association rules to correlations." SIGKDD Explorations, 1(2):46–48.

Ali, K., Manganaris, S., and Srikant, R. 1997. Partial classification using association rules. In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, pp. 115–118.

Anand, S., Hughes, J., Bell, D., and Patrick, A. 1997. Tackling the cross-sales problem using data mining. In Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 331–343.

Bixby, R.E., Fenelon, M., Gu, Z., Rothberg, E., and Wunderling, R. 1999. MIP: Theory and practice—Closing the gap. ILOG CPLEX Technical report.

Blischok, T. 1995. Every transaction tells a story. Chain Store Age Executive with Shopping Center Age, 71(3):50–57.

Böcker, F. 1978. Die Bestimmung der Kaufverbundenheit von Produkten. In Schriften zum Marketing, band 7.

Brijs, T., Vanhoof, K., and Wets, G. 2000. Reducing redundancy in characteristic rule discovery by using integer-programming techniques. Accepted for Publication in the Intelligent Data Analysis Journal.

Brin, S., Motwani, R., Ullman, J., and Tsur, S. 1997. Dynamic itemset counting and implication rules for market basket data. In Proceedings ACM SIGMOD International Conference on Management of Data, pp. 255–264.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. 1997. Discovering Data Mining: From Concept to Implementation. NJ: Prentice Hall.

Guillaume, S., Guillet, F., and Philippé, J. 1998. Improving the discovery of association rules with intensity of implication. In Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, pp. 318–327.

Hedberg, S. 1995. The data gold rush. BYTE, pp. 83–88.

Hruschka, H., Lukanowicz, M., and Buchta, C. 1991. Cross-category sales promotion effects. Journal of Retailing and Consumer Services, 6(2):99–106.

Kendall, M. and Stuart, A. 1979. The Advanced Theory of Statistics: Inference and Relationship. London: Charles Griffin and Company Ltd.

Kleinberg, J., Papadimitriou, C., and Raghavan, P. 1998. A microeconomic view of data mining. Data Mining and Knowledge Discovery Journal, 2(4):311–324.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A.I. 1994. Finding interesting rules from large sets of discovered association rules. In Proceedings of the Third International Conference on Information and Knowledge Management, pp. 401–407.

Liu, B. and Hsu, W. 1996. Post-analysis of learned rules. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), pp. 828–834.

Liu, B., Hsu, W., and Ma, Y. 1999. Pruning and summarizing the discovered associations. In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, pp. 125–134.
Mannila, H. 1997. Methods and problems in data mining. In Proceedings of the International Conference on Database Theory, pp. 41–55.
Merkle, E. 1981. Die Erfassung und Nutzung von Informationen über den Sortimentsverbund in Handelsbetrieben. In Schirften zum Marketing, band 11.
Padmanabhan, B. and Tuzhilin, A. 1999. Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems, Elsevier Science, 27:303–318.
Park, J., Chen, M., and Yu, Ph. 1995. An effective hash based algorithm for mining association rules. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 175–186.
Pessemier, E. 1980. Retail assortments—Some theoretical and applied problems. Technical Report, Marketing Science Institute Research Program.
Silberschatz, A. and Tuzhilin, A. 1996. What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering, 8(6):970–974.
Silverstein, C., Brin, S., and Motwani, R. 1998. Beyond market baskets: Generalizing association rules to dependence rules. Data Mining and Knowledge Discovery Journal, 2(1):39–68.
Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., and Mannila, H. 1995. Pruning and Grouping of Discovered Association Rules. In MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases, Heraklion, Crete, Greece, 1995.
Van der Ster, W. and van Wissen, P. 1993. Marketing & Detailhandel. Wolters-Noordhoff.
Viveros, M., Nearhos, J., and Rothman, M. 1996. Applying data mining techniques to a health insurance information system. In Proceedings of the 22nd International Conference on Very Large Data Bases, pp. 286–294.
Wang, K., Tay, S.H.W., and Liu, B. 1998. Interestingness-based interval merger for numeric association rules. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), pp. 121–127.
Zaki, M., Parthasarathy, S., Ogihara, M., and Li, M. 1997. New algorithms for fast discovery of association rules. In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, pp. 283–286.

Tom Brijs is a Ph.D. student at the Data Analysis and Modeling research group of the Limburgs Universitair Centrum Diepenbeek, where he obtained his master's degree in commercial engineer in business informatics in 1996. His research interests are primarily studying the use of association rules for retail marketing decision-making.

Gilbert Swinnen obtained his Ph.D. in applied economic sciences at the Universitaire Faculteiten Sint-Ignatius Antwerp (U.F.S.I.A.) and currently he is full professor of marketing research at the Limburg University Centre Diepenbeek. His research interests include retailing marketing, multivariate data analysis in marketing research and the use of data mining methods for the analysis of marketing databases.

Koen Vanhoof obtained his Ph.D. in computer science at the Catholic University of Leuven (K.U.L.) and currently he is full professor of business informatics at the Limburg University Centre Diepenbeek. His research interests include data mining, knowledge based systems, artificial intelligence and operations research.

Geert Wets obtained his Ph.D. in applied economic sciences at the University of Eindhoven (The Netherlands) and currently he is associate professor of business informatics at the Limburg University Centre Diepenbeek. His research includes data mining, decision tables, knowledge based systems, fuzzy set theory and verification and validation of knowledge based systems.