A Data Mining Framework for Optimal Product Selection in Convenience Stores

Tom Brijs, Gilbert Swinnen, Koen Vanhoof and Geert Wets
Department of Applied Economic Sciences
Limburg University Centre
B-3590 Diepenbeek, Belgium

***Abstract*-Previous research in the field of data mining has demonstrated that the technique of association rules is very well suited to find patterns in the purchase behaviour of customers. However, practitioners occasionally criticize that it is not straightforward to adopt the discovered knowledge for concrete retail marketing decision-making. This is partially due to the difficult integration of retail domain knowledge into the mining process which sometimes causes the discovered knowledge to be *sterile*. This paper makes an attempt at integrating category management knowledge into the knowledge discovery process in order to obtain more useful results, i.e. results that can better be used for concrete decision-making in retailing. More specifically, an integer programming model for product selection is proposed which takes into account cross-selling effects between products and also enables the retailer to integrate category management knowledge into the model. First results on real-world retail data demonstrate the success of the approach.**

I. 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 [1] 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. In this paper, it is defined as the understanding of all customers' purchasing behaviour as revealed through his or her sales transactions, i.e. *market basket analysis*.

Recently, the gradual availability of cheaper and better information technology has in many retail organisations resulted in an abundance of sales data. Hedberg [2] 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. The purpose of association rule discovery is to find items that imply the presence of other items in the same transactions, such as *diapers* $\Rightarrow$ *beer*, indicating that customers who buy diapers also tend to buy beer during the same shopping visit. While many researchers have contributed to the development of efficient association rule algorithms [3-8], literature on the use of this technique in concrete applications remains rather limited [9-11]. This partially depends on the fact that it is not always straightforward to convert the discovered knowledge into actionable commercial or marketing plans. Nevertheless, the widespread acceptance of association rules as a valuable technique to solve business problems will largely depend on its successful application on real-world data. This implies that patterns in the data are interesting only to the extent in which they can be used in the decision-making process of the enterprise to increase *utility*.

This paper deals with the issue of how association rules can be better integrated with domain-specific retailing knowledge in order to increase the utility of data mining results. More specifically, we will use the notion of *frequent itemsets* from association rule mining and integrate it into a micro-economic framework for product selection. At the same time, we want the framework to provide as many degrees of freedom to allow retailers to include domain-specific constraints that will increase the utility power of the proposed model.

The remaining part of the paper is organized as follows. Section 2 provides an overview of the technique of association rules. In section 3, we present the problem of measuring product interdependencies and introduce a product selection model based on the use of frequent itemsets. In section 4, we present the results of the empirical study. Finally, section 5 summarises our work and presents directions for future research.

II. Association Rules: Overview

A recent data mining technique for retail market basket analysis is *association rules*, introduced by Agrawal, Imielinski and Swami [3]. They provided the following formal description of this technique:

Let $I = \{i_1, i_2, \ldots, i_k\}$ be a set of literals, called items or also the product assortment of the retail store. Let D be a database of transactions, where each transaction T is a set of items such that $T \subseteq I$, i.e. T is a market basket. Associated with each transaction is a unique identifier, called its *TID*. We say that a transaction T *contains* X, a set of some items in I, if $X \subseteq T$. An *association rule* is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ holds in the transaction set D with *confidence* c if c% of transactions

search

in D that contain X also contain Y. The rule $X \Rightarrow Y$ has *support* s in the transaction set D if s% of transactions in D contain $X \cup Y$. Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than a user-specified minimum support (*minsup*) and minimum confidence (*minconf*).

The first step in generating association rules involves looking for so-called *frequent itemsets* in the data [12]. Indeed, the support of the rule $X \Rightarrow Y$ equals the frequency of the itemset $\{X, Y\}$. Thus by looking for frequent itemsets, we can determine the support of each rule.

Definition 1 (adapted from [12]) **Frequency of an itemset**
$s(X, D)$ represents the frequency of itemset X in D, i.e. the fraction of transactions in D that contain X.

Definition 2 (adapted from [12]) **Frequent itemset**
An itemset X is called frequent in D, if $s(X, D) \geq \sigma$ with σ the *minsup*.

A typical approach [12] to discover all frequent sets X is based on the insight that all subsets of a frequent set must also be frequent. This simplifies the discovery of all frequent sets considerably. Once all frequent sets are known, finding association rules is easy. Namely, for each frequent set X and each $Y \in X$, i.e. Y is a frequent subset of X, verify whether the rule $X \setminus \{Y\} \Rightarrow Y$ has sufficient confidence.

Definition 3 Confidence of an association rule
The confidence of an association rule $X \Rightarrow Y$ equals $s(\{X,Y\}, D)$ / $s(X, D)$.

To summarise, the technique of association rules produces a set of rules describing underlying purchase patterns in the data, such as *bread* $\Rightarrow$ *cheese* [support = 20% ; confidence = 75%]. Informally, the support of the association rule indicates how frequent the rule occurs in the data. The higher the support of the rule the more prevalent it is. Confidence is a measure of the reliability of an association rule.

III. Profset: a Product Selection Model

A. Problem Situation

Determining the *ideal* product assortment has been (and still is) the dream of every retailer. From the marketing literature [13], it is known that the optimal product assortment should meet at least 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 (unique selling proposition) and it is projected through 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 materialisation of the retailer's store formula. In many cases however, it is not individual products but rather product categories that are considered indispensable in order to comply with the store's image. Thinking in terms of product categories gets much more attention in recent years as a result of retailers to start thinking of product categories as being separate strategic business units (SBU's) [14]. For instance, for a prototypical convenience store, customers expect the store to carry at least one or more products from each of the following categories: milk, dairy products, bread, snack foods, meats, tobacco, paper products, soft drinks, beer and personal care items [15]. Absence of one or more of these product categories would cause basic expectations of customers about the store not to be met. We need to take this into account when constructing a product selection model. On the other hand, *added* products are chosen by the retailer to reinforce the store image and should be selected in order to maximise 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) [16]. 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.

The preceding qualitative considerations make clear that, to fulfil customers' expectations about the store's assortment, a product selection model should enable product and/or category constraint specifications to be easily added by the retailer. We will come back to this point in section III.C.3 where we elaborate on the formulation of the model.

Secondly, because retail organisations are profit seeking companies, the product assortment should be *quantitatively* appealing in terms of the profit that it generates for the retailer. This implies that both revenues and costs are required to build an accurate and realistic product selection model. In section III.C.1, these quantitative elements will be further defined.

Crucial in the above two criteria is the notion of 'product interdependencies'. Indeed, we believe that it is important to include *cross-selling effects* when selecting products for an 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.

B. Measuring Product Interdependencies: a Historical Overview

Since the idea of product interdependencies is crucial for the product selection problem, we believe that it is useful to provide a short literature overview on this topic. Moreover, the reader will notice that the use of frequent itemsets as an alternative method to measure product interdependencies can be better justified by examining the drawbacks of past techniques. In general, past techniques can be classified into two major categories: association coefficients and interaction parameters.

search

Association Coefficients

Already in the mid 70's and early 80's, in the marketing literature, Böcker [17] and Merkle [18] introduced a number of measures to investigate product interdependencies. Basically, association 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 ⇒ B and between B ⇒ C, then it is assumed that there also exists a relation between A ⇒ C. However, practical observations show that these assumptions are highly questionable. Furthermore, data storage problems are extremely cumbersome since calculating all association coefficients for some 5000 items in a small supermarket requires the construction of a (5000 x 5000)-matrix ! A similar idea as the one expressed by association coefficients is the Yule's Q-coefficient [19].

Interaction Parameters

A second family of measures for interdependence are the so-called interaction parameters that are frequently used in loglinear models to calculate joint purchase probabilities [20]. Although these models have a profound statistical background, they are limited in the number of products or categories they can handle. Mostly, only interactions between pairs of products or categories (first-order interactions) are included 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 individual products becomes problematic.

Frequent Itemsets: a Viable Alternative

Given the above drawbacks, we argue that frequent itemsets provide a viable alternative to the measurement of product interdependencies. First of all, because the measurement of interdependencies between products on the SKU[1]-level is empirically tractable. Secondly, because the frequent itemsets approach enables the discovery of higher-order interactions (interactions between more than two products). And finally, because problems with transitivity and symmetry are solved with the discovery of association rules. Indeed, association rules enable to distinguish between the confidence of the relationship A ⇒ B and B ⇒ A, i.e. symmetry is not assumed, and if A ⇒ B and B ⇒ C are supported, the association rules algorithm may still conclude that B ⇒ C does not meet the user-defined support and confidence thresholds, i.e. transitivity is not assumed.

Within the same body of data mining literature, a method to assess the *interestingness* [21] of association rules was introduced which enables easy interpretation in terms of the interdependency between products.

Definition 4 Interest

$s\,(X \Rightarrow Y)\,/\,(s\,(X) * s\,(Y))$

The nominator $s\,(X \Rightarrow Y)$ measures the observed frequency of the co-occurrence of the items in the antecedent (X) and the consequent (Y) of the rule. The denominator $s\,(X) * s\,(Y)$ 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 1 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.

TABLE I
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[2] effects between X and Y

Later on in this paper (see section IV.C.1) however, we will show that this measure alone is insufficient to determine the *real* interestingness of a product combination for the problem of product selection.

C. Construction of the PROFSET Model

According to the problem situation described above, a model should be constructed which is able to select a *hitlist* 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. More specifically, this background knowledge relates to category constraints specifying what categories, and how many or what products in each of them should (at least) be present in the final, optimal solution. It is the objective of the model to find the best set of products, i.e. the set of product that yields maximum profitability subject to the category and/or product constraints defined by the retailer. A solution that satisfies the above criteria will fulfil the requirements for a good product assortment, i.e. quantitative and qualitative attractiveness. In the PROFSET model, introduced in this paper, we implicitly take into account cross-selling effects by the use of frequent itemsets. Before specifying the microeconomic optimization model formally, we will first introduce the parameters and components of the model.

[1] SKU = Stock Keeping Unit (an individual product identification)

[2] Recall that substitutability indicates less than the expected level of mutual support.

search

1) *Model parameters*

Gross margin.

Let: T_j be a sales transaction generated at time j

SP_i be the selling price of product i

PP_i be the purchase price of product i

f_{ij} be the number of times product i was purchased in T_j

Definition 5 m_{Tj} is the gross margin generated by sales transaction T_j

$$\forall\, T_j:\; m_{Tj} = \sum_{i \in T_j} (SP_i - PP_i) * f_{ij}$$

Definition 6 M_X is the gross margin of frequent itemset X

$$\forall\, X:\; M_X = \sum_{j=1}^{\#\,transactions} m_j \quad \text{with} \quad \begin{cases} m_j = m_{Tj} \text{ if } X = T_j \\ m_j = 0 \text{ otherwise} \end{cases}$$

It is important to understand why X must equal T_j for m_j to be non-zero. The reason is that we will use the sum of all M_X to approximate the total profitability of the assortment. Now, suppose that $m_j \neq 0$ when $X \subseteq T_j$ instead of $X = T_j$ with $\{i_1, i_2\}$ a frequent itemset and $\{i_1, i_2, i_4\}$ a sales transaction. Clearly, $\{i_1, i_2\} \subseteq \{i_1, i_2, i_4\}$ but, because $\{i_1, i_2\}$ is frequent, it is known [2] that $\{i_1\}$ and $\{i_2\}$ must also be frequent[3]. Consequently, $\{i_1\} \subseteq \{i_1, i_2, i_4\}$ and $\{i_2\} \subseteq \{i_1, i_2, i_4\}$ and thus the gross margin generated by sales transaction $\{i_1, i_2, i_4\}$ will add to $M_{\{i1,i2\}}$, $M_{\{i1\}}$ and $M_{\{i2\}}$ even if i_4 is not selected for inclusion in the hitlist. Thus, if $m_j \neq 0$ when $X \subseteq T_j$, then a single sales transaction increases the M_X parameter of *all* the frequent itemsets that are contained in that transaction.

Thus, to summarise, 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. Consequently, X must be equal to T_j 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 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 C_i per product i can be obtained for all products.

2) *Model components*

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

Objective function.

The objective function represents the goal of the optimisation problem and therefore must reflect the microeconomic framework of the retail decision-maker. It is constructed in order to maximise the overall profitability of the hitlist. 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 $C_{i,k}$ associated with each individual product i,k, where the subscript i,k means the i-th product in product category k, contributes in a negative sense, but only if the product i,k is selected which is represented by a second type of boolean variable $Q_{i,k}$.

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,k are included in each frequent itemset X. This information can be obtained from association rule mining.

2. The *size* of the hitlist is specified by the *ItemMax* constraint.

3. One or more constraints related to category strategies developed by the retailer; these can be of a diverse nature. For instance, some categories of products mainly serve the purpose of *transaction building* and demand a high level of presence in the store while other categories may merely serve the purpose of *image building* such that the presence of only a few products of this category in the hitlist is sufficient.

3) *Model specification*

$$\text{Max } Z = \sum_{X=1}^{\#\,frequent\ sets} M_X * P_X - \sum_{i=1}^{\#\,prod.} \sum_{k=1}^{\#\,categ.} C_{i,k} * Q_{i,k}$$

s.t.

$$\forall\, X, \forall\, i,k \in X: Q_{i,k} \geq P_X \qquad (1)$$

$$\sum_{i=1}^{\#\,prod.} \sum_{k=1}^{\#\,categ.} Q_{i,k} = ItemMax \qquad (2)$$

$$\forall\, k: \sum_{i=1}^{\#\,prod.} Q_{i,k} \geq ItemMin_k \qquad (3)$$

with P_X and $Q_{i,k}$ booleans.

[3] Note that we use [..] to symbolize a frequent set and { ... } to symbolize a sales transaction.

search

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 however justified because it is highly probable that infrequent itemsets exist because of random purchase behaviour. Consequently, we claim that the objective function only measures the profit from structural, underlying purchase behaviour.

IV. Empirical Study

A. *Data Description*

The empirical study is based on a data set of 27148 sales transactions acquired from a fully-automated convenience 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, petfood, fruit, batteries, film supplies (camera, roll of film), which are displayed to the customer by means of an eight m2 window. More specifically, the product assortment of the store under study consisted of 206 different items, each of them assigned to one of 24 product categories by the retailer. Details about product categories and how many products are contained in each of them can be observed in table II.

TABLE II
Product Categories and Number of Products Included

Category Description	Number of Items	Category Description	Number of Items
Wine	7	Confectionery	12
Alcholic spirits	4	Divers	11
Dairy products	9	Candybars	9
Softdrinks + Fruit juices	23	Meat/salads	8
Bread	5	Hygiene products	11
Prepared meals	16	Snacks/appetizers	11
Beers	7	Cleaning products	2
General food items	18	Smokers' requisites	10
Baby products	2	Milky drinks	4
Canned food	9	Coffee/tea products	3
Chocolate items	15	Eggs	1
Biscuits	18	Petfood	2

The average sales transaction contains however only 1.4 different items because in this type of convenience store customers typically do not purchase many items during a single shopping visit. In fact, most of the items being sold are convenience and impulse products. With regard to the costs of each individual product in the assortment, detailed information on handling and inventory costs could not be obtained unfortunately so these will be considered equal for all products and therefore costs are not included in the model.

Basically, the empirical study involves two important phases. In the first phase, structural purchase behaviour under the form of frequent itemsets is discovered by using the data mining technique of association rules (section IV.B.). Then, in the second phase, the PROFSET method is used to select a hitlist of products from the assortment (see section IV.C.).

B. *Mining for Association Rules*

As the objective function in the PROFSET method requires frequent itemsets as input, frequent itemsets and association rules 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 being non-frequent, i.e. describing random purchase behaviour, 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 significantly 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 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, which will be explained in the next section.

C. *Product Selection (PROFSET)*

We believe that one of the strong points about the PROFSET model constitutes its ability to take into account cross-selling effects between products. This implies that the model selects products based on their positive interdependency with other products. To illustrate this, we will compare the results of our model with a heuristic that is frequently used by retailers which we will denote hereafter as the *product specific profitability heuristic*. The latter does not take cross-selling effects into account and merely considers the profitability generated by the product itself. Note that this may work counterproductive when putting together an 'optimal' assortment since products may be included that cannibalize each-other, although each of them individually may look interesting from the viewpoint of profitability. In contrast, the PROFSET model will try to maximally exploit cross-selling effects given some user-defined product and/or category constraints (i.e. retail domain knowledge).

search

Several parameter settings (*ItemMax* and *ItemMin$_k$*) were used to test PROFSET and compare it with the product specific profitability heuristic. Furthermore, since the presence of each product category was deemed necessary by the retailer to support the store's image, we forced the model to select at least one product (to be determined by PROFSET) from each category. This resulted in two important observations:

1. Given some retailer-specified category restrictions, PROFSET frequently selects products that are not top-sellers in their respective product category, i.e. they have a relatively low product specific profitability within their category but possess considerable cross-selling effects with products from the same or other categories that are contained in the optimal set.

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 optimal set.

Hereafter, both observations will be illustrated.

1) *Observation 1*

In order to make the comparison between PROFSET and the product specific profitability heuristic straightforward, we chose not to specify *basic* products (see section III.A.) in the model. Consequently, the model will be able to fully exploit cross-sales potential between items in the assortment, of course as long as category restrictions, specified by the retailer, are not violated (see bottom of previous page). For purposes of illustration, we chose *ItemMax* = 35, i.e. PROFSET must select the top 35 products from the assortment by taking into account cross-selling effects. This setting, for instance, might be appropriate to select a set of products to put into the centre of the convenience store's window in order to attract customer's attention and maximally promote cross-selling effects.

Table III ranks products from the smoker's requisites category in a descending order of product specific profitability and shows the products selected by PROFSET.

TABLE III
PRODUCT SELECTION IN SMOKER'S REQUISITES CATEGORY

Product	Product Specific Profitability (BEF)	Selection PROFSET
Marlboro	31030	X
L&M	22116	X
Drum tobacco	10353	
Belga	8892	X
Marlboro Light	8305	X
Bastos Filter	5819	
Boule D'Or	3426	
Cigarette paper	3258	
Barclay	1776	
Michel Green	400	

Table III shows that from the 35 products selected by PROFSET, 4 products were chosen from the smokers' requisites category. Furthermore, table III illustrates that Drum tobacco is not selected by PROFSET, even if it is known that Drum tobacco, on its own, generates 16% more profit than the Belga brand and 25% more than the Marlboro Light brand. The reason is that cross-selling effects between Belga or Marlboro Light and products from the same or other categories in the optimal set generates more profit in total than the combination Drum tobacco and other products in the hitlist. More specifically, replacing the Belga or the Marlboro Light brand by the Drum tobacco brand significantly decreases the profit generated by the optimal set by 3.1% and 2.7 % respectively. Replacing both the Belga and Marlboro Light brand by the combination Drum tobacco brand and cigarette paper, which is the next best cross-selling combination that is not included in the optimal set, places a less heavy burden on the profitability. However, the loss of profitability still mounts to 1.5 %.

The intensity of cross-selling effects can also be verified by examining the frequent sets and/or association rules generated in section IV.B. Careful analysis of the results showed that both the Marlboro Light and Belga brands have significantly higher cross-selling effects with beverage items than the Drum tobacco brand, as shown by the *support* of the frequent sets:

support {Marlboro Light and Beverages} = 0.27%
support {Belga and Beverages} = 0.20 %
support {Drum tobacco and Beverages} = 0.14 %

The above results illustrate that the statistical *interestingness* of a product combination (see definition 4) may not necessarily be a good measure to measure the *real* interestingness of associations in the case of a retailer. Indeed, when calculating the interestingness of the product combination Drum tobacco and cigarette paper (I = 76,7 >>1), both products seem to have high complementary effects with each other. However, because of their relatively low selling frequency with other products in the assortment, they were not selected by PROFSET. Instead, other products with higher complementary effects with other products in the assortment were chosen. This again shows that the micro-economic framework of the retailer ultimately determines the interestingness of product associations for the retailer.

2) *Observation 2*

One of the appealing properties of optimization models such as the integer programming model proposed in this paper is that the impact on total profitability caused by product assortment decisions can easily be assessed by means of sensitivity analysis. When for instance product $Q_{i,k}$ is deleted from the optimal set, and it is replaced by the best

search

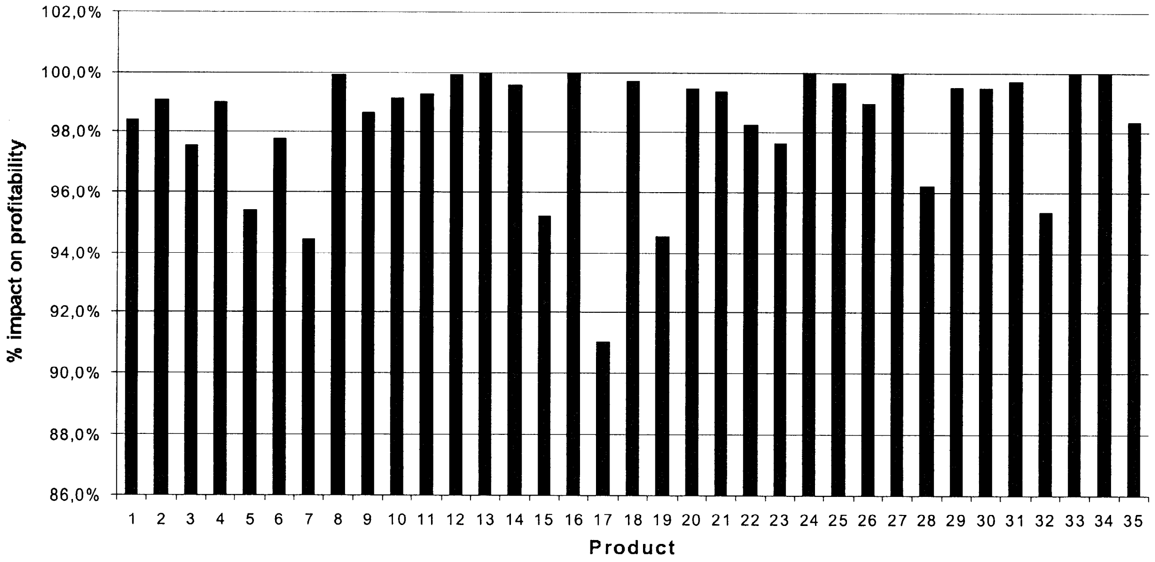

Figure 1. Profitability impact of product replacements

product $Q_{i,k}'$ outside the hitlist, its impact on profitability can easily be obtained from the optimisation model. Figure 1 illustrates this in a graphical way. One can observe that the impact on the profit of the optimal set fluctuates according to which product is being replaced. Whereas most products have only minor profit implications, some products (such as 5, 7, 15, 17, ...) should not be removed from the optimal set since they cause a heavy fallback in profitability of the hitlist. Furthermore, this kind of sensitivity analysis is useful for examining the implications of product removal operations on other products in the optimal set. Indeed, for this dataset, for two products their removal implied the deletion by PROFSET of another product from the hitlist. More specifically, the removal of product 22 (Chocolate milk drink) also implied the removal (by PROFSET) of product 13 (Kellog's variety pack) and the removal of product 23 (Sandwich cheese) also implied the removal of product 29 (Sandwich bacon) from the hitlist. Again, this illustrates that the PROFSET model takes into account cross-selling effects to execute an optimal product selection and that each alteration to the optimal set may have its repercussions on other items in the assortment.

V. Conclusions, Limitations and Future Research

A. 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 truely 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 can easily be carried out, enabling the retailer to quantitatively assess the profitability impact of product assortment decisions.

B. Limitations

The retailer should also consider the following limitations. Firstly, the presented model is deterministic in nature. This means that the model assumes that when for itemset $\{X, Y\}$ one of the items X or Y is not selected by the model, consequently all sales related to this itemset will be lost. This is of course too simplistic because 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, for instance as a result of customers switching over to substitute products.

C. 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, we have to confine ourselves to an analysis within or between categories.

search

Secondly, when sales transaction data from multiple stores with different product assortments but more or less the same underlying purchase behaviour 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. $X = T_j$), an alternative 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.

Acknowledgment

Tom Brijs is a research fellow of the Fund for Scientific Research Flanders (FWO-Vlaanderen).

References

[1] T. Blischok, "Every transaction tells a story", *Chain Store Age Executive with Shopping Center Age*, 71 (3), pp. 50-57, 1995.

[2] S. Hedberg, "The data gold rush", *BYTE*, October 1995, pp. 83-88.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", *Proceedings of ACM SIGMOD Conference on Management of Data*, 1993 (SIGMOD'93), pp. 207-216.

[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo, "Fast discovery of association rules", *Advances in Knowledge Discovery and Data Mining*: AAAI Press, 1996, pp. 307-328.

[5] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules", *Proceedings of the 20th International Conference on Very Large Databases*, 1994 (VLDB'94), pp. 487-499.

[6] S. Brin, R. Motwani, J. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data", *Proceedings ACM SIGMOD International Conference on Management of Data*, 1997 (SIGMOD'97), pp. 255-264.

[7] J. Park, M. Chen, and Ph. Yu, "An effective hash based algorithm for mining association rules", *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, 1995, pp. 175-186.

[8] M. Zaki, S. Parthasarathy, M. Ogihara, and M. Li, "New algorithms for fast discovery of association rules", *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 1997 (KDD'97), pp. 283-286.

[9] K. Ali, S. Manganaris, and R. Srikant, "Partial classification using association rules", *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 1997 (KDD'97), pp. 115-118.

[10] S. Anand, J. Hughes, D. Bell, and A. Patrick, "Tackling the cross-sales problem using data mining", *Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1997 (PAKDD'97), pp. 331-343.

[11] M. Viveros, J. Nearhos, and M. Rothman, "Applying data mining techniques to a health insurance information system", *Proceedings of the 22nd International Conference on Very Large Data Bases*, 1996 (VLDB'96), pp. 286-294.

[12] H. Mannila, "Methods and problems in data mining", *Proceedings of the International Conference on Database Theory*, 1997, pp. 41-55.

[13] W. Van der Ster, and P. van Wissen, *Marketing & detailhandel*, Wolters-Noordhoff, 1993.

[14] G. Cuomo, and A. Pastore, "A category management application in the frozen food sector in Italy: The Unilever-Sagit case", *Proceedings of the 10th International Conference on Research in the Distributive Trades*, 1999, pp. 225-233.

[15] R. Hasty, and J. Reardon, *Retail Management*, McGraw-Hill, 1997.

[16] E. Pessemier, "Retail assortments - some theoretical and applied problems", *Technical Report, Marketing Science Institute Research Program*, 1980.

[17] F. Böcker, "Die Bestimmung der Kaufverbundenheit von Produkten", *Schriften zum Marketing*, band 7, 1978.

[18] E. Merkle, "Die Erfassung und Nutzung von Informationen über den Sortimentsverbund in Handelsbetrieben", *Schirften zum Marketing*, band 11, 1981.

[19] M. Kendall, and A. Stuart, *The advanced theory of statistics: inference and relationship*. London: Charles Griffin and company Ltd, 1979.

[20] H. Hruschka, M. Lukanowicz, and C. Buchta, "Cross-category sales promotion effects", *Journal of Retailing and Consumer Services*, 6(2), 1991, pp. 99-106.

[21] C. Silverstein, S. Brin, and R. Motwani, "Beyond market baskets: generalizing association rules to dependence rules", *Data Mining and Knowledge Discovery Journal*, 2(1), 1998, pp. 39-68.