ASSORTMENT PLANNING USING DATA MINING ALGORITHMS

by

Ajlan Nihat Gün

Submitted to
The Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements
for the degree of
Master of Arts in Management Information Systems

Boğaziçi University

2008

The thesis of Ajlan Nihat Gün

is approved by:

Assistant Professor, Bertan Badur…….
(Thesis Advisor)

Assistant Professor, Sona Mardikyan…….

Dr, Halefşan Sümen …….

April 28, 2008

An abstract of the Thesis of Ajlan Nihat Gün for the degree of Master of Arts in Management Information Systems from The Institute for Graduate Studies in Social Sciences to be taken April 2008

Title: Assortment Planning Using Data Mining Algorithms

Assortment Optimization is not just selecting the best products according to the sales performance under a certain category, but also an execution method to apply retailers commercial strategy into market considering all strategies which retailer want to play. Regarding millions of data saved in databases and explosive growth of data leads to a situation in which it is increasingly difficult for retailers to understand the right information. To cope with this problem we are planning to use association algortihms to put in place data mining in product selection.

It should also be considered that selecting best and suitable products for assortment of retailer need not only sophisticated algorithms to take decisions but also business perspective to embed into decision system. In this study, we approach the assortment selection problem, by improving the PROFSET model and GENERALIZED PROFSET model, which is based on a microeconomic framework. We improved the basic model by introducing additional method of profit allocation over frequent item sets, constraints about categories and sold quantities. Finally we empirically test our model with sample retailer data. While doing this we will also take into consideration the retail industry characteristics and consumer and customer perceptions.

Sosyal Bilimler Enstitüsü'nde Yönetim Bilişim Sistemleri Yüksek Lisans derecesi için Ajlan Nihat Gün tarafından Nisan 2008'de teslim edilen tezin kısa özeti

Başlık: Veri Madenciliği Algoritmaları Yardımıyla Ürün Gamı Planlaması

Günümüzde ürün gamı optimizasyonu, belli kategoriye ait ürünler içerisinde satış performansı en iyi olanları bulmaktan ziyade perakendecilerin ticari stratejilerini pazarda uygulamaları için kullanılan bir metod haline gelmiştir. Veri tabanlarına her gün kaydedilen milyonlarca veriyi düşündüğümüzde gün geçtikçe çoğalan veriden gerekli bilgiyi çıkarmak perakendeciler için daha zor hale gelmektedir. Bu problemin çözümüne yardımcı olmak için ürün gamı seçiminde veri madenciliği uygulamalarını kullanmayı ve yoğun veriden bilgi edinmeyi amaçlamaktayız.

Şu da unutulmamalıdır ki perakendeciler en uygun ürün gamını seçerken sadece karmaşık algoritmalara güvenmemeli, iş mantığını da karar sisteminin içine yedirmelidir. Bu çalışmada biz bu konudaki önceki çalışmalardan PROFSET ve GENELLEŞTİRİLMİŞ PROFSET modellerine yeni geliştirmelerde bulunduk. Temel modeli sık alınan ürün kümesinin kar dağılım metodu, kategori başına sınırlamaları ve adetsel olarak bol satış yapan ürün kısıtlamalarında eklemeler ve düzeltmeler yaparak geliştirdik. Son olarak da yarattığımız modeli empirik olarak örnek bir perakende datası için test ettik. Tüm bunları yaparken perakende sektörü gerçekleri, tüketici ve müşteri algısını dikkatten kaçırmamaya özen gösterdik.

ACKNOWLEDGEMENTS

I would like to thank my dear family; my mother, my father and my sister. I am grateful to all others who helped me to revise parts of this thesis, particularly at the end of this project. I would also like to thank supervisors for their patience and full support during my thesis study.

I also would like to express my deep and sincere gratitude to my thesis advisor Bertan Badur for his support and valuable advises for my thesis.

CONTENTS

PREFACE ix

CHAPTER 1: INTRODUCTION.....1

CHAPTER 2: RELATED WORK 5
Category Management..... 6
Assortment Planning And Optimization.....10
Data Mining in Retail and Usage of Association Rules 13
PROFSET.....16

CHAPTER 3: PROBLEM DEFINITION.....21
Profit Allocation 23
Category Management Constraints 28

CHAPTER 4: EMPIRICAL STUDY 30

CHAPTER 5: CONCLUSION AND FUTURE RESEARCH.....41

REFERENCES 43

FIGURES

1. Algorithm of Profit Reallocation 25
2. Algorithm of Profit Maximization considering all constraints.....31
3. Code for optimization of profits for sample data.....40

TABLES

1. Summary of Association Mining Model in Clementine by product.....32
2. Results of Association Mining Model in Clementine by rules according to the product. 33
3. Iterative ranking results for first 45 products according to the last decision. 35
4. Summary of Association Mining Model in Clementine by subcategory.....37
5. Results of Association Analysis by rules according to the subcategories. 38
6. Sample table to show code of algorithm in Lingo 10.0 39

PREFACE

Compared to past decades, retailers' way of business is evolving. All well settled concepts from Supply Chain Management to Merchandising & Store operations have been changed. All these changes are evolved to support one single question. Which kind of products or services target consumers of retailers are demanding? Nowadays, to effectively answer this question "Category Management" is the new concept to be widely used. In detail, Assortment Planning is one of the key pillars of Category Management to understand and meet demand in the market.

On the other hand, in today's retailing realities, with the high turn-over and huge proliferation, the complexity of managing product assortment under a retailer product portfolio increases enormously. Compared to past, selecting products from a category or product-oriented point of view does not work considering high competition in retailing environment.

In an effective assortment planning process item selection is the key process to shape the vision of retailer for customers' insight. From a quantitative perspective item's performance and its cross-selling potential is very important for retailers. The difficulty is that the profit of one item not only comes from its own sales, but also reflect its influence on the sales of other items, which is called "cross-selling effect". Cross-selling effect also leads retailers to benefit from association mining.

Association mining, cross-selling effects, profit, basic and added product concepts are analyzed under PROFSET model and tested in a small retail chain data. PROFSET model also improved as GENERALIZED PROFSET model in category constraints and profit allocation area. Our approach will contribute to PROFSET not only adding new parameters to the decision making process but also differentiating profit allocation method of transactions over frequent item sets.

CHAPTER 1

INTRODUCTION

Compared to past decades, retailers are changing the way that they are doing their business. All main concepts from Supply Chain to Store operations have been changed. Although there are significant systematic effects such as the way how to keep data, transactions, how to analyze them, how to forecast the future, all these changes are evolved to support one single question. Which kind of products or services target consumers of retailers are demanding? Nowadays, to effectively answer this question "Category Management" is the new concept to be widely used. Category management (CM) has evolved to mean a process that involves managing product categories as business units and customizing them on a store-by-store basis to satisfy consumer demands. (Gnau et al., 1992)By another definition, CM describes a process whereby manufacturers and retailers come together and interact to create and manage strategies and operations for specific product categories – not just for individual brands (Araujo and Mouzas, 1998), (Dupre and Gruen, 2004). In scope of this term, one of the main levers to support CM process is Assortment Planning.

Marketing management and retailing textbooks highlight the importance of product assortment in achieving differentiation and satisfying the wants of target shoppers better than the competition (Kotler, 1997). It is obvious that no assortment planning process is capable of accounting for all of the marketing and operational implications of its decisions, due to limited data and the complexity of the task. (Cachon et al., 2007) In addition focus on the selection of products in a single category assuming store traffic is exogenous, *i.e.,* prices and variety within a category influences demand conditional on a store visit, but does not influence store choice. (Cachon et al., 2007) Due to reduction of variety in all categories based on single-category analyses, then the store becomes less attractive and some customers are likely to defect to other retailers, reducing store traffic. (Cachon et al., 2007) Considering effective Category Management principles, optimized Product Assortment should not only take into consideration best selling products within categories, but also should be realigned with the strategy of the retailer and reflect the choice and trend of consumers in a collaborative manner. That leads retailers to decide their product ranges not only according to qualitative factors but also according to quantitative constraints.

On the other hand, in today's retailing realities, with the high turn-over and huge proliferation, the complexity of managing product assortment under a retailer product portfolio increases enormously. Compared to past, selecting products from a category or product-oriented point of view does not work considering high competition in retailing environment. Since the approach should be more customer-oriented and strategy aligned retailers should understand other effects of each single products such as sales performance, inventory turnover, gross margin, cross selling effects, customer choice levers and local assortment opportunities.

In an effective assortment planning process item selection is the key process to shape the vision of retailer for customers' insight. From a quantitative perspective item's performance and its cross-selling potential is very important for retailers. The difficulty is that the profit of one item not only comes from its own sales, but also from its influence on the sales of other items, that is, the "cross-selling effect". (Su, 2002) Cross-selling effect also leads retailers to benefit from association mining.

Association mining, cross-selling effects, profit, basic and added product concepts are analyzed under PROFSET model (Brijs et al., 2004) and tested in a small retail chain data. PROFSET model also improved as GENERALIZED PROFSET model in category constraints and profit allocation area. (Brijs et al., 2001). Our approach will contribute to PROFSET not only adding new parameters to the decision making process but also differentiating profit allocation method of transactions over frequent item sets.

The rest of the thesis is built as follows. In Chapter 2, related work that attempts to solve the pillars of item selection according to customer oriented category management principles will be reviewed. To do this we divided the related work into Category Management, Assortment Planning and Optimization, Data mining in retail, Association Rules and PROFSET and GENERALIZED PROFSET, We reviewed all related works performed for these pillars one by one and specifically divide into different pillars Association Mining and PROFSET model since our suggested model will be customized version of GENERALIZED PROFSET. Then, in Chapter 3 we present an overview of our problem definition. We identify the assumptions, environment and data sources in our problem definition. In Chapter 4, we apply our model to the sample data and we highlight the differences with the previous studies to solve the item selection problems. Finally,

we conclude the thesis in Chapter 5 and also give possible directions for future research in this chapter.

CHAPTER 2

RELATED WORK

This chapter presents previously performed works related to the research herein. First section of chapter reviews the approach Category Management that attempts to manage retailers' categories as their brands. Second section of chapter provides a detailed look for Assortment Planning and Optimization which is a pillar of Customer oriented Category Management process. In the third section of chapter we explore Data Mining in Retail and usage of Association mining applications in previous studies. At last we will review PROFSET and related studies and approaches to PROFSET in fourth section of chapter. In this chapter, GENERALIZED PROFSET model, which is an improvement of the basic model, is introduced.

Category Management

Considering our main approach and basic question, “Which kind of products or services target consumers of retailers are demanding?” To reach this point retailers also try to find the answers of these questions. “Why do Consumers buy the category?” “Who buys the category?” “When do Consumers buy the category?” “How do Consumers buy the category?” “Where do Consumers buy the Category?” For sure most of the questions are out of scope of this study but they are really important to understand the vision of Category Management. CM is the big picture to leverage the target. As it is an evolving process since 1980s, for each decade the main focus is adapting itself according to the market conditions. CM concepts evolved as a key pillar in the development of retail sales strategy in the 1980s, and continued its growth through the 1990s with the availability of computerized sales data. In recent years, CM has become an important practice in enhancing business achievements by focusing on delivering increased value to the end consumer (Gruen and Shah, 2000), (Dupre and Gruen, 2004). As stated in the definition of Category Management Report, published by the Joint Industry Project on ECR, category management involves “the distributor/supplier process of managing categories as strategic business units, producing enhanced business results by focusing on delivering consumer value” (ECR, 1995). Retailers practicing category management focus on the category as the strategic business unit of analysis. Ensuring category leadership, increasing market share, revenues, and profitability are the key reasons cited by retailers practicing category management (Dhara et al., 2001). From this perspective Category Management seemed to be science-demanding insight of art for execution of the merchandising strategy of retailer. However, there is no limit for a

retailer to sell different categories at the same time. As resources are limited (direct monetary resources as spending budget or other resources such as workforce and shelf space), there should be a balance to optimize the distribution of resources and execute each category's strategy coherent with merchandising strategy.

At this point the need to differentiate the targets of categories is a must where comes "category role" concept to define. As stated by Dhara el. all (ECR, 1995) since consumer behavior and motivations can differ dramatically given what role the product plays in everyday life, the effectiveness of marketing actions should differ systematically across categories. These issues are key to any formal category management process where retailers must explicitly define the role that each category plays in the overall store portfolio. Effective category management requires that retailers understand where to allocate scarce marketing resources in order to get the biggest bang for the buck. (ECR, 1995) The key examples of category roles are "destination", "profit", "traffic builder" and "image creator". Although the roles of categories have been changing by market structure, sector, country and even by retailer destination categories are the categories of which makes majority of retailers' revenue. After defining role of categories, there comes building a strategy for each category by reviewing the category's past performance. The aim of building a strategy over a category is determining the specific actions that need to be taken to deliver the category roles and scorecard. "Building Traffic", "Building Transaction", "Profit focusing", "Cash Generating", "Image Creation" are relevant examples for category strategy. For example most of the grocery retailers are treating coke as a traffic & transaction builder to increase traffic inside the store and try to sell other profit generating product by the help of these types of categories.

After defining role of the categories, it is important to apply strategies managed by pillars. The strategy, which is defined according to the role of the category, leads the next level, “category tactics”. As stated by Lindblom and Okkonen (Lindblom and Olkkonen, 2006) Category Management strategies are applied by tactics which are divided into 4 main pillars:

. assortment planning;

. pricing;

. space allocation; and

. in-store promotional activity.

Retailers are usually assumed to be in charge of these four areas of CM decision making and, retailers who practice CM focus on the category as the strategic business unit of analysis. The key reasons cited by retailers for practicing CM are: “ensuring category leadership”, “ increasing market share”, “improving revenues” and “increasing profitability” (Gruen and Shah, 2000) , (Kurnia and Johnston, 2003). As we will focus on assortment planning as a matter of concept, it should be useful to give brief explanations and effects on assortment planning for other tactics. These tactics are also leading customer buying choices.

Pricing is one of the key factors why customer is buying a product from that retailer. As tactic, pricing is used to define pricing role of the categories according to the category role. Although each category tactic is shaped based on competition, pricing is the most important one. For instance, most of the traffic building categories are priced under or equal to the market conditions to retrieve traffic to the stores. Other pillar in category tactics is space allocation, which leads retailers to another optimization problem. Since there is a limited space inside the stores, all retailers should optimize their space resources according to their commercial strategy and

category roles. In addition, space is one of the main key performance indicators to take into consideration during category assessments. Usually retailers allocate more space destination categories compared to others. On the other hand, the average space usage of the category also affects the role of the category during strategy decision. (Selling white goods inside the stores require more investment than mobile phones) The last category tactics used by retailers to apply strategies are promotional activities. Although Gruen and Shah (Gruen and Shah, 2000) called this as in-store promotional activity, the term "in-store" does not reflect the real effort. These types of actions lead retailers to communicate through different categories and build synergy over categories. Since most of the promotional activities are referring cross selling opportunities, this pillar of category tactics is strictly related with assortment planning. Although assortment planning is the main process to select relevant items coherent with merchandising strategy, promotional activities are taken into consideration to catch cross-selling opportunities according to the strategy and category role. For instance independent from the sales performance cartridges should exist in range of a retailer (as accessories) if it is also selling printers. On the top of these entire tactics assortment planning is the most important one to focus on this study. All these tactics and assortment planning is targeting the success of one category. However, if collaborative retailer success is targeted all these tactics inside each category should also consider other categories success. The "Brand" of retailer would be built effectively if and only if all its sub-brands (categories) would be built coherent with the merchandising strategy.

Assortment Planning And Optimization

As Rajaram (Rajaram, 1999) stated assortment planning is the process conducted by a retailer to decide how many and which products to include in the product line and to determine the inventory levels of these products. The issues faced by the retailer in this process include estimating demand for each product, using this estimate to develop a profit function and choosing the best portfolio of products to maximize profits subject to budgetary and shelf space constraints. (Rajaram, 1999) Since this process also needs an optimization considering the available budget and space constraints, it has been evolved to an optimization problem. Regarding the category management tactics defining depth, size, price band and characteristics of the category is decided by assortment planning. Although the label of the concept is "assortment planning" the process is more assortment selection and item selection problem if the merchandise hierarchy structure is already defined according to the market conditions. Despite the first rule of customer centric category management is defining merchandise hierarchy of the products according to customer insights, all markets has a well-settled merchandise hierarchy. For example, Coke is always under Drinks category. On the other hand, it is useful to get the differentiating attributes from consumer surveys and analyze historical sales data according to this concept. In the marketing literature, different researchers from different perspectives have analyzed Assortment planning concept. Van Ryzin and Mahajan (Ryzin and Mahajan, 1999) and Kök and Fisher (Kök and Fisher, 2004) studied assortment selection in presence of substitutable products and considering inventory levels with enough store demand. Agrawal and Smith (Agrawal and Smith, 2003) added to the concept of market basket customers. Cachon et al. (Cachon et al., 2005) includes his

assortment planning study the assumption of non-buyer customer in case of wrong assortment. Chong et al. (Chong and Tang, 2001) present an empirically based modeling framework for managers to assess the revenue and lost sales implication of alternative assortments. Boatwright and Nunes (Boatwright and Nunes, 2001) tried to understand the effect of reduction in assortment without removing certain attributes.

Nevermore, assortment optimization has been lead category managers to review mainly sales performance of the items and rank them according to sales performance for many years. However, in high competition markets selling is not the solution to survive against competition. Retailers begin to consider also other effects like profit and handling cost of item (which leads inventory level) According to this consideration to select most sellable products and handling them in product assortment should be enough. On the other hand retailers should also take into consideration other factors. As studies show that consumers, most of which are not sure about the product while they are shopping inside the store, are requesting also assortment depth, retailers need to provide relevant variety inside offered products. So what must be done in an effective customer centric category management process as an assortment planning and optimization tactic is; try to offer products in each manner of customer requirement inside the same category. This idea also refers to perform customer researches and their parameters during assortment planning. As Cachon and Terwiesch (Cachon et al., 2005) stated even if a retailer chooses to implement a different analytical assortment planning process, that planning process must begin with a consumer choice model and some estimation of demand parameters. Simonson (Simonson, 1999) made a research to understand the effect of product assortment over buying preferences of consumers and grouped the effects in there psychological principles .One class of effects reflects the impact of product

assortment on the ease with which certain choices can be justified. A second category of effects are related to the ease of information processing, whereby consumers tend to prefer options with advantages that are more transparent given the manner in which the information is presented. And a third way in which product assortment influences preferences is by activating specific decision rules that consumers store in memory, such as the rule that variety is better than repetition. (Simonson, 1999)

Assortment selection and allocating limited retail shelf space is addressed in the literature. Bultez and Naert (Bultez and Naert, 1988) consider this problem by using marginal analysis of the profit function. They assume individual product demand is known and develop a generalized framework to allocate shelf space for an assortment with interacting demand. Based on this framework, they develop a practical allocation rule called Shelf Allocation for Retailers Profit (SHARP). (Bultez and Naert, 1988)

Another relevant concept with our study's subject is correlation between different items from categories to understand the effect of whole assortment of retailer. Since assortment planning and optimization target maximum profit for not only category manager but also retailer, the range should maximize retailers' earnings. Advancement in ways to process and store data (as complicated databases and specially data warehouses) lead retailers to perform market basket analysis, in which historical POS (Point of Sales) data is used to understand which items customers tend to buy together and because of that retailers would be able to better plan cross-merchandising, promotions and adjacencies. This process will be reviewed in more detail in following chapters and also PROFSET model will be reviewed to adjust our problem definition. After market-basket analysis the following

stage regarding assortment planning is deciding which type of products should be handled in retailers assortment together to increase sales synergy and which solution should be settled down to retrieve data more effectively and to solve the multiple buying needs of the consumer.

Data Mining in Retail and Usage of Association Rules

By definition data mining is a set of automated techniques used to extract buried or previously unknown pieces of information from large databases, using different criteria, which makes it possible to discover patterns and relationships. Ahmed, (Riaz, 2004) Han and Kamber (Han and Kamber, 2000) described data mining as the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Due to the enhancements of computer configurations the data mining technology is now more useful compared to past while answering critical and time-consuming questions. Data mining tools search databases for hidden patterns, finding predictive information that experts may miss because it was outside their expectations. (Riaz, 2004) Data mining usually yields five types of information in retail: Associations, Sequences (events are linked over time and these links are extracted), Classifications, (classification can help you discover the characteristics of customers who are likely to leave and provides a model that can be used to predict who they are) Clusters (the data-mining tool discovers different groupings with the data.), Forecasting (forecasting, is a different form of prediction. It estimates the future value of continuous variables — like sales figures — based on patterns within the data) (Riaz, 2004) Ahmet (Riaz, 2004) also emphasized the main usage of data mining in retail improving their

inventory logistics and thereby reduce their cost in handling inventory, identifying the demographics of its customers and the products that they buy which can be extremely beneficial in stocking merchandise in new store locations as well as identifying "hot" selling products in one demographic market that should also be displayed in stores with similar demographic characteristics. For nationwide retailers, this information can have a tremendous positive impact on their operations by decreasing inventory movement as well as placing inventory in locations where it is likely to sell.

Definition of association rules can be summarized by Brijs et al. (Brijs et al., 2004) as following:

"The technique of association rules produces a set of rules describing underlying purchase patterns in the data, like for instance bread 3 cheese [support = 20%; confidence = 75%]. Informally, support of an association rule indicates how frequent that rule occurs in the data. The higher the support of the rule the more prevalent the rule is. Confidence is a measure of the reliability of an association rule."

Although the other four types of information is also used for extracting useful knowledge we will focus on Association for Assortment optimization process. By another name, association rule mining (Agrawal et al., 2003) aims at understanding the relationships among items in transactions or market baskets. However, it is generally true that the association rules in themselves do not serve the end purpose of the business people. It is believed that association rules can aid in targets that are more specific. Recently, some researchers (Brijs et al., 1999) suggest that association rules can be used in the item selection problem with the consideration of relationships among items. Given the item selection considering cross selling effects, Wong et al., (Wong and Wang, 2005) model the item selection definition and

enhanced its previous model of Maximal-Profit Item Selection with cross-selling effect (MPIS) which is the problem of finding a set of J items with the consideration of the cross-selling effect such that the total profit from the item selection is maximized, where J is an input parameter. Wong also modeled the MPIS with the assumption of transaction history records is given for uncovering customer behaviors.

During these studies of association mining the challenging part of these relations extracted from business is the business applicability of the rules. As Cabena et al., (Cabena et al., 1997). declared, 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. Brijs et al., (Brijs et al., 2004) gave relevant instance about this issue. 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. So the new model for assortment optimization should also take into consideration business constraints of retailer to support retailers' image, also maximizing profit function for all categories. To perform this analysis the guidance methodology will be PROFSET to understand product interdependencies and selecting relevant items to the assortment to maximize the profit of retailer. In addition, the assortment optimization model in this study will be based on PROFSET.

PROFSET

To give more detail about the base methodology of assortment optimization according to the product interdependencies and profit maximization we will cite the relevant definitions and explanations about PROFSET. Mainly PROFSET model takes into account cross-selling effects by using frequent itemsets. PROFSET was developed to maximize cross-selling opportunities by evaluating the profit margin generated per frequent set of product s rather than per product. It uses the profitability per frequent set to determine the optimal selection of products in terms of maximal total profit. Below we will introduce the parameters and components of the model taken form the original source. (Brijs et al., 2004)

- ***Model Paremeters***

Gross margin:

Let: I = {i1, i2, . . . , im} be a set of items in a retail store (*i.e.,* the product assortment)

D be a set of transactions, where each transaction T ⊆ 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

m(T) is the gross sales margin generated by sales transaction *T*.

$$m(T) = \sum_{i \in T} (SP(i) - PP(i)) * f(i) \tag{1}$$

M(X) is the gross sales margin generated by frequent itemset *X*.

$$\sum_{T \in D} m'(T) \text{ with } \left\{ \begin{array}{l} m'(T) = m(T) \, if X = T \\ m'(T) = 0 \quad otherwise \end{array} \right\} \tag{2}$$

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. Since cost components of the function could be very difficult to obtain, for reasons of simplicity, it is assumed that a total cost figure $Cost_i$ per product i can be obtained for each product.

- ***Model Components***

The PROFSET optimization problem is operationalized by means of an integer programming model containing two important components: 1) Objective Function; 2) Constraints.

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 hitlist.

There are 3 main constraints;

1. Because the final decisions need to be taken on the product level instead of on the frequent item set level, it must be specified which products i are included in each frequent item set X. This information can be obtained from association rule mining.
2. Basic products (the products which should exist in the assortment mandatory) can be specified by forcing the model to select certain products.

3. The size of the hitlist is specified by the ItemMax constraint. (Maximum number of item inside assortment

- ***Model Specification***

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. Consequently, it is claimed that the objective function only measures the profit from structural, underlying purchase behavior.

- **Generalized Profset**

Although main idea of taking into consideration of frequent item sets while making assortment is raised by PROFSET model, some improvements are suggested with GENERALIZED PROFSET MODEL. Methods in allocation of profit and category management restrictions are two main improvements in GENERALIZED PROFSET MODEL. (Brijs et al., 2001) Although PROFSET is taking gross margin values for itemsets and improve the decision compared to association mining algorithms, there is still gap and PROFSET algorithms were not capable to handle supermarket data, rather than supermarket data it was suitable for convenience stores of which average number of items inside one shopping basket is very low. However, rather than taking into account transactions which are equal to frequent itemsets, GENERALIZED PROFSET support whole transaction which includes frequent itemsets. In addition, GENERALIZED PROFSET uses some constraints to support mission of retailers to support to provide wide range of products to their customers. (Brijs et al., 2001) So considering these differences from PROFSET, model parameters of GENERALIZED PROFSET is differentiating it self in 2 main areas;

The first one is about allocating gross margin to frequent itemsets. While PROFSET takes into account only transactions which are equal to frequent itemset,

GENERALIZED PROFSET also includes transactions which have frequent and non frequent itemsets in the same transaction as well. However, while allocating gross margin per transactions GENERALIZED PROFSET use probabilistic approach.

Definition: Let: F be the collection of all frequent subsets of a sales transaction T_J. Then X $\in$ F is called *maximal* denoted as $X_{max,}$ if and only if $\forall Y \in F : |Y| \leq |X|$

Using this definition, following rationale is adopted in GENERALIZED PROFSET to allocate the margin $m(T_J)$ of a sales transaction $T_{J.}$ (Brijs et al.,2001)

If there exists a frequent set $X = T_{J,}$ then $m(T_J)$ is allocated to *M(X)* in the PROFSET model. However, if there is no such frequent set, then one maximal frequent subset *X* will be drawn from all maximal frequent subsets according to the probability distribution Θ_{Tj} , with

$$\Theta_{Tj}(X_{max}) = \frac{\sup port(X\max)}{\sum_{Y\max \in Tj} \sup port(Y\max)}$$

After this, the margin m(X) is assigned to M(X) and the process is repeated for $T_J \backslash X$.

In summary:

For every transaction T_J **do {**

While (T_J contains frequent sets) **do {**

Draw X from al maximal frequent subsets

Using probability distribution Θ_{Tj} ;

M(X):= M(X) + m(X)

With m(X) the profit margin of X in T_J ;

$T_J := T_J \backslash X$;

}

}

Return all M (X); (Brijs et al.,2001)

Regarding second point, which GENERALIZED PROFSET algorithm give chance to users to define the minimum number of items per category. Considering these two improvements maximization function and the constraints for GENERALIZED PROFSET can be summarized as; (Brijs et al.,2001)

Let categories C_1, …., C_n be sets of items, L the set of frequent itemsets, and let P_X , $Q_i \in \{0,1\}$ be the decision variables for which the optimization routine must find the optimal values. P_X specifies whether an itemset X will positively contribute to the value of the objective function, and Q_i equals 1 as soon as any itemset X in which it is included is set to 1 ($P_X = 1$) by the optimization routine. Let $Cost_i$ be the inventory and handling cost of item i. And the objective of the following formula is to maximize all profits from cross-selling effects between products. (Brijs et al.,2001)

$$\text{Max } \left(\sum_{x \in L} M(X)Px - \sum_{c=1}^{n} \sum_{i \in Cc} Cost_{\,i}\, Q_i\right)$$

which is subject to the following constraints

$$\sum_{c=1}^{n} \sum_{i \in Cc} Q_{\,i} = \text{ItemMax} \qquad (1)$$

$$\forall X \in L, \forall i \in X : Qi \geq Px \qquad (2)$$

$$\forall Cc : \sum_{i \in Cc} Qi \geq ItemMin_{\,Cc} \qquad (3)$$

(Brijs et al.,2001)

CHAPTER 3

PROBLEM DEFINITION

Assortment optimization is one of the most important tasks, which drive sales. Most importantly, assortment should fit with retailers' commercial strategy which is not easily measured in a quantitative manner. On the other hand assortment should fit sales and profit targets of retailer to survive. The assortment should also be in accordance with retailers' constraints such as budget, space, variety in every category.

In this study, our problem is to find an optimized assortment which will maximize retailers' profit in a specified time period by considering various constraints faced by key business users together with cross-selling effects. Our approach is similar to the one used in the solution of the PROFSET and GENERALIZED PROFSET models. As it is stated, in fourth section of this chapter the PROFSET model suggested cross selling effect measurement and then according to basic constraints (basic product, maximum number of item) generates final product assortment by linear programming. (Brijs et al., 2004) In a further

improvement of PROFSET, GENERALIZED PROFSET model try to enhance constraint limitation from category management point of view (like product limits per category to support diversity) and also profit allocation (Brijs et al., 2001). In a similar manner, we suggest three improvements. First improvement is about profit allocation enhancement; second one is about introducing product constraints per category and third one is about volume generation. So our proposed approach will try to leverage better assortment distribution compared to the baseline solution. To reduce the complexity of the problem we take into consideration one list price and one product cost which also includes supply chain costs. So it is assumed that retailers are already purchased these items. The model will just help them in "what if" studies. Basically the problem is: retailers have limited space for product facing. Initially they were selling Q different items in their stores, now it is decided to standardize the store format and for that reason number of facing area decreased compared to past. So the new number of items to be faced in the store is M which is smaller than Q., In the common business practice category director should ask to their managers to have a meeting and decide which items should be faced for next months. As retailers give high importance to product profitability, without any consideration of cross-selling effects, the first M products should be selected sorted by gross margin in descending order.

We discuss our contributions to the assortment optimization process in the following subsections.

Profit Allocation

In the GENERALIZED PROFSET model, the profit; generated from products according to the sales and cost of good sold design does not reflect the retailer's needs and consumer demand. Sales of items in a basket could be easily affected by their sales. It is much more important to understand what the trigger of the specified sales transaction is. What are the effects of other items in sales transaction of a specified item? The best practice to obtain this information in retailer's database is asking to customers during shopping. However, this option does not fit retailing realities. It is nearly impossible to retrieve this information during shopping of a customer. So purchasing behavior of customer should be extracted from sales records according to predefined assumptions. Frequent item set approach is raised to solve this problem. Both PROFSET and GENERALIZED PROFSET models used this approach to allocate profit generated from non-frequent items to frequent-items. In this study we suggest a new distribution model for profit allocation, which is driven by frequent item sets. Firstly, we generated frequent item sets.

If D is a database of shopping baskets and X is a set of products (i.e. an item set), then the frequency of this item set X can be expressed as in Definition 1.

Definition 1. s(X, D) represents the frequency of item set X in D, i.e. the fraction of shopping baskets in D that contain X. Consequently, if the frequency of the item set X exceeds a user-defined frequency threshold σ, then this item set X is called frequent.

After identifying frequent item sets, profits per transactions should be allocated according to some criteria. In PROFSET whole profit is allocated to

frequent item sets and it is assumed that all profit is owed to frequent item set. (Brijs et al., 2004) In GENERALIZED PROFSET, allocation is made by taking into consideration support of frequent item set. However, it is also assumed that if there is more than one frequent item set in a transaction, one item set should be chosen according to the frequency of occurrence and profit should be allocated over relevant item set probabilistically. (Brijs et al., 2001)

Different from PROFSET and its GENERALIZED version, our approach is based on iteratively distributing portion of whole transaction's profit over frequent item sets. After scanning all records and identifying the transactions with specified frequent item set, our algorithm allocates a portion of the whole transaction's profit to distribute over frequent item set. So that for each transaction that contains frequent item sets, its whole profit is redistributed over items that are included in those frequent item sets. The sorting criteria among item sets are rule support of item sets. Before this operation, item sets should be sorted by rule support in a descending order. It is obvious that as number of frequent item sets increases iterations will cause time and performance problem. So in our empirical study we gave a threshold for rule support to perform this as a demo version. After last iteration, we obtained new version of gross margins per item and items are ready to be run in linear programming, for maximization of whole profit of retailer also considering other constraints. Our basic algorithm for profit reallocation is shown in figure 1.

(1) Find frequent item sets

By specifying minimum support, confidence; minimum and maximum number of items inside item sets

(2) Sort the frequent item sets in descending order according to their rule support

(3) For each item set X_i (each frequent itemset means new iteration) repeat

(a) Select transactions which includes X_i

(b) For each selected transactions

Reallocate profits from non-frequent items to frequent items

Figure. 1 Algorithm of Profit Reallocation

For each iteration of our algorithm, our suggested profit reallocation is described as follows.

During this explanation we will also illustrate our case by an example. Let we illustrate the definitions above with a simple example. Let we have 3 items apple, orange, banana, melon for all our item sets.

So by definition D = { apple, orange, banana, melon} and let X_1 = { apple, orange}

Definition 2. Let F be the collection all selected frequent item sets, T be collection all transactions and D be the collection of items in the assortment. And X_i the selected frequent item set in the iteration where X_i exists in transaction $T_{i,r}$ where $X_i \in F$ and $X_i \in D$. T_i is defined as transaction lists that includes item set X_i. And lines in transaction are defined $T_{i,r,q}$ which is composed of $X_{i,r,q}$ and $Y_{i,r,q}$.. Also $TG_{i,r,q}$ as $XG_{i,r,q}$ and $YG_{i,r,q}$ are the profits for each item in a transaction.

Let $Y_{i,r}$ be the item sets in transaction T_{ir} of which $Y_i \neq X_i$. On the other hand $X_i \cup Y_i = D$ and $X_i \cup Y_{i,r} = T_{i,r}$

There are also other transactions $T_{j,r}$ in data of which does not include selected frequent item set X_i . $T_{i,r}$ U $T_{i,r}$ = T

Till here i refers to frequent item set order, r refers to index of transactions in each transaction set, q refers to items in a transaction. For example, $T_{i,r,q}$ refers to the qth item in r^{th} transaction of transaction set Ti. Let also each frequent item set has a rule support of S_i and total gross margin (which can be formulized as sales revenue – cost of good sold) $G_{i,r.}$ And Let $M_{i,r}$ be the gross margin of $X_{i,r,}$ $XG_{i,r,q}$ be the gross margin of each item in $X_{i,r}$ and $YG_{i,r,q}$ be the gross margin of each item in $Y_{ir.}$

So by definition D = { apple, orange, banana, melon} and let X_1 = { apple, orange} Our transaction database Ti=1 includes T_{11}={ apple(40), orange(60), banana (100) }, T_{12}={ apple(40), orange (60)}, T_{13}={ apple(40), orange(60), melon (80)} The number in brackets shows the gross margin value per item for each transaction. So all transactions includes frequent itemset of X_1 = { apple, orange} in our example. To illustrate the definitions above and match with our example.

$T_{1,1,1}$= refers to apple which is 1st item of first transaction of first iteration.

$TG_{1,1,1}$= refers to apple's gross margin = 40

Y_1= refers to {banana, melon} in itemset D which illustrates non-frequent itemset for whole transactions of 1st iteration.

$XG_{1,1,apple}$= refers to apple's gross margin for 1st iteration's first transaction.

$YG_{1,1,banana}$= refers to banana's gross margin for 1st iteration's first transaction. Please consider that *banana* is not member of frequent itemset..

$G_{1,1}$= refers to 1st transaction of first iteration's total gross margin which is 40+60+100=200

$M_{1,1}$= refers to 1st transaction of first iteration's total gross margin for frequent itemsets in transaction, which is 40+60=100

Definition 3. For each transaction in which a frequent item set exist ($X_{i.}$) there should be an allocated value to be distributed over frequent item set, deduced from $YG_{i,r}$. So let we define this value as $TM_{i,r.}$

$$TM_{i,r} = G_{i,r.} * S_i \quad (3)$$

Till here we defined the value $TM_{i,r}$ to be reduced from $G_{i,r}$ - $M_{i,r}$ and distributed over $M_{i,r}$ according to the weight of $XG_{i,r,q}$ over $M_{i,r.}$ In parallel with this assumption $TM_{i,r}$ will be deduced from $YG_{i,r}$ according to the weight of $YG_{i,r,q}$ over $YG_{i,r.}$

If we continue to our sample case and let assume support of itemset X_1 ={apple, orange}=0.05 => S_1

$TM_{1,1} = G_{1,1} * S_1 = 200 * 0.05 = 10$ which refers to 10 will be reduced from $G_{1,1} - M_{1,1}$ (200-100=100) and distributed over $M_{1,1}$(100) according to the weight of $XG_{1,1,apple}$ over $M_{1,1.}$ (40/100=%40) In parallel with this assumption $TM_{1,1}$ (10) will be deduced from $YG_{1,1}$ according to the weight of $YG_{1,1,banana}$(100) over $YG_{1,1}$(100) (100/100=%100)

Definition 4. After these findings we will have a pure profit to distribute again, $P_{i,r}$. This pure distribution value depends to the item's label, if $T_{i,r,q}$ is a frequent item in X_i. The formulation is; *Pir* with

$$\left\{ \begin{array}{lll} Pir & = & Gir \\ Pir & = & Gir - TMir \end{array} \right. \quad if \quad \left. \begin{array}{lll} Tirq & = & Xirq \\ Tirq & = & Yirq \end{array} \right\} \quad (4)$$

Definition 5. After defining pure distribution value we will calculate the result of iteration. For each iteration i (the number of selected item sets according to the support) $M_{i,}$ starting from 0 as of beginning profit of line will be iterated and changed

item set by item set and $TG_{i,r,q}$ will be updated at the end of first iteration as shown equation 5. So the formulation of $TG_{i,r,q}$ for i^{th} iteration will be

$$TG_{i,r,q} \text{ with } \begin{cases} (TGirq \;\; / \; Mirq \;\; * \; TMirq \;\;) + TGirq & if \; TGirq = XGirq \\ TGirq \;\; / \; Gir \;\; * \; Pir & \quad\;\; TGirq = YGirq \end{cases} \quad (5)$$

After first iteration, $TG_{1r,q}$ should be replaced with $TG_{2,r,q}$ and iteration should continue until the rule support threshold for frequent item set is achieved.

For our case example, $TM_{1,1} = 10$ will be distributed over $XG_{1,1,apple}$ (10 * 40/100 = 4) and $XG_{1,1,orange}$ (10 * 60/100 = 6). So the new gross margins for transaction T1 for iteration 2 will be;

T_{21}={ apple(44), orange(66), banana (90) }

As other transactions will also be updated like this , we will have new distributed gross margins.

Category Management Constraints

One of the limitations of the GENERALIZED PROFSET model was not considering category diversity over the whole assortment. As we just described cross selling effect over items, different category offerings attract customers to the store and help for other categories to be sold. Although this limitation is also removed in GENERALIZED PROFSET model by adding some constraints, the model does not suggest a qualitative formulation to select category dependency. We suggest using association mining to gain insight about dependent categories. Instead of selecting transactions and items to extract dependency, the association analysis is done at the category level rather then the item level using higher minimum support values to gain insight about which categories are purchased together.

Hence, we will obtain frequent category sets that will be included into the model as number of minimum items per category. In our model we formulated it as if there is a frequent category set, each component of this category set should contain coherent number of items..

In addition to category based constraints as used in GENERALIZED PROFSET and "basic product" constraint used in PROFSET model, our model also offers another set of constraints which makes sense in retail environment. Till here the quantities sold in one transaction for a specific item is not considered. However, huge volumes in terms of quantities generate traffic inside the store and also increase turnover of items which is an important KPI (key performance indicator) for every retailer. Because of this reason we performed a Pareto Analysis for quantities sold per item and put constraints for items which generate 40% of whole sales in terms of quantity sold.

After considering the constraints above the last objective function should be defined before applying all findings in an empirical environment.

CHAPTER 4

EMPIRICAL STUDY

Our empirical study is based on a data set of 142,208 market baskets obtained from 3 stores of an electronic retailer in one year. Total number of transaction-item combination is 692,340 and average number of items per basket is 4.86. These transactions are performed by 6130 different customers. There are 72 different items in 5 categories and 22 subcategories. As stated above our problem is that the retailer decided to standardize all its stores into one unique format which will decrease the maximum area of facing and decreases the itemmax parameter to 45 where itemmax is the maximum number of items in retailers' assortment. So the retailer should decide which items should be selected in their assortment. Without any business constraint the model should sort items by their generated profit and then select the top 45 items for their assortment. However, regarding cross selling effects, basic products, category dependency and volume generating products constraints our profit allocation and linear programming formulation is applied to this dataset. The main pseudo code for problem solution can be reviewed in Figure 1.;

(1) Select most profitable solution over 72 items from previous assortment model

The ranking of items according to the gross margin has been analyzed above and the last ranking is defined after 5th iteration.

(2) The limitation for number of products in new store environment is 45.

(3) The best selling items in terms of quantity which makes %20 of whole sold quantity volume should not be discarded after analysis.

(a) Use the results of Pareto Analysis made as an extra analysis and define items which makes %20 of whole quantity volume.

(4) The correlated subcategories according to the association mining results should include similar number of products not to lose its dependency.

(a) Use the results of Association mining for subcategories to identify relevant subcategories and equalize the number of items inside each subcategory.

(5) Review the results.

Figure. 2. Algorithm of Profit Maximization considering all constraints

To avoid complexity and see the results in a limited demo version we chooses our critical support limit as 7.3% and minimum confidence limit as 40%. The summary of the association analysis by product can be found in Table 1. The other relevant results of Association Analysis can be found in Table 2. There are 2 limits for frequent item sets. Minimum number of items in a frequent item sets is 2 and that of maximum number is 5. After performing association analysis in SPSS Clementine (http://www.spss.com) we obtained 5 different frequent item sets and we performed 5 different iterations to obtain reallocated profit per item. After Performing Pareto Analysis for quantities sold per item and defining category dependencies by the help

of association algorithm, we obtained constraints of our linear programming. We will just highlight four points to illustrate the improvement of our model. To find frequent item set and association rules we used SPSS Clementine 11. To solve profit allocation problem we used PLSQL functionalities and to solve linear programming Lindo 9.0 (http://www.lindo.com) is used.

Table 1. Summary of Association Mining Model in Clementine by product.

Analysis
Number of Rules: 40
Number of Valid Transactions: 142,208
Minimum Support: 10.016%
Maximum Support: 15.529%
Minimum Confidence: 40.064%
Maximum Confidence: 78.835%
Minimum Lift: 2.619%
Maximum Lift: 8.23%
Minimum Deployability: 0.0%
Maximum Deployability 0.0%
Minimum Rule Support: 0.0%

Table 2. Results of Association Mining Model in Clementine by rules according to the product.

Consequent	Antecedent	Support %	Confidence %	Rule Support %
48 = T	40 = T	15.127	66.214	10.016
40 = T	48 = T	15.298	65.475	10.016
30 = T	33 = T	12.336	78.835	9.725
33 = T	30 = T	15.529	62.625	9.725
30 = T	31 = T	13.079	58.181	7.609
31 = T	30 = T	15.529	48.999	7.609
48 = T	30 = T	15.529	47.274	7.341
30 = T	48 = T	15.298	47.989	7.341
116 = T	119 = T	10.401	70.428	7.325
25 = T	23 = T	10.112	72.121	7.293
117 = T	119 = T	10.401	69.157	7.193
26 = T	23 = T	10.112	65.285	6.602
131 = T	130 = T	10.897	60.166	6.557
40 = T	30 = T	15.529	41.27	6.409
30 = T	40 = T	15.127	42.367	6.409
114 = T	119 = T	10.401	61.443	6.391
133 = T	130 = T	10.897	57.43	6.258
42 = T	40 = T	15.127	40.112	6.068
132 = T	130 = T	10.897	54.933	5.986
120 = T	130 = T	10.897	54.462	5.935
130 = T	120 = T	10.963	54.137	5.935
131 = T	120 = T	10.963	53.733	5.891
113 = T	120 = T	10.963	53.451	5.86
48 = T	33 = T	12.336	47.033	5.802
33 = T	31 = T	13.079	42.895	5.61
31 = T	33 = T	12.336	45.477	5.61
133 = T	120 = T	10.963	50.071	5.489
132 = T	120 = T	10.963	49.846	5.465
28 = T	23 = T	10.112	53.143	5.374
30 = T	40 = T and 48 = T	10.016	52.57	5.266
113 = T	130 = T	10.897	47.364	5.161
24 = T	23 = T	10.112	50.848	5.142
118 = T	119 = T	10.401	49.429	5.141
40 = T	33 = T	12.336	41.31	5.096
115 = T	119 = T	10.401	45.731	4.756
27 = T	23 = T	10.112	44.284	4.478
42 = T	40 = T and 48 = T	10.016	44.012	4.408
48 = T	120 = T	10.963	40.064	4.392
41 = T	40 = T and 48 = T	10.016	42.867	4.294
33 = T	40 = T and 48 = T	10.016	42.004	4.207

First observation is about frequent item sets constraint. Although Item 119 (CD-R with Cases) was ranked as 50th based on generated profit after 5th iteration it is selected for assortment since it is included in frequent items.

Second observation is about comparison of ranked items before iterations and after iterations. For example *item id 30* (Mouse Pad) is ranked as 53rd without any iteration and generated profit was 51,156.09 NTL.(new Turkish Lira) However, after distributing profit in other items real profit generated by this item would increase to 416,471.39. In Table 3. ranking of first 45 products (acc. To the last iteration) for each iteration can be found.

Table 3. Iterative ranking results for first 45 products according to the last decision.

PRO ID	Prod Desc	GMR	Itr 5 Rank	Itr 0 Rank	Itr 1 Rank	Itr 2 Rank	Itr 3 Rank	Itr 4 Rank
18	Envoy Ambassador	3,757,039.95	1	1	1	1	1	1
17	Mini DV Camcorder with 3.5" l LCD	2,196,249.87	2	2	2	2	2	2
14	17" LCD HDTV Tuner	1,818,258.11	3	3	3	3	3	3
13	5MP Telephoto Digital Camera	1,499,316.22	4	4	4	4	4	4
21	18" Flat Panel Graphics Monitor	1,039,482.98	5	5	5	5	5	5
20	Home Theatre Package with DVD-Audio/Video Play	810,926.64	6	6	6	6	6	6
15	Envoy 256MB – 40GB	697,056.71	7	7	7	7	7	7
29	8.3 Minito Speaker	612,275.01	8	9	8	8	8	8
28	Unix/Windows 1-user pack	594,422.37	9	8	9	9	9	9
130	Model A45H Black Image Cartridge	450,834.85	10	11	12	12	11	10
26	SIMM- 32MB PCMCIAII card	449,920.99	11	10	11	11	10	11
40	O/S Documentation Set – English	431,023.05	12	18	10	10	12	12
25	SIMM- 8MB PCMCIAII card	421,118.73	13	12	13	13	13	13
30	Mouse Pad	416,471.39	14	53	52	26	17	14
33	PCMCIA modem/fax 19200 baud	359,520.59	15	24	24	14	14	15
120	DVD-R Disc with Jewel Case, 4.9 GB	302,692.32	16	13	14	15	15	16
24	PCMCIA modem/fax 2880 baud	283,971.33	17	14	15	16	16	17
127	Model CD132 Tricolor Ink Cartridge	256,417.96	18	15	16	17	18	18
129	Model NM High Yield Toner Cartridge	234,260.31	19	16	17	18	19	19
16	Y Box	230,458.61	20	17	18	19	20	20
123	DVD-R Discs, 4.9GB, Pack of 5	190,637.34	21	21	20	20	21	21
138	256MB Memory Card	186,689.33	22	19	19	21	22	22
37	Envoy External 8X CD-ROM	181,482.59	23	20	21	22	23	23
48	Keyboard Wrist Rest	178,516.14	24	49	28	30	30	24
35	External 8X CD-ROM	147,675.58	25	22	22	23	24	25
36	Envoy External 6X CD-ROM	140,463.25	26	23	23	24	25	26
128	Model SM263 Black Ink Cartridge	131,237.87	27	26	25	25	26	27
133	Model K8822S Cordless Phone Battery	117,535.57	28	27	29	28	27	28
140	Endurance Racing	114,365.37	29	28	27	29	29	29
32	Multimedia speakers- 5" cones	113,664.41	30	25	26	27	28	30
135	S272M Extended Use w/l Phone Batt.	110,739.38	31	29	30	31	31	31
19	Laptop carrying case	109,269.76	32	30	32	33	33	33
118	OraMusic CD-R, Pack of 10	108,914.95	33	31	31	32	32	32
34	External 6X CD-ROM	98,241.81	34	32	33	34	34	34
137	256MB Memory Card	93,117.94	35	35	35	36	37	35
27	Multimedia speakers- 5" cones	92,928.93	36	36	37	37	36	36
41	O/S Documentation Set – Russian	91,430.30	37	33	34	35	35	37
132	Model C97B Cordless Phone Battery	89,250.10	38	38	40	41	39	38
42	O/S Documentation Set – Italian	88,848.22	39	34	36	38	38	39
148	Xtend Memory	87,983.25	40	39	38	40	41	40
39	Internal 16X CD-ROM	87,184.05	41	37	39	39	40	41
116	CD-RW, High Speed Pack of 6	83,590.84	42	56	56	55	56	56
31	1.44MB External 3.5" Diskette	76,558.66	43	60	60	62	42	42
45	O/S Documentation Set - Kanji	75,873.80	44	40	41	42	43	44
131	Model K38L Cordless Phone Battery	75,722.04	45	41	43	43	44	43

In the third observation we obtained from our constraints is about traffic generating items. According to the ranking of items for fifth iteration *item 23* (External 101-key keyboard) is ranked in 47th which will not be selected for assortment of retailer. However, this item makes huge sales volumes (20,381 pieces, 2.17% of all sales quantity) we will include this item for our assortment.

The fourth observation to mention regarding our model is subcategory constraint to be introduced for the assortment. According to our association mining for subcategories, subcategories 2056 (Recordable DVD Discs) and 2036 (printer supplies) seemed to exist in the same basket. The Clementine outputs can be reviewed in **Table 4** and **Table 5**. However in our last assortment range (iteration 5) there seemed to be 2 kinds of Recordable DVD Discs against 4 kinds of Printer Supplies. To leverage this we also give a constraint that number of items in subcategory 2056 should be at least 3. Due to this constraint *item 124* (DVD-RW Discs, 4.9GB) is added to assortment.

Table 4. Summary of Association Mining Model in Clementine by subcategory

Analysis
Number of Rules: 42
Number of Valid Transactions: 142,208
Minimum Support: 10.397%
Maximum Support: 39.029%
Minimum Confidence: 40.328%
Maximum Confidence: 79.287%
Minimum Lift: 1.039%
Maximum Lift: 7.621%
Minimum Deployability: 0.0%
Maximum Deployability 0.0%
Minimum Rule Support: 0.0%

Table 5. Results of Association Analysis by rules according to the subcategories.

Consequent	Antecedent	Support %	Confidence %	Rule Support %
2051 = T	2034 = T	20.301	77.53	15.74
2034 = T	2051 = T	39.029	40.328	15.74
2051 = T	2031 = T	21.104	73.597	15.532
2051 = T	2054 = T	18.231	69.691	12.705
2036 = T	2056 = T	20.303	60.741	12.332
2056 = T	2036 = T	19.879	62.034	12.332
2051 = T	2032 = T	21.552	53.362	11.501
2055 = T	2056 = T	20.303	51.209	10.397
2056 = T	2055 = T	21.507	48.342	10.397
2031 = T	2034 = T	20.301	46.997	9.541
2034 = T	2031 = T	21.104	45.209	9.541
2051 = T	2033 = T	11.482	79.287	9.104
2056 = T	2042 = T	13.332	66.702	8.893
2042 = T	2056 = T	20.303	43.8	8.893
2034 = T	2031 = T and 2051 = T	15.532	57.149	8.876
2031 = T	2034 = T and 2051 = T	15.74	56.395	8.876
2036 = T	2042 = T	13.332	64.903	8.653
2042 = T	2036 = T	19.879	43.527	8.653
2055 = T	2036 = T	19.879	41.426	8.235
2013 = T	2014 = T	11.69	69.995	8.182
2031 = T	2054 = T	18.231	42.795	7.802
2032 = T	2054 = T	18.231	41.996	7.656
2055 = T	2042 = T	13.332	55.024	7.336
2036 = T	2055 = T and 2056 = T	10.397	66.493	6.913
2055 = T	2036 = T and 2056 = T	12.332	56.059	6.913
2042 = T	2036 = T and 2056 = T	12.332	55.768	6.877
2031 = T	2054 = T and 2051 = T	12.705	53.177	6.756
2054 = T	2031 = T and 2051 = T	15.532	43.499	6.756
2052 = T	2033 = T	11.482	56.859	6.528
2034 = T	2033 = T	11.482	55.181	6.336
2054 = T	2032 = T and 2051 = T	11.501	54.552	6.274
2032 = T	2054 = T and 2051 = T	12.705	49.38	6.274
2042 = T	2055 = T and 2056 = T	10.397	59.858	6.223
2034 = T	2032 = T and 2051 = T	11.501	51.477	5.92
2031 = T	2033 = T	11.482	51.439	5.906
2031 = T	2032 = T and 2051 = T	11.501	50.975	5.863
2054 = T	2042 = T	13.332	42.84	5.711
2043 = T	2014 = T	11.69	48.622	5.684
2034 = T	2054 = T and 2051 = T	12.705	43.901	5.578
2051 = T	2042 = T	13.332	40.556	5.407

The last part of the analysis is to code LP (lineer programming) algorithm according to the pseudo code in Figure 2. Since we could not obtain full version of licence for Lindo systems, we can not achieve 50 integer variable limit (we have 72 items) we

will just show in a limited environment, how algorithm works. We will just select 12 products (after profits are allocated over in previous steps) and the algorithm will try to find most valuable 10 products according to given constraints. Below in Figure 3. the code of algorithm is shown. Before to show the algorithm Table 1. show the selected sample and its indent.

Table 6 Sample table to show code of algorithm in Lingo 10.0

PRO ID	Prod Desc	GMR	Sub Category
18	Envoy Ambassador	3,757,039.95	A
17	Mini DV Camcorder with 3.5" Swivel LCD	2,196,249.87	A
14	17" LCD w/built-in HDTV Tuner	1,818,258.11	D
13	5MP Telephoto Digital Camera	1,499,316.22	B
21	18" Flat Panel Graphics Monitor	1,039,482.98	B
20	Home Theatre Package with DVD-Audio/Video Play	810,926.64	B
15	Envoy 256MB – 40GB	697,056.71	B
29	8.3 Minitower Speaker	612,275.01	C
28	Unix/Windows 1-user pack	594,422.37	C
130	Model A45H Black Image Cartridge	450,834.85	C
30	Mouse Pad	416,471.39	C
24	PCMCIA modem/fax 28800 baud	283,971.33	A

Let's say that, according to the our sample, A and B subcategories are labeled as dependent to each after association mining and due to this fact the number of items for A and B subcategories must be equal. In addition, *product id 30* is the best selling item in terms of quantity volume. This means, Mouse Pad (ID 30) is a traffic generator item and must be inside the assortment.

!First of all we define the integer variables for each item which are 12 items;
X18 (Envoy Ambassador) ,
X17 (Mini DV Camcorder with 3.5" Swivel LCD),
X14 (17" LCD w/built-in HDTV Tuner);
X13 (5MP Telephoto Digital Camerax);
X21 (18" Flat Panel Graphics Monitor);
X20 (Home Theatre Package with DVD-Audio/Video Play);
X15 (Envoy 256MB – 40GB);
X29 (8.3 Minitower Speaker);
X28 (Unix/Windows 1-user pack);
X130 (Model A45H Black Image Cartridge);
X30 (Mouse Pad);
X24 (PCMCIA modem/fax 28800 baud);

```
The items after X shows the code for each item in our selected
sample.
```

Although there are other ways to code this linear programming structure we will use gross margin values as coefficients of items. So each item in our sample use their reallocated gross margin as their coefficients;

```
MAX=  3757039.94857049*X18 +  2196249.86739742*X17+
      1818258.10805928*X14+   1499316.21893539*X13+
      1039482.97594381*X21+   810926.642458826*X20+
      697056.710823968*X15+   612275.012244315*X29+
      594422.366951353*X28+   450834.854220575*X130+
      416471.387938747*X30+   283971.331751726*X24;
```

Constraint shows that max number of items which need to be selected is 10 so in our 12 item database sample it is said that maximum 10 item could be selected in our assortment. So the algorithm try to find maximum contribution in terms of gross margin;

X18 (Envoy Ambassador) + X17 (Mini DV Camcorder with 3.5" Swivel LCD) + X14 (17" LCD w/built-in HDTV Tuner) + X13 (5MP Telephoto Digital Camerax) + X21 (18" Flat Panel Graphics Monitor) + X20 (Home Theatre Package with DVD-Audio/Video Play) + X15 (Envoy 256MB – 40GB) + X29 (8.3 Minitower Speaker) + X28 (Unix/Windows 1-user pack) + X130 (Model A45H Black Image Cartridge) + X30 (Mouse Pad) + X24 (PCMCIA modem/fax 28800 baud) =10;

After pareto analysis it is shown that X30 (*Item id* is 30, Mouse Pad) is a traffic generator item and it hould be in assortment;
X30 (Mouse Pad)=1;

To support associated subcategories, the number of items inside the assortment for each subcategory should be equal. The retailer must offer exact number of items in its assortment from these subcategories;

(X13 (5MP Telephoto Digital Camerax) + X21 (18" Flat Panel Graphics Monitor) + X20 (Home Theatre Package with DVD-Audio/Video Play) + X15 (Envoy 256MB – 40GB)) – (X18 (Envoy Ambassador) + X17 (Mini DV Camcorder with 3.5" Swivel LCD) + X24 (PCMCIA modem/fax 28800 baud) = 0;

Figure 3. Code for optimization of profits for sample data.

CHAPTER 5

CONCLUSION AND FUTURE RESEARCH

In this study, we approach the assortment selection problem, by improving the PROFSET and GENERALIZED PROFSET models, which are based on a microeconomic framework. We improved the basic model by introducing an additional method of profit allocation over frequent item sets, constraints about categories and sold quantities. Considering limitations of these models we proposed mainly three improvements to these models. One of them is the profit allocation method for frequent item sets. Instead of distributing all profits to frequent item sets or distributing profit based on a probabilistic approach we allocate acceptable portion of total profit of a transaction to frequent item sets. Since allocation is performed iteratively for each frequent item sets in a descending order by rule support, it is

assumed that total profit is allocated in a better way than allocating all profit to one item set in one transaction as GENERALIZED PROFSET suggests. The other improvement that we suggest is using association mining to decide category dependency. The third and the last improvement is considering volume generation of items independent from their sales revenue. By the help of this constraint retailers will keep volume generating items (in terms of quantity rather than revenue) in their assortment. After definition of improvements we applied our model for an electronic retailer having 3 stores, 72 items in its assortment. And the main problem is reducing number of items in assortment to 45. We also illustrated remarkable results considering our assumptions and constraints.

To improve suggested model, frequent item sets could be generated in a different manner. As there are three different stores and transactions have been consolidated, there is no distinction for different stores. In future researches frequent item sets could be defined store by store. After each store defined their own frequent item sets, variance analysis can be performed and frequent item sets could be consolidated in company level. Another improvement direction could be in the customer dimension. If retailer could store valuable customer data and perform valuable segmentations, this parameter could be added into frequent item sets analysis.

REFERENCES

Agrawal N., & S.A. Smith. (2003).Optimal retail assortments for substitutable items purchased in sets. *Naval Research Logistics. 50 (7)* 793-822.

Agrawal, R., T. Imielinski, & A. Swami. (2003). Mining association rules between sets of items in large databases. In Buneman, P., and Jajodia, S., (eds.). Proceedings of ACM SIGMOD Conference on Management of Data, 1993 (SIGMOD93), 207-2 16.

Ahmed, Syed Riaz. (2004). Applications of Data Mining in Retail Business. Proceedings of the International Conference on Inf. Technology:Coding and Computing(ITCC'04), vol. 2,455-489

Araujo, Luis & Mouzas, Stefanos. (1998). Manufacturer-retailer relationships in Germany: The institutionalization of category management. In: Network Dynamics in International Marketing, In Naude´, P. & Turnbull, P. (Ed), 211-232. Elsevier Science Ltd.

Boatwright, P. & J.C. Nunes. (2001). Reducing assortment: An attribute-based approach. Journal of Marketing, 65 (3), 50-63

Brijs, T., G. Swinnen, K. Vanhoof, & G. Wets. (1999). Using association rules for product assortment decisions: A case study. In Proc. of ACM SIGKDD, 254–260.

Brijs, T., B. Goethals, G. Swinnen, K. Vanhoof & G. Wets. (2001). A Data Mining Framework for Optimal Product Selection in Retail Supermarket Data: The Generalized PROFSET Model, Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining, August 20-23, Boston MA (USA), 300-304.

Brijs T, G. Swinnen, K. Vanhoof & G. Wets. (2004) Building an Association Rules Framework to Improve Product Assortment Decisions, Data Mining and Knowledge Discovery, 8, 7–23.

Bultez A., & P. Naert. (1988). S.H.A.R.P.: Shelf allocation for retailer's profit. Marketing Science, Vol. 7, No. 3 (Summer, 1988), 211-231.

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

Cachon, Gérard P., C, Terwiesch & Yi Xu. (2005). Retail Assortment Planning in the Presence of Consumer Search. Manufacturing & Service Operations Management. Vol. 7, No. 4, Fall 2005, 330–346.

Cachon, G., & A.G. Kok. (2007). Category management and coordination in retail assortment planning in the presence of basket shopping consumers. Management Science. 53(6). 934-951.

Chong, J-K., T-H. Ho, & C.S. Tang. (2001). A modeling framework for category assortment planning. Manufacturing & Service Operations Mgmt. 3(3) 191-210.

Dhara S. K., S. J. Hochb, & N. Kumarc. (2001). Effective category management depends on the role of the category. Journal of Retailing, 77 (2001) 165–184.

Dupre, K. & T.W.Gruen,. (2004). The use of category management practices to obtain a sustainable competitive advantage in the fast-moving-consumer-goods industry. Journal of Business & Industrial Marketing, Vol. 19 No. 7, 444-459.

ECR Best Practices Operating Committee. (1995). Category management report – enhancing consumer value in the grocery industry. Joint Industry Project on Efficient Consumer Response, USA.

Gnau, K., T. Richardson, & J. Dippold. (1992). Nielson Category management: Positioning your organization to win. Chicago, NTC Business Books/American Marketing Association.

Gruen, T.W. & R.H. Shah,. (2000). Determinants and outcomes of plan objectivity and implementations in category management relationships. Journal of Retailing, Vol. 76 No. 4, 483-510.

Han J. & M. Kamber. (2000). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Series.

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

Kotler, Philip. (1997). Marketing Management: Analysis, Planning, Implementation, and Control, Englewood Cliffs, NJ: Prentice Hall.

Kök, A.G., & M.L. Fisher. (2004). Demand estimation and assortment optimization under substitution: methodology and application. Operations Research, 55(6), 1001–1021.

Kurnia, S. & R.B. Johnston,. (2003). Adoption of efficient consumer response: key issues and challenges in Australia, Supply Chain Management: An International Journal, Vol. 8 No. 3, 251-262.

Lindblom, A., & R. Olkkonen,. (2006). Category management tactics: an analysis of manufacturers' control. International Journal of Retail & Distribution Management Vol. 34 No. 6, 2006, 482-496.

Rajaram, Kumar. (2001). Assortment planning in fashion retailing: methodology, application and analysis. European Journal of Operational Research 129 (2001) 186-208.

Simonson, Itamar. (1999). The Effect of Product Assortment on Buyer Preferences. Journal of Retailing, Volume 75(3), 347–370.

Su, T. M.Y., (2002) Item Selection By "Hub-Authority" Profit Ranking, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,652-657

Van Ryzin, G., & S. Mahajan,. (1999). On the relationship between inventory costs and variety benefits in retail assortments. Management Science, 45 ,1496-1509.

Wong R.C. F. A.W. & K. Wang. (2005). Data Mining for Inventory Item Selection with Cross-Selling Considerations. Data Mining and Knowledge Discovery, 11, 81–112.

.