

**A PROPOSED ALGORITHM FOR THE DERIVATION OF CONSUMER  
PROFILE FROM MINIMAL TRANSACTIONAL DATA**

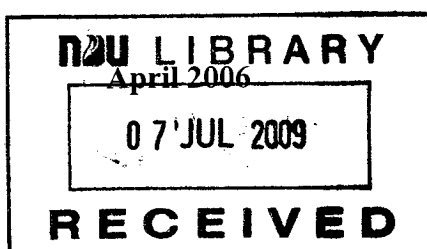
**By**

**Françoise Mhanna**

**A Thesis Study**

**Submitted in Partial Fulfilment of  
The Requirements for the degree of Master in  
Computer Science  
(Computer Information Systems)**

**Department of Computer Science  
Faculty of Natural and Applied Sciences  
Notre Dame University  
Zouk Mosbeh, Lebanon**

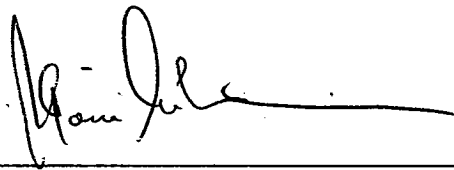


**A Proposed Algorithm For The Derivation Of Consumer  
Profile From Minimal Transactional Data**

**By**

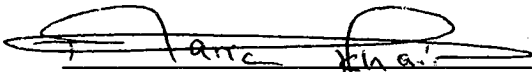
**Françoise Mhanna**

**Approved by:**



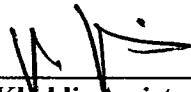
---

**Dr. Mario Missakian: Assistant Professor of Computer Science,  
Advisor.**



---

**Dr. Marie Khair: Associate Professor of Computer Science,  
Member of Committee.**



---

**Dr. Khaldoun ElKhaldi: Assistant Professor of Computer Science,  
Member of Committee.**



---

**Dr. Holem Saliba: Assistant Professor of Mathematics,  
Member of Committee.**

**Date of Thesis Defense: April 12<sup>th</sup>, 2006**

## **Acknowledgements**

I dedicate this study to my family, especially my father and friends, for they have been of a great support to me all along the road.

Special thanks to Mr. Joseph Mattar who back me up on both technical and moral levels.

Needless to say, my deepest gratitude to my university NDU and all the staff members particularly Dr. Mario Missakian for his supervision and guidance as well as the company I work for which is Profiles Software for its encouragement.

## **Abstract**

The 80/20 Rule says that 20% of the customers produce 80% of the sales; this rule indicates the existence of hidden sales potentials that must be revealed. Those hidden sales potentials can only be discovered by building a customer profile. A smart customer profile finds the high potential targets, creates CRM strategies, and starts programs to sell the hidden targets. Shifting just a small percentage of the customers whom are not generating profit to the top group of customers generating profit adds significantly sales growth increases profits. Customer profile puts the full picture together to build sales and profits and maximizes the marketing ROI (Return on Investment).

Stating the importance of having a clear customer profile, the question of how to build a clear, relevant and comprehensive customer profile is raised.

In this thesis, an algorithm is suggested, which if adapted ends up providing the user with a demographic profile of his customers and a clear distribution of the customers according to profitability. This distribution is based on the measurement of customer's LTV (Lifetime value) and Loyalty.

In addition to the distribution which will result from the application of this algorithm, the resulting multidimensional valuable metric can constitute consistent data to apply data mining techniques and get important results.

## List of figures

Figure 1: Rating of five users in five restaurants using the collaborative approach .....	17
Figure 2: The ratings of a user for the restaurants according to their description .....	18
Figure 3: Demographic information on the users who rated a restaurant.....	19
Figure 4: A decision tree classifying transactions into five groups .....	25
Figure 5: A neural network with two hidden layers .....	26
Figure 6: Example of an import cube .....	29
Figure 7: Profile parameters with their acceptable values .....	38
Figure 8: Percentage values resulting for each parameter .....	40
Figure 9: Customer C0001 tentative profile .....	41
Figure 10: Customer C0001 final demographic profile.....	42
Figure 11: Distribution of the values factors over the loyalty scale .....	46
Figure 12: Five Loyalty factors, with their corresponding weighted indices .....	47
Figure 13: Customer Loyalty Value distributed over 5 ranges.....	48
Figure 14: Customer C0001 transactions results concerning loyalty parameters scale values.....	48
Figure 15: Loyalty ranking result for customer C0001 .....	49
Figure 16: Customer LTV ranking .....	50
Figure 17: Distribution of the customer according to their Loyalty and LTV values .....	51
Figure 18: Database “CustProfil” created in SQL server 2000 .....	55
Figure 19: Demographic profile parameters, types and their factor .....	56
Figure 20: Demographic profile parameters values.....	56
Figure 21: Analysis services connection to an SQL server database.....	57
Figure 22: Cube “CustProfilingThesis” in Analysis Services .....	57
Figure 23: Selection of “FACT” table for cube “CustProfilingThesis” .....	58
Figure 24: Choosing the cube “CustProfilingThesis” measure parameter .....	58
Figure 25: List of cube “CustProfilingThesis” used dimensions.....	59
Figure 26: Cube “CustProfilingThesis” .....	59
Figure 27: Result in percentage for each parameter .....	60
Figure 28: Demographic profile tentative results .....	61
Figure 29: Weighted values for each loyalty parameter .....	62

Figure 30: Creation of new cube LoyaltyLtvCube .....	63
Figure 32: Number of customer per Loyalty ranking/LTV ranking .....	64
Figure 33: list of clients per each Ltv-Loyalty combination .....	64
Figure 34: Creation of a mining model using a wizard .....	65
Figure 35: Dimensions for the created data mining model.....	65
Figure 36: Result of the data mining model processing .....	66
Figure 37: Clusters probability percentages of combinations .....	66

## Table of Contents

### Chapter 1: INTRODUCTION AND PROBLEM DEFINITION

1.1 Introduction to the general problem .....	7
1.2 Problem definition .....	7
1.3 Research Objectives .....	7
1.4 Research approach .....	8
1.5 Thesis organization .....	9

### Chapter 2: BACKGROUND AND DEFINITIONS

2.1 Definition of the basic concepts .....	10
2.1.1 Customer Relationship Management .....	10
2.1.2 Enterprise Resource Planning .....	11
2.1.3 CRM v/s ERP .....	12
2.1.4 Customer Profiling .....	12
2.1.4.1 Demographic Profiling .....	13
2.1.4.2 Geographic Profiling .....	14
2.1.4.3 Behavioural Profiling .....	14
2.1.4.4 Customer Loyalty .....	15
2.1.4.5 Customer Lifetime value .....	15
2.2 Techniques used to build a customer profile .....	16
2.2.1 Collaborative Filtering Approach .....	17
2.2.2 Content Based Approach .....	18
2.2.3 Demographic Based Approach .....	19
2.2.4 Geographic Based Approach .....	19
2.3 Data Mining .....	20
2.3.1 What's Data Mining? .....	21
2.3.2 Data Mining Algorithms .....	22
2.3.2.1 Statistical Algorithms .....	22
2.3.2.2 Artificial Intelligence .....	22
2.3.2.3 Cluster Analysis .....	23
2.3.2.4 Apriori Algorithm .....	23
2.3.2.5 Decision Trees .....	24
2.3.2.6 Neural Networks .....	25
2.4 Mining data using SQL server 2000 based on Analysis Services .....	27
2.4.1 Analysis Services in SQL server 2000 .....	27
2.4.2 Analysis Services Decision Cubes .....	28
2.4.3 Analysis Services Data Mining Techniques .....	31
2.4.4 SQL Server Data Mining Algorithms .....	31
2.5 Background and Previous studies .....	32

### Chapter 3: A new customer profiling algorithm based on transactional data

3.1 Introduction to the new approach in extracting customer's profile .....	36
3.2 Extracting the Demographic Profile .....	36
3.2.1 Preparation Phase .....	37
3.2.2 Definition of the customer profile template .....	37
3.2.3 Definition of the range of values for each profile parameter .....	37

3.2.4 Product categorization according to the range of values defined for each of the parameters.....	39
3.2.5 Extracting the customer demographic profile.....	39
3.2.5.1 <i>Extracting the tentative profile</i> .....	39
3.2.5.2 <i>Extracting the real profile based on sophisticated rules</i> .....	41
3.3 Measurement of Customer Loyalty and Customer Lifetime Value.....	42
3.3.1 Determination of loyalty factors .....	43
3.3.2 Determination of Customer Lifetime Value factors .....	44
3.3.3 Measurement's scale definition and rank of factor's values accordingly .....	45
3.3.4 Assigning weight for each factor to extract loyalty .....	47
3.3.5 Extracting the scaled values.....	47
3.3.6 Customer Loyalty measurement .....	48
3.3.7 Customer Lifetime value measurement .....	49
3.4 Building the loyalty and LTV metric.....	51
3.5 What are the benefits of this metric, when and how to use it? .....	52

## **Chapter 4: Application of the suggested algorithm on real data**

4.1 Introduction.....	53
4.2 Data Description .....	53
4.3 Cleaning and preparing the data .....	54
4.4 Applying the suggested Model .....	55
4.4.1 Definition of the customer profile template.....	55
4.4.2 Definition of the range of values of each template .....	56
4.4.3 Product categorization according to the range of values defined for each of the profile parameter.....	57
4.4.4 Extracting Customer Profile .....	57
4.4.4.1 <i>Extraction of the customers' tentative profile</i> .....	60
4.4.4.2 <i>Extraction of the final customer profile</i> .....	61
4.4.4 Extracting the Customer Loyalty ranking.....	62
4.4.5 Extracting the Customer LTV ranking .....	62
4.4.6 Building the resulting matrix: .....	63
4.5 Data mining.....	64

## **Chapter 5: Conclusion**

5.1 Advantages and Disadvantages of the new algorithm .....	67
5.2 Possibility of extensions and future work.....	68
5.3 Conclusion of the main contributions in this thesis.....	68



# CHAPTER 1: INTRODUCTION AND PROBLEM DEFINITION

## 1.1 Introduction to the general problem

Face to the complexities, difficulties and high standards of customer needs, all enterprises should either unify what they have and transform into services for their customers or to systematize their needs in order to deal with their customers. [6]

Customer Relationship Management (CRM) helps companies improve the profitability of their interactions with customers. An important objective of CRM is to achieve a one-to-one relationship rather than a one-to-everyone relationship. This one-to-one relationship is known as personalization. In other words, Personalization is the ability to provide content and services tailored to individuals on the basis of knowledge about their demographic and behavioral attributes. [2] Customer loyalty and customer life-time value are one of the most important aspects of CRM that determine future relation with the customers.

This thesis addresses these issues by developing an approach which uses information learned from customers' transactional histories to construct accurate comprehensive individual profiles. One part of the profile demographically describes the customer while the other part gives the analytical point of view.

## 1.2 Problem definition

In order to achieve personalization, Answers on the following questions should be provided.

- How to provide personal recommendations based on a comprehensive knowledge of who customers are, how they behave, and how similar they are to other customers.[2]
- How to extract this knowledge from the available data and store it in a clear, comprehensive, readable and simple way. [2]

## 1.3 Research Objectives:

The research objective is to provide a model allowing a company to:

- Build up the customer identity without going into the hassle of direct contact with customers by asking questions and inquiring about the information they need.

- Segment customers by loyalty rate and customer lifetime value using the minimal data available.
- Extract both the customers' demographic and analytical identity using only the customers' historical sales transactions

#### 1.4 Research Approach

This study in its general format consists of two parts:

- Theoretical part
- Practical part

The theoretical part suggests a model to extract two faces of customer profiling; demographic profile and analytical profile. The analytical profile tackles both the customer loyalty and customer lifetime value. The suggested model is generic and is applicable in different lines of business. This model needs a direct interaction from the user only in its preparation phase in order to reach a useful and meaningful profile for a company's future projects. If adapted, this model can provide the company with an up to date profile about each of its customers with very low expenses and without going into the hassle of contacting the customers.

The model consists of three phases:

Phase 1: Extracting the customer demographic profile, this is being realized in two phases

- The preparation phase where the interaction of the user is needed to define the demographic parameters which are consistent with the company's line of business and future projects.
- The extraction phase based on simple and sophisticated rules according to the parameters set in the preparation phase

Phase 2: Measure the loyalty of each customer according to predefined parameters.

Phase 3: Measure the customer lifetime value according to predefined parameters.

Once the three phases are successfully completed, in order to extract the essence of the work done, a matrix which dimensions are the customer loyalty values and customer lifetime values is to be built. This matrix provides the user with a clear distribution of the customers in terms of loyalty and customer lifetime value, and allows also a detailed analysis in terms of demographic parameters.

As stated before, the suggested model has been tested by applying it on a real case study. The data used consists of the sales transactions of a sport equipment stores. The results of the application proved that the model is consistent.

### **1.5 Thesis organization**

The thesis in its complete format consists of five chapters. In addition to the first chapter containing the thesis's general introduction, the problem definition, the objectives and the research approach, four other chapters present, discuss, analyze, provide solution and test the problem in details. Chapter 2 consists of the background of the thesis; it contains definitions and details about the techniques used. Chapter 3 describes in detail the new suggested approach to be adopted for the extraction of the demographic and analytical customers profile from historical transactions. In chapter 4, the suggested model is applied in a real life scenario for evaluation and approval of its validity. Finally, Chapter 5 contains the conclusion and proposes possible future work.

## CHAPTER 2: BACKGROUND AND DEFINITIONS

### 2.1 Definition of the basic concepts

#### 2.1.1 Customer Relationship Management

Customer Relationship Management (CRM) is an integrated approach to identifying, acquiring, and retaining customers. By enabling organizations to manage and coordinate customer interactions across multiple channels, departments, lines of business, and geographies, CRM helps organizations maximize the value of every customer interaction and drive superior corporate performance.

In order to build long term mutually beneficial relationship with their customers, many companies try to establish and improve connections with their customers via CRM system. CRM is a comprehensive approach to improve the relationships with all kinds of customers, link back office functions (financial, operation, logistic and human resource) with customers (via such as internet, email, sales, direct mail, call centre, advertising, fax etc.)

When CRM is fully and successfully implemented, it is a cross-functional, customer driven technology-integrated business process management strategy which maximizes relationships and encompasses the entire organization.

CRM is not an event or a technology, or even an application or a process. Ideally, CRM is a comprehensive strategy that integrates all areas of business that touch the customer – though mainly, it is limited to marketing, sales, customer service and field support — through the integration of people, process and technology.

To be successful, CRM requires acquiring and distributing knowledge about one's customers across the enterprise, to balance costs, revenue and profits with customer satisfaction. Obviously, business processes and key technologies are required to optimize CRM strategies.

CRM has the following benefits:

- It can extend the capability to the customer for self-service and Internet applications.
- It can attract existing and new customers through personalized communications and improved targeting.

- It can integrate customer and supplier relationships.
- It can construct metrics to analyze common and unique customer patterns.
- Organizationally, CRM is a strategic focus on the behavior of, and communication with, the customer.
- Technologically, CRM is based on the use of data mining to identify customer profiling, preferences and behavior.
- In business processes, CRM is the use of this data to improve efficiencies and effectiveness in marketing, sales and support.
- CRM is a commitment to drive customer satisfaction and shareholder satisfaction simultaneously. Such action implies allocating scarce resources to provide a seamless, high-quality experience for a company's most valuable customers, and shedding the least desirable customers.

‘To be successful and deliver value with CRM you have to connect all the dots (Dhore)’. [15] This means first identifying the business challenge, setting very specific business goals that would meet that challenge, and then building the strategy to achieve those goals, including incorporating all relevant systems, groups and processes. [15]

### **2.1.2 Enterprise Resource Planning (ERP)**

Enterprise Resource Planning (ERP) aims to obtain functional integration between the main business areas of an organization. ERP was intended to provide the connectivity and the common data models needed to link and coordinate the disparate functional areas within the organization, such as product planning, purchasing, logistics finance, etc.

ERP was originally envisioned as a “one-size-fits-all” modularized software approach to the management of most core business activities. Today companies use ERP to manage product planning, purchasing and logistics, inventory management, production, vendor management, customer service, finance, human resources and many other basic business activities. In addition, companies may also utilize best of

breed applications for many of their specific needs. All of these systems must be integrated into and work with the core ERP system. [17]

### 2.1.3 CRM v/s ERP

As mentioned above, ERP and CRM both try to provide connections between all different areas of an organization. In fact, most of the nowadays successful ERP vendors pay attention on CRM markets, such as Seibel, SAP etc., and try to build relation between CRM and ERP.

The major difference between ERP and CRM is that ERP focuses on building foundation with tightly integrated back office functions, while CRM tries to link front and back office application to maintain relationships with customers to optimize customer satisfaction and profitability. Although ERP is not required for a CRM system, it is beneficial for a CRM system if there is underlying infrastructure such as ERP.

One of the major points in achieving a fruitful Customer relationship management is by knowing the customers, this can be achieved through customer profiling. [13]

### 2.1.4 Customer Profiling

Customer profiling is the act of describing customers by their attributes, such as age, income and lifestyles.

Customer profiling provides a basis for marketers to “communicate” with existing customers in order to offer them better services and to retain them. This is done by assembling collected information on the customer such as:

- **Demographic information** describe characteristics of populations and include age, gender, cultural background and ethnic, education, occupation, income, religion, marital status, children, life style, socioeconomic status, and so on.
- **Geographic information** includes various classifications of geographic areas, for example, zip code, state, country, region, climate, population, and other geographical census data.

- **Behavioral information** include product usage rate and end, benefit sought, decision making units, ready-to-buy stage, and so on. This information can be extremely useful for marketing purposes.
- **Loyalty information:** Recency (time since the last purchase), Frequency, customer ranking and monetary values.
- **Customer lifetime value** [3]

Depending on a company's target goal, decision makers must define the customer profile which will be relevant to achieve this target. A simple customer profile is a file containing at least his name, address and phone number.

Customer profiling is one of the best prospecting tools. Applying profiling techniques allows a full exploitation of the customer's data buying patterns and behaviour, and helps gain a greater understanding of consumer motivation. Customer Profiling helps to dramatically increase response rates of the marketing campaigns by micro targeting the right customer with the right product. Businesses today are using profiling to reduce fraud, to anticipate demand, to increase new customer acquisition and customer loyalty. Customer profiling is also used to develop lifelong relationships with customers by anticipating and fulfilling their needs. Consumers appreciate a personal touch and something they can act on. In other words, customer profiling is about: who is the customer? What the customer does? And is the customer loyal to the organization?

#### 2.1.4.1 Demographic Profiling

One of the many customer features that can be used for demographic profiling are:

- **Age:** What is the predominant age group of the target buyers? How many children and what ages are in the family?
- **Gender:** Will be needed when targeting customers with certain products related to a specific gender
- **Cultural and ethnic:** What languages do they speak? Does ethnicity affect their tastes or buying behaviour?
- **Economic conditions, income and/or purchasing power:** What is the average household income or purchasing power of the customers? Do

they have any payment difficulty? How much or how often does a customer spend on each product?

- For acquired customer, shopping frequency, frequency of complaints, degree of satisfaction, preferences may be used to build a purchase profile.
- Values, attitudes, beliefs. What is the customers' attitude toward the kind of product or service?
- Life cycle: How long has the customer been regularly purchasing products?
- Knowledge and awareness: How much knowledge do customers have about a product or service, or industry?
- Lifestyle: How many lifestyle characteristics about purchasing are useful? [3]

#### **2.1.4.2 Geographic Profiling**

As described before, geographic profiling is about customer geographic belonging. Where does he live, in which country, state, region etc...? [8]

#### **2.1.4.3 Behavioral Profiling**

Demographic data is used to describe customer segments (profiling), but it is much less effective if not accompanied by customer behaviour profile. Not all 45- to 55-year-olds with a household income between \$50,000 and \$75,000 have the same purchase interests and spending habits. So having the demographic data along with customer behaviour will form the back-bone for an accurate customer profiling.

Behavioural data goes beyond knowing that a customer has purchased a certain product. It involves capturing customer events and actions over time and using these stored interactions to determine typical behaviour and deviations from that behaviour.

Customer analytics exploit customer behavioural data to identify unique and actionable segments of the customer base. These segments may be used to increase targeting methods. Ultimately, customer analytics enable effective and efficient customer relationship management (CRM). The analytical techniques vary based on



objective, industry and application; however the technique the most used is the segmentation technique:

- Segmentation techniques: segment groups of the customer base that have similar spending and purchasing behaviour. Such groups are used to enhance the predictive models as well as improve offer and channel targeting.

A behavioural profile can only be extracted from the customer transactional history. The better the customer behavior is understood before jumping into full-blown CRM, the more likely the final CRM solution will have the right functionality - build or buy.

Knowing the customer behavior will allow the prediction how the customer may react on newly introduced items or offers, will help how to best target the customer and what may be the response on marketing campaigns and may even assist on predicting future sales and profits. [7]

#### **2.1.4.4 Customer Loyalty**

The main purpose of relationship marketing is customer retention and loyalty. A number of researchers indicate that customer loyalty is a key ingredient of firms' profitability due to the high cost of acquiring new customers. It is not surprising, therefore, that customer retention and the management of relationships with customers have become a major issue and a key objective in modern retailing. To conclude, loyalty equals profit. [27]

#### **2.1.4.5 Customer Lifetime Value**

Customer Lifetime Value (CLV) permits the evaluation of the customer value over time. Lifetime Value is the profit expected to receive from a customer discounted over time. If the Lifetime Value is increased this means additional profits for the organization are generated.

The lifetime value of a customer can be also described as being the expected cash flow from each online visitor over the lifetime of that relationship. Lifetime value can be calculated in three parts:

- The sum of the expected lifetime revenues of a customer

- The lifetime cost of the customer, including acquisition cost and operating expense
- The operating-cost reduction due to online self-service

Recently, quite a bit has been written about measuring the lifetime value of a customer, and there are a few noteworthy observations about it as the basis for determining return on investment (ROI):

- It is important to focus continually on new-customer acquisition, the conversion of new visitors to buyers, and the repeat frequency of existing buyers. The greater that each of these values are, the greater the lifetime value of a customer.
- As expected, the costs of new customer acquisition are substantial, which suggests that existing customers are responsible for near-term profits, and new customers will only contribute in the future. This simply means that it is not economical to build the business only on first-time buyers, and that a continued online customer relationship is critical to short- and long-term profits. [14]

## **2.2 Techniques used to build a Customer Profile**

As mentioned previously customer profiling goal is to provide personal recommendations based on a comprehensive knowledge of who the customers are, how they behave, how similar they are to other customers, and how to extract this knowledge from the available data and store it in customer profiles.

Also building a customer profile helps to find new customers for a business. It will extract people and/or businesses that match the profile of the current customers. This provides a list of prospective customers, who could have already bought products similar to the company's products, have a need for the company's product or are more inclined to buy the company's product or service.

As mentioned before depending on the company goal, one has to select what is the profile relevant to achieve this goal.

Concerning demographic information, most rely on information supplied by the customer while behavioral and loyalty ranking are to be extracted from the transactional data history.

Many techniques has been used to build-up customer profile, we'll list some of the most commonly used one.

- The collaborative-filtering
- The content based approach (some systems integrate the two methods both the collaborative-filtering and the content based approach.
- Demographic Based approach
- Geographic approach.

### 2.2.1 Collaborative-Filtering Approach

Collaborative filtering compares customers according to their preferences. Therefore, a database of user's preferences must be available. The preferences can be collected either explicitly (explicit rating) or implicitly (implicit rating). In the first case the user's participation is required. The customer explicitly submits his/her rating of the given item. Such rating can, for example, be given as a score on a rating scale from 1 to 5. The implicit ratings, on the other hand are derived from monitoring the user's behaviour.

Fig.1 below is an example; it gives the ratings of 5 restaurants by 5 users. A "+" indicates that the user liked the description of the restaurant and a "-" indicates that the user did not like the restaurant.

	Karen	Lynn	Chris	Mike	Jill
Kitima	-	+	+	+	-
Marco Polo	+	+	+	+	+
Spiga	+	-	+	-	+
Thai Touch	-	+	-	+	-
Dolce	+	-	+	-	?

Fig.1: Rating of five users in five restaurants using the collaborative approach

To predict the rating that Jill would give to Dolce, we can look for users that have a similar pattern of ratings with Jill. In this case, Karen and Jill have identical tastes and one might want to predict that Jill would like Dolce because Karen does. A more general approach would be to find the degree of correlation between Jill and other users. Rather than relying on just the most similar user, a weighted average of the recommendations of several users can be found. The weight given to a user's rating would be found by degree of correlation between the two users.

The collaborative filtering process can be divided into two phases:

- The model generation phase
- The recommendations phase

Algorithms which tend to skip the first phase are the so called memory-based approaches. The preferences database is a huge user-by-item matrix constructed from the data on hand. A matrix element represents user's rating of item. Memory based approaches search the matrix for relationships between customers and items. Model-based approaches, on the other hand, use the data from the matrix to build a model that enables faster and more accurate recommendations. The model generation is usually performed offline over several hours or days.

When dealing with collaborative filtering, two fundamental problems of collaborative filtering have to be taken into account:

- The scarcity of the data
- The scalability problem

The first problem, which we encounter when the matrix is missing many values, can be partially solved by incorporating other data sources, by clustering customers and/or items, or by reducing the dimensionality of the initial matrix. The last two techniques also counter the scalability problem. This problem arises from the fact that the basic nearest neighbour algorithm fails to scale up its computation with the growth of the numbers of users and the number of items. Some of the approaches for countering the two problems are described. [22]

### **2.2.2 Content based approach**

Content-based methods make recommendations by analyzing the description of the items that have been rated by the user and the description of items to be recommended. They go as far as determining why a customer has preferred some shops to other shops of the same nature

Fig. 2 shows an example with 5 restaurants and 5 words that appear in descriptions of the restaurants. Jill's ratings on these pages are also shown in the table.

	noodle	shrimp	basil	exotic	salmon	Jill
Kitima	Y	Y	Y	Y	Y	-
Marco Polo		Y	Y			+
Spiga	Y		Y			+
Thai Touch	Y	Y		Y		-
Dolce		Y	Y		Y	?

Fig. 2: the ratings of a user for the restaurants according to their description

A variety of algorithms have been proposed for analyzing the content of text documents and finding regularities in their content that can serve as the basis for making recommendations. Many approaches are specialized versions of classification learners, in which the goal is to learn a function that predicts which class a document belongs to. Other algorithms would treat this as a regression problem in which the goal is to learn a function that predicts a numeric value.

### 2.2.3 Demographic Based Approach

Demographic information can be used to identify users that like a certain object. One might expect to learn the type of person that likes a certain restaurant. Similarly LifeStyle Finder attempts to identify one of the 62 pre-existing clusters to which a user belongs and to tailor recommendations to users based upon information about others in this cluster. LifeStyle Finder enters into dialog with the user to help categorize the user.

Fig. 3: Represents a demographic information on the users who rated a restaurant together with the ratings of the users for that restaurant

	gender	age	area code	education	employed	Dolce
Karen	F	15	714	HS	F	+
Lynn	F	17	714	HS	F	-
Chris	M	35	714	C	T	+
Mike	F	40	714	C	T	-
Jill	F	10	714	E	F	?

Fig. 3: demographic information on the users who rated a restaurant

Since Jill demographic profile is more likely to Karen and Lynn then it is most possible that she will like the 'Dolce' restaurant. [22]

### **2.2.4 Geographic Based Approach**

The profiling process begins by creating a spreadsheet with customer addresses. GIS software uses this spreadsheet to map the location of each customer's address. The initial mapping process shows the distribution of every customer home and provides a picture of the areas where customers are clustered. While the map provides a general view, the GIS can also be used to calculate numbers about those areas and distances that generate the most customers. Typically, geographic profiling is conducted in two ways: by drive time segments and by smaller geographic areas such as counties. These calculations provide insight into how far customers are willing to travel as well as the areas that produce the most customers.

Not only is GIS useful in determining the geographic origin of customers, but also their demographic composition. By knowing customer addresses, demographic information can be obtained about the neighbourhood where they live. Pre-defined neighbourhoods, such as census block groups or zip codes, have robust demographic information associated with their boundaries. Having demographic information about a neighbourhood means we can use the premise that birds of a feather flock together. That is, knowing something about a customer's neighbourhood also means knowing information about the residents. [8]

## **2.3 Data Mining**

Data mining is the process of extracting valid, authentic, and actionable information from large databases. It is not the process of extracting specific data but instead deriving information that the data as a whole can provide. The real power of data mining is that it can go beyond the obvious to finding hidden patterns someone would otherwise not think to look for in large databases.

Data mining has been used for several CRM purposes like:

- Census & Survey Analysis
- Customer Profiling
- Customer Retention
- Customer Segmentation
- Database Marketing
- Deviation Detection
- Direct Mail Marketing

- Direct Marketing
- Fraud Detection
- Insurance Risk Analysis
- Marketing Research
- Market Segmentation [3]

### 2.3.1 What is Data Mining?

Data mining is a set of computer-assisted techniques designed to automatically mine large volumes of integrated data for new, hidden or unexpected information, or patterns. Data mining is sometimes known as knowledge discovery in databases (KDD).

In recent years, database technology has advanced in stride. Vast amounts of data have been stored in the databases and business people have realized the wealth of information hidden in those data sets. Data mining then become the focus of attention as it promises to turn those raw data into valuable information that businesses can use to increase their profitability.

Data mining can be used in different kinds of databases (e.g. relational database, transactional database, object-oriented database and data warehouse) or other kinds of information repositories (e.g. spatial database, time-series database, text or multimedia database, legacy database and the World Wide Web).

Therefore, data to be mined can be numerical data, textual data or even graphics and audio.

The capability to deal with voluminous datasets does not mean data mining requires huge amount of data as input. In fact, the quality of data to be mined is more important. Aside from being a good representative of the whole population, the data sets should contain the least amount of noise -- errors that might affect mining results. There are many data mining goals have been recognized; these goals may be grouped into two categories -- verification and discovery. Both of the goals share one thing in common -- the final products of mining processes are the discovered patterns that may be used to predict the future trends.

In the verification category, data mining is being used to confirm or disapprove identified hypotheses or to explain events or conditions observed. However, the limitation is that such hypotheses, events or conditions are restricted by

the knowledge and understanding of the analyst. This category is also called top-down approach. Another category, the discovery, is also known as bottom-up approach. This approach is simply the automated exploration of hitherto unknown patterns. Since data mining is not limited by the inadequacy of the human brain and it does not require a stated objective, inordinate patterns might be recognized. However, analysts are still required to interpret the mining results to determine if they are interesting.

In recent years, data mining has been studied extensively especially on supporting customer relationship management (CRM) and fraud detection. Moreover, many areas have begun to realize the usefulness of data mining. Those areas include biomedicine, DNA analysis, financial industry and e-commerce. However, there are also some criticisms on data mining shortcomings such as its complexity, the required technical expertise, the lower degree of automation, its lack of user friendliness, the lack of flexibility and presentation limitations. Data mining software developers are now trying to mitigate those criticisms by deploying an interactive development approach. It is expected that with the advancement in this new approach, data mining will continue to improve and attract more attention from other application areas as well. [4]

### **2.3.2 Data Mining Algorithms**

As mentioned above, there are plenty of algorithms in use to mine data. Due to the scope limitation, this section is focused on the most frequently used and widespread recognized algorithms that can be indisputable as data mining algorithms; neither pure statistical, nor database algorithms. The examples include Apriori algorithms, decision trees and neural networks. Details of each algorithm are as follows:

#### **2.3.2.1 Statistical Algorithms**

The distinction between statistics and data mining is indistinct as almost all data mining techniques are derived from the statistics field. It means statistics can be used in almost all data mining processes including data selection, problem solving, result presentation and result evaluation.

Statistical techniques that can be deployed in data mining processes include mean, median, variance, standard deviation, probability, confidence intervals,



correlation coefficient, non-linear regression, chi-square, Bayesian theorem and Fourier transforms. [4]

### **2.3.2.2 Artificial Intelligence**

Artificial intelligence (AI) is the scientific field seeking for ways to locate intelligent behavior in a machine. It can be said that artificial intelligence techniques are the most widely used in mining process. Some statisticians even think of data mining tool as an artificial statistical intelligence. The capability of learning is the greatest benefit of artificial intelligence and which is most appreciated in the data mining field.

Artificial intelligence techniques used in data mining processes include neural network, pattern recognition, rule discovery, machine learning, case-based reasoning, intelligent agents, decision tree induction, fuzzy logic, genetic algorithm, brute force algorithm and expert system. [4]

### **2.3.2.3 Cluster Analysis**

Cluster analysis addresses segmentation problems. The objective of this analysis is to separate data with similar characteristics from the dissimilar ones. The difference between clustering and classification is that while clustering does not require pre-identified class labels, classification does. That is why classification is also called supervised learning while clustering is called unsupervised learning.

As mentioned above, sometimes it is more convenient to analyze data in the aggregated form and allow breaking down into details if needed. For data management purpose, cluster analysis is frequently the first required task of the mining process. Then, the most interesting cluster can be focused for further investigation. [4]

### **2.3.2.4 Apriori Algorithms**

Apriori algorithm is the most frequently used in the dependency analysis cases. It attempts to discover frequent item sets using candidate generation for Boolean association rules. Boolean association rule is a rule that concerns associations between the presence or the absence of items.

The steps of Apriori algorithms are as follows:

- (a) The analysis data is first partitioned according to the item sets.

- (b) The support count of each item set (1-itemsets), also called candidate, is performed.
- (c) The item sets that could not satisfy the required minimum support count are pruned. Thus creating the frequent 1-item sets (a list of item sets that have at least minimum support count).
- (d) Item sets are joined together (2-itemsets) to create the second-level candidates.
- (e) The support count of each candidate is accumulated.
- (f) After pruning unsatisfactory item sets according to minimum support count, the frequent 2-itemsets is created.
- (g) The iteration of (d), (e) and (f) are executed until no more frequent k item-sets can be found or, in other words, the next frequent k-item-sets contains empty frequent.
- (h) At the terminated level, the Candidate with maximum support count wins.

By using Apriori algorithms, the group of item sets that most frequently come together is identified. However, dealing with large amounts of transactions means the candidate generation, counting and pruning steps needed to be repeated numerous times. Thus, to make the process more efficient, some techniques such as hashing (reducing the candidate size) and transaction reduction can be used. [4]

### **2.3.2.5 Decision Trees**

Decision tree is a predictive model with tree or hierarchical structure. It is used most in classification and prediction methods. It consists of nodes, which contain classification questions, and branches, or the results of the questions. At the lowest level of the tree -- leaf nodes -- the label of each classification is identified.

The structure of decision tree is illustrated in fig. 4.

Typically, like other classification and prediction techniques, the decision tree begins with an exploratory phase. It requires data sets with labels to be fed. The underlying algorithm will try to find the best-fit criteria to distinguish one class from another. This is also called tree growing. The major concerns are the quality of the classification problems as well as the appropriate number of levels of the tree. Some leaves and branches need to be removed in order to improve the performance of the decision tree. This step is also called tree pruning.

On the higher level, the predetermined model can be used as a prediction tool. Before that, the testing datasets should be fed into the model to evaluate the model's performance. Scalability of the model is the major concern in this phase.

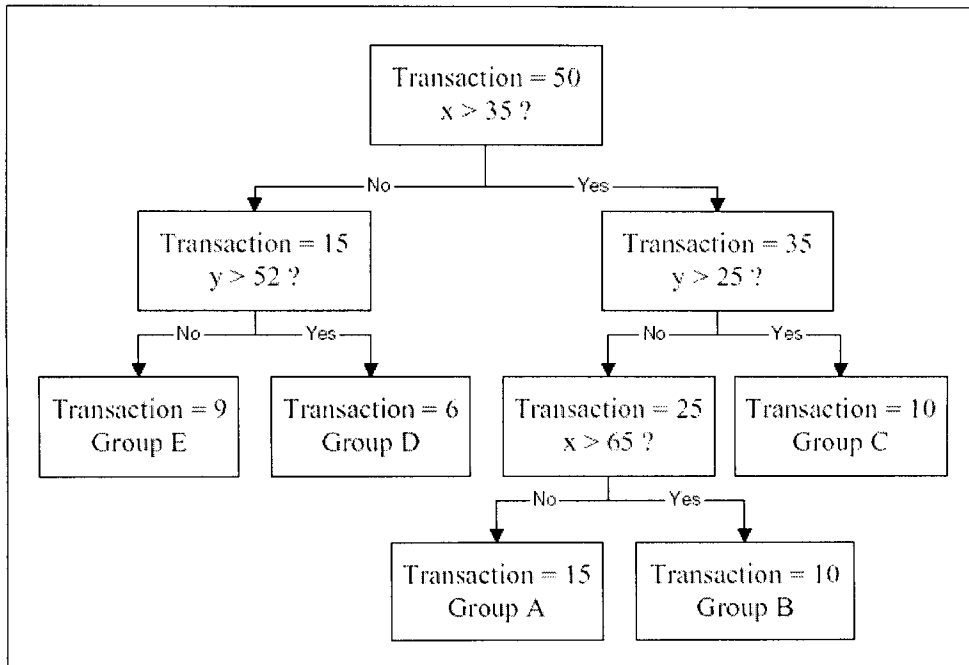


Fig. 4: A decision tree classifying transactions into five groups

The fundamental algorithms can be different in each model. Probably the most popular ones are Classification and Regression Trees (CART) and Chi-Square Automatic Interaction Detector (CHAID). For the sake of simplicity, no details of these algorithms are provided; only perspectives of them are provided.

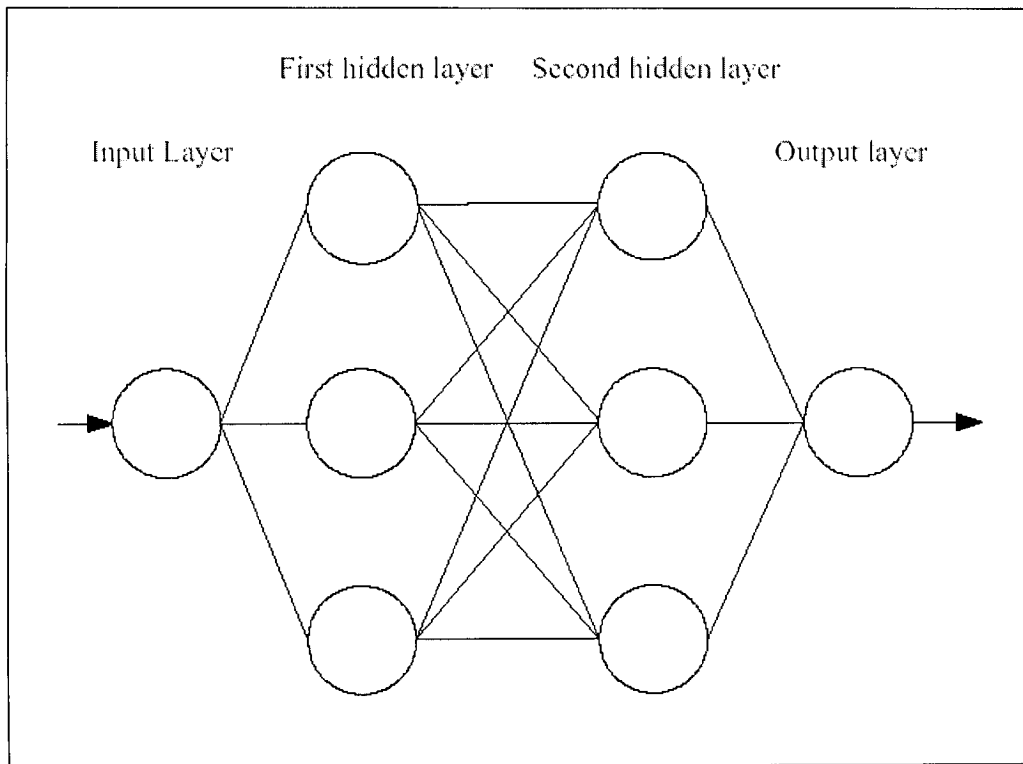
CART is an algorithm developed by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone. The advantage of CART is that it automates the pruning process by cross validation and other optimizers. It is capable of handling missing data and it sets the unqualified records apart from the training data sets.

CHAID is another decision tree algorithm that uses contingency tables and the chi-square test to create the tree. The disadvantage of CHAID compared to CART is that it requires more data preparation process. [4]

### 2.3.2.6 Neural Networks

Nowadays, neural networks, or more correctly the artificial neural networks, attract the most interest among all data mining algorithms. It is a computer model based on the architecture of the brain. To put it simply, it first detects the pattern from

data sets. Then, it predicts the best classifiers. And finally, it learns from the mistakes. It works best in classification and prediction as well as clustering methods. The structure of neural network is shown in fig. 5.



**Fig. 5:** A neural network with two hidden layers

As shown in fig. 5, neural network is comprised of neurons in input layer, one or more hidden layers and output layer. Each pair of neurons is connected with a weight. In the cases where there are more than one input neurons, the input weights are combined using a combination function such as summation. The most well known neural network learning algorithm is Back-propagation. It is the method of updating the weights of the neurons. Unlike other learning algorithms, back-propagation algorithm works, or learns and adjusts the weight, backward which simply means that it predicts the weighted algorithms by propagating the input from the output.

Neural networks are widely recognized for their robustness; however, the weakness is their lack of self-explanation capability. Though the performance of the model is satisfactory, some people do not feel comfortable or confident to rely irrationally on the model.

It should be noted that some algorithms are good at discovering specific methods while some others are appropriate for many types of methods. The choice of algorithm or set of algorithms used depends solely on a user's judgement. [4]

## **2.4 Mining data using SQL server 2000 based on Analysis Services**

Since the case study will be performed using data mining based on SQL server 2000 Analysis Services, an overview of these tools will be presented:

### **2.4.1 Analysis services in SQL server 2000:**

Microsoft SQL Server 2000 Analysis Services is a middle-tier server for online analytical processing (OLAP) and data mining. The Analysis Services system includes a server that manages multidimensional cubes of data for analysis and provides rapid client access to cube information. Analysis Services organizes data from a data warehouse into cubes with pre-calculated aggregation data to provide rapid answers to complex analytical queries. Analysis Services also allows to create data mining models from both multidimensional (OLAP) and relational data sources. Data mining models can be applied to both types of data. PivotTable Service, the included OLE DB compliant provider, is used by Microsoft Excel and applications from other vendors to retrieve data from the server and present it to the user, or create local data cubes for offline analysis.

By supporting various data and storage models, Microsoft SQL Server 2000 Analysis Services helps creating and maintaining a system that meets the organization's needs.

Microsoft SQL Server 2000 Analysis Services provides a scalable architecture to address a variety of data warehousing scenarios.

Microsoft SQL Server 2000 Analysis Services works with other components and programs to ensure enterprise-level robustness.

Rapid access to data warehouse data is provided by Microsoft SQL Server 2000 Analysis Services. Data from the data warehouse is extracted, summarized, organized, and stored in multidimensional structures for rapid response to end user queries.

Analysis Services also provides architecture for access to data mining data. This data can be sent to the client in either a multidimensional or relational form.

Analysis Services and PivotTable Service provide the capability to design, create, and manage cubes and data mining models from data warehouses and to provide client access to OLAP data and data mining data. The Analysis server manages the data; PivotTable Service works with the server to provide client access to the data. [26]

### **2.4.2 Analysis Services Decision Cubes:**

Cubes are the main objects in online analytic processing (OLAP), a technology that provides fast access to data in a data warehouse. A cube is a set of data that is usually constructed from a subset of a data warehouse and is organized and summarized into a multidimensional structure defined by a set of dimensions and measures.

A cube provides an easy-to-use mechanism for querying data with quick and uniform response times. End users use client applications to connect to an Analysis server and query the cubes on the server. In most client applications, end users issue a query on a cube by manipulating the user interface controls, which determine the contents of the query. This spares end users from writing language-based queries. Precalculated summary data called aggregations provides the mechanism for rapid and uniform response times to queries. Aggregations are created for a cube before end users can access it. The results of a query are retrieved from the aggregations, the cube's source data in the data warehouse, and a copy of this data on the Analysis Server, the client cache, or a combination of these sources. An Analysis Server can support many different cubes, such as a cube for sales, a cube for inventory, a cube for customers, and so on.

Every cube has a schema, which is the set of joined tables in the data warehouse from which the cube draws its source data. The central table in the schema is the fact table, the source of the cube's measures. The other tables are dimension tables, the sources of the cube's dimensions. A cube is defined by the measures and dimensions that it contains. For example, a cube for sales analysis includes the measures `Item_Sale_Price` and `Item_Cost` and the dimensions `Store_Location`, `Product_Line`, and `Fiscal_Year`. This cube enables end users to separate `Item_Sale_Price` and `Item_Cost` into various categories by `Store_Location`, `Product_Line`, and `Fiscal_Year`.

Each cube dimension can contain a hierarchy of levels to specify the categorical breakdown available to end users. For example, the `Store_Location`

dimension includes the level hierarchy: Continent, Country, Region, State\_Province, City, Store\_Number. Each level in a dimension is of finer granularity than its parent. For example, continents contain countries, and states or provinces contain cities. Similarly, the hierarchy of the Fiscal\_Year dimension includes the levels Year, Quarter, Month, and Day.

Dimension levels are a powerful data modeling tool because they allow end users to ask questions at a high level and then expand a dimension hierarchy to reveal more detail. For example, an end user starts by asking to see Item\_Cost values of products for the past three fiscal years. The end user may notice that 1998 Item\_Cost values are higher than those in other years. Expanding the Fiscal\_Year dimension to the Month level, the end user sees that Item\_Cost values were especially high in the months January and August. The end user may then explore levels of the Store\_Location dimension to see if a particular region contributed significantly to the high Item\_Cost values, or may expand into the Product\_Line dimension to see if Item\_Cost values were high for a particular product group or product. This type of exploration, known as drilldown, is common in client applications. Fig.6, considers the following Imports cube, which contains two measures, Packages and Last, and three dimensions, Route, Source, and Time.

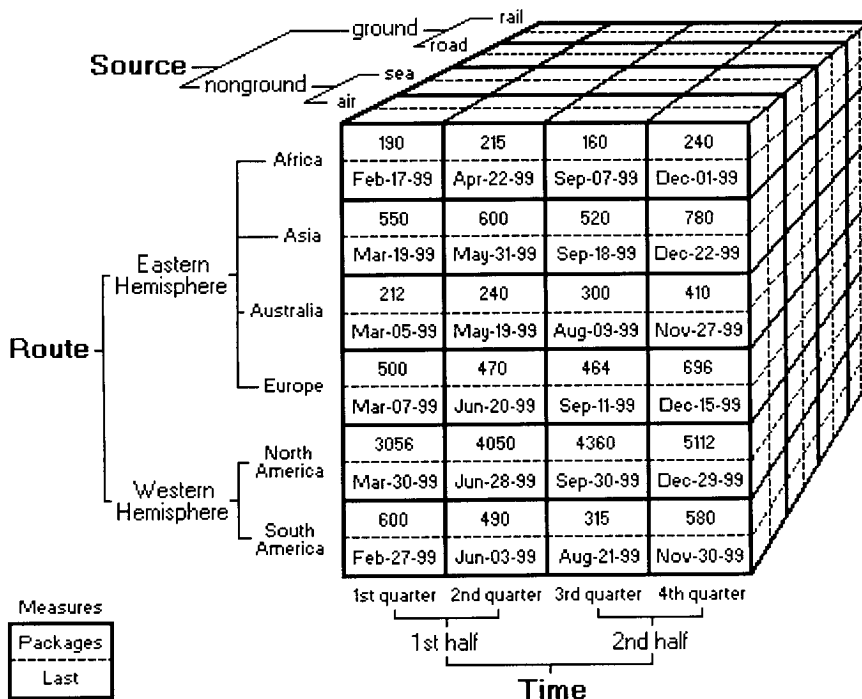


Fig. 6: Example of cube, Import cube

The smaller alphanumeric values around the cube are the members of the dimensions. Example members are ground, Africa, and 1st quarter.

The values within the cube represent the measures. Example measures are Packages: 190 and Last: Feb-17-99. These values exist for all cells in the cube but are shown only for those in the foreground. (In a real cube, the words Packages and Last would not appear in the cube cells, but they are shown here to distinguish the measures. In a real cube, measures are separated within a special dimension called the Measures dimension.)

The Packages measure represents the number of imported packages, and it aggregates by the Sum function. The Last measure represents the date of receipt, and it aggregates by the Max function. The Route dimension represents the means by which the imports reach their destination. The Source dimension represents the locations where the imports are produced. The Time dimension represents the quarters and halves of a single year.

End users of a cube can determine its measures' values for each member of every dimension. This is possible because measure values are aggregated by the members.

A cube can contain up to 128 dimensions, each with thousands or millions of members, and up to 1,024 measures. A cube with a modest number of dimensions and measures usually satisfies the requirements of end users.

Cubes are immediately subordinate to the database in the object hierarchy. A database is a container for related cubes and the objects they share. A database must be created before the creation of a cube. In the object hierarchy, the following objects are immediately subordinate to the cube:

1. Data sources: A cube has a single data source. It can be selected from the data sources in the database or created during cube creation. A cube's dimensions must have the same data source as the cube, but its partitions can have different data sources.
2. Measures: A cube's measures are not shared with other cubes. The measures are created when the cube is created. A cube can have up to 1,024 measures.
3. Dimensions: A cube's dimensions are either shared with other cubes in the database or private to the cube. Shared dimensions can be created before or during cube creation. Private dimensions are created when the cube is created.



Although the term cube suggests three dimensions, a cube can have up to 128 dimensions.

4. Partitions: A single partition is automatically created for a cube when the cube is created. If Analysis Services for SQL Server 2000 Enterprise Edition is installed, after the cube creation a cube, additional partitions in the cube can be created.
5. Cube roles: Every cube must have at least one cube role in order to provide access to end users. Cube roles are derived from database roles, which can be created before or after cube creation. Cube roles are created after cube creation.
6. Commands: Commands are optional. Commands are created after cube creation [26]

### **2.4.3 Analysis Services Data Mining Techniques**

Microsoft SQL Server 2000 Analysis Services did introduce a new feature, data mining, integrates significant data analysis and prediction capabilities into Analysis Services. PivotTable Service enables clients to interact with these new data mining features.

PivotTable Service supports data mining by providing support services that are very similar to the services it provides for online analytical processing (OLAP).

Two data mining algorithms are included with Analysis Services: Microsoft Decision Trees and Microsoft Clustering. The decision trees algorithm is based on the notion of classification. The clustering algorithm uses an expectation-maximization method to group records into clusters (or segments) that exhibit some similar, predictable characteristic. [26]

### **2.4.4 SQL server 2000 Data Mining Algorithms**

Data mining technology analyzes data in relational databases and OLAP cubes to discover information of interest. The data mining features of Microsoft SQL Server 2000 Analysis Services are incorporated in an open and extensible implementation of the new OLE DB for Data Mining specification. SQL Server 2000 includes data mining algorithms developed by Microsoft Research.

Central to the data mining process, data mining algorithms determine how the cases for a data mining model are analyzed. Data mining model algorithms provide

the decision-making capabilities needed to classify, segment, associate and analyze data for the processing of data mining columns that provide predictive, variance, or probability information about the case set.

Many data mining algorithms are goal-oriented; given a case set, a data mining algorithm will predict something about the case, usually an attribute of the case itself. Most algorithms require a training set of cases where the attributes to be predicted are already known, at which point the algorithm constructs a data mining model capable of predicting these attributes for cases in which the attributes are unknown.

Data Mining Algorithm Providers:

In SQL server 2000 the following data mining algorithms are used.

- Decision Trees:

A decision tree is a form of classification shown in a tree structure, in which a node in the tree structure represents each question used to further classify data. The various methods used to create decision trees have been used widely for decades, and there is a large body of work describing these statistical techniques.

- Clustering:

Like decision trees, clustering is a well-documented data mining technique. Clustering is the classification of data into groups based on specific criteria. The topic discussing the Microsoft Clustering algorithm goes into greater detail regarding the details of clustering as a data mining technique. [26]

## **2.5 Background and Previous Studies**

So many studies have been performed on the CRM subject, and customer profiling especially has been tackled from all sides.

All studies were considering one parameter at a time of this subject, some were considering the demographic parameter, others the behavioral or the analytical side like customer loyalty or customer lifetime value, what makes the suggested algorithm special and unique is that this algorithm gives a global view about a customer and its relationship with the company, it starts by building the demographic profile of the customer, then continues to analyze the loyalty and lifetime value.

In addition to this, the proposed metric segregates the customer into segments allowing an analysis on the behavioral side of the client.

All other studies were relying on huge amounts of data gathered from so many applications. The suggested algorithm relies only on the sales transactions which are available at any site of business. From this restricted amount of information, a user can extract and build a very clear global image about customers.

Concerning building the demographic profile of the customers, most of the studies rely on the direct contact with the customer through phone, emails... using questionnaires templates... others have decided on the demographic profile of a customer by using similarities with other defined customer profiles.

The suggested algorithm which extracts customer demographic profiles is much reliable due to many facts:

- The customer profile is built without going into the hassle of contacting customers; which makes it costly effective.
- This algorithm keeps the customer profile up to date; it does not need a permanent contact with the customer.
- It is directly related to the need of the company since parameters are user defined according to the company's line of business.

Gediminas Adomavicius and Alexander Tuzhilin in their study 'Using data mining methods to build customer profiles' [2] suggested an algorithm that they called 1:1Pro to extract the behavioral customer profile out of the historical transactions. The factual data which is basically the demographic info about the customer was directly retrieved from the customer file. In the algorithm suggested the demographic data is not available in the customer files and is being extracted from the transactional data.

In the behavioral part, they were relying on the data mining association and classification techniques. Those techniques did generate a set of rules many of which although they were from the statistics point of view acceptable they were trivial, spurious, or just not relevant to the application. In order to extract the acceptable rules, they had to rely on human experts to dissociate the needed info. The analytical side of the customer profiling was not tackled in the 1:1Pro algorithm.

France Leclec in his study 'Quantifying Customers' [13] segments customers by behavior and generates quantitative measures of loyalty based on that behavior.

This approach has important advantages:

First, it is based on actual purchases, rather than on age, zip code, gender, ethnicity or other demographic.

Second, it is based on data that already exists in order entry or accounting systems, so results are available in real time.

Third, every transaction adds to the data set so learning is continuous.

Fourth, the parameters of loyalty (size of purchase, frequency, Recency, duration, product categories) create a holistic picture of the customer and produce segmentation that illuminates behavior patterns which may otherwise go unnoticed.

D. R. Mani in his study 'Lifetime value Modeling, the most valuable metric' [9] which is built on transactions history, suggested the following LTV Formula:

$$\text{Pr(Active)} * \text{Risk} * [\text{Cross-Sell/Up-Sell+Product Profitability}] * \text{Persistency Index} - \text{Marketing Expenses}$$

Following is a listing of terms used in the above formula:

### **Pr(Active)**

Pr(Active) is the probability of becoming an active account. This value is a probability derived from a predictive model. Using data from a past campaign, logistic regression is used to create a model. (Logistic regression is a statistical technique that uses continuous values such as age and income to predict a binary outcome such as 'active' or 'non-active' account status.)

This component combines response, approval and activation. The target for the model becomes those accounts that responded, were approved and activated vs. all others.

### **Risk**

Risk is an index value. Each value represents an adjustment to the average for the risk of a death claim. These values are taken from historical research provided. It is also possible to use a model to predict this amount.

### **Cross-sell/Up-sell**

Cross-sell/Up-sell is an additional amount representing average net revenues over a five-year period. This value is calculated using linear regression on historic data. (Linear regression is similar to logistic regression, except that the amount that is predicted is a continuous value.)

### **Product Profitability**

Product Profitability is the net revenue amount supplied. This, in combination with the cross-sell/up-sell revenues, represents the core value for the Lifetime Value measure.

### **The Persistency Index**

The Persistency Index is an adjustment based on the type of payment plan. If a customer signs up for an automatic deduction from his checking account to pay his premium, he has much higher persistency.

### **Marketing Expense**

Marketing Expense is the amount per piece.

Jeffrey Pease in his study 'Customer Value Management, New Techniques for Maximizing the Lifetime Profitability of your Customer Base' [16] when explaining about the 3 "Rs" (Right Customers, Right Relationship and Right Retention) of the CVM cycle, in the paragraph talking about 'Right Relationship' provided a formula to simplify the view of customers LTV.

$$\text{LTV} = \text{purchase size} * \text{frequency} * \text{duration}$$

## CHAPTER 3: A CUSTOMER PROFILING ALGORITHM BASED ON TRANSACTIONAL DATA

### 3.1 Introduction to the new approach in extracting customers' profile:

In general, extracting or building customer profile does rely on information supplied by the customer himself. As stated before, the customer himself should submit some information about himself like customer name and address; however other components in customer profile can be extracted from customer transactions.

Using this new approach, a company will be able, out of the purchasing transactions, to:

- Extract customer's demographic profile
- Measure each customer's loyalty
- Measure each customer's Lifetime value
- Determine and focus on customers with highest Loyalty and LTV
- Predict the demographic profile of the prospects of best interest for the company to target and acquire

Extracting customer profiles based on the customer transactions means that the sales data is to be used.

Sales data is divided into two segments, static data and transactional data:

Static data: consists mainly of the customer files and the product files.

- Customer file contains an ID for each customer, name, address, ...
- Product file contains the product code, description, cost, selling price...

Transactional data: consists of records of the customer's purchases during a specific period. A purchase record includes the purchase date, product purchased, amount paid, item cost at that date.

### 3.2 Extracting the demographic profile:

As stated before, a customer profile consists of many segments. The demographic profile segment is the basic one in which a detailed idea of the customer identity could be built.

### **3.2.1 Preparation phase**

The first phase in retrieving a customer demographic profile is the preparation phase. During this phase, the company has to set the information needed in the customer demographic profile to build up the customer's identity. The preparation phase is very essential in the thesis. In other words, the company has to set its goals and information needed regarding the customer demographic profile definition.

The preparation phase consists of three steps which are:

- Definition of the customer profile template
- Definition of the range of values for each profile parameter
- Product categorization according to the range of values defined for each of the profile parameters

### **3.2.2 Definition of the customer profile template**

Usually, little information about the customer is available in the company's database. It could be the customer ID (credit card number, name), email and phone number.

For a company looking into building a meaningful and complete profile about the customers, it must define the parameters that are essential and related to the company's line of business, taking into consideration its future projects and goals.

Those parameters differ from one company to another based on a company's line of business.

- For a store selling sports equipment, the information needed in customer demographic profile will be: age, gender, hobby ....
- For a company selling cars and automotive equipments, the information needed in customer demographic profile will be: age, gender, marital status, income...

### **3.2.3 Definition of the range of values for each profile parameter**

In order to build a customer demographic profile, the company has to decide on the parameters that are essential in the creation of the customers identity.

Each one of these parameters could have two to more values.

Some of these parameters could have in their values the "ALL" value (because some of the products, when assigned to these parameters, could be generic to all the

values and not a specific only one. In the next section, this idea will further be elaborated.

Parameters of the demographic profile can, in their turn, be divided into three types:

- Type A: Parameters that accept one value and only one value. A good example of this will be the “Income” parameter, where a person can have one value in this parameter; either his income is “High” or “Average” or “Low”.
- Type B: Parameters with one value but where the “ALL” value can be accepted. For Example, the parameter “Gender”, items can be assigned specifically to the Male or Female gender; but some items can be used by both, so the “ALL” value can be acceptable in such case
- Type C: Parameters that can have more than one value. A good example of this will be the parameter “Hobby”; a person could have more than one hobby.

So after defining a company’s demographic profile template, the type and range of values for each parameter should be defined.

Parameters	Type	Values
Gender	B	ALL
		Female
		Male
Age	B	ALL
		Kids
		Adults
Marital Status	B	ALL
		Single
		Married
Income	A	High
		Average
		Low
Hobbies	C	Football
		Golf
		Basketball
		Tennis

Fig. 7: Profile parameters with their acceptable values



### 3.2.4 Product categorization according to the range of values defined for each of the profiles parameters

Once the customer profile template and the range for each parameter is defined, product categorization should be done.

Product categorization should be performed by assigning a value for each combination of product, profile parameter.

One essential point to mention, as described before, is that in some cases the value to be assigned for (products, profile parameter) is generic and not specific. That is why the *ALL* parameter should be assigned in the values for all parameters.

An example to clarify this point is that of a product used by both males and females; the value to be assigned in this case will be *ALL* rather than *MALE* or *FEMALE*.

Consider the case of a golf ball. Assigning values for the golf ball and each of the profile parameters we agreed on previously could result in:

- Gender: ALL
- Age: Adult
- Marital Status: All
- Income : Average to High
- Hobbies: Golf

The case of an expensive female swimming suit for pregnancy could generate other results such as:

- Gender: Female
- Age: Adult
- Marital Status: Married
- Income : High
- Hobbies: Swimming

In other words, at this stage, while both the female swimming suit and the golf ball belong to category *HIGH* concerning the income, Concerning the parameter gender, the golf ball belongs to category *ALL* and the swimming suit belongs to category *FEMALE*.

### 3.2.5 Extraction of the Customer Demographic Profile

The extraction of the customer’s demographic profile will go through two phases:

- Phase 1: extracting tentative profile
- Phase 2: extracting the real profile

#### 3.2.5.1 Extracting the tentative profile:

Based on the customers’ purchase transactions, the customer demographic profile will be built based on the items purchased by the customer.

Each item purchased will be replaced by its corresponding value defined for the profile template parameters.

The quantity purchased is the factor considered in this phase.

The calculation results in extracting for each (parameter, value) example: (gender, Male) its corresponding percentage form the total purchase. So, at this stage a percentage is calculated for each value of the different parameters constituting the customer demographic profile. Having done this, the value with the highest percentage (thus considered the value for this parameter) can be extracted for each parameter.

In fig.8 the transactions analysis of customer C0001 generated the percentage values for each of the profile parameters. For the parameter AGE, 83% of the items purchased by C0001 correspond to adults, 2% can be used by all ages and 15% of them are for Kids. The tentative profile of Customer C001 is represented in fig.9; it is based on the highest percentage corresponding to each parameter.

<b>AGE</b>	ALL	2.00%
	ADULT	83.00%
	KID	15.00%
<b>GENDER</b>	ALL	55.00%
	MALE	23.00%
	FEMALE	22.00%
<b>MARITAL STATUS</b>	ALL	49.00%
	MARRIED	42.00%
	SINGLE	9.00%
<b>INCOME</b>	High	2.00%
	LOW	20.00%
	AVERAGE	78.00%
<b>HOBBIES</b>	Golf	5.00%
	FOOTBALL	55.00%
	BASKETBALL	29.00%
	TENNIS	11.00%

Fig.8: Results in percentage for each parameter

<b>AGE</b>	ALL	2.00%	ADULT
	ADULT	83.00%	
	KID	15.00%	
<b>GENDER</b>	ALL	55.00%	ALL
	MALE	23.00%	
	FEMALE	22.00%	
<b>MARITAL STATUS</b>	ALL	49.00%	ALL
	MARRIED	42.00%	
	SINGLE	9.00%	
<b>INCOME</b>	ALL	0.00%	AVERAGE
	LOW	20.00%	
	AVERAGE	78.00%	
	HIGH	2.00%	
<b>HOBBIES</b>	ALL	5.00%	FOOTBALL
	FOOTBALL	55.00%	
	BASKETBALL	29.00%	
	TENNIS	11.00%	

Fig.9: Customer C0001 tentative profile

- At this stage, parameters corresponding to type A (accepting one and only one value) are resolved by assigning to them the highest percentage value. In the fig.9 the parameter *INCOME* is assigned the value *AVERAGE* since it has the highest percentage rate.

### 3.2.5.2 Extracting the real profile based on sophisticated rules:

The term tentative is used because this current profile is still far from completion; parameters of type B and C are still not resolved. In order to complete this profile, more sophisticated rules adopted, other than the simple previous one where the highest percentage of the value was used.

Two types of rules should be set according to the type of parameter.

- Parameters of type B (accepting generic value). For this type of parameters the user has to define a dominant percentage value. The highest parameter value having a percentage equal or higher than this specified percentage is considered otherwise the generic value is assigned to this parameter.
- Parameters of type C (accepting more than one value), an acceptable value should be defined, in other words the user has to define the acceptable percentage of a value to be considered. Justification: a

client whose hobby related purchase resulted in 30% for *Football*, 55% for *Basketball*, 6% for *Golf* and 9 % for *Tennis* means that his major hobby is *Basketball* while the hobby *football* is to be taken into consideration because its value is important.

In fig.10 the Customer C0001's final profile is represented. By considering 65% as dominant factor and since parameter age is a parameter of type B, Adult is the result for Age because it has 83% as the resulting values. Hobbies is a parameter that accepts more than one value, considering 20 % as acceptable value, the result is *Football* and *Basketball* since both have results greater than 20%.

<b>AGE</b>	ALL	2.00%	ADULT
	ADULT	83.00%	
	KID	15.00%	
<b>GENDER</b>	ALL	55.00%	ALL
	MALE	23.00%	
	FEMALE	22.00%	
<b>MARITAL STATUS</b>	ALL	49.00%	ALL
	MARRIED	42.00%	
	SINGLE	9.00%	
<b>INCOME</b>	HIGH	2.00%	AVERAGE
	LOW	20.00%	
	AVERAGE	78.00%	
<b>HOBBIES</b>	Golf	5.00%	FOOTBALL, BASKETBALL
	FOOTBALL	55.00%	
	BASKETBALL	29.00%	
	TENNIS	11.00%	

Fig. 10: Customer C0001 final demographic profile

To summarize, using the customer transactions and based on product categorization, we established first the customer tentative customer profile. More sophisticated rules have been set. Those rules were adopted to move the tentative profile to amore solid final profile.

### 3.3 Measurement of customer loyalty and customer lifetime value:

The second factor to be tackled when building a customer profile is the measurement of the customer loyalty (LV) and customer lifetime value (LTV).

Customer loyalty and LTV should be ranked based on a range of values rather than judged. Whether a customer is loyal or not, or whether a deal with a customer is profitable or not has to be measured.

The Reason for this is that such crucial information should be clear for management to get a comprehensive idea about the customers' relation with the company and to establish future projects, plans and goals. Ranking customer according to LTV and LV, provide flexibility in evaluating the evolution of the relationship between the organisation and the customers. Such ranking identify the customers constituting the source of profit for the organisation.

Consider a ranking for customer loyalty of scale up to 5, this means customers ranked 3/5 on loyalty scale should be addressed in a different way than a customer ranked as 1/5. Customer ranked 3/5 is generating more profit to than the one ranked 1/5.

Measurement of the customer loyalty and LTV basically consists of four steps:

- Step 1: Determination of the factors based on which the customer loyalty and the LTV will be measured.
- Step 2: Measurement of scale definition and ranking factors' values accordingly
- Step 3: Assigning a weight for each factor
- Step 4: Measuring the customer loyalty and customer lifetime value

### **3.3.1 Determination of Loyalty factors**

France Lelec in his study "Quantifying Customers" states that loyalty parameters Size of purchase, Frequency, Recency and Duration create a holistic picture of a customer and illuminates hidden patterns. [12]

Measuring customer loyalty cannot rely on a single factor; using one factor for such ranking will not result in accurate customer loyalty ranking.

Imagine using the total amount of purchase for a customer to determine loyalty, such a method can easily fall into traps. The reason for this is the following:

A customer purchased only once from a company in the last 2 years. The total amount of this single purchase was huge. Using only the factor "total amount of purchase" to calculate loyalty, results into placing this customer at high loyal rank. looking deeply into this customer's records we can recognize that it is not the case, this customer is not loyal to the company he only purchased once during the 2 past years.

So, to determine the customer loyalty, we should rely on many factors and not just one.

Since the only data we have is historical transactions, loyalty indices on which we can build our loyalty studies are:

- Monetary: Total amount purchased
- Frequency: Number of purchases
- Quantity: Number of product purchased
- Retention time: duration of customer relationship with the company
- Recency time: time spent since that last purchase

### **3.3.2 Determination of Customer Lifetime Value factors**

Customer lifetime value being the current and potential commercial benefit of the relationship with a customer calculated overtime can be divided into two segments:

- Historical lifetime value
- Predictive lifetime value

In order to come out with a formula for the calculation of LTV, two studies were considered. The first study is 'Customer Value Management, New Techniques for Maximizing the Lifetime Profitability of your Customer Base' by Jeffery Pease [16] and the second one is for D. R. Mani, 'Lifetime value Modeling, the most valuable metric' [9]

In the first study Jeffery Pease suggests the following as formula to calculate a simplified view of Customer Lifetime Value

$$\text{LTV} = \text{Purchase Size} * \text{Frequency} * \text{Duration}$$

D. R. Mani stresses on the importance of the profit factor in the calculation of LTV when he stated 'Lifetime value is the net present value of all future contributions to overhead and profit expected from a new customer. In simpler terms, how much a customer is worth to you today, given how much profit she/he will generate in the future'. [16]

In this study the factors considered for the calculation of the lifetime value of a customer are three:

- Frequency: Number of purchases
- Retention time: duration of customer relationship
- Profit

The values Frequency and Retention time are shared with the loyalty factor

However there is one factor still missing in order to be able to calculate the customer LTV it is the **Profit**.

The Profit achieved during the whole relationship period with the customer can be defined as follows:

Profit = Monetary (total purchase)-total expenses.

Due to the limited info available, since we only have the purchasing transactions, total expenses will be restricted to items' cost.

### 3.3.3 Measurement's Scale definition and rank of factor's values accordingly

Due to the variety in the unit of measurement of these factors (total amount purchased is measured by amount, number of purchases by number, Recency time by month and year) and because of the wide range of those values (amount purchased might reach thousands of dollars); a unique unit of measurement should be adapted in the evaluation of each factor. In order to achieve this unification, a scale for such measurement should be set.

Scale of values can for example be ranged from [0.....10], or a wider range according to the users' case.

Having defined the scale of our measurement, the next step is to assign the scale values to the parameters used in the calculation of loyalty. This is achieved by distributing the results of the parameters in ranges over the scale

In fig.11 an interval scale ranged from 1 to10 is considered, values of each parameters are distributed according to the range of values.

	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	Scale 6	Scale 7	Scale 8	Scale 9	Scale 10
Monetary: Total amount purchased	< 1000\$	> 1000\$ and < 5000\$	> 5000\$ and < 10000\$	...	...	...	...	...	...	...
Frequency: Number of purchases	< 5	> 5 and < 10	> 10 and < 25	...	...	...	...	...	...	...
Number of product purchased	< 15	> 15 and < 35	> 35 and < 65	...	...	...	...	...	...	...
Retention time	< 1 month	> 1 month and < 3	> 3 months and < 8	...	...	...	...	...	...	...
Recency time	> 1 year	< 1 year and > 11 months	< 11 months and > 10	...	...	...	...	...	...	...
Benefit	< 10 %	> 10 % and < 20%	> 20% and < 30%	...	...	...	...	...	...	...

Fig. 11: Distribution of the values factors over the loyalty scale

Each company according to its activity has to determine these ranking per factor. Imagine a supermarket and retail clothes shop.

The “Recency” factor, for a supermarket may depend on weeks and days, due to its daily activity while for retail clothes shop the Recency factor may depend on months and seasons.



### 3.3.4 Assigning weight for each factor to extract loyalty

As stated previously, the customer loyalty factors considered are the following:

- Monetary: Total amount purchased
- Frequency: Number of purchases
- Quantity: Number of product purchased
- Retention time: duration of customer relationship
- Recency time: time since that last purchase

Having defined the ranking per factor, a company should also define a weighted value for each factor in order to build up a global ranking loyalty factor. Each factor weight may differ from one company to another according to the company activity.

For a supermarket, the frequency parameter is as much important as the monetary value.

Recency time for a car dealership is not that much important, because the car lifetime is high. A customer may take more than 5 years before buying a new car.

Loyalty Parameters	Weighted Value
Total amount purchased	8
Frequency: Number of purchases	11
Number of product purchased	5
Retention time	7
Recency time	9
<b>Total</b>	<b>40</b>

Fig. 12: Five Loyalty factors, with their corresponding weighted indices

In fig. 12, summing up all weighted values resulted 40, which is the base for determining customer loyalty.

### 3.3.5 Extracting the scaled values

Once all definitions are set, extracting the value for each factor out of the purchasing transactions can be easily done.

Each value extracted should be replaced by the scaled value (SV) according to the scale adopted.

### 3.3.6 Customer Loyalty measurement

The customer loyalty value is the sum of each scaled value \* the weight of the corresponding factor.

$$\sum(SV_i * W_i)$$

The Max Loyalty value constitutes the basis for customers' loyalty ranking. In fig. 12, the total of weighted values is 40. Based on this total of weighted values and the number of scale ranges [0...10], a perfect customer loyalty ranking will be 400. The company defines loyalty ranking based on this total.

Considering a ranking range of 1 up to 5, this means the 400 being the largest value should be distributed over 5. The representation is clarified in fig. 13.

<b>Ranking Five</b>	<b>350 to 400</b>
<b>Ranking Four</b>	<b>260 to 349</b>
<b>Ranking Three</b>	<b>171 to 259</b>
<b>Ranking Two</b>	<b>60 to 170</b>
<b>Ranking One</b>	<b>0 to 59</b>

Fig.13: customer Loyalty Value distributed over 5 ranges

Customer loyalty increases when moving from rank 1 toward the highest rank.

Taking Customer C0001 as an example, his results were:

Total amount purchased is 110,000 USD. This means SV for parameter total amount purchased is 4. Similar approach applied on the other parameters in Fig. 14, the other parameters results are displayed.

Loyalty Parameters	Scale over 10
Total amount purchased	4
Frequency: Number of purchases	7
Number of product purchased	5
Retention time	6
Recency time	7

**Fig. 14:** Customer C0001 transactions results concerning loyalty parameters scale values

Multiplying the weighted factors defined in fig. 12 by the scale values of each parameter, the result of total amount purchase becomes  $4 * 8=32$ . Results of the other parameters detailed in fig. 15.

Loyalty Parameters	Results
Total amount purchased	$4 * 8= 32$
Frequency: Number of purchases	$7 * 11= 77$
Number of product purchased	$5 * 5= 25$
Retention time	$6 * 7= 42$
Recency time	$7 * 9 = 63$
<b>Total</b>	<b>239</b>

**Fig. 15:** Loyalty ranking result for customer C0001

Considering the fig.13, the customer loyalty ranking, this customer will be placed in Ranking “3”, which is relatively a good ranking in customer loyalty

### 3.3.7 Customer Lifetime Value measurement

Once loyalty is measured, the next factor to compute is the Customer Lifetime value.

The LTV formula consists of the following:

$$\text{LTV} = \text{Profit (SV)} * \text{frequency (SV)} * \text{duration (SV)}$$

Since the LTV consists of two segments **historical** and **predictive**, this formula will be developed as to combine the two values; it becomes:

$$\text{LTV} = \text{Historical LTV} + \text{Predictive LTV}$$

$$\text{Historical LTV} = \text{Historical Profit (SV)} * \text{Historical frequency (SV)} * \text{duration (SV)}$$

Concerning the Predictive LTV:

Since marketers cannot know the duration of the relationship with the customer until it is over, the user should specify the period for which the prediction study will be based on. This period will be referred to by the term **prediction duration (Pduration)**.

Starting from the fact that the relationship with the client will keep on being stable for the coming period (Prediction period) and by integrating this prediction duration constant (Pduration) to the formula, the Predictive lifetime formula to be adapted becomes:

$$\text{Predictive LTV} = (\text{Historical Profit (SV)} * \text{Pduration}) / \text{duration} + (\text{Historical frequency (SV)} * \text{Pduration}) / \text{duration} + \text{Pduration}$$

In the calculation of the lifetime value, only historical LTV will be considered. Predictive LTV will be neglected in this study.

A perfect score in customer life time value will consists of

$$\text{LTV} = \text{Max Profit (SV)} * \text{Max Frequency (SV)} * \text{Max Duration (SV)}$$

Taking as example fig. 11, the perfect LTV score will result in:

$$10 * 10 * 10 = 1000$$

LTV has also been ranked depending on results factor, the highest the score the more LTV is profitable for a customer.

In fig. 16, we are considering a ranking up to 10.

<b>Customer LTV ranking</b>	
Rank Ten	950 – 1000
Rank Nine	850 – 949
Rank Eight	700 – 849
Rank Seven	500 – 699
Rank Six	421 – 499
Rank Five	351 – 420
Rank Four	281 – 350
Rank Three	200 – 280
Rank Two	100 – 200
Rank One	0 -100

Fig.16: Customer LTV ranking

Having done all of this, customer demographic profile will have been defined; customer loyalty and customer LTV will have been determined based on customer transactions.

Now that customer loyalty and LTV are determined, comes the question of how to get a clearer view about the distribution of the customer according to those parameters? Since those parameters are tightly linked to profitability, How can we get a distribution of the customers according to profitability?

### 3.4 Building the loyalty and LTV metric

A metric containing the distribution of clients by loyalty and LTV should be built. The dimensions of this metric are the LTV and the loyalty ranks. Each cell of this metric contains the number of clients falling in the corresponding range of Loyalty and LTV.

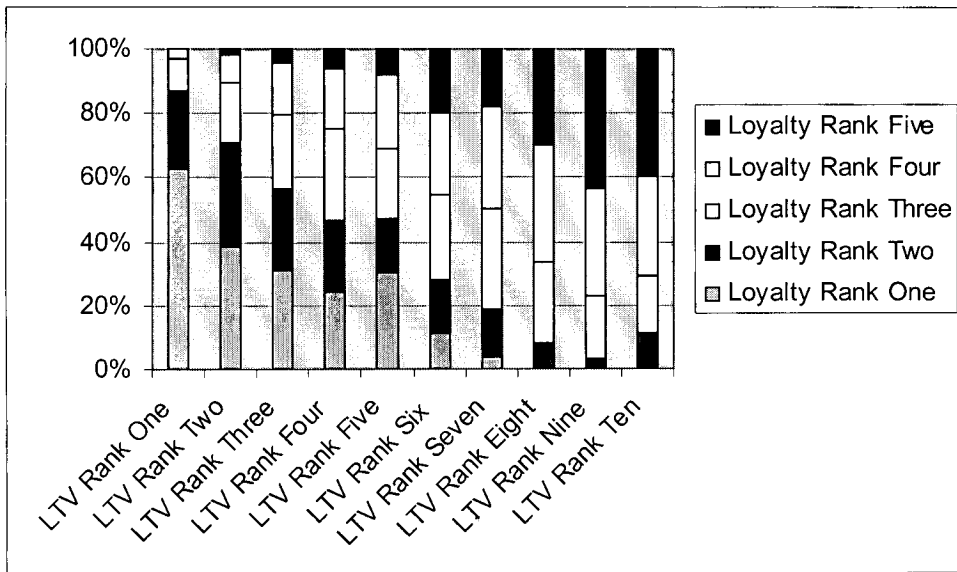
In fig 17 represented below, we can say that cells which range from [LTV rank five, LTV Rank ten] and from [Loyalty rank three and Loyalty rank five] are the key cells in the company's business. Those cells represent the customers that are generating profit for the organization. Those are the most important customers that need great care.

The scope of clients falling in the range from [LTV rank one, LTV rank four two] and from [Loyalty rank one, Loyalty rank two] are of less importance for the company.

	Loyalty Rank One	Loyalty Rank Two	Loyalty Rank Three	Loyalty Rank Four	Loyalty Rank Five
LTV Rank One	2850	1099	456	126	9
LTV Rank Two	2569	2134	1256	567	120
LTV Rank Three	1678	1345	1231	876	234
LTV Rank Four	1345	1234	1567	999	356
LTV Rank Five	1750	987	1235	1342	467
LTV Rank Six	459	657	1090	1023	786
LTV Rank Seven	125	456	987	989	567
LTV Rank Eight	26	234	857	1234	986
LTV Rank Nine	17	112	878	1456	1879
LTV Rank Ten	4	344	567	976	1243

Fig. 17: Distribution of the customer according to their Loyalty and LTV values

Representation of this table in a graph clarifies the Loyalty percentage for each LTV rank.



### **3.5 What are the benefits of this metric, when and how to use it?**

**Understand the customers:** Out of this matrix, a company can easily extract the hidden sales potentials and high potential targets. It can also figure out the percentage of customers producing profit.

**Retrieve the most valuable and loyal clients:** The target should be the most profitable client; this means clients falling in cell LTV Rank Ten and Loyalty rank five. Using data mining clustering techniques on the demographic data of these clients, we can extract which combination of demographic parameters is the most profitable. Examples (gender: male, hobby: football constitutes 60 percent) of the most valuable clients. For this range of clients the company can launch a rewards program.

**Launch a retention program:** The targeted customers could be those ranked four concerning the loyalty factor, and ranked eight and nine concerning LTV.

**Launch a new prospect-targeting plan and find the best sales prospects:** having extracted the profile of the most valuable customer, using the collaborative filtering technique we can invest on prospects having similar parameters.

**Lower marketing expenses:** Since the customers are very well segmented and categorized, this makes the advertising effort more targeted and more focused leading to efficiency.

## CHAPTER 4: APPLICATION OF THE SUGGESTED ALGORITHM ON REAL DATA

### 4.1 Introduction

In chapter 3 a new approach for extracting customers profile out of historical transactions has been defined. In order to test the validity of this approach, real data is needed. The data needed to test the model is the sales transactions along with the clients file and the products being sold. Such data is very confidential since:

- It reveals the real costs of some items along with the profit achieved by a company
- It reveals the clients of the company, which is very valuable to competitors.

Finding such data was not an easy issue, but it got solved on one condition which is confidentiality.

The test has been conducted on data belonging to store selling sports equipment.

This department store has 17 branches spread all over Lebanon. It has thousands of customers and a huge daily sales volume.

It has a wide variety of sports items covering almost all different age, gender, hobbies...

### 4.2 Data Description

This store has 16 branches and a main branch. Each of the 16 branches has between 1 to 3 personal computers. In the main branch, where the regional office is located, there are 13 personal computers.

In the main server is in the main branch, and all PCs in all branches communicate with this server.

Software wise, The ERP solution used in this store is provided by Profiles Software SARL.

The store uses 2 applications in order to carry out with its work:

- PIMS2: known also as Profiles Integrated management system can be considered as the back-office application used by the store to manage its solid data. In this system, all needed information concerning clients,



their daily transactions, products, products categories and so on... are available and will be used to carry out this study.

- SwiftPos: is a point of sale software spread over the store branches. Using this software, all daily sales to different customers are performed stored and communicated to the back-office application located in the main branch

The data submitted contains a lot of tables allowing the store to control the whole ERP tasks such as the inventory in the warehouses, the suppliers and customers corresponding transactions..... What are needed to test this approach are only the sales transactions, the clients and the products. With the help of their IT we succeeded to extract the needed tables for the purposes of this research.

The data extracted consists of:

Transactions tables: containing the sales transactions of the shop. It consists of two tables: transactions headers and transactions details.

- Transactions header (named StkDoc): it contains the client code, the transaction date and the total amount purchased, and so many other fields. Each record in the transaction header table is related to one or several records in the transactions details table.
- Transactions details (named StkLin): It contains the items purchased along with the quantity and the price. Each record is related to one line in the transaction header table.

Products table (named StkItems): It contains the items code, description, along with the cost

Clients table (named AuxilAct): It stores the client code, name, address, phone number and email along with other information which are irrelevant to our study.

### **4.3 Cleaning and preparing the data**

Before applying the proposed approach on the available data, this data should be first cleaned and prepared.

- Inconsistent data formats, data encoding geographic spellings, abbreviations and punctuation have been resolved

- Unwanted fields were stripped out because they are meaningless for the work to be carried out
- As stated above the transactions tables consists of two related tables header and lines. Those two tables were combined into one table allowing the access to those tables easier

#### 4.4 Applying the suggested model

Once the data has been prepared and cleaned, the application of the suggested model can start.

In order to carry out with this study, a database in *SQL server 2000* has been created. The database created was called *CustProfil*, represented in fig.18. It includes the tables extracted from *PIMS2* that are needed to this study.

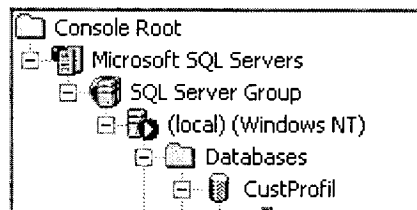


Fig.18: Database "CustProfil" created in SQL server 2000

As stated in chapter 3, the first step in the suggested approach is the preparation phase. The preparation phase consists of three steps

- Definition of the customer profile template
- Definition of the range of values for each profile parameter
- Product categorization according to the range of values defined for each one of the profile parameters

##### 4.4.1 Definition of the customer profile template

Being a sports equipments store, the profile information needed for the business is:

- **Age:** The age is an important parameter for sports equipment. Not all ages are allowed to practice all kinds of sports and equipment sold in general depends on the age of the user.
- **Gender:** Categorizing the customers by gender is essential for this store especially that many sports and exercises depends on gender.

- **Income:** Some of the sports equipments are very expensive and it is essential to know who are interested in such equipments
- **Hobbies:** Since the selling sports equipment is based on the customer hobbies, so the hobby parameter is the most important parameter to extract.
- **Marital Status:** Such parameter is also important since it categorizes some items.

#### 4.4.2 Definition of the range of values of each template

A new table named *ProfilDef* has been created; this table includes the different parameters needed to extract the customer demographic profile. In this table each existing profile parameter has been assigned an ID, description and type. Concerning types and as already discussed in chapter three represented in fig.19, it can be one out of three.

- Type A: Parameters that accept one value and only one value. A good example of this will be the “Income” parameter, where a person can have one value in this parameter; either his income is “High” or “Average” or “Low”.
- Type B: Parameters with one value but where the “ALL” value can be accepted. For Example, the parameter “Gender”, items can be assigned specifically to the Male or Female gender; but some items can be used by both, so the “ALL” value can be acceptable in such case
- Type C: Parameters that can have more than one value. A good example of this will be the parameter “Hobby”; a person could have more than one hobby.

	ProfileCode	ProfileType	Factor
▶	AGE	B	60
	GENDER	B	60
	HOBBIES	C	20
	INCOME	A	0
	MARITALSTATUS	B	60

Fig. 19: Demographic profile parameters, types and their factor

Another table *ProfilParam*, has been also created. This table stores the profiles parameters along with all the acceptable values. This is clarified in fig. 20.

	catid	ValueDesc
1	AGE	ADULT
2	AGE	KID
3	GENDER	FEMALE
4	GENDER	MALE
5	GENDER	UNISEX
6	HOBBIE	ATV
7	HOBBIE	BALLET

Fig. 20: Demographic profile parameters values

#### 4.4.3 Product categorization according to the range of values defined for each of the profile parameter

For the purpose of categorizing the products according to the parameter values, a table *ProdCateg* has been created, this table is one of the key essential tables for our study. In this table each product is related to a profile parameter value.

The whole preparation phase was carried out by the store manager which has a solid knowledge about each item.

#### 4.4.4 Extracting Customer Profile

Once the database was created in *SQL server*, the first step was to extract the customer's tentative profiling.

*The Analysis services* utility of *SQL Server 2000* was used to achieve this goal. Below are in details the steps followed to extract the customer tentative profiling.

##### Step 1:

- A new Database in Analysis services has been created
- In this new Database, a DataSource has been created and connected to the SQL database, this connection is represented in fig.21

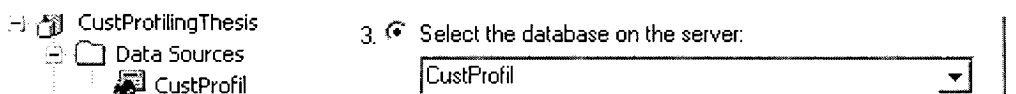


Fig. 21: Analysis services connection to an SQL server database

- Once connection is established and tested, a cube has been created and designed to generate the requested results. Cube creation is represented in fig.22.

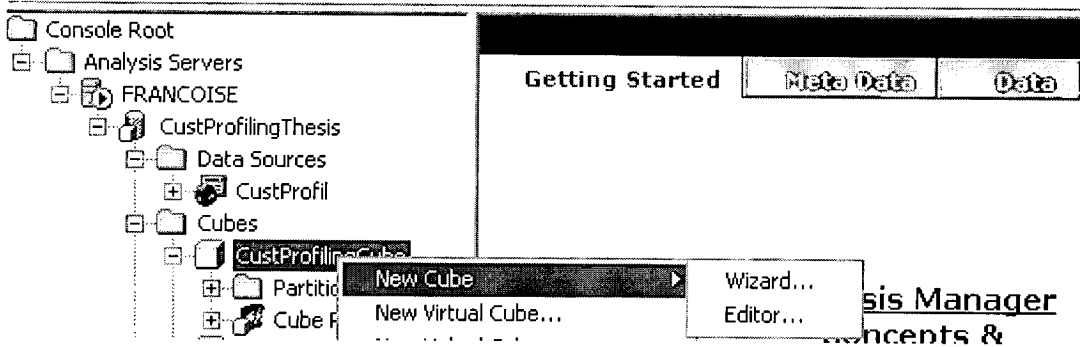


Fig. 22: Cube “CustProfilingThesis” in Analysis Services

Creation of a cube can be fulfilled by using wizard or an editor. The following details how it is done using the wizard:

- A fact table is the first to be selected. The *fact table* is the table which contains the information to be measured to build up the cube. It can be the quantity sold or the amount sold. In other words, the fact table should be the table which holds the transactions of the customer. In this case, the fact table is the *stklin* table because it holds all the customer sales transactions.

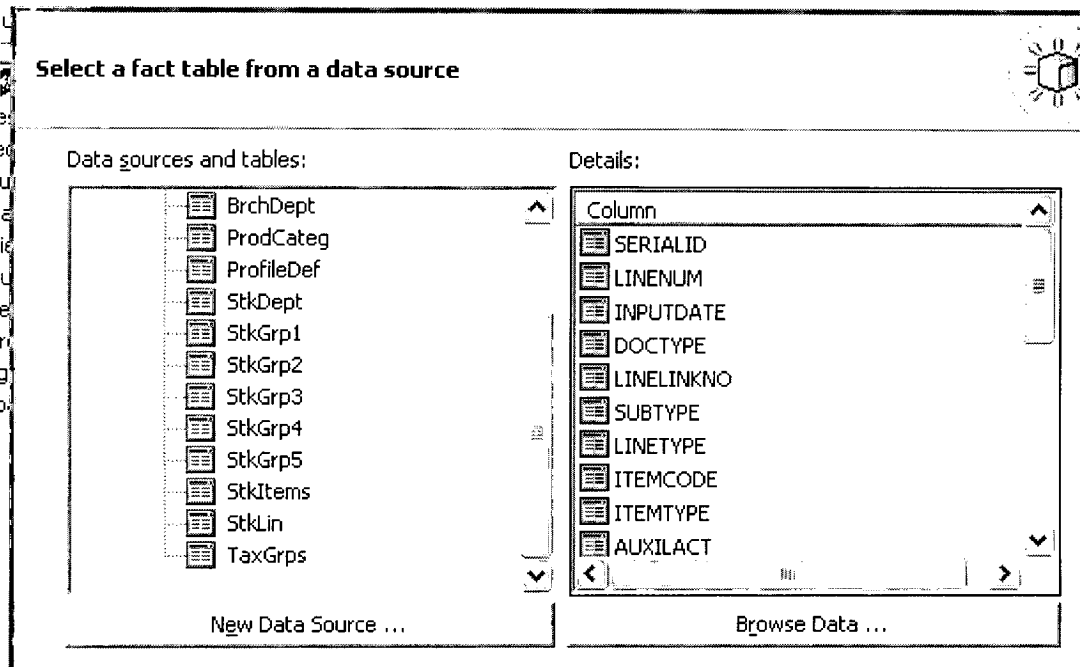


Fig 23: Selection of “FACT” table for cube “CustProfilingThesis”

- Once the fact table is chosen, the measurement parameters on which to build the results should be chosen. In this study the field *Q1Qty*

which holds the number of quantity sold for each item per customer was the one chosen as *measure parameter*

Cube measures:	
Measure name	Source column
#Q1qty	Q1QTY

Fig. 24: choosing the cube “CustProfilingThesis” measure parameter

- Once the cube measures are chosen, the user must choose the cube dimensions. By cube dimensions, we mean the tables needed to conduct the study. If we are to extract our customer profile then the fact table should be the table which contains the transactions, and dimension tables should be the table containing the customers, the table of products and the table which hold the customer profile parameters is shown in figure 25.

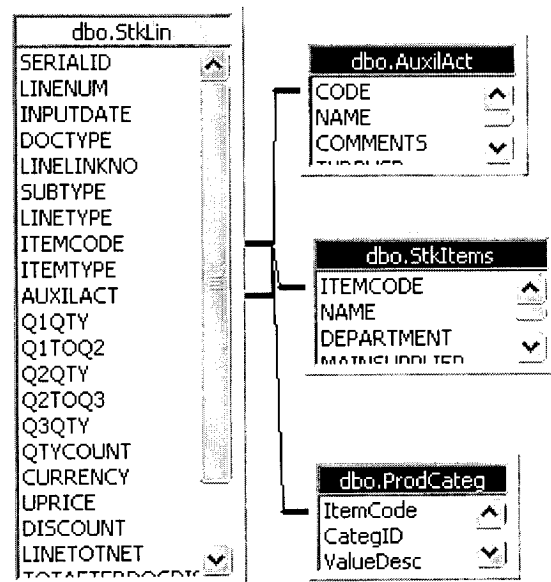


Figure 25: list of cube “CustProfilingThesis” used dimensions

The first result generated by the cube is the total value per parameter for each customer.

In fig. 26, the total quantity purchased corresponding to each value in each parameter is extracted. Client *CCB-BAIN* has purchased a total of 418 pcs. For the *Age* profile parameter, 273 out of 418 are related to *Adult* products, 32 out of 418 are related to *KID* products and so on.

Code	- Categ Id	Value Desc	MeasuresLevel Q1qty
CCB-BAIN	- AGE	AGE Total	418.00
		ADULT	273.00
		ALL	113.00
		KID	32.00
	- GENDER	GENDER Total	418.00
		ALL	285.00
		FEMALE	56.00
		MALE	77.00

Fig. 26: Cube “CustProfilingThesis”

#### 4.4.4.1 Extraction of the customers’ tentative profile

Once the results are extracted, calculated members should be added in order to extract the cube’s result. What is meant by calculated members?

- The measure parameter on which the cube bases its calculation has been set. In our example it is the quantity purchased by the customers.
- To achieve this, calculation members should be added to sub-class the measured parameter. In a calculated member, we can use different methods to build up a customer member (SQL command, source code...). The thesis’ interest is to build up a customer profile based on profile parameters where each profile parameter has also pre-defined members. The first calculation member will yield to the percentage of each profile-parameter member. The command used to build up the percentage value per profile-parameter member (see Appendix A-1)

In fig.27 for customer *CCB-BAIN* the total of quantity purchased is 418, for parameter *AGE*, 273 out of 418 corresponds to the value *ADULT* in this parameter which corresponds to a percentage of 65.31.

Code	- Categ Id	Value Desc	MeasuresLevel	
			Q1qty	CategPct
CCB-BAIN	- AGE	AGE Total	418.00	100
		ADULT	273.00	65.31
		ALL	113.00	27.03
		KID	32.00	7.66
	- GENDER	GENDER Total	418.00	100
		ALL	285.00	68.18
		FEMALE	56.00	13.40
		MALE	77.00	18.42

Fig. 27: Result in percentage for each parameter

- Once percentages of all parameter values are extracted, another calculated member is added to extract the max value of each parameter. This new calculated member will be called *TempoResult* (Appendix A-2)

In fig. 28 after applying this new calculated member, client *CCB-BAIN* will have the value ADULT for parameter AGE, the value ALL for parameter Gender...

Code	- Categ Id	Value Desc	MeasuresLevel		TempoResult
			Q1qty	CategPct	
CCB-BAIN	- AGE	AGE Total	418.00	100	ADULT
		ADULT	273.00	65.31	
		ALL	113.00	27.03	
		KID	32.00	7.66	
	- GENDER	GENDER Total	418.00	100	ALL
		ALL	285.00	68.18	
		FEMALE	56.00	13.40	
		MALE	77.00	18.42	
	+ HOBBIE	HOBBIE Total	418.00	100	SWIMMING
	+ INCOME	INCOME Total	418.00	100	LOW
+ MARITALSTATUS	MARITALSTATUS Total	418.00	100	SINGLE	

Fig. 28: demographic profile tentative results

#### 4.4.4.2 Extraction of the final customer profile:

The next step in this thesis is the extraction of the final customer profile.

As previously stated, each of the profile parameters corresponds to a parameter type (type A, B or C); check fig. 19.

- Parameter of type 'A' which is the 'INCOME' has no factor value. This mean the member with the highest percentage in this parameter will be considered as the result of this parameter.



- Parameters of type 'B' which are 'AGE', 'GENDER' and 'MARITALSTATUS' have a dominant factor. This means even if one of the parameters member has the highest percentage, this member cannot be considered as the result of this parameter unless its percentage value is equal or above its parameter factor value. If none of the members percentage did exceed this factor, the member 'ALL' will be considered as a value for this parameter
- Parameters of type 'C' which is the income, have what is called an acceptable factor. Each member having a percentage equal or greater than this factor will be considered as valid value for these parameters. In conclusion such parameters may have many values.

In order to extract the customer final profile, additional conditions are to be added. How can this be to achieved?

Calculated members holding the parameters type and condition are created, these calculated members are extracted from table *ProfileDef*.

Once these values are added to the cube, a new calculated member which bases its calculation on the *TempoResult* calculated member and by applying the conditions of parameters types will result in the final customer profile. (Appendix A-3)

In the example in fig. 28 the *Age*, *Gender*, *Income*, *MaritalStatus* will keep the same value assigned in the tentative profile, but the *HOBBY* parameter's results will be "*Swimming + Football*".

#### 4.4.4 Extracting the Customer loyalty ranking

Once the customer profile has been extracted using *Analysis Services*, *Decision Cube* the user can start the work on the loyalty ranking per customer.

A table containing the parameters weights has been created fig. 29

The extraction of customer Loyalty ranking is done as follows:

- The value of each parameter is extracted for each client, example client *CCB-BIN* purchased amount was 9000USD
- Once the value of each parameter is extracted, the scale value will be extracted depending on the range number where the parameter value

all into. example client *CCB-BIN* scale value of the purchased amount was 1

- Scale value is multiplied by the parameter weighted value defined in fig. 29
- The sum of the scale values multiplied by their weighted values generates the customer loyalty score.
- According to the resulting score, the customer loyalty ranking is defined and updated in the client record.

The extraction of loyalty ranking has been done using “SQL server stored procedure” (see Appendix B-1)

	Weighted value
Total amount purchased	8
Frequency: number of purchases	11
Number of products purchased	5
Retention time	7
Recency time	9
total	40

fig.29: Weighted values for each loyalty parameter

#### 4.4.5 Extracting the Customer LTV ranking

Once customer Loyalty rankings extracted, comes next the calculation of the LTV ranking per customer.

The extraction of customer Loyalty ranking is done as follows:

- The value of each parameter is extracted for each client, example client *CCB-BIN* profit amount was 5%
- Once the value of each parameter is extracted, the scale value will be extracted depending on the range number where the parameter value all into. example client *CCB-BIN* scale value of the profit percentage was 1
- The scale values of the three parameters are multiplied to generate the customer LTV score.
- According to the resulting score, the customer LTV ranking is defined and updated in the client record.

This was done also using “SQL server” stored procedures (see Appendix B-2)

#### 4.4.6 Building the resulting matrix:

The next step is the creation of a matrix containing the distribution of the customers according to the LTV and loyalty values.

To achieve this, a new decision cube is created, LoyaltyLtvCube which is shown in fig. 30.



Fig. 30: creation of new cube LoyaltyLtvCube

The dimensions of this decision cube are the LtvRanking, LoyaltyRanking as represented in fig. 31.

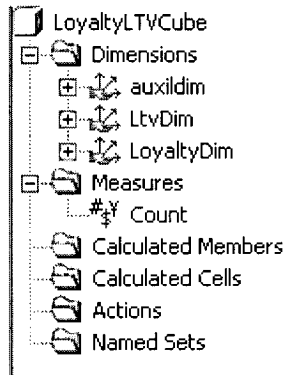


Fig. 31: cube LoyaltyLtvCube and its different dimensions

For each combination of LTV-Loyalty ranking, the number of clients corresponding to each combination is extracted, representation is clear in fig.32.

auxildim		All auxildim
Ltvranking	Loyaltyranking	MeasuresLevel Count
2	1	101.00
	2	35.00
	3	160.00
	4	
	5	

Fig. 32: number of customer per Loyalty ranking/LTV ranking

Expanding the cube displays the clients corresponding to this combination

fig.33.

Ltvranking	Loyaltyranking	Code
1	1	CCB-AVAL
		CCB-BAIM
		CCB-BAIN
		CCB-BEAS
		CCB-BEBA
		CCB-BLUE
		CCB-CAST
		CCB-CHAB
		CCB-CHAM
		CCB-CHVI
		CCB-CLAB
		CCB-CLUB
		CCB-COLL
		CCB-COMI
		CCB-COUN
		CCB-CSGH

Fig. 33: list of clients per each LTV-Loyalty combination

#### 4.5 Data Mining

As stated before, the resulting matrix constitutes a consistent data for data mining to help reveal some hidden and useful patterns. In order to prove this, data mining is performed on the decision cube created. The Data mining used is based on Analysis services. A new mining model is created using the wizard.

The selected source type is OLAP data since the data mining will be based on the cube, representation is in fig. 34.

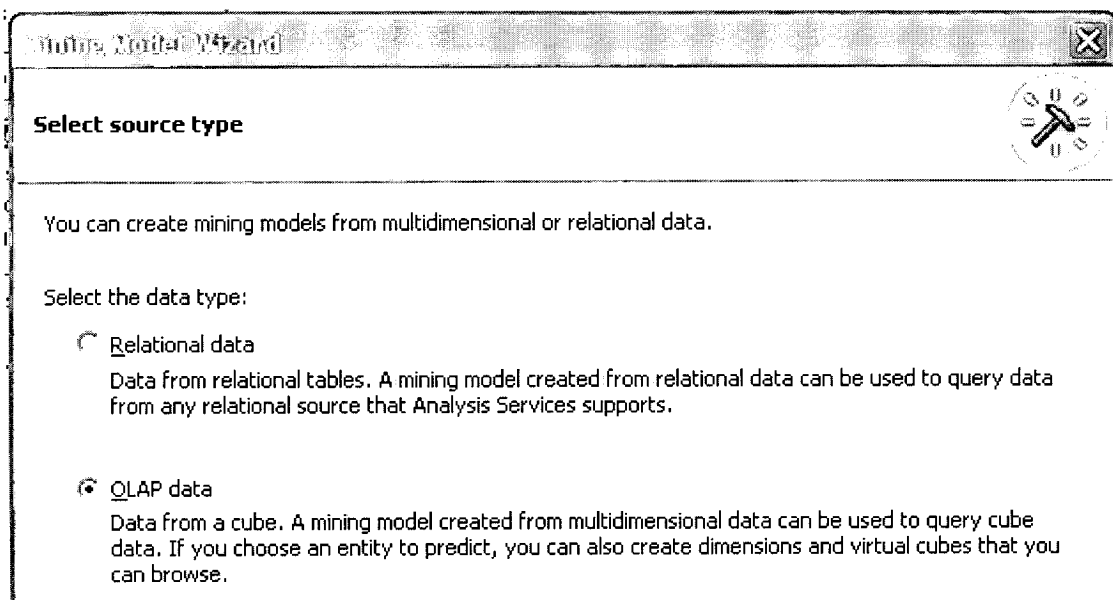


Fig. 34: Creation of a mining model using a wizard

SQL server 2000 provides only two data mining techniques: Clustering and decision trees. Clustering technique was the one selected since the objective is to get a general idea on how the customers' demographic parameters react to each other, example: males have a high percentage for the hobbies football and Basket ball, and the Hobby Ski goes only with average and high income.

The next important step is to specify the mining model calculation dimensions

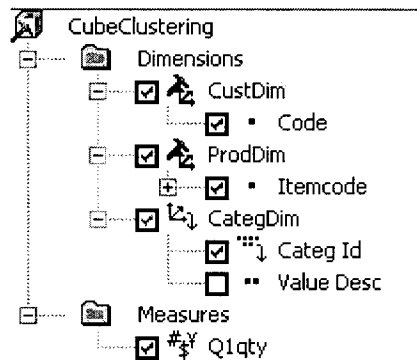


Fig. 35: Dimensions for the created data mining model

Since the result to get is the most common profiles which occur together, the calculations were based on CustDim, ProdDim and CategDim choosing the Categ ID. The number of clusters set to be generated was 10 as it is appearing in fig. 36.

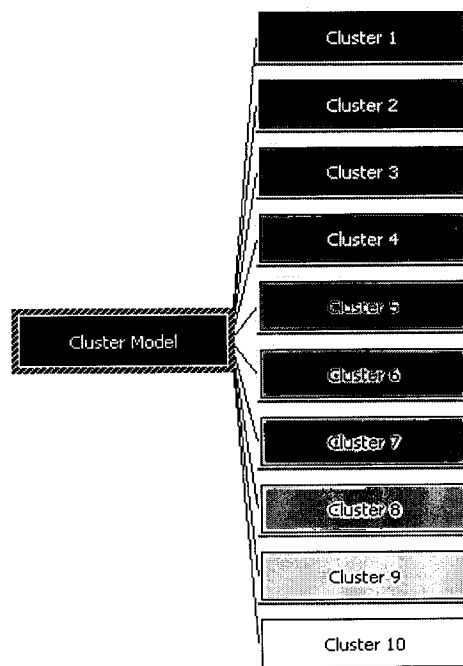


Fig.36: result of the data mining model processing

Each cluster contains different percentage figures for the parameter values. These clusters allow determining which parameter values are to be present together in a demographic profile.

On each cluster, the data mining will give results related to this cluster data. This way, we can determine the most common demographic profile for each cluster. These results are clearly shown in fig. 37.

Attributes		
Totals   Histogram		
Value	Cases	Probability ▲
KID	0	0.00%
LOW	0	0.00%
MALE	0	0.00%
MARRIED	0	0.00%
MARTIAL-AR	124	0.37%
MOTOR-SPOR	42	0.13%
NOTDEFINED	1612	4.77%
RUNNING	219	0.65%
SCOOTER	34	0.10%
SINGLE	0	0.00%

Fig. 37: Cluster probability percentages of combinations

## CHAPTER 5: CONCLUSION

### 5.1 Advantages and Disadvantages of the proposed algorithm

Like all proposed algorithms, this algorithm has a lot of advantages but it also has weaknesses and disadvantages.

#### Advantages

- This algorithm provides the user with a general profile about his customer; in addition to the demographic profile, it helps in the analysis of the customer from the loyalty and lifetime points of view
- The preparation phase of this algorithm is mainly user definable. Users suggest the parameters they need and their values. This makes the suggested algorithm somehow generic; it can be applied to all lines of business.
- Since this algorithm needs user intervention in its preparatory stages to define the needed profiles parameters, the results generated are straight forward and perfectly related to the user needs.
- As stated before, the result of this algorithm is a cube containing all info about the customer with no redundant and meaningless data. Thus, all the results from mining this cube are meaningful and relevant to the subject with no trivial and spurious results.
- The matrix that constitutes the results of this algorithm can be helpful in marketing, management decisions, financial decisions, etc., because it provides a summary for the customer's distribution according to profitability and demographic parameters.
- This algorithm presents a technique to extract consumer's profiles out of their purchasing transactions without going into the hassle of direct contact with the customer, in addition to being expense wise very effective, it provides a continuous up to date profile for all the customers.

#### Disadvantages

- Time Consuming: This algorithm needs time especially in its preparation phases. Categorizing the items according to the

demographic parameters needs manual work, which naturally is time consuming.

- Accuracy: The demographic profile extracted shows some weaknesses in accuracy especially in the extraction of the customer demographic profile.

## **5.2 Possibility of extensions and future work**

- Enhancement can be done on the demographic profile extraction part, and on the definition of the range of values of each profile template. While many types of parameters can be defined, only three have been used.
- Work on data mining can be extended, one example was provided by using the clustering technique to mine the resulting cube and to categorize the customers' behavior according to their demographic parameters. Using other techniques can be helpful also to reveal other hidden patterns which can be useful in decision making, future projects and improvement of the contact with clients.

## **5.3 Conclusion of the main contributions in this thesis**

A lot of work has been done on the customer profiling topic, and several algorithms and techniques have been suggested to extract customers' profiles. What makes this algorithm different from the others is that with only historical transactions we ended up extracting a complete demographic and analytical profile of the customers. The results of this algorithm are that apart from being useful to understand the distribution of the customers according to profitability parameters, they can provide important basis for a company's future improvement projects.



## References

- [1] Gediminas Adomavicius and Alexander Tuzhilin, "Integrating User Behavior and Collaborative Methods in Recommender Systems", May 1999, <http://www.patrickbaudisch.com>, 2005
- [2] Gediminas Adomavicius and Alexander Tuzhilin, "Using data mining methods to build customer profiles", IEE computer, February 2001, <http://www.cse.buffalo.edu>, 2005
- [3] Catherine Bounsaythip, Esa Rinta-Runsala, "Overview of data mining for customer Behavior Modeling", Finland, June 2001, <http://virtual.vtt.fi/inf/julkaisut/muut/2001/customerprofiling.pdf>, 2005
- [4] C.Hall, "data mining Tools, Techniques, and Services", April 1999, <http://www.cutter.com/itgroup/reports/datatool.html>, 2005
- [5] Colleen Cunningham, II-Yeol Song, Peter P.Chen, "Data Warehouse Design to Support Customer Relationship Management Analyses", 2004 <http://www.ececs.uc.edu/~dolap04/DOLAPdocs/papers/cunningham.pdf>, 2005
- [6] Chin-Chun Lin, "A study of knowledge Management Based Customer Relationship Management-The case of Software Service Provider", 2002, [http://thesis.lib.cycu.edu.tw/ETD-db/ETD-search/view\\_etd?URN=etd-0827103-141927](http://thesis.lib.cycu.edu.tw/ETD-db/ETD-search/view_etd?URN=etd-0827103-141927), 2005
- [7] Christopher S. Andrews, Dialogos, Inc, "A Methodology for Customer Segmentation Using Existing Product Category Schemes", Boston, 2000, <http://www2.sas.com/proceedings/sugi25/25/st/25p257.pdf>, 2006
- [8] Derek J. Paulsen, "Geographic Profiling Hype or Hope? Preliminary Results into the Accuracy of Geographic Profiling Software", [http://www.ucl.ac.uk/jdi/downloads/pdf/second\\_mapping\\_conference\\_papers/D\\_Paulsen.pdf](http://www.ucl.ac.uk/jdi/downloads/pdf/second_mapping_conference_papers/D_Paulsen.pdf), 2005
- [9] D.R. Mani, "Lifetime Value Modeling, the most valuable metric", 2006, <http://72.14.203.104/search?q=cache:PZQWtuSyX8UJ:road.uww.edu/road/peltierj/Databse%2520Mkt%2520Fall%25202004/LTV/Lifetime-Value%2520Modeling%2520fall%25202004.ppt+++%22Lifetime+Value+Modeling,+the+most+valuable+metric%22&hl=en&ct=clnk&cd=3>, 2006
- [10] D.R. Mani , James Drew, Andrew Betz, Piew Data, "Statistics and Data Mining Techniques for Lifetime Value Modeling", 1999, <http://portal.acm.org/citation.cfm?doid=312129.312205>, 2006
- [11] Eric W. Ganther "Demand Response Reference Design Final Report", June 2004, [http://openami.org/twiki/pub/Main/RDPrinciples/CIEE\\_CEC\\_DR\\_REFERENCE\\_DESIGN\\_FINAL\\_REPORT\\_06\\_15\\_04.pdf](http://openami.org/twiki/pub/Main/RDPrinciples/CIEE_CEC_DR_REFERENCE_DESIGN_FINAL_REPORT_06_15_04.pdf), 2005

[12] France Lelec, "Quantifying Customers", 2002,  
[http://www.systemdynamics.org/conf2004/SDS\\_2004/PAPERS/403YEON.pdf](http://www.systemdynamics.org/conf2004/SDS_2004/PAPERS/403YEON.pdf), 2006

[13] Hui Liu, "Case management system", Eindhoven, November 2005,  
[http://www.gosps.com/downloads/Case\\_Mgmt\\_practice.pdf](http://www.gosps.com/downloads/Case_Mgmt_practice.pdf), 2006

[14] Hwang, Hyunseok, Jung, Taesoo and Suh, "An LTV Model and Customer Segmentation Based on Customer Value: a case study on the wireless telecommunication industry", 2004,  
[http://www.feb.ugent.be/Fac/Research/WP/Papers/wp\\_04\\_282.pdf](http://www.feb.ugent.be/Fac/Research/WP/Papers/wp_04_282.pdf), 2006

[15] Injaz J.Chen, Karen Popovich, Cleveland Ohio, USA, "Understanding Customer Relationship Management (CRM) people, process and Technology", 2003,  
<http://stafweb.uum.edu.my/hartini/crm.pdf>, 2005

[16] Jeffrey Pease, "Customer Value Management, New Techniques for Maximizing the Lifetime Profitability of Your Customer Base", 2003,  
[http://www.csn.no/whitepapers/cvm\\_wp.pdf](http://www.csn.no/whitepapers/cvm_wp.pdf), 2006

[17] Jhon W.White, "Making ERP Work the way your business works, a new approach improves ERP simplicity, Productivity and ROI", 2004,  
<http://www.cio.com/sponsors/fuego.pdf>, 2006

[18] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl, "Explaining Collaborative Filtering Recommendations", USA, 2000,  
<http://www.grouplens.org/papers/pdf/explain-CSCW.pdf>, 2005

[19] Karla Loria, Thomas Kuaku Obeng, "Customer Relationship Management Implementation, A case study of two service companies", January 2005,  
[http://www.ies.luth.se/home/users/timfos/Thesis2/Seminars\(Jan05\)/CRM%20Implementation.pdf](http://www.ies.luth.se/home/users/timfos/Thesis2/Seminars(Jan05)/CRM%20Implementation.pdf), 2006

[20] K.Thearling, "data mining and customer relationship", 2000,  
[http://shamsgroup.com/pdfs/muse/archive/CRM\\_Reg02.pdf](http://shamsgroup.com/pdfs/muse/archive/CRM_Reg02.pdf), 2005

[21] Matt Kures, Bill Ryan, Greg Lamb "Customer Profiling and Prospecting Analysis: for the Door Country Lodging Industry", May 2001,  
<http://www.uwex.edu/ces/cced/Doorsum.pdf>, 2005

[22] Michael J.Pazzani, "A Framework for Collaborative, Content Based and Demographic Filtering", California, 1999,  
<http://www.ics.uci.edu/~pazzani/Publications/AIREVIEW.pdf>, 2006

[23] M.Van Satten, "User interfaces for personalized information systems", January 2003, <https://doc.freeband.nl/dscgi/ds.py/Get/File-28132/uipresinfosys.pdf>, 2005

[24] Ross B. Garber, "Strategies for E-Business Applications", January 2000, <http://www.acs.org.au/Certification/Documents/EBus/2001EB1-DotComCrash.pdf>, 2005

[25] Rosset, Saharon, Neumann, Einat, Eick, Uri, Vatnik, Nurit, "Customer Lifetime Value Models for Decision Support", 2003, <http://www-stat.stanford.edu/~saharon/papers/ltv-journal.pdf>, 2006

[26] Seth Paul, Nitin Gautam, Raymond Balint, "Preparing and mining data with Microsoft SQL server 2000 and Analysis services", May 2004, [www.sqldatamining.com](http://www.sqldatamining.com), 2005

[27] Valoris A. Hawkes, "Measuring Customer Loyalty", London, 2000, [http://www.fessel.at/de/download/PRSNT/2\\_Measuring\\_Customer\\_satisfaction\\_and\\_Loyalty\\_in\\_CEE\\_Zeh.pdf](http://www.fessel.at/de/download/PRSNT/2_Measuring_Customer_satisfaction_and_Loyalty_in_CEE_Zeh.pdf), 2006

[28] WWW.Profiles-Software.com

## Appendix A

In this Appendix is listed the code of the measures used in decision cube (Fig 21)

### A-1 Measure CategPct

Extract the percentage for each value of each profile parameter, SUM is used to extract the total quantity used for each value

```
IIF (ISLEAF ([CategDim].CURRENTMEMBER), [Measures].[Q1Qty] /  
SUM([CategDim].CURRENTMEMBER.SIBLINGS, [Measures].[Q1Qty]) * 100,  
100)
```

### A-2 Measure TempoResult

Extract the maximum percentage of each parameter values, it will result in having for each parameter the value holding the maximum percentage

```
IIF (ISLEAF ([CategDim].CURRENTMEMBER), IIF ([Measures].[CategPct] >  
0, IIF ([Measures].[CategPct] = MAX([CategDim].  
CURRENTMEMBER.SIBLINGS, [Measures].[CategPct])), [CategDim].  
CURRENTMEMBER.DATAMEMBER.NAME, ""), "", "")
```

### A-3 Measure FinalResult

Extract the final values for each parameter after applying conditions on these value depending on each parameter type

```
IIF (ISLEAF ([CategDim].CURRENTMEMBER), IIF  
([CategDim].CURRENTMEMBER.PARENT.Type = "B", IIF  
([Measures].CategPct > [CategDim].CURRENTMEMBER.PARENT.Factor,  
MAX({[CategDim].CURRENTMEMBER.SIBLINGS, "ALL"}), IIF ([CategDim].  
CURRENTMEMBER.PARENT.Type = "C",  
{[Measures].CategPct > [CategDim].CURRENTMEMBER.PARENT.Factor, ([CategD  
im], CURRENTMEMBER.SIBLINGS, ""))), IIF  
([CategDim].CURRENTMEMBER.PARENT.Type = "A", Max([CategDim].  
CURRENTMEMBER.SIBLINGS, " "))))
```

## Appendix B

In this appendix are listed the stored procedures used to update the clients Loyalty and LTV ranking

### **B-1 Updating Customer loyalty ranking:**

This procedure will update the field *LoyaltyRanking* in *AuxilAct* table

```
CREATE PROCEDURE UpdateAuxilLoyalty AS

DECLARE @code char(16), @LoyaltyRanking char(10),
        @TotPrch money, @TotPrchRank integer, @TotPrchRankedWeight integer,
        @Frequency money, @FreqRank integer, @FreqRankWeight integer,
        @ItemProdNbr money, @ItemProdNbrRank integer, @ItemProdNbrRankWeight integer,
        @RetentionDate datetime, @RecencyDate datetime, @RecencyRankWeight integer,
        @RetentionDateRank integer, @RecencyDateRank integer, @RetentionRankWeight integer,
        @LoyaltyWeightTotVal integer, @CustTotWeightVal integer, @LoyaltyRank integer,
        @DateDiff money

-- first open a cursor of current clients, so select all clients is needed

DECLARE auxiliary_cursor CURSOR FOR
SELECT code FROM AuxilAct
ORDER BY code

-- using another procedure, get the Total of all loyalty factors weight, it will be used to determine client
ranking
SET @LoyaltyWeightTotVal = ([dbo].GetLoyaltyTotWeight())

-- scan the client cursor now and update each client loyalty ranking
OPEN auxiliary_cursor
-- get record
FETCH NEXT FROM auxiliary_cursor
INTO @code
-- store current code into a paramter which will be used for Update later

WHILE @@FETCH_STATUS = 0
BEGIN
-- get this client parmaters to evaluate Loyalty Ranking values
SELECT @TotPrch = SUM(LineTotnet), -- total of purchases values
        @Frequency = COUNT(DISTINCT(serialid)), -- number of purchases
        @ItemProdNbr = COUNT(DISTINCT(itemcode)),-- number of products purchased
        @RetentionDate = MIN(inputdate),-- since when first transaction time
        @Recencydate = MAX(inputdate) -- last transaction time
FROM stklin WHERE auxilact = @code

-- get this client rank value according to his purchases value
SET @TotPrchRank = ([dbo].GetParamRankAccordingToVal ('PCHAMNT', @TotPrch))
-- now multiply this purchase amount rank * weight of the purchase amount paramter
SET @TotPrchRankedWeight = @TotPrchRank * ([dbo].GetLoyaltyWeightPerType('PCHAMNT'))

-- get difference in days between current date and time of first transaction, will be considered as
retention
```

```

SET @DateDiff = DATEDIFF(month, getDate(), @RetentionDate)
-- get this client rank value according to his retention value
SET @RetentionDateRank = ([dbo].GetParamRankAccordingToVal ('RETENTION', @DateDiff))
-- now multiply this retention rank * weight of the retention parameter
SET @RetentionRankWeight = @RetentionDateRank *
([dbo].GetLoyaltyWeightPerType('RETENTION'))

-- get difference in days between current date and time of last transaction, will be considered as recency
SET @DateDiff = DATEDIFF(MONTH, GETDATE(), @RecencyDate)
-- get this client rank value according to his recency value
SET @RecencyDateRank = ([dbo].GetParamRankAccordingToVal ('RECENCY', @DateDiff))
-- now multiply this retention rank * weight of the recency parameter
SET @RecencyRankWeight = @RecencyDateRank *
([dbo].GetLoyaltyWeightPerType('RECENCY'))

-- get this client rank value according to his frequency value
SET @FreqRank = ([dbo].GetParamRankAccordingToVal ('PCHNBR', @Frequency))
-- now multiply this frequency rank * weight of the parameter
SET @FreqRankWeight = @FreqRank * ([dbo].GetLoyaltyWeightPerType('PCHNBR'))

-- get this client rank value according to his number of products purchased value
SET @ItemProdNbrRank = ([dbo].GetParamRankAccordingToVal ('PRODNBR', @ItemProdNbr))
-- now multiply this number of products purchased rank * weight of the parameter
SET @ItemProdNbrRankWeight = @ItemProdNbrRank *
([dbo].GetLoyaltyWeightPerType('PRODNBR'))

-- once all values have been fetched, calculate this client final scale value
SET @CustTotWeightVal = (@ItemProdNbrRankWeight + @FreqRankWeight +
@RecencyRankWeight + @RetentionRankWeight + @TotPrchRankedWeight)

-- once scale value is calculated, get this client rank in Loyalty parameter
SET @LoyaltyRank = ([dbo].GetClientLoyaltyRank(@CustTotWeightVal))

-- set the calculated loyalty rank value in client table
UPDATE AuxilAct SET LOYALTYRANKING = @LoyaltyRank WHERE code = @code

-- continue doing the same for other clients
  FETCH NEXT FROM auxiliary_cursor
  INTO @code
END

CLOSE auxiliary_cursor
DEALLOCATE auxiliary_cursor
GO

```

## **B-2 Updating Customer LTV ranking:**

This procedure will update the field *LTVRanking* in *AuxilAct* table  
CREATE PROCEDURE UpdateAuxilLTV AS

```

DECLARE @code char(16),
        @Profit money, @ProfitRank integer,
        @Frequency money, @FreqRank integer,
        @RetentionDate datetime, @RetentionDateRank integer,
        @CustTotWeightVal money,

```

```

        @LTVRank integer,
        @DateDiff money

-- first open a cursor of current clients, so select all clients is needed

DECLARE auxiliary_cursor CURSOR FOR
SELECT code FROM AuxilAct
ORDER BY code

OPEN auxiliary_cursor

-- scan the client cursor now and update each client LTV ranking
FETCH NEXT FROM auxiliary_cursor
INTO @code

WHILE @@FETCH_STATUS = 0
BEGIN

-- get this client parmeters to evaluate Loyalty Ranking values

-- get its profit first
SELECT @Profit = ( (SUM(LineTotnet) - SUM(ABS(Q1Qty)*AvgValCur)) /
SUM(ABS(Q1Qty)*AvgValCur)) * 100, -- profit done with this client
        @Frequency = COUNT(DISTINCT(serialid)), -- number of purchases
        @RetentionDate = MIN(inputdate) -- since when first transaction time
FROM stklin WHERE auxilact = @code

-- get this client rank value according to his profit value
SET @ProfitRank = ([dbo].GetParamRankAccordingToVal ('PROFIT', @Profit))

-- get difference in days between current date and time of first transaction, will be considered as
retention
SET @DateDiff = DATEDIFF(month, @RetentionDate, getDate())
-- get this client rank value according to his retention value
SET @RetentionDateRank = ([dbo].GetParamRankAccordingToVal ('RETENTION', @DateDiff))

-- get this client rank value according to his frequency value
SET @FreqRank = ([dbo].GetParamRankAccordingToVal ('PCHNBR', @Frequency))

-- get this client LTV weight value
SET @CustTotWeightVal = @FreqRank * @RetentionDateRank * @ProfitRank

-- once scale value is calculated, get this client rank in LTV parameter
SET @LTVRank = ([dbo].GetClientLTVRank (@CustTotWeightVal))

-- set the calculated LTV rank value in client table
UPDATE AuxilAct SET LTVRANKING = @LTVRank WHERE code = @code

FETCH NEXT FROM auxiliary_cursor
INTO @code
END

CLOSE auxiliary_cursor
DEALLOCATE auxiliary_cursor
GO

```

### **B-3 GetLoyaltyTotWeight procedure:**

This function will return the total weighted value of all loyalty parameters (fig 29)

```
CREATE FUNCTION GetLoyaltyTotWeight ()
RETURNS integer
AS
BEGIN
DECLARE @Value integer

SELECT @Value = SUM(WeightValue) FROM LoyaltyDef
RETURN(@Value)
END
```

### **B-4 GetLoyaltyWeightPerType procedure:**

This function will return the weighted value of a loyalty parameter (fig 29)

```
CREATE FUNCTION GetLoyaltyWeightPerType (@LoyaltyID nchar(20))
RETURNS integer
AS
BEGIN
DECLARE @Value integer

SELECT @Value = WeightValue FROM LoyaltyDef WHERE LoyaltyID = (@LoyaltyID)
RETURN(@Value)
END
```

### **B-5 GetRankAccordingToVal procedure:**

This function will return the rank value of a parameter (fig 11)

```
CREATE FUNCTION GetParamRankAccordingToVal (@ParamID nchar(20), @RankValue float)
RETURNS integer
AS
BEGIN
DECLARE @Value integer
SET @Value = 1
SELECT @Value = RankValue FROM ParamRangeDef WHERE (@ParamID = ParamID)
and ( (@RankValue >= RangeMin) and (@RankValue <= RangeMax))

RETURN(@Value)
END
```

### **B-6 GetClientLoyaltyRank procedure:**

This function will return the loyalty rank depending on client loyalty value (fig 13)

```
CREATE FUNCTION GetClientLoyaltyRank (@LoyaltyValue float)
Returns integer
AS
BEGIN
```



```

DECLARE @Value integer
SET @Value = 1
SELECT @Value = RankValue FROM LoyaltyRank WHERE (@LoyaltyValue >= RangeMin) and
(@LoyaltyValue <= RangeMax)

RETURN(@Value)
END

```

### **B-7 GetClientLTVRank procedure:**

This function will return the LTV rank depending on client LTV value (fig 13)

```

CREATE FUNCTION GetClientLTVRank (@LTVValue float)
RETURNS integer
AS
BEGIN
DECLARE @Value integer
SET @Value = 1
SELECT @Value = RankValue FROM LTVRank WHERE (@LTVValue >= RangeMin) and
(@LTVValue <= RangeMax)

RETURN(@Value)
END

```