A Method for Membership Card Generation Based on Clustering and Optimization Models in A Hypermarket

Chen Xiaojun, Premlal Bhatraoi
This thesis is submitted to the School of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author(s): Chen Xiaojun
Address: Kungsmarksvägen 101, Lgh-1066, 371 44 Karlskrona
E-mail: chenheejun@hotmail.com

Author(s): Premlal Bhattrai
Address: Lindblomsvägen 96, Lgh-1104, 372 33 Ronneby
E-mail: battraiPrem@gmail.com

University advisor(s):
Johan Holmgren (Postdoctoral)
School of Computing
Blekinge Institute of Technology (BTH)
ABSTRACT

Context: Data mining as a technique is used to find interesting and valuable knowledge from huge amount of stored data within databases or data warehouses. It encompasses classification, clustering, association rule learning, etc., whose goals are to improve commercial decisions and behaviors in organizations. Amongst these, hierarchical clustering method is commonly used in data selection preprocessing step for customer segmentation in business enterprises. However, this method could not treat with the overlapped or diverse clusters very well. Thus, we attempt to combine clustering and optimization into an integrated and sequential approach that can substantially be employed for segmenting customers and subsequent membership cards generation. Clustering methods is used to segment customers into groups while optimization aids in generating the required membership cards.

Objectives: Our master thesis project aims to develop a methodological approach for customer segmentation based on their characteristics in order to define membership cards based on mathematical optimization model in a hypermarket.

Methods: In this thesis, literature review of articles was conducted using five reputed databases: IEEE, Google Scholar, Science Direct, Springer and Engineering Village. This was done to have a background study and to gain knowledge about the current research in the field of clustering and optimization based method for membership card generating in a hypermarket. Further, we also employed video interviews as research methodologies and a proof-of-concept implementation for our solution. Interviews allowed us to collect raw data from the hypermarket while testing the data produces preliminary results. This was important because the data could be regarded as a guideline to evaluate the performance of customer segmentation and generating membership cards.

Results: We built clustering and optimization models as a two-step sequential method. In the first step, the clustering model was used to segment customers into different clusters. In the second step, our optimization model was utilized to produce different types of membership cards. Besides, we tested a dataset consisting of 100 customer records consequently obtaining five clusters and five types of membership cards respectively.

Conclusions: This research provides a basis for customer segmentation and generating membership cards in a hypermarket by way of data mining techniques and optimization. Thus, through our research, an integrated and sequential approach to clustering and optimization can suitably be used for customer segmentation and membership card generation respectively.

Keywords: Data mining, Hierarchical clustering, Fuzzy clustering, Optimization model, Membership card
ACKNOWLEDGEMENTS

First, I would like to express my earnest gratitude and thankfulness to my project supervisor Mr. Johan Holmgren for pondering his inestimable guidance, perseverant support and advises throughout the course of our thesis project. I am deeply indebted to my supervisor for his persistent support and advices during the course of our project, which helped me to make necessary improvements and thus helped completing our project. Next, I am deeply indebted to Mr. Yizhu Shi (the CEO of Wal-Mart in Shanghai) for sharing valuable information and domain knowledge.

Further, I am grateful my friends and colleagues at BTH far and away and who stand by me at times of need. My special thanks are also to Shanghai Second Polytechnic University for providing me a chance to be an exchange student for Master’s programme in BTH.

The last but not the least, I am very much appreciative of my family back home in China who are a constant source of inspiration during the term of my study here in Sweden. Their love, encouragement and enduring support is highly valued.

In a nutshell, I thank one and all who are directly or indirectly involved at different stages of my works and helped me accomplish this task.

-- Chen Xiaojun
To begin with, I would like express my profound tribute to my project supervisor Mr. Johan Holmgren for providing me the opportunity to work on this intriguing project. I am deeply indebted to my supervisor for his persistent support and advices during the course of our project, which helped me to make necessary improvements and thus helped completing our project.

Thanks are due to my friends and colleagues at BTH for their unconditional support and encouragement due to which I am able to accomplish up all the challenges to complete our project. Further, my special thanks and appreciation goes to the library staffs for helping us obtain relevant materials. I am also very much grateful to Mr. Yizhu Shi for his selfless services in sharing valuable information.

Last but not least, I would like to thank my parents, back home in Nepal for being a constant source of inspiration and guidance that kept me well focused and motivated.

Conclusively, my thesis would not have been completed without the grace of The Almighty God, who bestow upon me, the courage to face all the complexities and complete this project successfully.

--Premlal Bhattrai
# CONTENTS

**ABSTRACT** .......................................................................................................................... I

**ACKNOWLEDGEMENTS** ......................................................................................................... II

**INTRODUCTION** ....................................................................................................................... 1

**CHAPTER 1: BACKGROUND** .................................................................................................... 3

1.1 DATA MINING ....................................................................................................................... 3
1.2 CLUSTERING .......................................................................................................................... 4
    1.2.1 How Is Clustering Used? ................................................................................................. 5
1.3 OPTIMIZATION ...................................................................................................................... 6
    1.3.1 Linear Programming Model Usage .................................................................................. 6
1.4 RELATED WORK .................................................................................................................... 7

**CHAPTER 2: PROBLEM DEFINITION AND GOALS** ................................................................. 10

2.1 MOTIVATION ........................................................................................................................ 10
2.2 RESEARCH QUESTIONS ....................................................................................................... 11
2.3 GOALS / RESULTS ............................................................................................................... 12

**CHAPTER 3: METHODOLOGY** ............................................................................................... 13

3.1 RESEARCH PROBLEM FORMULATION .............................................................................. 13
3.2 LITERATURE REVIEW ....................................................................................................... 14
    3.2.1 Search Strategy for Literature Review ......................................................................... 14
3.3 INTERVIEWS FOR DATA COLLECTION AND APPROACH EVALUATION ....................... 15
3.4 DEVELOPMENT ................................................................................................................... 15
3.5 PROOF-OF-CONCEPT IMPLEMENTATION ......................................................................... 15

**CHAPTER 4: INTERVIEW AND DATA PREPARATION** ............................................................... 16

4.1 INTERVIEW DESIGN ............................................................................................................. 16
4.2 INTERVIEW ANALYSIS ......................................................................................................... 16
4.3 DATA PREPARATION ............................................................................................................ 18
    4.3.1 Data Collection .............................................................................................................. 18
    4.3.2 Data Processing ............................................................................................................. 19

**CHAPTER 5: A METHOD FOR CARD GENERATION** ................................................................. 20

5.1 CLUSTERING MODEL DESIGN ............................................................................................. 20
    5.1.1 Design Principle ............................................................................................................. 20
    5.1.2 Data Standardization ....................................................................................................... 21
    5.1.3 Building Fuzzy Similarity Matrix ..................................................................................... 22
    5.1.4 Clustering ....................................................................................................................... 22
5.2 OPTIMIZATION MODEL DESIGN .......................................................................................... 23
    5.2.1 Design Principle ............................................................................................................. 23
    5.2.2 Simplify Problem ............................................................................................................. 24
    5.2.3 Modeling ......................................................................................................................... 25
5.3 PROOF-OF-CONCEPT IMPLEMENTATION DESIGN ............................................................ 26
    5.3.1 Design Principle ............................................................................................................. 26
    5.3.2 Implementation Flow ...................................................................................................... 26

**CHAPTER 6: PROOF-OF-CONCEPT IMPLEMENTATION** ......................................................... 29

6.1 PROGRAM FOR CLUSTERING MODEL ................................................................................. 29
6.2 VARIABLE DEFINITION ....................................................................................................... 29
LIST OF FIGURES

Figure 1: Data mining process .................................................. 3
Figure 2: An example of clustering ............................................ 5
Figure 3: Research process flow diagram ................................ 13
Figure 4: Work flow in general .................................................. 20
Figure 5: Design flow in clustering model ............................... 21
Figure 6: Design flow in optimization model ............................ 24
Figure 7: Implementation flow for clustering model .................. 27
Figure 8: Implementation flow for optimization model ............... 28
Figure 9: Result of Formula 2 (part) ......................................... 31
Figure 10: Result of Formula 3 (part) .......................................... 32
Figure 11: Result of Formula 4 (part) ......................................... 33
Figure 12: Result of objective function ................................... 37
Figure 13: Assessment flow of optimization model .................... 39

LIST OF TABLES

Table 1: The characteristics of customers .................................. 18
Table 2: Membership application form ...................................... 19
Table 3: Membership card information ..................................... 19
Table 4: Customers’ consumption records ................................ 19
Table 5: Customers’ consumption records (addendum) ............... 19
Table 6: Customer information (after data extraction and cleaning) 19
Table 7: Customers’ consumption records (after data extraction and cleaning) .......................................................... 19
Table 8: Example set for testing .............................................. 30
Table 9: The result of clustering model ..................................... 35
Table 10: The parameters of optimization model ....................... 35
Table 11: Preliminary result of membership card generation ....... 37
Table 12: Comparison between Hierarchical clustering and Fuzzy clustering ................................................................. 38

LIST OF FORMULAE

Formula 1 (Matrix 1): Initial data matrix .................................. 21
Formula 2: Standard deviation transformation ........................ 21
Formula 3: Range transformation .......................................... 22
Formula 4: Correlation coefficient method ............................... 22
Formula 5: Transitive closure method ................................... 22
INTRODUCTION

The tremendous amount of information embedded in huge databases belonging to enterprises has spurred a great interest in the areas of data mining [1]. With the recent development of data mining technology, many service industries such as hypermarkets, insurances and hotels accumulate giant amounts of data [2]. This has caused a problem of how to effectively utilize and manage such data [3]. From enterprises’ perspective, mining useful data by an efficient data mining method is an economical way of increasing profits [3] [4].

Data mining includes classification, clustering and association rule learning, etc., whose final objectives are to improve commercial decisions and behaviors in different business organizations [2]. Among various types of data mining, higher management in a hypermarket often choose clustering to reduce search space to a set of most important attributes for customer segmentation [5]. In doing so, the management can identify customers’ characteristics and behaviors that can be exploited in order to avoid customer attrition phenomenon (i.e., loss of customers) [5] [6].

Furthermore, optimization models could potentially be used to analyze a given problem scenario (i.e., economic benefit) and to solve decision problems (e.g., generating membership cards) for a single goal—to maximize profit in a hypermarket [4] [7]. Additionally, it may include both single and multiple objective functions, and analysis of the problem may require individual considerations of the separate objectives [8].

However, in a hypermarket, optimization model usually contains the control of profit and cost but combines little function related to data mining [8] [10]. This is because it would be difficult for the management to consider plenty of customer information thoroughly while generating membership cards thereby limiting the application of data mining and optimization model [9] [10]. Therefore, we believe that the higher management has lost sight of impact of data mining during optimization modeling. Hence incorporating data mining technique with optimization model would be suitable to solve such kind of commercial problem.

Our thesis is aimed at developing a clustering and optimization based method for generating membership cards in a hypermarket by using a two-step sequential approach: first, we build a clustering model for customer segmentation, and then we create an optimization model for card generation.

Chapter 1 illustrates the background concerning data mining, clustering, optimization and how these can be used in the real market. First, we describe the background knowledge about data mining. Second, we illustrate the concept of clustering, some important types of clustering methods and how clustering can be used in customer analysis. Besides, we also discuss about optimization: its application domain and the objective of employing optimization model in market. Finally, the chapter ends with the discussion of related work concerning the above techniques.
Chapter 2 deals with the problem definitions and project goals. We describe the current scenario of data mining technology, clustering method and optimization model in our field of study (i.e., in a hypermarket). Further, we display our research questions and objectives.

Chapter 3 introduces research methodology. We start with research design as guidance to get a firsthand knowledge about which methodology is appropriate to our research field. Additionally, we collect data models and fulfill evaluation during video interviews with Wal-Mart (in Shanghai). Finally, in order to test and validate the clustering and optimization models, we make a proof-of-concept implementation of our approach.

Chapter 4 describes interviews and data preparation in detail. It helps us to acquire the knowledge about the current situation of customer segmentation and characteristics, membership card information from the perspective of the Wal-Mart.

Chapter 5 explains about the method for membership card generation. In this chapter, we present the procedures of clustering and optimization modeling in detail. Clustering model is used for segmenting customers, and the result of which is applied as an input to optimization model for generating membership cards. In addition, we design a proof-of-concept implementation for testing and validating our method for customer card generation.

Chapter 6 focuses on proof-of-concept implementation, which shows the preliminary results of the two models by testing the example dataset. The descriptions of the proof-of-concept implementation, variables definition in programs and implementation environment are elaborated. Finally, the outcome from the clustering and optimization models as our preliminary results will be eventually sent to Wal-Mart for evaluation of our methodological approach.

Chapter 7 provides the validity assessment, which helps us to review and summarize on our work. In this chapter, we assess our two models individually, provides theoretical argument as to why our approach is good in solving the problem and present feedback from the Wal-Mart about our method via video interviews. This chapter concludes with the description validity threats and ways to mitigate these threats.

Finally, chapter 8 concludes and discusses our thesis work and implies for future works that need to be done because of the gaps arising in our work.
CHAPTER 1: BACKGROUND

This chapter describes the background related to data mining, clustering, optimization and shows how these could be used in the real market. It begins with the background knowledge about data mining, introduces the concept of clustering, some important types of clustering methods, and how clustering can be used in customer analysis. In addition, the chapter also briefly mentions knowledge about optimization model and its usage in market. It ends with the related works performed in each of the above field.

1.1 Data Mining

Briefly, data mining is a product from the crossover of multiple subjects combining database technique, artificial intelligence, machine learning and statistics, etc., [3]. There are various types of data mining approaches including classification, clustering and association rule learning. One of the concerned objectives is to improve commercial decisions and behaviors in different domains such as hypermarkets, insurances and hotels [2]. Additionally, data mining can better be defined as an information process technology, which extracts data concealed in a large quantity of incomplete, noisy, vague and random practical application data and refines into knowledge. Generally, a data mining process involves three steps: Selection & Transformation, Modeling and Validation & Model Use as shown in Figure 1 below [5] [11].

![Figure 1: Data mining process [5]](image-url)
From enterprises’ perspective, mining useful data by an efficient data mining method is an economical way of increasing profits [3] [4]. Data mining technology can help enterprises accumulate and analyze diverse ranges of information pertaining to customer, marketing, sales, services, and the internal enterprise management favorably and precisely [12]. This is to say that data mining technology can effectively aid analyzing customer behavior and market tendencies, duly providing products and services to the targeted customers and enhancing customers’ satisfaction [12].

1.2 Clustering

In the words of Steven Pinker [13], “An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe, it has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand”. Indeed, one of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification [6]. The idea of sorting similar things into categories evidently dates back to primitive times. For instance, the early man must have been able to realize that many individual objects shared certain properties such as being edible or non-edible [6].

Clustering is defined as numerical methods of classification and probably the preferred generic term for procedures that seek to uncover groups in data [6]. It is an unsupervised learning process [14]. In other words, the manner of grouping data that shares similar characteristics into a group is called clustering or unsupervised classification [15].

There are various methods for clustering such as hierarchical, partitioning, subspace and fuzzy, whose results are concerned with exploring data sets to assess whether or not they can be summarized meaningfully in terms of a relatively small number of groups or clusters of objects or individuals. Each object in a cluster resembles one another while differs in some respects from objects in other clusters. The following figure shows three different colored squares into three different clusters [6].
Furthermore, clustering aims at segmenting a given set of data or objects into subsets, groups, clusters or classes based on the properties of homogeneity and heterogeneity. In non-fuzzy clustering (such as hierarchical clustering), data objects are divided into crisp clusters, which means each example belongs to exactly one cluster. Whereas, fuzzy clustering is a multivariate statistical analysis method, it defines the similarity of examples with fuzzy boundaries by using mathematical approaches, thereby objectively segmenting these examples. In fuzzy clustering, the first step is to eliminate the data dimensions; then build fuzzy similarity matrix so that it establishes fuzzy clustering relations and obtains the transitive closure; finally, define the threshold value in order to segment the examples into different clusters (detailed in Chapter 5). Besides, the examples may belong to more than one cluster and each of the examples is associated with a similarity coefficient that signifies in which the examples belong to different clusters [48].

In real world application, there is sometimes no sharp boundary between clusters. In such scenario, fuzzy clustering can suitably be employed. It takes similarity coefficient ranging from zero to one instead of crisp assignment of examples to clusters as the case of non-fuzzy clustering [47] [48].

1.2.1 How Is Clustering Used
In most applications of clustering, segmentation of data is sought in which each individual or object belongs to a single cluster, and the complete set of clusters contains all individuals [6]. However, in certain circumstances, overlapping clusters may provide a more acceptable solution. This is to say that there is no need of grouping such overlapped data into any single cluster [6].

The general problem, addressed by cluster analysis, appears in many disciplines: Biology, Botany, Geography, Marketing, Image processing, Archaeology and so on [6]. The data is
first preprocessed and then the features, based on which clustering is done, are selected [14]. For instance, in marketing research, dividing customers into homogeneous groups is one of the strategies of marketing [4]. A market researcher may, for example, ask how to group consumers who seek similar benefits from a product so the researcher can communicate with them better [6]. Or a market analyst may be interested in grouping financial characteristics of companies so as to be able to relate them to their stock market performance [4] [6].

1.3 Optimization

The field of optimization belongs to the domain of applied mathematics and encompasses the use of mathematical models and methods to find the best alternative in decision-making situations. The main purpose is to get decision support by using a mathematical model for real world problems [7].

Optimization problems occur in all areas of sciences and engineering, arising whenever there is a need to minimize (or maximize) an objective function that depends on a set of variables while satisfying some constraints [16]. For instance, the management in a hypermarket wants to maximize its profits while at the same time trying to satisfy certain constraints such as throughput, selling price and production efficiency [7]. Additionally, applications of optimization include constrained and unconstrained optimization, combinatorial optimization, stochastic optimization, multi-objective optimization, etc., [17]. It also covers linear programming (LP) model, complexity theory, approximations and error analysis, and goal programming model to name a few [17]. Before designing an optimization model, it is important to consider the form and mathematical properties of the objective functions, constraints, and decision variables [8] [16]. For example, the objective function may be linear or nonlinear, convex or non-convex, etc.; the decision variables might be continuous or discrete; the feasible region might be convex or non-convex [8]. These differences each affect how the model can be solved, and thus optimization models are classified according to these differences [8].

1.3.1 Linear Programming Model Usage

In a Linear Programming (LP) model [7], each part is structured by the mathematical framework of LP problems. An LP problem in its general form can be written as:

\[
\begin{align*}
\text{min or max} (z) &= \sum_{j=1}^{n} c_j x_j \\
\text{subject to:} & \sum_{j=1}^{n} a_{ij} x_j \leq \text{or } \geq b_i, \quad i = 1, 2, \ldots, m \\
& x_j \geq 0, \quad j = 1, 2, \ldots, n
\end{align*}
\]

where \( c_j \) is the objective function coefficient for variable \( x_j \), \( a_{ij} \) is the constraint coefficient for variable \( x_j \) in constraint \( i \), and \( b_i \) is the right hand side coefficient in constraint \( i \).
For example, in a market, the higher management considers the costs and profits of $n$ products so as to maximize profits, subject to $m$ restrictions on the costs [8] [18]. The model then is to choose $x_1, x_2, ..., x_n$ so as to:

$$\text{Maximize } f(x_1, x_2, ..., x_n)$$

Subject to $g_m(x_1, x_2, ..., x_n) \leq b_m$

and

$$x_1 \geq 0, x_2 \geq 0, ..., x_n \geq 0.$$  

In this model, the objective function $f(x_1, x_2, ..., x_n)$ measures the profit, and the constraints of them are represented by the inequalities $g_i(x_1, x_2, ..., x_n) \leq b$ for $i = 1$ to $m$.

Furthermore, the validation of optimization model is usually implemented by synthetic or real data through experiments [19]. This phase includes the work to verify that the solution is correct based on the formulated optimization model and to validate that the model describes the problem accurately enough [7].

1.4 Related Work

Though a lot of research has been performed in the field of cluster analysis and clustering methods which have been used in customer segmentation, very few studies are focused towards membership card generation in a hypermarket. In [20], Liu and Luo employed the k-means and k-medoids methods for clustering analysis on customers in a departmental store. The k-means used the average value in the group, where clustering was done by choosing $k$ objects each representing the average of each group with the shortest distance. Then, computed the average of each group, and regrouped all objects by the new average value according to the distance. The process was repeated until the new groups remained unchanged. In k-medoids, the object at the center of the group represented each group. However, the clustering process in k-medoids was similar to k-means. Besides, in [21], Shin and Sohn used three clustering methods: K-means, Fuzzy K-means and Self-organizing Map to find graded stock market brokerage commission rates of two different transaction modes (representative assisted and online trading system). Among the three clustering methods and based on the empirical results, authors maintained fuzzy K-means cluster analysis as the most competent one. However, the data used for clustering contained relatively short history of customers’ transactions, meaning less amount of customers’ information, and primarily relied on the cumulative transaction. Other cues such as customers’ consumption, and customers’ income level were not considered for better segmentation. Moreover, Wang [22] proposed hybrid approach that incorporates kernel induced fuzzy clustering techniques called Robust Possibilitic C Means (RPCM) and Robust Fuzzy C Means (RFCM) to limit outliers$^1$ for effective customer segmentation. RPCM was used to detect outliers while RFCM was utilized.

$^1$ Outliers are clustering errors caused by some data which violate the clustering rule (i.e., data which lie beyond the specialized range of data value) [22].
to segment objects (both used kernelized distance method\(^2\)). Their results showed a decrease in outliers. Further, the paper does not present any scheme for membership card generation for customers. On the contrary, in our proposed method, fuzzy clustering adopts Standard Deviation Transformation and Range Transformation formulae to eliminate any possible outliers. Furthermore, we attempt to proceed one-step further from customer segmentation and that is to generate membership card by building an optimization model.

Additionally, Zhang et al. [23] used data mining classification algorithms C5.0 and CART to model customers’ membership card classification. Customers’ income level and number of children were used as the two main attributes that affect their choice of cards. However, our project is based in Shanghai, China (one child policy), so we do not consider customers’ number of children for generating membership cards. Instead, customers’ consumption as an important customer attribute is considered for membership generation in our project.

Further, traditional clustering such as hierarchical clustering method could well handle distinct groups, however, they could not treat with the overlapped or diverse clusters very well [22]. Thus, fuzzy clustering can substantially be employed for segmenting complex customer profiles because fuzzy clustering takes examples that may belong to more than one cluster and each of the examples is related with a similarity coefficient that indicates in which the examples belong to different clusters. Moreover, optimization model from the hypermarket’s point of view is about the control of profit and cost rather than taking into account the constraint and objective functions (based on mathematical approach). The management usually finds difficult in considering the large number of the customer information due to lack of suitable customer segmentation method. Therefore, we believe that integrating fuzzy clustering with optimization as a sequential approach to our problem domain would fulfill our objective.

Additionally, Chang et al. [24] employed cluster analysis to cluster loyal customers possessing similar personal backgrounds and purchasing behavior. While they used similarity analysis to measure the similarity between potential customers who have never before purchased products and the loyal customers using squared Euclidean distance, which is the most commonly-used method to calculate distance between two points. Whether or not a potential customer falls in the range of a group of loyal customers, the customer is determined by his or her purchasing behavior. Guha et al. [1] proposed Clustering Using Representatives (CURE), which employed a hierarchical method that took a middle point between the all point-extremes and the centroid. However, no papers provide any method for generating customer membership cards.

Because, to the best of our effort and knowledge, we do not find employing any integrated approaches to clustering method and optimization model in customer segmentation and membership card generation in a hypermarket. As such, we provide some examples to

\(^2\) Kernelized distance method is a method to calculate distance between each object and the cluster center in order to define which object belongs to which cluster [22].
describe the combined approaches in other areas of research. For instance, Lee et al. in [35] used a combination of constraint optimization formulation and hierarchical clustering technique in a two dimensional auto-regressive modeling technique for texture characterization problem. They used constraint optimization formulation to estimate the auto-regressive (AR) model coefficient and hierarchical clustering to obtain the final coefficient estimation. Meanwhile, a two-dimensional auto-regressive model was used for description of image field and texture characterization where a different set of two-dimensional auto-regressive model coefficients exhibited each individual texture. The autoregressive (AR) model assumed a local interaction between image pixels in which pixel intensity was a weighted sum of neighboring pixel intensities. Additionally, Wong et al. [36] proposed an image clustering algorithm using particle swarm optimization. They used population based stochastic optimization technique modeled by social behavior of bird flocks in which the algorithm maintained a population of particles, where each particle represented a potential solution to the optimization problem. Whereas Bifulco et al. in [37] presented a methodology based on a process that generated multiple clustering solutions using global optimization. They generated a number of different clustering solutions by exploiting the global optimization algorithm based on Controlled Random Search (CRS). The solutions so obtained were then clustered by using hierarchical method. Bing et al. [34] proposed a weighted fuzzy clustering algorithm to optimize the multicast routing in the overlay network. Explanatorily, multicast services involve one-to-many or many-to-many communications and can be provided as a basic network service or as an application-layer service such as peer-to-peer file sharing and video conferencing. In addition, an overlay network can be thought of as a virtual network of links and nodes built on top one or more existing network [46] [49], whose objective is to implement a network service that is unavailable in the existing network. Because of its setup, overlay multicast networks find the problems that manage their resource usage [46]. The interface bandwidth management outlines a major cost that constraint simultaneous multicast session supported by the overlay network, and hence overlay multicast network uses routing algorithms to optimize its use [34] [46] [49]. Bing et al. performed simulation and analyzed the evaluation of the multicast routing performance. Note that an overlay multicast is an application level multicast. It is based on a set of distributed Multicast Service Nodes (MSN) and provides multicast services for end users. Their research was based on routing optimization for better clustering performance, which was a multi-objective optimization with Non-Linear Programming characteristics.

Because, as mentioned above, an integrated approach of clustering method and optimization model is found to yield good results in other domains, we believe that integrating fuzzy clustering method with optimization model would in a way produce also achieve good results for customer segmentation and membership card generation in hypermarket (Wal-Mart in Shanghai).
CHAPTER 2: PROBLEM DEFINITION AND GOALS

2.1 Motivation

Many service industries such as hypermarkets, insurance and services industries accumulate giant amounts of data, which has caused a problem of how to effectively utilize and manage data, because data mining is an exploratory data analysis, trying to discover useful patterns in data that are not obvious to the data user [5] [26] [27].

Usually, the first task of a data mining process consists of summarizing the object information stored in a database in order to better understand its content, which is done by means of statistical analysis or query-and-reporting techniques [5]. Then, more complex operations are involved such as to identify models including supervised learning (the desired output is known and implicated) or unsupervised learning (the output is not considered and the method learns by itself only from input attributes) [5].

However, many companies realize the poor quality of their data collection only when a data mining analysis is started on it [27]. This is because the classical scenario is popularized as follows: a company firmly believes that there might be valuable information in the data they gather, then starts by building a long-term repository (a data warehouse) to store as much data as possible (e.g. a hypermarket systematically records all purchase and customer information) [5] [18]. In the meantime, they lose a sight of developing efficient methods for mining these data [28].

Among various types of data mining, higher management in a hypermarket is used to choose clustering in the data selection-preprocessing step due to the property of learning unsupervised similarities between objects and reducing the search space to a set of most important attributes for customer segmentation [5]. Additionally, the most frequently used clustering method is the hierarchical method, which identifies a certain number of groups with similar objects; it may be used in combination with the nearest-neighbor rule, which classifies any new object in the group most similar to it depending on a threshold value [1] [5]. By threshold value, it means that each element in a clustering model is defined by a numerical value 0 or 1. Thus, depending on the threshold value, each element is assigned to exactly one cluster [5]. In such a way, the management can identify characteristics and behaviors that can be exploited in order to reduce the customer attrition phenomenon (i.e., the loss of customers) by searching for customers that exhibit characteristics typical of someone who is likely to leave for a competing hypermarket during customer segmentation in the market place [5] [6]. However, this will cause a threshold value selection problem [29].

The clustering capability of hierarchical method is mainly determined by its threshold value [29]. This is to say that segmenting or merging customer information by their similarity is limited in performance with selection of threshold value, which determines the number of clusters [29]. Nevertheless, the threshold value is usually defined as either 0 or 1, which
forms the maximum mutual information [6] [29]. In this way, data clustering identifies the sparse and the congested places, and hence discovers the overall distribution patterns of the dataset [25]. In other words, with the increase of customer information, the customer information in each cluster tends to reach saturation level, because the properties of customers become more indistinct, thus the differences between clusters tend to be ambiguous [30].

The last but not the least, if an optimization model is to analyze a given problem scenario (i.e., economic benefit) and to solve a decision problem (e.g., generate membership cards) in a hypermarket, the control of profit and cost is considered [7]. Normally, the higher management defines restrictions as the costs of products and surrender part of the profits (i.e., products discount) so that it creates the constraint or cost functions; when the management formulates the objective functions, the selling prices of both products and membership cards have to be considered [4] [8]. Thus, they produce an optimization model for maximizing the economic benefit in terms of sales strategy (membership card generation).

However, in a hypermarket, optimization model usually consists of the profit improving based functions in its objective and constraint functions but combines little functions related to data mining [8] [10]. Because it would be prohibitively difficult for the management to consider plenty of customer information thoroughly while generating membership cards, they desert this procedure and prefer the traditional method of generating optimization model [9]. As such, it will lead to limit the application of optimization model [10]. Therefore, we believe that, the higher management has lost sight of impact of data mining during optimization modeling and hence such kind of commercial optimization model which needs the results of data mining for decision making is suitable for the solution with incorporation between optimization and data mining technique.

2.2 Research Questions

The main research question is:

*How can clustering and optimization be used for generating membership cards in a hypermarket?*

We divide our main research question into the following sub-questions:

**RQ1:** What is an appropriate method for customer segmentation based on their membership information?

**RQ2:** What is a suitable model for generating membership cards according to the business requirements?
2.3 Goals / Results

Our thesis project aims at developing a methodological approach for customers segmentation based on their characteristics in order to define membership cards attributed to mathematical optimization model in a hypermarket.

The project has several sub-objectives as follows:

- Design a clustering model for segmenting customers into clusters based on their characteristics;
- Create an optimization model for generating membership cards;
- Test and validate our models through a proof-of-concept implementation;
- Evaluate each of the models and our methodological approach.

Our study could substantially make it possible to utilize and manage customer information by way of clustering and optimization model. Thus, it would aid higher management to provide better services for the customers in a hypermarket.
CHAPTER 3: METHODOLOGY

A research methodology is a plan or strategy to conduct a research work in a scientific way. It links methods to outcome, which defines how to develop research activity and what measurement should be utilized to advance the research [31] [32]. In other words, the applied and realistic steps by means of which we found answers to our research questions constituted the research methodology. The general steps in a research process are shown in the figure below.

![Research process flow diagram](image)

**Figure 3: Research process flow diagram**

3.1 Research Problem Formulation

The research in our thesis work revolved around the problems and issues relating to a hypermarket, where the higher management wants to classify customers and generate membership cards based on certain customer information in order to provide better service there by maximizing the profit. Normally, higher management emphasizes the costs and profits of products and analyses of sales situation without mining further into customer information from the customer database when generating customer membership cards.
Therefore, the problems of segmenting customers and generating appropriate membership cards for customers formed our major tasks. We can gather the requirements via interviews, develop a method for segmenting customers and generating membership cards, and test and validate the approach using a proof-of-concept implementation.

3.2 Literature Review

This is an essential preliminary task in order to acquaint us with the available body of knowledge in the area of interest. Additionally, literature review as a first step helped us to gain a first-hand background knowledge, to identify the scope and purpose of the research and to find out appropriate papers.

Furthermore, reviewing the literature helped us in the following ways [33]:

- **Bring clarity and focus on our research problem**: Literature review helped us better understand the research area to visualize research problem.

- **Enhance methodology**: It benefited us to be in a better position to choose a methodology capable of delivering valid answer(s) to our research question(s).

- **Expand researchers’ knowledge**: Literature review ensured us to study extensively our intended subject area in order to conduct research study. It helped us to learn how our findings conform to the existing body of knowledge.

3.2.1 Search Strategy for Literature Review

Search strategy intends to find past and current, published and unpublished research papers. We used online databases provided by the library of Blekinge Institute of Technology. Five databases namely Google Scholar, IEEE Explore, Science Direct, Springer and Engineering Village were used to conduct searches. We selected these five databases because of their popularity, familiarity, exportability, coverage and advanced search facilities. In addition, we formulated and used specifically the following key words: data mining, clustering, optimization, optimize, membership, customer, and consumer (the Search Table for Literature review is shown in Appendix C).

Besides, we studied the abstracts of the related papers to find their relevancy and usefulness to our research. We extracted only the useful and relevant articles and studied thoroughly to grasp better knowledge and understanding of the subject. This gave a general assessment into the current research in the field of clustering and optimization based method for membership card generating in a hypermarket. However, articles with little or no relevance were simply discarded. Further, the parameters for our inclusion/exclusion criteria were languages, time frame and the scope of the articles. We read the articles in English and Chinese ranging from the year 1996 to 2011 because selecting recent papers would provide us with the latest available knowledge in our research domain. We also believed that recent papers would cover very important views, concepts, methods, techniques, etc.
3.3 Interviews for Data Collection and Approach Evaluation

Interviews can be of different types, which includes structured, semi-structured and unstructured, etc. We employed unstructured interview because it was a flexible and more casual method of data collection and there was no need to follow specific interview guidelines [31]. This would further help to get a broader understanding of clustering method and optimization model used for membership card generation in a hypermarket.

Furthermore, Wal-Mart Stores, Inc. has become the world’s largest private employer and retailer for over forty years and topped the Fortune 500 list and has been among the most valuable brands for many years now. Therefore, it is quite worthy for us to be a research object. During video interviews with the manager of Wal-Mart (in Shanghai), we asked several questions related to our research. Meanwhile, the required data were collected according to the relevant business requirements (the details of which will be discussed in Chapter 4). This would help us design a clustering model for customer segmentation and an optimization model for membership card generation. Further, we would send these two models with preliminary results to Wal-Mart Shanghai for evaluation of our methodological approach.

3.4 Development

We designed a clustering model and an optimization model according to the requirement. The development of clustering model involves three steps: data standardization, building fuzzy similarity matrix and clustering. Furthermore, development of optimization model includes simplifying problem, modeling and solution. (The details of these two models’ development will be described in Chapter 5.)

3.5 Proof-of-Concept Implementation

To test and validate the suggested method for customer card generation, we made a proof-of-concept implementation of our approach. Meanwhile, we substituted a certain amount of processed data into clustering model, which consists of five different phases including standard deviation transformation, range transformation, correlation coefficient method, transitive closure method and observation method. Further, on substitution of the results of clustering model to optimization model to obtains preliminary result. It may be noted that an optimization model consists of objective functions and constraint functions. (The details of which would be discussed in Chapter 5 and 6).
CHAPTER 4: INTERVIEW AND DATA PREPARATION

This chapter begins with the description of the video interviews with Mr. Yizhu Shi (the CEO of Wal-Mart in Shanghai) in order to grasp business requirements and data collection. Following the description of interviews, we illustrate how the gathered data are to be processed and prepared for designing clustering model and optimization model.

4.1 Interview Design

Interviews have been used extensively for data collection across all the disciplines of social sciences and in educational research [38]. In our project, the purpose of the interview is to acquire the knowledge about the current situation of Wal-Mart (in Shanghai) in terms of customer segmentation and membership card information (from Wal-Mart’s perspective). To achieve our target, we prepare the following targeted questions to Mr. Yizhu Shi:

Q1: Briefly represent the membership card policy in Wal-Mart Shanghai.
Q2: How is your membership card policy generated?
Q3: What is the current situation of customer segmentation (or classification)?
Q4: What are the characteristics of customers in each partition (or group)?)
Q5: According to the current situations about membership card policy and customer segmentation, what is your perspective about them?

Through these questions, we can define the needed data attributes of customer information and grasp the internal problems required to be solved. This will help us to develop an approach for customers clustering based on their characteristics in order to define membership cards by a mathematical optimization model. However, it may be noted here that we conducted interviews in Chinese, and translated and summarized into English.

4.2 Interview Analysis

The responses to the targeted questions are as follows:

Q1: Briefly represent the membership card policy in Wal-Mart Shanghai.

In Wal-Mart Shanghai, there are four types of membership cards: Diamond, Platinum, Gold and Classic memberships. And the membership cards are adopted on a prepayment method.

The price of Diamond Membership card is 2500 CNY (1 SEK≈1 CNY). The customers with this card have the privilege to obtain a 20% discount on all products including discounted products (double discount) with the exception of Liquor and Cigarettes.

The price of Platinum Membership card is 2000 CNY. The members holding this card can get a 15% discount on all products including discounted products (double discount) except for Liquor and Cigarettes.
The price of Gold Membership card is 1500 CNY. The members holding this card can get a 10% discount on all products including discounted products (double discount) except for Liquor and Cigarettes.

The price of Classic Membership card is 500 CNY. The members holding this card can get a 5% discount on all products except for Liquor, Cigarettes and already discounted products.

Furthermore, the validity of each membership card is up to 30 days (counted from the day of recharging); otherwise, the members cannot get any discount on their purchases after the validity period.

Q2: How is your membership card policy generated?
In general, to identify the value of each membership card, we make a control of the products’ costs and profits. Meanwhile, we define restrictions as the costs of products and surrender part of the profits (i.e., products discount). Additionally, the selling prices of both products and membership cards have to be considered. Typically, the stocking price (cost) of products is 50%-65% of its selling price in ordinary retail businesses. In this case, the reason why we extend such membership services is to set our sights on attracting more customers.

Q3: What is the current situation of customer segmentation (or classification)?
Originally, segmentation of customers depends on the types of membership cards they select. However, with the increasing number of customers, the types of membership cards have not been relatively improved. Therefore, some customers are not satisfied with the existing status of membership because these membership cards may not be tailored to each customer’s need. Therefore, we are considering how to improve customer segmentation (4-6 groups among customers could be a good idea).

Q4: What are the characteristics of customers in each partition (or group)?
Most of the customers work or live near the Wal-Mart store. They choose different membership cards and have diverse characteristics as depicted in Table 1 shows below.

<table>
<thead>
<tr>
<th>Type of memberships</th>
<th>Characteristics</th>
<th>Average consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diamond</td>
<td>Higher frequency, Housewives, Big family income</td>
<td>About 2200 CNY per month</td>
</tr>
<tr>
<td>Platinum</td>
<td>Medium frequency, High/Medium income, office workers (30-40 years old)</td>
<td>About 1600 CNY per month</td>
</tr>
<tr>
<td>Gold</td>
<td>Medium frequency; Medium income, office workers (25-30</td>
<td>About 1200 CNY per month</td>
</tr>
<tr>
<td>Classic</td>
<td>Medium / Low frequency, Medium / Low income, office / retired workers, students,</td>
<td>About 600 CNY per month</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------------------</td>
<td>-------------------------</td>
</tr>
</tbody>
</table>

Table 1: The characteristics of customers

Q5: According to the current situations about membership card policy and customer segmentation, what is your perspective about them?

In recent years, we found that there is a slight (even negative) increase in the number of customers. There are two reasons for this. First, in the early stage of extending membership cards, we did not have enough customer information in our database. Hence, the membership card policy emphasized only on costs and profits of products, and was formulated through analyses of sales situation and market evaluation. Second, we did not actively classify the increasing customers but acquiescently assigned them to one of the groups according to which type of membership cards they choose.

In other words, we formulated the membership cards first, and then segmented the customers. This causes a problem that, gradually, the customers may not satisfy with the current membership card policy. However, if we reclassify the customers by their different characteristics such as demands, consumption capacities and so on, it would be prohibitively difficult to consider a large amount of customer information thoroughly. Therefore, we must think of a solution that will improve the membership services with respect to both customer segmentation and membership card policy in order to enhance customers’ satisfaction and consequently attract more customers.

From the above interview, we find that customer satisfaction level decreases with increase in the number of customer because the prices of existing membership cards tend to deviate from the monthly consumptions of customers. Therefore, implementing our research objective (develop an approach for defining membership cards based on clustering and optimization models in a hypermarket) needs data attributes basically including monthly consumption, salary/income, name, age and marital status.

4.3 Data Preparation

4.3.1 Data Collection

In this section, we require to collect the targeted data from Wal-Mart (in Shanghai) in order to develop clustering and optimization model. From the interview analysis, we are able to gather customer information in the membership application form, membership cards information and the customers’ consumption records by email, which are vital for our thesis project.

However, the collected data contain some individual privacy and in Chinese Language. In order to fix the problem of customer’s privacy, we have replaced customers’ names and marked them as (i.e., customer1, customer2). Additionally, because the quantity of raw data is
too large to display wholly, we illustrate the properties of these data as shown in Table 2 – Table 5.

<table>
<thead>
<tr>
<th>Name</th>
<th>ID</th>
<th>Gender</th>
<th>Age</th>
<th>Marital status</th>
<th>Occupation</th>
<th>Salary</th>
<th>Telephone Number</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer 1</td>
<td>...</td>
<td>M / F</td>
<td>e.g.</td>
<td>Yes / No</td>
<td>e.g.</td>
<td>e.g.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>46</td>
<td></td>
<td></td>
<td>Business</td>
<td>10000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

... ... ... ... ... ... ...

Table 2: Membership application form

<table>
<thead>
<tr>
<th>Card Types</th>
<th>Price</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diamond</td>
<td>2500</td>
<td>20% discount on all products including discounted products (double discount) with the exception of Liquor and Cigarettes</td>
</tr>
<tr>
<td>Platinum</td>
<td>2000</td>
<td>15% discount on all products including discounted products (double discount) except for Liquor and Cigarettes</td>
</tr>
<tr>
<td>Gold</td>
<td>1500</td>
<td>10% discount on all products including discounted products (double discount) except for Liquor and Cigarettes</td>
</tr>
<tr>
<td>Classic</td>
<td>500</td>
<td>5% discount on all products except for Liquor, Cigarettes and discounted products.</td>
</tr>
</tbody>
</table>

Table 3: Membership card information

<table>
<thead>
<tr>
<th>ID</th>
<th>Membership Status</th>
<th>Average Consumption (monthly)</th>
<th>Frequency (monthly)</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>Diamond / Platinum / Gold / Classic</td>
<td>2326 CNY</td>
<td>e.g. 5</td>
<td>List</td>
</tr>
</tbody>
</table>

... ... ... ... ... ...

Table 4: Customers’ consumption records

Table 5: Customers’ consumption records (addendum)

4.3.2 Data Processing
The collected data from Wal-Mart (in Shanghai) needs to be refined and processed before utilization. Indeed, the purpose of data processing is to structure analyzable data among raw data, which includes the following chain of sequences [39] [40] [41]:

- **Data Extraction:** This is the first step in data processing. It involves isolating only the relevant data from the huge amount of information stored in the database and requires complex scrutiny to identify the needed data. In our project, the data we need are customer’s name, ID, salary and monthly consumption as the conditions of customer segmentation.
Data Cleaning: Data cleaning comprises of eliminating redundant data and duplication of records such as spelling errors and domain inconsistencies. We believe that there are small amounts of incomplete customer data and hence cannot be utilized. For instance, some customers may not have enough purchase records and such information is short of certain accountability. Therefore, we have to remove them in order to ensure the quality of the data.

Data Conversion: Data conversion involves switching of one type of data to another type (e.g. numeric to symbolic or vice versa), defining new attributes, and undoing with noises etc. In our case, when we analyze different data tables (matrices), each element in these tables will be transformed by using different formulae during cluster modeling.

Data Summarization: Data summarization is the proposition of data resulting in a smaller set with accumulated information, whose scheme is to generate concise and essential description of the data set. In this case, we should summarize the customer data (such as average consumption) in each partition after clustering. With this, we can apply these summarized data on testing optimization model.

Data Loading: This is the final step in data processing. It loads the resulting data from each step into database. We should employ this process every time we execute each of the data processing phases.

Following data extraction and data cleaning processes, the raw data now has been reform as usable information for customer clustering as shown in the table below.

<table>
<thead>
<tr>
<th>Name</th>
<th>ID</th>
<th>Gender</th>
<th>Age</th>
<th>Marital status</th>
<th>Consumption Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer 1</td>
<td>...</td>
<td>M</td>
<td>46</td>
<td>Yes</td>
<td>List</td>
</tr>
<tr>
<td>Customer 2</td>
<td>...</td>
<td>F</td>
<td>38</td>
<td>Yes</td>
<td>List</td>
</tr>
</tbody>
</table>

Table 6: Customer information (after data extraction and cleaning)

<table>
<thead>
<tr>
<th>Year/Month</th>
<th>Customer1</th>
<th>Customer2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010/1</td>
<td>1865 CNY</td>
<td>967 CNY</td>
<td>...</td>
</tr>
<tr>
<td>2010/2</td>
<td>2307 CNY</td>
<td>1108 CNY</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Average</td>
<td>2198.5 CNY</td>
<td>1306.6 CNY</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 7: Customers’ consumption records (after data extraction and cleaning)
CHAPTER 5: A METHOD FOR CARD GENERATION

In this chapter, we present a clustering and optimization based method for membership card generation in a hypermarket. A flow chart for our approach is shown in the figure below.

5.1 Clustering Model Design

5.1.1 Design Principle

In a hierarchical clustering, data are not segmented into a particular number of classes or clusters at a single step. Instead, the clustering consists of a series of partitions, which may run from a single cluster containing all individuals, to \( n \) clusters each containing a single individual. Hierarchical clustering techniques may be subdivided into agglomerative methods, which proceed by a series of successive fusions of the \( n \) individuals into groups, and divisive methods, which separate the \( n \) individuals successively into finer groupings [6].

The clustering capability of hierarchical method is mainly determined by the selection of its threshold value that is used to be defined by either zero or one. As such, data clustering identifies the sparse and the congested places, and hence discovers the overall distribution patterns of the dataset [25]. Furthermore, in a real environment (a hypermarket), there are many ambiguous concepts. For example, the monthly consumption of a customer can be “high or low”; chocolate can be “delicious” or “distasteful”. These concepts are difficult to classify by a certain number and the perspectives vary from person to person. Therefore, in order to solve the issues with such equivocal concepts, we utilize fuzzy clustering in our thesis.

In fuzzy clustering, we first calculate the similarity coefficient of any two objects, and then keep it being a threshold value ranging between 0 and 1. Based on this, we build the similarity matrix and consequently form the fuzzy clustering relationship [42]. The procedure of cluster modeling mainly includes following steps as the figure shows below.

Figure 4: Work flow in general
5.1.2 Data Standardization

First, we transform the initial data (in the Table 6 and 7) as data matrix: assume customer data set $U = \{x_1, x_2, \ldots, x_n\}$ to be the clustered objects, and each one has $m$ indices, which describe its properties as $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}, (i = 1, 2, \ldots, n)$. Thereupon, we get the initial data matrix as follows:

\[
\begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1m} \\
  x_{11} & x_{12} & \cdots & x_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nm}
\end{pmatrix}
\]

**Formula 1 (Matrix 1):** Initial data matrix

In the matrix, $x_{nm}$ indicates the $m^{th}$ property of the $n^{th}$ initial data.

Furthermore, in a practical problem, different data have different dimensions that illustrate values of different properties. In order to compare with these dimensions, we have to appropriately transform the initial data in such a way that it compresses the data into the region between 0 and 1. Typically, the following transformation methods are employed in order to perform fuzzy clustering to the desired results [42]. We use Standard deviation transformation to eliminate the initial data dimensions. However, this method may not compress the data in the expected region of 0 to 1. Therefore, we make use of Range Transformation.

*Standard Deviation Transformation:*

\[
x'_{ik} = \frac{|x_{ik} - \overline{x_k}|}{s_k}, \quad (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m)
\]

**Formula 2:** Standard deviation transformation

where $\overline{x_k} = \frac{1}{n} \sum_{j=1}^{n} x_{jk}$ presents the average value of each column in the matrix, and $s_k = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \overline{x_k})^2}$ describes the coefficient of standard deviation.
Although this method can eliminate the dimensions of initial data, $x'_{ik}$ may not be in the region $[0,1]$. We will further describe this in Chapter 6.

**Range Transformation:**

$$x'^*_{ik} = \frac{x'_{ik} - \min_{1 \leq e \leq n} \{x'_{ik}\}}{\max_{1 \leq e \leq n} \{x'_{ik}\} - \min_{1 \leq e \leq n} \{x'_{ik}\}}, \quad (k = 1, 2, \ldots, m)$$

**Formula 3:** Range transformation

This formula is based on Formula 2. Obviously, there must exist $0 \leq x'^*_{ik} \leq 1$, and it eliminates the dimensions. We will prove this method further in the next chapter.

### 5.1.3 Building Fuzzy Similarity Matrix

In customer data set $U = \{x_1, x_2, \ldots, x_n\}$, $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$, $(i = 1, 2, \ldots, n)$, we need to define similarity coefficient between any two objects so as to build fuzzy similarity matrix. Assume $r_{ij}$ is the similarity of $x_i$ and $x_j$, $r_{ij} = R(x_i, x_j)$, $(j = 1, 2, \ldots, n)$. Additionally, the strategies of defining $r_{ij}$ can be learnt from Similar Coefficient Method and Distance Method. However, there are several formulae such as Angle Cosine Method, Correlation Coefficient method and Hamming Distance Method, which have good performance in building fuzzy similarity matrix [42].

In our project, we choose Correlation Coefficient Method (Formula 4) because the result of this formula illustrates a certain index of correlation degree between elements in a fuzzy similarity matrix [42]. By index of correlation degree, it means the customer’s consumption ability, which will be utilized in our optimization model.

$$r_{ij} = \frac{\sum_{k=1}^{m} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{m}(x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{m}(x_{jk} - \bar{x}_j)^2}} \quad (k = 1, 2, \ldots, m)$$

**Formula 4:** Correlation coefficient method

where $\bar{x}_i$ and $\bar{x}_j$ are the average values of relevant columns, $x_{ik}$ is each element in the matrix.

### 5.1.4 Clustering

In this section, we need to establish fuzzy clustering relations through the fuzzy matrix $R$. Although the fuzzy matrix consists of similar coefficients between any two elements, it may not constitute fuzzy clustering relations [43]. Therefore, we have to reform accordingly the fuzzy matrix as fuzzy equivalent matrix. However, it may be noted that Transitive Closure and Boolean Matrix method can be used to establish fuzzy clustering relations. In our project, the elements of the fuzzy matrix are transformed from customer’s monthly consumption in Table 7, where the data are equivalent. Therefore, we select Transitive Closure Method.

The purpose of Transitive Closure Method is to compute the fuzzy equivalent matrix $R^*$ via the fuzzy matrix $R$ (namely, transitive closure $t(R) = R^*$). According to the theorem [43], assume that $R$ has $n$ fuzzy similarity relations (elements of the matrix $R$), there must exist a minimal natural number $k \leq n$, such that $t(R) = R^k$ and constantly $R^L = R^k$ ($L$ is a natural number, $n \geq L \geq k$). This is to say that we can obtain the transitive closure $t(R)$ of $R$ within $n$ computations thereby getting a fuzzy equivalent matrix. Additionally, to increase computational speed, we utilize Square Method to calculate: 

$$(R \circ R) = R^2, (R^2 \circ R^2) = R^4, \ldots, R^k, R^L$$

in turn, until $R^k = R^L$, then $t(R) = R^k$.

**Formula 5:** Transitive closure method

Finally, we define the threshold value (noted as $\lambda$) which means that we can now determine the number of clusters. However, in the clustering analysis, “How many clusters are most
suitable?” This is a very confusing question which experts remain unsuccessful in their bid to find a perfect solution. Nonetheless, it is an inevitable question.

Primarily, there are two approaches to determine the threshold value $\lambda$ as follows [42]:

- **Statistical Method**: This method gives the exact threshold value $\lambda$ and focuses on the performance of clustering. It can be achieved by utilizing the quotient of two object distances to obtain an exact $\lambda$. Meanwhile, the numerator is the distance between any two classes (columns) and the denominator is the distance between any two examples in the same class. In addition, this method obtains a bigger value of $\lambda$. In other words, there are many partitions to be clustered with slight distances (differences). However, in our project, too many partitions are likely to cause trouble and that it is impossible to generate so many membership cards for each partition in a hypermarket. Therefore, in order to avoid such situation, we choose another approach (observation method, which will be discussed in the succeeding step) to find the threshold value $\lambda$.

- **Observation Method**: Using observation method, the threshold value $\lambda$ can be obtained in two ways. One way is to use trial and error method, where we first find a maximal element $T$ ($0 < T < 1$) as $\lambda$ in a given fuzzy equivalent matrix $R^*$. If the number of clusters is bigger than what we expect, then choose the second biggest element as $T$ and assigned as $\lambda$. The process is repeated in the same manner until the best result is obtained. However, this is a lengthy and time-consuming process because one needs to use the value of each element from the fuzzy matrix until an appropriate/expected/desired value of $\lambda$ is obtained. Therefore, we prefer to use another method in our thesis project, which is to ask an expert to know the number of clusters thereby defining the value of $\lambda$. We inquire an expert from the Wal-Mart (Mr. Yizhu Shi) to determine the number of clusters via the video interview.

Thus, observation method best suit our study to find the threshold value $\lambda$. Eventually, using appropriate formulae, steps and procedures, our clustering model has been completed.

### 5.2 Optimization Model Design

#### 5.2.1 Design Principle

In this section, we present how optimization problem in membership card generation can be expressed mathematically and formulated as an optimization model. The purpose is to provide an understanding in how to reason when an optimization model is formulated based on clustering results. An important step in the modeling process is to identify and define decision variables and to introduce a suitable notation.

On the basis of LP principle (described in Chapter 1.3), now we start to design the optimization model according to the design flow as the Figure 6 shows below.
5.2.2 Simplify Problem

Wal-Mart (in Shanghai) is a large-scale supermarket consisting of many fixed customers (members). These customers are segmented into $n$ clusters with different consumption characteristics by the clustering model (described in Chapter 5.1). In each cluster, there is an average similarity coefficient $s_i (i = 1, 2, \ldots, n)$, which is an average value of the elements in $i^{th}$ cluster extracted from fuzzy equivalent matrix (shown in Chapter 5.1) and represents the coefficient of customers’ consumption ability. In addition, we also consider the average values about customers’ monthly consumption in $n$ clusters. Furthermore, from the interview, we grasp an important condition: “the stocking price (cost) of products is 50%-65% of its selling price in ordinary retail businesses.” In other words, retailers can earn at least 35% from every product. Therefore, we define the profit coefficient $p$ as 0.35. All parameters are as follows:

1) The number of clusters is $n$;
2) The average similarity coefficients are denoted by $s_i$;
3) The average value of customers’ monthly consumption in each cluster is denoted as $c_i$;
4) The profit coefficient is $p$.

Meanwhile, it is evident from the interview that the price of membership card should not be greater than the customer’s monthly consumption $c_i$ because the customers will not get any discount on their purchases after the validity period (30 days). Therefore, we assume the price of membership card $x_i$ as a variable. Additionally, in order to maximize the profit, the customer’s monthly consumption $c_i$ needs to be increased depending on $s_i$. Furthermore, in order to decide on the discount for each membership card, we define $d_i$ as another variable. In this way, we suppose two variables as follows:

1) $x_i = \text{price of membership card for } i^{th} \text{ cluster, } i = 1, 2, \ldots, n.$
2) $d_i = \text{how much discount (percentage) the customer can get in the } i^{th} \text{ cluster, } i = 1, 2, \ldots, n.$
5.2.3 Modeling

After problem simplification, the problem of membership card generation becomes simple to understand, which is to maximize the Wal-Mart’s benefit with the consideration of the cost in monthly consumption using the supposed membership card.

First, we structure the objective function as follows:

$$\max \left\{ \sum_{i=1}^{n} (p \times x_i) - \sum_{i=1}^{n} (c_i \times d_i) \right\}$$

where $\sum_{i=1}^{n} (p \times x_i)$ means the profit that Wal-Mart gets from the customers who pay for the membership cards, and $\sum_{i=1}^{n} (c_i \times d_i)$ is the cost that Wal-Mart bears because of the provided discount.

The constraints on this objective function are the customers’ average monthly consumption ($c_i$) and the certain value of monthly purchase depending on the customer’s consumption ability ($c_i \times s_i$). Therefore, we can write the constraint function as:

$$s.t. \quad x_i \leq c_i$$

$$c_i \times (1 - d_i) \geq c_i \times s_i$$

$$x_i > 0$$

$$0 < d_i < 1$$

$$i = 1, 2, ..., n$$

where $x_i \leq c_i$ implies that the price of membership card for $i^{th}$ cluster should not be greater than the monthly consumption $c_i$ of customers in $i^{th}$ cluster, because customer cannot enjoy any discount on their purchase after the validity period (30 days). Furthermore, in order to enhance the profit maximization, we supposed that there is a certain potentiality that the customers’ monthly consumption $c_i$ can be increased depending on their consumption abilities. Explanatorily, the customers’ consumption ability coefficients ($s_i$) of different clusters extracted from fuzzy equivalent matrix are between 0 and 1, which means the more the coefficient $s_i$ tends to 0, the greater will be such potentiality $(1 - s_i)$ [44]. Consequently, we can state this as:

$$x_i \leq c_i \times (1 + (1 - s_i))$$

where $(1 + (1 - s_i))$ means how much space the customer’s monthly consumption ($c_i$) can be increased.

Further, $c_i \times (1 - d_i) \geq c_i \times s_i$ implies that the monthly consumption $c_i$ after discount $d_i$ should not be less than customer’s consumption ability. For instance, if a customer can afford a product priced for 100 SEK, then the price of the product should not be less than 100 SEK after discount. In addition, to guarantee the cost minimization, we assume that there is a certain potentiality that the customers’ consumption ability ($s_i$) can be boosted. However, we considered that there exists a linear relation between $d_i$ and $s_i$, which means the more consumption ability coefficient is increased (this is to say that the coefficient $s_i$ more tends to 0), the more discount is declared [44]. Therefore, in order to keep the balance between $d_i$ and $s_i$, we need to enlarge the customer’s consumption ability coefficient $s_i$ appropriately. Thus, we can write this as:

$$c_i \times (1 - d_i) \geq c_i \times s_i (1 + \frac{1 - s_i}{3})$$

where $(1 + \frac{1 - s_i}{3})$ shows a rising range of the customers’ consumption ability ($s_i$).

Eventually, we obtain a solution of optimization model which consists of two parts: profit and cost problems. For one thing, the profit problem illustrates how to express the total profit and is defined as summation of the profit on the sales of membership cards. For another, cost problem presents the total cost and is interpreted as accumulation of the cost on the discount. Additionally, both of these exist as a linear relation as follows:
\[
\max \left\{ \sum_{i=1}^{n} (p \times x_i) - \sum_{i=1}^{n} (c_i \times d_i) \right\}
\]

s.t.
\[
x_i \leq c_i \times (1 + (1 - s_i))
\]
\[
c_i \times (1 - d_i) \geq c_i \times s_i (1 + \frac{1 - s_i}{3})
\]
\[
x_i > 0
\]
\[
0 < d_i < 1
\]
\[
i = 1, 2, ..., n
\]

where \(c_i, s_i\) and \(p\) are known numbers (parameters) and hence we obtain \(x_i\) and \(d_i\). Further, we will test and validate this solution in Chapter 6 and 7.

5.3 Proof-of-Concept Implementation Design

5.3.1 Design Principle

Foremost, the purpose of creating a proof-of-concept implementation is to analyze the performance of clustering and optimization models that are being implemented in our project. The formulae, the resulting data of each phase and the utilization of these resulting data will be practiced and analyzed in detail. Practically, the experimental data is valuable for both researchers and users (higher management in hypermarket). For the researchers of clustering and optimization, the empirical data will validate how well the theoretical strategies were implemented and where and even how they need to be improved or refined. As to the users, the experimental data can be seen as preliminary results, which seems more important because the data will be regarded as a standard to evaluate the performance of customers segmentation and membership cards generation.

5.3.2 Implementation Flow

In the first place, the customer information (monthly consumption) as input data is designed for initial matrix in the clustering model. Meanwhile, we extract a part of the processed data for proof-of-concept implementation instead, because clustering is an unsupervised learning process and hence we need not test all the data but implement the model with a certain number of data [14] [15].

Furthermore, these data are reprocessed by using corresponding formulae: Standard Deviation Transformation, Range Transformation, Correlation Coefficient Method, Transitive Closure Method and Observation Method (Clustering). Eventually, we obtain the clustering result. Remarkably, the inputs and outputs of these formulae are inherited with each other as the Figure 7 shows below.
Moreover, prior to implement our optimization model, it is important to analyze the clustering result. We, therefore, extract and calculate the following required values that can serve as input to our optimization model.

- Number of clusters
- Average similarity coefficients in each cluster
- Average value of customer’s monthly consumption in each cluster
- Profit coefficient

Based on this, we first substitute the conditions into constraint functions in order to obtain the unknown numbers (\(x_i\) and \(d_i\) shown in Chapter 5.2), and then substitute these unknown numbers into objective function (the implementation flow is displayed as the Figure 8 below). Thus, we suggest the membership card generation according to the result of optimization model after the result description.
Figure 8: Implementation flow for optimization model
CHAPTER 6: PROOF-OF-CONCEPT IMPLEMENTATION

In this chapter, we first illustrate the program (codes) for clustering model. Second, we describe the variables definition in programs and implementation environment. Furthermore, the functions and the effects of the programs regarding the formulae are detailed in implementation analysis, and then we obtain a result of clustering. In addition, to test the optimization model, we substitute the related coefficients and the result of clustering into optimization model thereby obtaining the result. Finally, we combine the outcomes of clustering and optimization models as the preliminary result of our solution, which will eventually be used for membership card generation.

6.1 Program for Clustering Model

Our clustering model covers a number of formulae that are responsible for different tasks at different stages of modeling. Besides, because of large computations, we develop program for the implementation of formulae (all source codes are provided in Appendix B). The following are the formulae that are utilized in each step of our modeling processes:

◆ In order to fulfill data standardization, we apply the following two formulae:
  ➢ Standard Deviation Transformation (Formula 2)
  ➢ Range Transformation (Formula 3)
◆ Next, in order to build fuzzy similarity matrix, we use:
  ➢ Correlation Coefficient method (Formula 4)
◆ Similarly, to compute the fuzzy equivalent matrix, we employ:
  ➢ Transitive Closure method (Formula 5)
◆ Finally, to cluster customers:
  ➢ Observation Method (define the threshold value λ for clustering)

6.2 Variable Definition

In the proof-of-concept implementation, the program used for clustering is divided into five parts (as the Figure 7 shows): Standard Deviation Transformation, Range Transformation, Correlation Coefficient method, Transitive Closure method and Observation Method. The variables in these programs are concluded the two types of variables related to the proof-of-concept implementation: (The details of the variables are described in Appendix E.)

A. Independent Variables:
✓ Input data, such as initial matrix and coordinate of the target element;
✓ Size of initial matrix;

B. Dependent Variables:
✓ Result and temporary matrices;
✓ Sizes of these result matrices;
✓ Numerators and denominators;
✓ Average values;
✓ Minimal and maximal values
✓ Flag.

6.3 Implementation Environment

Because the main objective of this implementation is to calculate the result matrices in each phase via programs according to the related formulae, the major tasks include a wide variety of pure computations. Therefore, we choose Console Application based on Windows 32-bit as our implementation environment. Apart from this, we utilize the classical C/C++ on Visual Studio 2008 as the language and platform for the implementation respectively.
6.4 Implementation Analysis
6.4.1 Standardizing Data
We first extract the initial data from the Table 6 and 7 as an example set shown by Table 8 (detailed in Appendix A) below:

<table>
<thead>
<tr>
<th>Year/Month</th>
<th>Customer1</th>
<th>Custom2</th>
<th>…</th>
<th>Customer100</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010/1</td>
<td>908.6</td>
<td>4074.94</td>
<td>…</td>
<td>1165.5</td>
</tr>
<tr>
<td>2010/2</td>
<td>1489.3</td>
<td>3689.5</td>
<td>…</td>
<td>1963.88</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>2010/12</td>
<td>1435.89</td>
<td>4156.8</td>
<td>…</td>
<td>3166</td>
</tr>
</tbody>
</table>

Table 8: Example set for testing

Additionally, we import the data of Table 8 into the data matrix (Matrix 1):

\[
\begin{pmatrix}
  x_{11} & x_{12} & \ldots & x_{1m} \\
  x_{n1} & x_{n2} & \ldots & x_{nm}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
  908.6 & 4074.94 & \ldots & 1165.5 \\
  1489.3 & 3689.5 & \ldots & 1963.88 \\
  1435.89 & 4156.8 & \ldots & 3166
\end{pmatrix}
\]

where \( x_{11} = 908.6, x_{12} = 4074.94, \ldots, x_{1100} = 1165.5, x_{21} = 1489.3, \ldots, x_{121} = 1435.89, \)

and \( n = 12, m = 100. \)

Second, we import this initial data matrix into Standard Deviation Transformation formula (Formula 2) as the Figure 9 shows below:

\[
x_{ik}' = \frac{x_{ik} - \overline{x_k}}{s_k}, \quad (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m) \quad \text{(Formula 2)}
\]

\[
s_k = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \overline{x_k})^2}
\]
Because of the nature of the formula on computation, there are many elements beyond the region between 0 and 1 (as the circled elements in the Figure 9), despite eliminating the dimensions of initial data. Therefore, we use Range Transformation formula (Formula 3) for making up with such drawbacks. We input the result of Standard Deviation Transformation into Formula 3 and obtain the result as Figure 10 displays below.

\[
x_{ik}'' = \frac{x_{ik}' - \min\{x_{ik}'\}}{\max\{x_{ik}'\} - \min\{x_{ik}'\}}, \quad (k = 1, 2, \cdots, m) \quad (\text{Formula 3})
\]
Through the computation of Formula 3, there must exist $0 \leq x_{ik}^{(l)} \leq 1$ and the dimensions of the result matrix have been eliminated.

### 6.4.2 Constructing Fuzzy Similarity Matrix

In this section, we need to define similarity coefficient between any two objects so as to build fuzzy similarity matrix $r_{ij} = R(x_i, x_j)$ by using Correlation Coefficient method (Formula 4). First of all, we import $r_{ij}$ into Formula 4, and then the result is spread out as in Figure 11.

$$R = r_{ij} = \frac{\sum_{k=1}^{m}(x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})}{\sqrt{\sum_{k=1}^{m}(x_{ik} - \bar{x_i})^2} \cdot \sqrt{\sum_{k=1}^{m}(x_{jk} - \bar{x_j})^2}} (k = 1, 2, ..., m) \quad \text{(Formula 4)}$$
After computation, the size of the result matrix $r[i][j]$ should become $[m][m]$, because the required matrix whose similarity coefficients are used for internal comparison with each other has to meet the conditions of Reflexivity and Symmetry. Explanatorily, Reflexivity means that the matrix $r[i][j]$ satisfies $R_{ij} = 1$; and Symmetry is subject to $R_{ij} = R_{ji}$.

In this way, the values along with the diagonal of the matrix $r[i][j]$ are 1.000 and the values of the $i$th row are equal to those of the $i$th column with respect to Figure 11.

6.3 Building Fuzzy Equivalent Matrix

In this section, we need to establish fuzzy clustering relations through the fuzzy similarity matrix $R$. Although the matrix $R$ contains similar coefficients between any two elements and satisfies Reflexivity and Symmetry, it may not constitute fuzzy clustering relations because the matrix $r[i][j]$ should become $[m][m]$, because the required matrix whose similarity coefficients are used for internal comparison with each other has to meet the conditions of Reflexivity and Symmetry. Explanatorily, Reflexivity means that the matrix $r[i][j]$ satisfies $R_{ij} = 1$; and Symmetry is subject to $R_{ij} = R_{ji}$.

After computation, the size of the result matrix $r[i][j]$ should become $[m][m]$, because the required matrix whose similarity coefficients are used for internal comparison with each other has to meet the conditions of Reflexivity and Symmetry. Explanatorily, Reflexivity means that the matrix $r[i][j]$ satisfies $R_{ij} = 1$; and Symmetry is subject to $R_{ij} = R_{ji}$.

Figure 11: Result of Formula 4 (part)
the matrix $R$ possibly does not meet the condition of Transitivity. This is to say that $R$ should be calculated as $(R \circ R) = R^2, (R^2 \circ R^2) = R^4, \ldots, R^k, R^L$ in turn until $R^k = R^L$, then the Transitive Closure of $R$ is written as $t(R) = R^k$. Therefore, we utilize Formula 5 for obtaining the $t(R)$.

First, we calculate $(R \circ R) = R^2$, then $(R^2 \circ R^2) = R^4$, the rest can be done as the same manner. Meanwhile, we compare the results of computations with each other in order to identify whether there is some difference among them or not. If so, the computation will return to obtain the Transitive Closure (shown in the source code).

In contrast, $R^2 \neq R^4 \neq R^8 \neq R^{16} \neq R^{32} = R^{64}$. Therefore, according to the theorem [43], we believe that,

$$R^{32} = R^{64} = R^L \ (L > 64, L \ is \ a \ natural \ number)$$

Consequently, the Transitive Closure $t(R)$ is $R^{32}$.

### 6.4.4 Clustering Customers

In this section, we define the threshold value (noted as $\lambda$) by using observation method mentioned in chapter 5. During the proof-of-concept implementation, we first set $\lambda = 0.971$, which obtains 91 clusters wherein customers $\{C8, C17, C18, C19, C24, C29, C47, C60, C68, C78\}$ come under one cluster, while each of the remaining 90 clusters contains one customer each (i.e., $\{C1\}$, $\{C2\}$ and so on). Obviously, this clustering result can not achieve our requirement. Therefore, we continue to define the value of $\lambda$ as follows:

- **Set $\lambda = 0.95$**
  
  There are 78 clusters: $\{C8, C13, C15, C16, C17, C18, C19, C21, C23, C24, C25, C29, C32, C43, C47, C60, C61, C68, C69, C74, C76, C78, C95\}$ and each of the other customer are segmented in each of the remaining clusters.

- **Set $\lambda = 0.92$**
  
  We have 50 clusters: $\{C8, C13, C15, C16, C17, C18, C19, C21, C23, C24, C25, C29, C32, C43, C47, C60, C61, C68, C69, C74, C76, C78, C95\}, \{C8, C15, C16, C25, C29, C32, C68, C76, C95, C6, C9, C11, C26, C27, C28, C31, C33, C34, C36, C39, C41, C42, C48, C49, C50, C54, C58, C65, C67, C71, C73, C75, C89, C91, C93, C97, C98, C100\}$ and each of the remaining customers occupy the remaining individual cluster.

- **Set $\lambda = 0.96$**
  
  We get 34 clusters: $\{C8, C13, C15, C16, C17, C18, C19, C21, C23, C24, C25, C29, C32, C43, C47, C60, C61, C68, C69, C74, C76, C78, C95\}, \{C8, C15, C16, C25, C29, C32, C68, C76, C95, C6, C9, C11, C26, C27, C28, C31, C33, C34, C36, C39, C41, C42, C48, C49, C50, C54, C58, C65, C67, C71, C73, C75, C89, C91, C93, C97, C98, C100\}, \{C5, C7, C10, C30, C35, C44, C45, C46, C51, C56, C59, C60, C70, C82, C84, C87, C96\}$ whereas each of the remaining clusters contain an individual customer.

- **Set $\lambda = 0.850$**
  
  There are 8 clusters: $\{C8, C13, C15, C16, C17, C18, C19, C21, C23, C24, C25, C29, C32, C43, C47, C60, C61, C68, C69, C74, C76, C78, C95\}, \{C8, C15, C16, C25, C29, C32, C68, C76, C95, C6, C9, C11, C26, C27, C28, C31, C33, C34, C36, C39, C41, C42, C48, C49, C50, C54, C58, C65, C67, C71, C73, C75, C89, C91, C93, C97, C98, C100\}, \{C5, C7, C10, C30, C35, C44, C45, C46, C51, C56, C59, C60, C70, C82, C84, C87, C96\}$ whereas rest of the clusters contain an individual customer each.

- **Set $\lambda = 0.830$**
  
  We obtain 5 clusters: $\{C8, C13, C15, C16, C17, C18, C19, C21, C23, C24, C25, C29, C32, C43, C47, C60, C61, C68, C69, C74, C76, C78, C95\}, \{C8, C15, C16, C25, C29, C32, C68, C76, C95, C6, C9, C11, C26, C27, C28, C31, C33, C34, C36, C39, C41, C42, C48, C49, C50, C54, C58, C65, C67, C71, C73, C75, C89, C91, C93, C97, C98, C100\}, \{C5, C7, C10, C30, C35, C44, C45, C46, C51, C56, C59, C60, C70, C82, C84, C87, C96\}$.
Furthermore, during proof-of-concept implementation, we find that there are a certain number of customers overlapping between Cluster 1 and Cluster 2. These are C8, C15, C16, C25, C29, C32, C68, C76 and C95. However, we solve this ambiguity by finding the similarity coefficients from the overlapped customers’ records. The similarity coefficients show that the overlapped customers tend more towards Cluster 1. Therefore, according to fuzzy clustering principle [48], we can classify these customers in Cluster1.

Consequently, the result of clustering is shown in Table 9 below and the clustering figure is displayed in Appendix D:

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster1</td>
<td>{C8, C13, C15, C16, C17, C18, C19, C21, C23, C24, C25, C29, C32, C43, C47, C60, C61, C68, C69, C74, C76, C78, C95}</td>
</tr>
<tr>
<td>Cluster2</td>
<td>{C6, C9, C11, C26, C27, C28, C31, C33, C34, C36, C39, C41, C42, C48, C49, C50, C54, C58, C65, C67, C71, C73, C75, C89, C91, C93, C97, C98, C100}</td>
</tr>
<tr>
<td>Cluster3</td>
<td>{C5, C7, C10, C30, C35, C44, C45, C46, C51, C56, C59, C60, C70, C82, C84, C87, C96}</td>
</tr>
<tr>
<td>Cluster4</td>
<td>{C1, C3, C4, C12, C14, C20, C22, C37, C38, C40, C52, C53, C55, C62, C64, C66, C72, C77, C79, C80, C83, C85, C86, C88, C90, C92, C94}</td>
</tr>
<tr>
<td>Cluster5</td>
<td>{C2, C57, C81, C99}</td>
</tr>
</tbody>
</table>

| Table 9: The result of clustering model |

### 6.4.5 Collecting the Conditions (for Optimization Model)

On the basis of clustering result and interviews, we need to collect the following parameters:
- The average similarity coefficients: $s_i$;
- The average value of customers’ monthly consumption in each cluster: $c_i$;
- The profit coefficient: $p$.

Through computation and extraction from the Transitive Closure $R^{32}$, Table 9 and the example set (in Appendix A), we obtained the required parameters as shown in Table 10 below.

<table>
<thead>
<tr>
<th>$s_i$</th>
<th>$c_i$</th>
<th>$p$</th>
<th>Number of clusters (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.922</td>
<td>1456.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.885</td>
<td>1820.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.862</td>
<td>1959.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.828</td>
<td>2299.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.801</td>
<td>2585.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Table 10: The parameters of optimization model |

Further, the variables are as follows:
- $x_i =$ price of membership card for $i^{th}$ cluster, $i = 1, 2, ..., n$.
- $d_i =$ how much discount (percentage) the customer can get in the $i^{th}$ cluster, $i = 1, 2, ..., n$.

### 6.4.6 Substituting the Conditions into Optimization Model

The optimization model we have designed is presented as:

$$\max \left\{ \sum_{i=1}^{n} (p \cdot x_i) - \sum_{i=1}^{n} (c_i \cdot d_i) \right\}$$
We first substitute the parameters (shown in Table 10) into the constraint functions in order to obtain the variables $x_i$ and $d_i$.

\[
x_i \leq c_i \times (1 + (1 - s_i))
\]

\[
c_i \times (1 - d_i) \geq c_i \times s_i (1 + \frac{1 - s_i}{3})
\]

\[
x_i > 0
\]

\[
0 < d_i < 1
\]

\[
i = 1, 2, ..., n
\]

Using the same method for obtaining $x_1$, we can figure out $x_2, x_3, x_4$ and $x_5$ as follows:

\[
x_2 \leq 2029.30
\]

\[
x_3 \leq 2230.41
\]

\[
x_4 \leq 2695.56
\]

\[
x_5 \leq 3100.22
\]

Further,

\[
c_1 \times (1 - d_1) \geq c_1 \times s_1 (1 + \frac{1 - s_1}{3})
\]

\[
(1 - d_1) \geq s_1 (1 + \frac{1 - s_1}{3})
\]

\[
(1 - d_1) \geq 0.922 \times (1 + \frac{1 - 0.922}{3})
\]

\[
\Rightarrow d_1 \leq 0.05
\]

Using the same method for obtaining $d_1$, we can figure out $d_2, d_3, d_4$ and $d_5$ as follows:

\[
d_2 \leq 0.08
\]

\[
d_3 \leq 0.10
\]

\[
d_4 \leq 0.12
\]

\[
d_5 \leq 0.15
\]

Additionally, we substitute $p, c_i, x_i$ and $d_i$ ($i = 1, 2, ..., n$) into the objective functions as follows:

\[
\max \left\{ \sum_{i=1}^{n} (p \times x_i) - \sum_{i=1}^{n} (c_i \times d_i) \right\}
\]

\[
(p \times x_1) - (c_1 \times d_1) \approx 476.62
\]

\[
(p \times x_2) - (c_2 \times d_2) \approx 564.60
\]

\[
(p \times x_3) - (c_3 \times d_3) \approx 584.65
\]

\[
(p \times x_4) - (c_4 \times d_4) \approx 667.45
\]

\[
(p \times x_5) - (c_5 \times d_5) \approx 697.23
\]

\[
\text{Total} \approx 2990.55
\]

where the results show how much profit the Wal-Mart can make from a customer’s monthly purchase in $i$-th cluster as the figure shows below ($¥$ means CNY, and 1 CNY $\approx$ 1 SEK).
6.4.7 Result Description
With the results of clustering model and optimization model, we can obtain the preliminary result of membership card generation shown in Table 11.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Card types</th>
<th>Price</th>
<th>Discount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>1569.81</td>
<td>5%</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>2029.30</td>
<td>8%</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>2230.41</td>
<td>10%</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>2695.56</td>
<td>12%</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>3100.22</td>
<td>15%</td>
</tr>
</tbody>
</table>

Our preliminary result produces five different clusters and hence five different types of cards. These are card types A, B, C, D and E corresponding to clusters 1, 2, 3, 4 and 5 respectively. For instance, Cluster 1 has card (type) A and the cost of the card is 1569.81 CNY. The discount given for this card is 5%. In other words, customers falling in Cluster 1 will get card (type) A which will cost them 1569.81 CNY with a discount of 5% on every purchase they make with the card.
CHAPTER 7: VALIDITY ASSESSMENT

Assessment of any research work carries two main objectives, which help to learn and summarize what has been researched [45]. In our project, we designed a two-step sequential method (clustering model + optimization model) in order to mitigate the problem of membership card generation in a hypermarket. First, in order to assess our clustering model, we compare fuzzy clustering and hierarchical clustering and evaluate that our model complies with the requirements. Furthermore, we work to verify that the solution is correct based on the formulated optimization model and to validate that the model describes the problem accurately enough. Finally, we send our models with preliminary result to Wal-Mart in Shanghai for evaluation of our two-step sequential method.

7.1 Clustering Model Assessment

As we mentioned in Chapter 2, the most frequently used clustering method is hierarchical clustering. However, with this method, the number of clusters cannot be predefined, because the cluster(s) which includes divisions (by divisive method) or fusions (by agglomerative method), once made, are irrevocable so that when two or more objects are joined or split, they cannot subsequently be separated or combined. This is one reason as to why the objects, at first, are reasonably segmented into different clusters, but with the increase in the number of customers’ data, the differences between these objects thus become more and more indistinct. The last but not the least, the similarity of data in hierarchical clustering cannot represent a certain index of correlation degree (customers’ consumption abilities), because the hierarchical method is based on initial data which cannot be transferred by any formulae. These two main factors limit the performance of hierarchical clustering method.

However, in our fuzzy clustering model the threshold value ($\lambda$) is flexible and adjustable accordingly. This is to say that we can define the number of clusters by threshold value ($\lambda$) according to the requirement (4-6 clusters) of the hypermarket. This can be achieved by adjusting and assigning different threshold values until the objects can be segmented into the required number of clusters. By doing so, the quality of clustering will not be affected even if there is increase in customer data. Furthermore, in our project, the similarity coefficients in fuzzy clustering illustrate customers’ consumption abilities, which are used as input to optimization model in order to generate membership cards.

The following table shows the functional comparison between hierarchical and fuzzy clustering in terms of our case. From the table, it is relevant that hierarchical clustering method is incapable of providing any input for the optimization model while fuzzy clustering complies with the requirement of our project and thus is preferred over hierarchical method.

<table>
<thead>
<tr>
<th>Functions</th>
<th>Hierarchical Clustering</th>
<th>Fuzzy Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predefine the number of clusters (as an input to optimization model)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Separate objects from a cluster</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Combine objects from different clusters</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Use coefficients for optimization model (as input to optimization model)</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 12: Comparison between Hierarchical clustering and Fuzzy clustering

7.2 Optimization Model Assessment

The solution of the optimization model must be evaluated by transferring back into its preceding phases as the red arrows in Figure 13 shows below.
In the first step of assessing our optimization model, we transfer the obtained solution back into optimization model, which includes the price of membership card $x_i$ and the discount for each membership card $d_i$ from different clusters. Accordingly, the objective of our optimization model is to figure out the prices of membership card and the relevant discount in each cluster with the constraints on the customers’ average monthly consumption ($c_i$) and the certain value of monthly purchase depending on the customer’s consumption ability ($c_i \times s_i$). Meanwhile, we have appropriately increased the customers’ average monthly consumption and their consumption ability coefficients $s_i$ in order to achieve the goal of objective function (which is to maximize the Wal-Mart’s benefit with the consideration of the cost in monthly consumption using the supposed membership card). Therefore, the solution is found to be correct based on the formulated optimization model.

Furthermore, we mapped back the solution into the simplified and real problems. From Wal-Mart’s view, customers’ consumption is the profit source, which means if customers purchase using the amount in the supposed membership cards, then the amount prepaid for membership cards can be seen as the profit earned by Wal-Mart ($p \times x_i p$ is profit coefficient described in Chapter 5.2). Hence, defining the price of membership card $x_i$ is one of considerable problems in our optimization model. For another, in order to attract more customers and popularize membership cards, the Wal-Mart provide different level of discount ($d_i$) for customers in different clusters, which implies the discount is the surrendered part of the profit. In other words, Wal-Mart will bear the cost for the provided discount in customers’ consumption ($c_i \times d_i; c_i$ is customers’ average monthly consumption illustrated in Chapter 5.2). In this way, computing the discount ($d_i$) for customers in different clusters is another considerable problem. Consequently, the solution in optimization model can describe the problem accurately enough.

7.3 Assessment for combination of two models

After assessment of these two models individually, we would like to validate how the combination of clustering and optimization models would suffice to our research in a very good way based on research from other domains. To cite as an example, Bing et al. [34]
proposed a weighted fuzzy clustering algorithm to optimize the multicast routing in the overlay network (detailed in Chapter 1.4). Their research was based on routing optimization for better clustering performance, which was a multi-objective optimization with Non-Linear Programming characteristics. The problem of major cost, resources usage and multicast session were effectively optimized. Another example by Bifulco et al. in [37] presented methodology based on a process that generated multiple clustering solutions using global optimization. Further, Wong et al. [36] showed image clustering using particle swarm optimization techniques.

The above research showed that an integrated approach to clustering and optimization model was found to have a good effect in solving problems in other domains. Therefore, we believed that it could also in a way contribute to our research domain and help to solve the existing problem faced by Wal-Mart (in Shanghai).

### 7.4 Assessment by Wal-Mart

Following the proof-of-concept implementation of our method, the models with preliminary result were sent to Wal-Mart (in Shanghai). In addition, we made a video interview with Mr. Yizhu Shi for validating our method and models (clustering and optimization models). In the meantime, we presented the details of design principles and modeling procedures. On subsequent assessment by the sales and marketing departments from the Wal-Mart, our current solution could substantially improve the issues pertaining to customer segmentation and membership card generation. However, if this solution would be formally come into service, we need to complete some future works (i.e., structure the supporting database, application interface and so on).

### 7.5 Validity Threats

Our research identifies a number of validity threats. It comprises threats related to the conduct of literature review and threats related to conducting interviews. As already mentioned in the research methodology section that all the papers used for our research were based on five identified databases only and though chances of missing some very important papers from other databases, which we ignored were slim but this threat cannot be ruled out. Another important and obvious threat to our research is that our may not be a true representation of a real world. In order to mitigate this threat, we made a video interview with Mr. Yizhu Shi for validating our method and models (clustering and optimization models). Besides, we presented the details of design principles and modeling procedures of our models. Following assessment by the sales and marketing departments from the Wal-Mart (in Shanghai), our current solution could substantially improve the issues pertaining to customer segmentation and membership card generation. Further, we also validated our models by mapping back the solution into the optimization model, simplified and real problem to verify its correctness based on the formulated optimization model and to validate that the model described the problem.

In addition, we also considered threat of reliability of the interview results. Some questions could provide space for processing error due to the ambiguity of the question design structure. However, in order to mitigate this threat, we designed and asked open-ended questions that provided the interviewee to respond more objectively according to his understandings of business requirements and prevailing situation of the subject concerned rather than imposing questions that gives space for random choice or guessing response from the options provided. Because our interviews were based on only one person, threats of missing some important aspects of business requirements, complete customer information, and current problems with the hypermarket cannot be ignored. However, to alleviate this threat, we chose our interviewee as the top management personnel (Manager of Wal-Mart) who had all the requisite knowledge and information at his disposal, and was well aware of the current situation of the hypermarket.
Further, our selected methods and models might not be applicable to every business set-ups and projects, which is another threat we anticipated. However, in order to overcome this threat, we ensured that the strategy and method of our clustering model could potentially be applied to other projects except for the optimization model because optimization model may differ while considering its objective and constraint functions depending upon the situation. Given this effect, we cannot make claim of the results to be the true representation of every business organizational practices. Because of the significance of our research, we tried to prevail over all these factors to make our study more convincing.
CHAPTER 8: DISCUSSION AND CONCLUSION

8.1 Discussion
Following literature review, we were convinced that there existed a gap in customer segmentation and membership card generation in hypermarkets. It was evident from the interview that there was a problem in the existing policy of membership card generation in Wal-Mart (Shanghai). This would result in the phenomenon of customer attrition (i.e., loss of customer). Through literature review, we found that integrating clustering method with optimization model was found to have a very good effect in other areas of research; we posit that it would also deliver better results in our research domain. Therefore, we set our goal accordingly. Our goal was to develop a methodological approach for customer segmentation based on customers’ characteristics and to subsequently define membership cards attributed to mathematical optimization model in a hypermarket. This led us to formulate our research questions. To answer our research questions, we proposed and designed a two-step sequential method; first customer information (here customers’ consumption) were analyzed, and were segmented into different clusters. Following customer segmentation, the clustering results were used as inputs to optimization model in order to generate membership cards. Fuzzy clustering method was used to segment customers into different cluster. Our models with preliminary results were validated individually and our solution sent to Wal-Mart for evaluation with feedback that our current solution could substantially improve the issues pertaining to customer segmentation and membership card generation.

Every research work has its limitations and ours is not an exception to it. During the course of our research work, we encountered with many obstacles that limit our research work. First, our literature review was conducted using papers from five identified databases only and though the chance missing other important papers from other databases is very remote, we simply ignored and carried with the available resources at our disposal. Second, because our interviews were conducted in Chinese, some communication gap, misinterpreting and misunderstanding about some concepts between us and the interviewee (Mr. Shi) could have existed and hence research bias, however small, cannot be ruled out. Besides lack of a very good English vocabulary in us could be another potential threat as we might missed out proper translation to English. In addition, some questions could provide space for processing error due to the ambiguity of the question design structure. Third, because our interviews were based on only one person, the interviewee’s responses might not be the true representation of the actual problem.

Further, there could be alternative approaches to our method for clustering customer and generating membership cards. Besides, there is room for improvement in our method such as programming, testing solutions by means of other formulae, and implementation to test large number of customer data records.

Finally, lack of validity from an expert (third party) is another limitation of our research. However, validating our solutions from a third party expert will be our future task.

8.2 Conclusion
The primary goal of this research work was to develop a methodological approach for customers segmentation based on their characteristics and to define membership cards attributed to mathematical optimization model in a hypermarket. We achieved our goal by using a two-step sequential approach: first, we build a clustering model for customer segmentation, and then created an optimization model for membership card generation by taking input from the clustering result.
In our research project, we found fuzzy clustering as an appropriate method for customer segmentation. Fuzzy clustering method made use of Deviation Transformation and Range Transformation formulae to compress data (customers’ monthly consumption) dimensions into a range between 0 and 1 such that these data could be compared with one another thereby eliminating incomplete and redundant data. Further, Correlation Coefficient Method was used for building fuzzy similarity matrix, where the similarity coefficients represented the customers’ consumption abilities (which would be the input for our optimization model). In addition, we could predefine the number of clusters by setting threshold value ($\lambda$) according to the hypermarket’s requirement (4-6 clusters). This answered our RQ1.

In order to answer RQ2, it was learnt from the interview analysis that the existing management finds burdensome to consider wholly the vast amount of customer information. This had caused deviation in the existing method used for generating membership cards from the demands of the growing customers’ needs. Therefore, optimization model was developed based on the clustering result, which not only provided certain economic benefits for the management but also catered to the customers’ consumption abilities. To achieve our goal, we collected the needed conditions (parameters) and supposed the required outcomes (variables) for structuring the objective function. Further, the constraint functions corresponded to customers’ monthly consumption. Thus, with the formulation of optimization model and the input from the clustering result, we were able to generate suitable membership cards. Eventually, optimization model was executed to get the result for RQ2.

To sum up, because the clustering is an unsupervised learning process, we need not test all the data but implement the model with a certain number of data [15]. Therefore, we extracted 100 records as an example set for our implementation. Thus, the preliminary outcome from the clustering and optimization models produced five clusters and five types of membership cards. Finally, we performed the validity assessment for clustering and optimization models respectively. Besides, validating these two models, our preliminary result was sent to Wal-Mart for evaluation of our methodological approach. The feedback we received was that our methods and solution was well received by Wal-Mart.
**FUTURE WORK**

Our research opens up many new scopes for future work. Concerning future research, working with the objects (customers) in the overlapped clusters will be one of our future work. First, we would extract the customers from the overlapped clusters; do re-analyze their similarity coefficients in order to generate appropriate cards for such special customer groups. In other words, there is feasibility and possibility to generate a new membership card type for the customers that fall under overlapped clusters. Currently we could only assign the customers from the overlapped clusters to the nearest cluster based on the tendency of their similarity coefficients, which may not be the best-case scenario.

In future, it would be possible to test our solution for more generalization of results and issues pertaining to customer segmentation and membership card generation. Additionally, it could also be possible to address with multiple customer attributes such as monthly salary, gender, within this domain.

In addition, because we lacked a suitable supporting database, much time was spent on processing raw data. Therefore, designing and structuring an entire database that could help input data conveniently for our models would be our next further work. Moreover, if our clustering and optimization models would formally come into service, then designing the supporting application (software) interface would remain another future work.

Besides, our method can also preferably be used in other research domains (in the future research) such as image segmentation. Our sequential method of clustering and optimization method can be further extended into image segmentation. First, clustering method can be used to segment one or more pictures by gray levels, textures and chromatic aberration while a mathematical method based optimization model can be employed subsequently for rebuilding an image model depending on the above properties, thereby establishing 3D model and/or keeping the picture(s) clearer and more distinct.

Another important area of research where our method can be implemented is in geological and plant classification. In other words, plants can be classified according to the geography of a land. This can be done by implementing our method to segment land into different areas based on the type of soil, climatic conditions, land textures, etc., and subsequently segmenting plants by their categories such as phyla, classes, orders, family, genus, and species; finally, build an optimization model for computing the matching degree between land and plants.
REFERENCES


## APPENDIX A: EXAMPLE SET FOR TESTING

<table>
<thead>
<tr>
<th>Date</th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
<th>COL5</th>
<th>COL6</th>
<th>COL7</th>
<th>COL8</th>
<th>COL9</th>
<th>COL10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX B: SOURCE CODE
/*The header files*/
// stdafx.h : standard system files.

#pragma once
#include "targetver.h"
#include <stdio.h>
#include <tchar.h>
#include <iostream>
#include <math.h>
#include <conio.h>

// TODO: Refer to the needed header files.

/*The main program*/
// Experiment.cpp : Define the enter point of console application.

#include "stdafx.h"
using namespace std;

/*Define the spaces of matrices and arrays*/
double sdt[12][100]; //define the space of result matrix (Standard Deviation Transformation) depending on the initial matrix.
double rt[12][100]; //define the space of result matrix (Range Transformation) according to the initial matrix.
double x[100], x1[100]; //define space for the average values x[] and x1[] of columns,
double s[100]; //s[] is the denominator of the formula(Standard Deviation Transformation).
double r[100][100]; //define the space of fuzzy similarity matrix depending on the columns of initial matrix.
double tc[100][100]; //define the space of transitive closure depending on the fuzzy similarity matrix.
double temp_tc[100][100]; //define the space of temp tc for clustering.
double tc_min[100], tc_max; //define an array and a variable for getting the minimal and maximal values during obtaining transitive closure.

/*Define the initial matrix*/
double init_matrix[12][100] =
{908.6, 4074.94, 1206.28, 984.27, 3269.29, 3516.93, 1297.78, 1150.6, 1318.27, 2713.6, 1714.92, 2160.45, 992.98, 4199.95, 1017.78, 1008.93, 826.82, 1095.27, 1132.93, 3169.27, 2289.32, 922.91, 490.8, 988.81, 1168.97, 1180.29, 3269.35, 1563.25, 1100.93, 4416.93, 872.93, 1015.93, 3269.27, 27, 1350.93, 2997.82, 2880.27, 1299.1, 1822.27, 1247.27, 3398.6, 4200.27, 4037.93, 907.27, 760.6, 1003.49, 1042.9, 2541.93, 2315.6, 927.93, 2158.27, 3205.3, 3100.38, 1213.67, 2500.58, 430.5, 157, 3524.33, 753.7, 1158.34, 1325.67, 2721.27, 1722, 1863.7, 1000.33, 2707.48, 1025, 2016.3, 3, 12783.3, 1102.86, 1140.8, 3176.51, 2296.82, 787.72, 2376.78, 3570.77, 1108.28, 1024.5, 3880, 3276.78, 3570.77, 1108.28, 1024.5, 3880, 1023.28, 676.77, 958.36, 505.86, 3887.66, 4306.5, 1829.6, 1254.52, 1406, 507.33, 3045.54, 214.42, 768, 1010, 1650.65, 549.42, 323, 235.72, 1165.6, 1489.3, 3689.5, 1213, 991, 2490, 2523.9, 9, 985.8, 2893, 1789, 2720.4, 1721.75, 1477.2, 1097, 3206.7, 2024.9, 1015.77, 1833.68, 1102, 2139.77, 3176, 2296, 2329.7, 1497.91, 295.68, 675.73, 3187.3}
3213.5, 3996.32, 3489.8, 1578.8, 3190.67, 3905.76, 420.78, 1072.2, 2390, 3765, 1636, 2081
. 76, 298.78, 3121.42, 939.930, 4, 748.33, 1016.9, 1054, 2090.67, 2210.7, 844.33, 412.32, 1910.3, 2090.24
01.66, 4190.62, 884.6, 2022.3, 4338.74, 794.537, 74, 4190.6, 1872.78, 3419.2, 1801.8, 1820.58, 2743.72,
1168.6, 4320, 3121.72, 4959.8, 1828.67, 3682, 3234.9, 964.97, 3463.36, 1237, 849.86, 2079.92, 1795,
1426.67, 952, 2260, 2559.7, 592.38, 1835,
1056.42, 1521, 374, 437.62, 960.9,
2845.18, 870, 546.9, 1729.65, 1861.11,
3224.18, 2539.67, 988.9, 1799.28, 2867,
1748.58, 1744, 596, 2252.58, 970,
351.59, 2841, 3809.94, 1927, 623, 2067,
601.33, 2307.72, 1135, 2470.4, 3924.3,
3226.69, 2517.58, 2580, 2267.48, 2756.72, 2667,
1799.33, 2649.99, 2834, 3552, 2060.77, 3156,
2368.1, 3590.76, 654.6, 971.87, 1800.7,
2128.87, 1285.34, 387, 3792.4, 2701.27,
1702.53, 3371.7, 980.51, 1787.53, 1005.53,
319.86, 814.5, 1082.44, 1120.53, 3156.9,
2276.83, 910.57, 878.53, 976.48, 1156.53,
1167.87, 1556.8, 1550.81, 1088.56, 2904.53,
860.53, 3003.53, 3256.87, 553.53, 985.42,
1867.92, 1286.7, 1809.87, 1234.9, 1386.2,
4187.87, 2025.53, 894.94, 748.24, 991,
1030.58, 1529.53, 3303.2, 2015.5, 4145.8, 465.78, 1509,
394, 973.92, 3258.63, 563.6, 3077.45,
3490.8, 876.98, 1890, 1704.33, 1084.7,
3256.3, 2618, 3007.42, 2500.75, 2316.41,
1084.8, 1122.65, 1658.92, 2278.67, 912.39,
2080.29, 978.34, 1158.3, 269.7, 2258.67,
2552.62, 4090.44, 2706.6, 862.4, 3705.45,
3258.84, 3940.28, 1987, 1869.67, 2888.72,
1811.54, 1236.67, 3388, 1189.6, 3027.6,
896.76, 750, 992.99, 1032, 331.4, 1305,
917, 2147.67,
790.67, 395.9, 2790.57, 525.8, 1289.7, 1806,
1899.8, 1051.5, 1219, 2614.5, 1615.87, 2061,
1893.88, 3100.99, 918.8, 909.89, 727.83, 996,
1033.9, 3070, 2190.18, 3453.76, 1391.88,
889.76, 369.88, 1081, 3170, 1464,
1001.92, 4317.86, 2673.86, 916.8, 3170,
851.8, 1528.66, 3781, 3200, 3223, 2548,
2699.56, 2401, 4938.98, 808, 2661.5,
1304.38, 2443.83, 1442.89, 1216.39, 728.83,
2059, 908.6, 4074.94, 1206.28, 984.27, 3269.29, 3516.93, 1297.78, 1150.6, 1318.27, 271,
3.6, 1714.92, 2160.45, 992.98, 4199.95, 1017.78, 1008.93, 826.82, 1095.27, 1132.93, 3169.27, 2289.32, 9,
93, 2160.45, 992.98, 4199.95, 1017.78, 1008.93, 826.82, 1095.27, 1132.93, 3169.27, 2289.32, 9,
834.33, 1102.86, 1140.8, 3176.51, 2296.82,
930.33, 498.52, 996.51, 776.33, 787.72,
3276.78, 3570.77, 1108.28, 1024.5, 3880,
1023.28, 676.77, 958.36, 505.86, 3887.66,
4306.5, 1829.6, 1254.52, 1406, 507.33,
3045.54, 214.42, 768, 1010, 1650.65,
549.42, 323, 235.72, 1165.6, 2368.1,
654.6, 971.87, 1800.7, 2128.87, 387,
3792.4, 2701.27, 1702.53, 3371.7, 980.51,
1787.53, 1005.53, 319.86, 814.5, 1082.44,
1120.53, 3156.9, 2276.83, 1550.81, 3303.2,
976.48, 1156.53, 1167.87, 3003.53, 3256.87,
1088.56, 2904.53, 2701.27, 1285.34, 1550.81,
538.53, 985.42, 1867.92, 1285.34, 1809.87,
1234.9, 1386.2, 4187.87, 2025.53, 894.94,
748.24, 991, 1030.58, 3256.87, 3003.53,
1120.53, 3156.9, 2276.83, 1550.81, 3303.2,
1448.09, 1971.18, 1396.89, 2949.54, 2987,
2015.5, 4145.8,

560.6, 987, 1355.72, 1133.66, 2807, 1821,
2066, 1300, 2491.6, 1609.43, 1864.32,
2309.99, 1142.28, 3349.4, 1158.33, 1092.44,
3276.38, 3244.69, 3682.33, 4118.67, 2438.39,
1072.45, 640.6, 3138.4, 910.57, 878.53,
918.6, 1712.61, 1250.81, 1550.81, 3303.2,
1088.56, 2904.53, 860.53, 3003.53, 3256.87,
538.53, 985.42, 1867.92, 1285.34, 1809.87,
1234.9, 1386.2, 4187.87, 2025.53, 894.94,
748.24, 991, 1030.58, 3256.87, 3003.53,
1120.53, 3156.9, 2276.83, 1550.81, 3303.2,
1448.09, 1971.18, 1396.89, 2949.54, 2987,
2015.5, 4145.8,
int n=12,m=100;//copy the size of init_matrix in order to calculate the further loops directly.

/*Note: sdt = Standard Deviation Transformation; rt = Range Transformation; cc = Correlation Coefficient; nume = numerator; deno = denominator; tc = Transitive Closure*/

int _tmain(int argc, _TCHAR* argv[])
{
    /*Define the variables*/
    int i,j,k,i1,j1;
    double sdt_nume=0,sdt_deno=0;//nume is numerator, deno is denominator;
    double min,max;//define the minimal and maximal values for Range Transformation.
    double cc_nume=0,cc_deno1=0,cc_deno2=0,cc_deno=0;//cc_deno1 and cc_deno2 are the parts of denominator in the cc formula.
    int R_num=2;//after getting transitive closure,first loop is R^2, so define the initial value is 2.

    /*Calculate the average value of every column in init_matrix for sdt formula*/
    for(i=0;i<m;i++)
    {
        for(k=0;k<n;k++)
        {
            x1[i]+=init_matrix[k][i];
        }
        x1[i]=x1[i]/n;
    }

    /*Calculate the value of the denominator of sdt formula.*/
    for(i=0;i<m;i++)
    {
        for(k=0;k<n;k++)
        {
            s[i]+=(init_matrix[k][i]-x1[i])*(init_matrix[k][i]-x1[i]);
        }
        s[i]=sqrt(s[i]/n);
    }
}
/*Calculate the elements one by one according to the sdt formula.*/
for(i=0;i<m;i++)
{
    for(k=0;k<n;k++)
    {
        sdt_nume=fabs((init_matrix[k][i]-x1[i]));
        sdt_deno=s[i];
        sdt[k][i]=sdt_nume/sdt_deno;
    }
    sdt_nume=sdt_deno=0;
}

/*Compute the minimal and maximal values of sdt matrix for Range Transformation*/
min=max=sdt[0][0];//initially, assume the first element is min and max separately.
for(i=0;i<m;i++)
{
    for(k=0;k<n;k++)
    {
        if(min>sdt[k][i])
            min=sdt[k][i];
    }
}//find the minimal element from "sdt" matrix.
for(i=0;i<m;i++)
{
    for(k=0;k<n;k++)
    {
        if(max<sdt[k][i])
            max=sdt[k][i];
    }
}//find the maximal element from "sdt" matrix.

/*Calculate "rt" matrix based on "sdt" matrix.*/
for(i=0;i<m;i++)
{
    for(k=0;k<n;k++)
    {
        rt[k][i]=(sdt[k][i]-min)/(max-min);
    }
}

/*Print the Standard Deviation Transformation matrix and Range Transformation matrix*/
cout<<"The Standard Deviation Transformation matrix is:\n"<<endl;
for(i=0;i<n;i++)
{
    for(k=0;k<m;k++)
    {
        printf("%6.3f",sdt[i][k]);
    }
    printf("\n");
cout<<endl<<"The Range Transformation matrix is:\n"<<endl;
for(i=0;i<n;i++)
{
    for(j=0;j<m;j++)
    {
        printf("%.3f",rt[i][j]);
    }
    printf("\n");
    //getchar();
}

/*Calculate the average value of every colomn of rt matrix*/
for(i=0;i<m;i++)
{
    for(k=0;k<n;k++)
    {
        x[i]+=rt[k][i];
    }
    x[i]=x[i]/n;
}

/*Calculate the elements one by one according to the Correlation coefficient formula.*/
for(i=0;i<m;i++)
{
    for(j=0;j<m;j++)
    {
        for(k=0;k<n;k++)
        {
            cc_nume+=fabs((rt[k][i]-x[i])*(rt[k][j]-x[j]));
            cc_deno1+=(rt[k][i]-x[i])*(rt[k][i]-x[i]);
            cc_deno2+=(rt[k][j]-x[j])*(rt[k][j]-x[j]);
            cc_deno=sqrt(cc_deno1*cc_deno2);
            r[i][j]=cc_nume/cc_deno;
        }
        //However, the size of result matrix r[i][j] should become [m][m].
        //This is because the required matrix has to achieve Reflexivity and Symmetry.
        cc_nume=cc_deno1=cc_deno2=cc_deno=0;
    }
}

/*Print the fuzzy similarity matrix R. (because it is produced by correlation coefficient formula, we can also name it as correlation coefficient matrix)*/
cout<<endl<<"The Fuzzy Similarity matrix R is:\n"<<endl;
for(i=0;i<m;i++)
{
    for(j=0;j<m;j++)
    {
        printf("%.3f",r[i][j]);
    }
    printf("\n");
}
/*Obtain the Transitive Closure*/
A: for(i=0;i<m;i++)
    { 
        for(j=0;j<m;j++) 
            { 
                for(k=0;k<m;k++)
                    { 
                        if(r[i][k]<=r[i][j])//find the minimal values from ith rows and kth columns.
                            tc_min[k]=r[i][k];
                        else
                            tc_min[k]=r[k][j];
                    } 
                tc_max=tc_min[0];//assume the first value in tc_min[k] is maximal.
                for(k=0;k<m;k++)
                    { 
                        if(tc_max<tc_min[k])
                            tc_max=tc_min[k];
                    } //find the maximal value from the tc_min[k].
                tc[i][j]=tc_max;
                //the elements in tc[i][j] are the maximal values from the seeked minimal values in ith rows and kth columns.
            }
    }

/*Print R^* (R^2,R^4,...,R^k,R^L) until R^k = R^L*/
cout<<endl;
cout<<"R^"<<R_num<<"is:
<endl;
for(i=0;i<m;i++)
    { 
        for(j=0;j<m;j++)
            { 
                printf("%6.3f",tc[i][j]);
            }
        cout<<endl;
    }

/*Match  R^*s with each other, if there is any difference, compute the next R^*.*/
for(i=0;i<m;i++)
    { 
        for(j=0;j<m;j++)
            { 
                if(r[i][j]!=tc[i][j])
                    { 
                        for(i1=0;i1<m;i1++)
                            { 
                                for(j1=0;j1<m;j1++)
                                    { 
                                        r[i1][j1]=tc[i1][j1];
                                    }
                        R_num*=2;//count the R^*
                        goto A;
                    }
            }
    }
}
/*Define the threshold value and start clustering*/
int targ_x,targ_y,f;
for(int z=0;;z++)
{
    cout<<"Enter 10 for quit, others for the Threshold Value:"<<endl;
    cin>>f;
    if(f==10) return(0);
    cout<<"Set the Threshold Value(enter the coordinate i.e., x y):"<<endl;
    cin>>targ_x>>targ_y;
    for(i=0;i<m;i++)
    {
        for(j=0;j<m;j++)
        {
            if(tc[i][j]>=tc[targ_x][targ_y])
                temp_tc[i][j]=1;
            else
                temp_tc[i][j]=0;
        }
    }
    //cout<<endl;
    for(i=0;i<m;i++)
    {
        for(j=0;j<m;j++)
        {
            printf("%2.0f",temp_tc[i][j]);
        }
        cout<<endl;
        //getchar();
    }
    getchar();
}
APPENDIX C: LITERATURE REVIEW

Search Strings:

<table>
<thead>
<tr>
<th>Strings No.</th>
<th>Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Clustering OR Cluster</td>
</tr>
<tr>
<td>2</td>
<td>Optimization OR Optimize OR Optimisation OR Optimise</td>
</tr>
<tr>
<td>3</td>
<td>Customer OR Consumer OR Membership</td>
</tr>
<tr>
<td>4</td>
<td>Hypermarket OR Supermarket OR Hyper-market OR Super-market</td>
</tr>
<tr>
<td>5</td>
<td>Data mining</td>
</tr>
</tbody>
</table>

Search Results:

<table>
<thead>
<tr>
<th>Steps</th>
<th>Total number of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iteration 1 (After using search strings)</td>
<td>539</td>
</tr>
<tr>
<td>Iteration 2 (Selection by titles)</td>
<td>122</td>
</tr>
<tr>
<td>Iteration 3 (Selection by abstract and conclusion)</td>
<td>43</td>
</tr>
</tbody>
</table>

Searched databases:

1. IEEE = 53
2. SpringerLink = 99
3. Google Scholar = 147
4. Science Direct = 64
5. Engineering Village (Compendex) = 95
6. Engineering Village (Inspec) = 81
APPENDIX D: CLUSTERING FIGURE

\[ \lambda = 0.971 \]

\[ \lambda = 0.950 \]

\[ \lambda = 0.920 \]
APPENDIX E: VARIABLES FOR PROGRAM

The programs used for clustering contain a number of variables that need to be defined as follows:

- **Standard Deviation Transformation**
  - `double sdt[n][m]`: define the space of result matrix of Standard Deviation Transformation depending on the initial matrix.
  - `double sdt_nume` and `sdt_deno`: define the numerator and denominator of Standard Deviation Transformation formula (Formula 2).
  - `double s[m]`: define an array s for calculating the denominator of the Standard Deviation Transformation formula (Formula 2).

- **Range Transformation**
  - `double rt[n][m]`: define the space of result matrix of Range Transformation according to the initial matrix.
  - `double min` and `max`: define the minimal and maximal values for Range Transformation.

- **Correlation Coefficient method**
  - `double r[m][m]`: define the space of result matrix of Correlation Coefficient method (fuzzy similarity matrix) depending on the columns of initial matrix.
  - `double x[m]`: define the average values of every column of rt matrix.
  - `double cc_nume`, `cc_deno1`, `cc_deno2` and `cc_deno`: define the numerator and denominators of Correlation Coefficient method (Formula 4)

- **Transitive Closure method**
  - `double tc[m][m]`: define the space of transitive closure depending on the fuzzy similarity matrix.
  - `double temp_tc[m][m]`: define the space of temp tc for clustering.
  - `tc_min[m]` and `tc_max`: define an array and a variable for getting the minimal and maximal values during obtaining transitive closure.
  - `int R_num`: define a flag for the loops of transitive closure, where the first loop is $R^2$, so define 2 as the initial value of it.

- **Observation Method**
  - `int targ_x` and `targ_y`: define the coordinate of the target element which is set as threshold value in tc matrix.
  - `int f`: define the control during obtaining transitive closure.

- **Others**
  - `double init_matrix[n][m]`: define the space of initial matrix
  - `double x1[m]`: define the average value of every column in init_matrix
  - `int n` and `m`: define the rows and columns of initial matrix