# A Data Mining Framework for Target Marketing

Daniela Stan Raicu
School of Computer Science,
Telecommunication and Information Systems (CTI)
DePaul University
243 South Wabash Avenue
Chicago, Illinois 60604-2302
dstan@cs.depaul.edu

**Abstract**

In this paper we propose a theoretical data mining framework for automatic gathering of consumer data for companies interested in developing product prototypes based on customer needs and preferences; we also provide automatic methods for discovering the relationships between customers' preferences and the product's physical characteristics. The current framework is presented for the automotive industry, but it can be successfully implemented for any other applications where the customers' preferences are taken into account for product design.

**Keywords** : Data mining, market segmentation

## 1. Introduction

In today's society of high demanding customers and increasing competitiveness, it has become essential to apply customer-focused approach to product development. In order to avoid the initial investment of producing a product without being certain that it will be successful, a customer-focused approach will first discover what are the customers' preferences, needs, and behavioral characteristics with respect to the product's concept and only then, the concept may be developed into a real product. This approach improves the chances of a particular product in being successful and also provides valuable insight into understanding the emotional and rational benefits derived from a particular product.

Using data mining techniques, we propose an approach that will provide the decision makers with a-priory knowledge about customers' preferences and needs. Since there are many different kinds of customers with different kinds of needs and preferences, the proposed approach is meant to be a tool for performing *market segmentation:* divide the total market, choose the best segments, and design strategies for profitability serving the chosen segments better than the company's competitors do. In order to produce superior value and customer satisfaction, the company needs information at every level of a product's life cycle from the initial concept to the final product. Fortunately, increasing information requirements have been met by an explosion of information technology; using improved information systems, data can be gathered in great quantities. In fact, today's managers sometimes receive too much data, in which case the real problem is how to separate the useful and relevant nuggets of information from the mountains of data. As an analyst points out: "Running out of information is not a problem, but drowning in it is" [1]. Often, important information arrives too late to be useful, or on-time information is not accurate enough. The proposed system will have incorporated a *method to collect data* that will be relevant, accurate, current, and unbiased. Since human beings are very good at perceiving structure in visual forms, our system will create, through a *visualization approach***,** an immersive virtual reality environment to display the data in various forms; as a consequence, a decision maker will be able to move through this environment discovering hidden relationships between product characteristics and customer needs and preferences.

The current approach is described for product development in auto industry, but it can be successfully implemented for any other applications where it is necessary to find the correlations between the customer feelings or perceptions and the physical characteristics of a product.

## 2. The Data Mining Framework

### 2.1. Domain Understanding

Any data mining application should start by understanding the business goals of the application since the blind application of data mining techniques without the requisite domain knowledge often leads to the discovery of irrelevant or meaningless patterns.

In order to understand the target customers of an automotive company, it would be helpful to examine the relationships between the *vehicle image/attributes* and the *customer emotional benefits* that are tied to psychological needs, personality traits, and personal values. The proposed approach will enable us to understand more completely how product specific characteristics relate to customer needs and the benefits a customer hopes to obtain from them. For many people, cars, homes,

restaurants, and vacations provide emotional benefits as well as rational benefits. However, for a wealthy person who has everything, the emotional benefits provided by status, prestige and superiority of an expensive automobile ("we are what we have") could outweigh rational benefits such as gas economy, lower maintenance and insurance costs, and resale value.

Therefore, it will be beneficial to have a tool that will help us to respond to questions such as: *What and how many of the personality attributes used to describe the customer might be shaped by the vehicle's image? What kind of vehicle this customer or group of customers will buy?*

Table 1 gives few vehicle attributes and Table 2 gives few customer anticipated feelings that are used to describe the proposed approach; however, many other attributes can be added by business and information technology people, and psychologists involved in this stage of the data mining process.

Table 1: Vehicle image/attributes

| | | |
|---|---|---|
| Adventurous | Aggressive | Family-oriented |
| Fast | Luxurious | Modern |
| Muscular | Powerful | Sophisticated |

Table 2: Customer Feelings

| | | |
|---|---|---|
| Accomplished | Anonymous | Confident |
| Cutting-edge | Free to do anything | In-control |
| Sporty | Tough | |

Fig.1 illustrates how the mapping of the vehicle attributes to the customer feelings may be used to communicate the emotional benefits of a certain style of vehicle. For example, if the automotive company will advertise a sport car design, the commercial can say *"if you want to feel sporty, free to do anything, in control and confident, this car will be the best vehicle for you"*.

**Target Customer Emotional Benefits**

**Personality Profile:**

- Sporty
- Free to do anything
- In control
- Confident

Figure 1: Sport car advertisement by emotional benefits

Therefore, the final goal of our approach when applied to automotive industry is to automatically find the associations/mappings between customer feelings and vehicle attributes.

**2.2. Data Selection**

The second step of our approach calls for targeting a database or selecting a subset of fields to be used for the data mining. The following issues should be considered in developing a plan for collecting data efficiently: evaluation of existing data sources, specification of research approaches, and data gathering (contact methods, sampling plans, and instruments).

We propose the *survey research as a method to collect data*. One of the advantages of the survey research is flexibility because it can be used to obtain many different kinds of information in many different situations. Furthermore, depending on the survey design, it may also provide information quicker at a lower cost compared to manual processing. The survey will be in the form of a questionnaire that is very flexible as there are many ways to ask questions. In preparing the questionnaire, only the questions contributing to the research objectives will be asked. The questions are *closed-ended*, as they include all possible answers. In designing the survey, we also make sure that the questions are simple, direct and arranged in a logical order. The first question should create interest if possible, and difficult or personal questions should be asked last so that respondents do not become defensive; the demographics (age, income, education etc) are the last set of questions.

Instead of a mail questionnaire, we propose the *computer interviewing* in which respondents sit down at a computer, read questions from a screen, and type their own answers into the computer at their own leisure. The beauty of our approach consists of its multiple benefits.

As a first benefit, the respondents' answers are automatically stored in a database. Second, the survey is posted on the web and it can be accessible by an unlimited number of people. Filling out the survey becomes a non-time consuming task even for a busy person: the survey is on the web and it is accessible for anybody at any time; the submission of the completed survey requires only a 'click on' action executed by respondent, action possible through an interactive survey implementation. Third, the computers might be located at different locations such as auto shows, dealerships, or retail locations. The biggest benefit is the collection of more relevant data since people present at those locations are most likely willing to answer correctly to the questions because they are interested in automobiles. The approach can be implemented such that the data is gathered from numerous computers at different locations and stored in a unique and global database. As a fourth benefit, same survey format will be accessible to different categories of people: expert people (such as car designers) or people less familiar with auto domain characteristics. The large number of respondents and their diversity give more reliability on the results than small samples.

### 2.3. Data Cleaning and Preprocessing

This stage is the most time-consuming stage of the entire data mining process. Data is never clean and in a form suitable for data mining. There are few typical data corruption problems in business databases such as duplication of the records, missing data fields, and presence of outliers.

The preprocessing step involves integrating data from different sources and making choices about representing or coding certain data fields that serve as inputs to the data discovery stage. Such representation choices are needed because certain fields may contain data at level of details not considered suitable for the pattern discovery stage. For example, it may be counter-productive to represent the actual date birth of each customer to the data discovery stage. Instead, it may be better to group customers into different age groups and the chosen age groups should have some significations for the research goal.

It is important to remember that the preprocessing stage is a crucial step. The representation choices made at this stage have a great bearing on the kinds of the patterns that will be discovered by the next stage of data discovery.

### 2.4. Discovering Patterns: Market Segmentation

Because there are so many ways we are each different, it should not be surprising that we would differ in our needs for automobiles. While there are many factors/variables that contribute to these differences, we are considering the following factors for presenting our data mining framework: vehicle image (Table 1), customer anticipated feelings (Table 2), and demographics (such as age, sex, income, occupation, education etc). The demographic factor plays an important role in the proposed analysis. For example, consider how customer needs and preferences for an automobile change as one moves demographically from college student to management trainee; changes in income, occupation, and educational status each contribute to a changing set of customer needs for a variety of products such as an automobile.

Many other variables can be incorporated. For example, it will be interesting to add a geographic variable (world region, country region, city or metro size, etc) to our data to allow geographic segmentation; this will offer the company the possibility of localizing their products, advertising, promotion, and sales efforts to fit the needs of individual regions, cities, and even neighborhoods.

There is no single way to segment a market. We have to try different segmentation variables, alone and in combination, to see which give the best segmentation opportunities. In addition to what variables should be used, there is a question of what segmentation methodology is more appropriate for the research objectives.

We propose three different techniques to perform market segmentation:

- *Clustering:* this approach implies data grouping or partitioning
- *Association*: this approach seeks to establish associative relationships between different variables in the database
- *Visualization:* this approach consists of providing the user with an immersive virtual reality environment so that the user can move through this environment discovering hidden relationships.

**Clustering approach**

Since it is not known a-priori the number of market segments and each customer to which segment is most likely to belong, we will be approaching the market segmentation by using an unsupervised learning algorithm, called k-means clustering (k stands for the number of groups or segments) [2]. That is, the customers will be grouped in k segments such that each group contains the customers similar with each other in terms of the variables selected for clustering and, the difference between clusters is maximized. The market segmentation process should start first, with grouping customers having similar behavior and needs, and then discover which demographics make the customers distinct. In order to measure the similarity between two customers, there is necessary to define a similarity metric. Irrespective of the measure being used, a small value between two customers implies high-similarity and vice-versa a large value implies low similarity.

While we described the segmentation as a process of grouping similar costumers, the process is never perfect. That is, even when customers share, let us say, common feelings, there are still differences in demographics, that cannot be fully addressed within a segment and further segmentation needs to be performed within that segment. To deal with this situation we propose a hierarchical clustering approach for the market segmentation such that at every level of the hierarchy k-means clustering maybe applied.

**Association approach**

This approach to pattern discovery seeks to establish relationships between different items in the database. The association approach is very useful when one has an idea of different associations that are being sought out. Since we know that we want to find the associations between *vehicle attributes* and *anticipated feelings*, we believe that this approach will produce useful results for the decision makers.

Two methodologies can be used for the association approach:

- *Fuzzy and Rough Sets*: these methods incorporate the vagueness and imprecision that is common in everyday life. The imperfection dealt in fuzzy sets is with respect to objects within the same segment; in contrast, rough sets deal with imperfections between groups of objects in different segments.
- *Latent Semantic Association*: this method finds the hidden associations between different kinds of data [3]; this method will be very relevant to our project if we decide to incorporate in the database some other vehicle image attributes such as color and shape.

**Visualization approach**

We propose Multidimensional Scaling (MDS) approach [4] to visualize the relationships between customers with respect to their preferences and needs. The approach relies on a projection from a high dimensional space to a low-dimensional space (two or three dimensions) to uncover similarities among customers. The MDS mapping from a high dimensional to a low-dimensional space preserves the inter-point distances as much as possible, such that the mapping of objects corresponds to the perception of the human eye.

We will be using MDS to map both the relationships between variables for a given segment and the relationships between customers for a specific market. The proposed visualization tool will allow identifying the gaps between vehicle attributes and customer personality attributes and preferences. This will help the decision-maker understand the weakness of the current vehicle conception and create the premises for innovations by filling in the gap between product concept and the consumer preferences.

**3. Results Evaluation, Interpretation, and Knowledge Discovery**

To test how well the identified segments perform when predicting preferences for new customers, two approaches can be considered: train and test error estimation, and cross validation.

After the prediction accuracy is verified by one of the above methods, the segments will be evaluated by the business people in order to determine the usefulness of the segments. The evaluation of usefulness of the market segments should be made by the business team with respect to the following characteristics [5]:

- Substantiality (segment size): The market segments are large or profitable enough to serve.
- Measurability (segment profile): The market segments can be identified and measured in terms of data already available. The segment identification is very important: segments that are based on meaningful differences in customer needs but lack clear segment identification will fail because the segment identity will not be known and an actionable marketing strategy cannot be developed.
- Actionability: Effective programs can be designed for attracting and serving the segments. The market attractiveness depends on *market opportunity, competitive environment, and market access.*

If a segment fits the company's objectives, the company must decide whether it possesses the skills and resources needed to succeed in that segment. If the company lacks the strengths needed to compete successfully in a segment and cannot readily obtain them, it should not enter that segment. Even if the company possesses the required strengths, it needs to employ skills and resources superior to those of the competition in order to really win a market segment. Once the company has decided what segments to enter, it must decide on its *market positioning strategy* – on which positions to occupy in its chosen segments.

**4. Conclusions and future work**

In this paper we proposed a theoretical data mining framework for automatic gathering of relevant and unbiased data, and for automatic discovering of the relationships between customer profiles and vehicle's image characteristics. As a result, the initial investment of producing a vehicle without being certain that it will be satisfying people's needs will be eliminated. Discovering a-priori segments of people being interested in a certain vehicle will also help the managers focus their advertising, promotion, and sales efforts on those categories of people and thus, the time and costs will be significantly reduced.

**5. References**

[1] R. Tetzeli. "Surviving the information overload," Fortune, July, pp. 60-64, 1994

[2] A.K. Jain and R.C. Dubes. "Algorithms for Clustering Data," Prentices Hall Advanced Reference Series, 1998

[3] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. "Indexing by latent semantic analysis," Journal of American Society for Information Science, vol. 41, pp. 391-407, 1990

[4] T. Cox and M. Cox. "Multidimensional Scaling," Chapman & Hall, 1994

[5] P. Kotler and G. Armstrong. "Marketing: An Introduction," Prentices Hall, Fourth Edition, 1997