Mining Unstructured Data using Artificial Neural Network and Fuzzy Inference Systems Model for Customer Relationship Management

P.Isakki alias Devi[1] and Dr.S.P.Rajagopalan[2]

[1] Department of Master of Computer Applications, Guru Nanak College
Chennai, Tamilnadu – 600 042, India

[2] School of Computer Science & Engineering, Dr.M.G.R.University
Chennai, Tamilnadu – 600 095, India

Abstract

Data warehouse and mining are able to provide the structure to record whole customer's information, detecting important customers systematically, the change of identifying the individual and valuable customers. "Customer Relationship" is one of the most important factors to construct the core of competitiveness, especial in service industries for running business forever. Therefore, the objective of this paper is to apply the data warehouse and data mining technologies to analyze the customers' behavior in order to form the right of customers' profile. This could provide the best service model owing to the enounced of customer-orientation and making more effective marketing strategy.

Keywords: *Content analysis method , Data warehouse , Data mining, CRM, Artificial Neural Network and Fuzzy Inference Systems (ANFIS) Model.*

1. Introduction

When managing thousands of customers, business will have difficulty sustaining the rising costs created by interactions among people. However, if all customer data is inserted into a database, the resulting records will provide a detailed profile of these customers and their interactions with one another, and will be an important resource for businesses that wish to probe customer data, customer needs, and customer satisfaction levels. Data mining uses transaction data to gain a better understanding of customers and effectively discover hidden knowledge through the insertion of business intelligence into the process of customer relationship management. Data warehousing is useful and accurate for assembling a business's dispersed heterogeneous data and providing unified convenient information access technique. Data mining technology can be used to transform hidden knowledge into manifest knowledge. An integrated CRM system is extremely flexible. It can adjust customer needs throughout a product's life cycle, and to analyze and actively monitor customer preferences. Therefore, one of the best competitive strategies is the successful utilization of Information Technology to swiftly and effectively integrate business knowledge and provide the business with timely quality decision support.

The most common forms of customer interaction are as follows: (1) Face-to-face interaction with retail personnel; (2) Calls to customer service centers and conversations with customer service representatives; (3) Comments on company websites; and (4) Feedback expressed through e-mail. Customer data harvested through these methods is usually unstructured; however, most data mining technologies can only handle structured data. Therefore, during customary data warehousing processes, unstructured data is not taken into account and much valuable customer information is lost.

This paper uses content analysis to transform unstructured textual content into structured data. Unstructured data has no identifiable structure. It includes bitmap images/objects, text and other data types that are not part of a database. It is a generic label for describing any corporate information that is not in a database. It can be textual or non-textual. Textual unstructured data is generated in media like email messages, PowerPoint presentations, Word documents, collaboration software and instant messages [1]. Non-textual unstructured data is generated in media like JPEG images, MP3 audio files and Flash video files. Some current technologies used for content searches on unstructured data require tagging entities such as names or

applying keywords and meta tags. Therefore, human intervention is required to help make the unstructured data machine readable.

The systematic application of the coding principles of content analysis can produce derived variables and objectively quantify unstructured textual content. These construct a more complete customer data platform for data mining analysis and the extraction of hidden individualized knowledge for optimizing marketing strategies.

2. Design Methodology

The information flow from data collection to useable knowledge is shown in Figure 1.

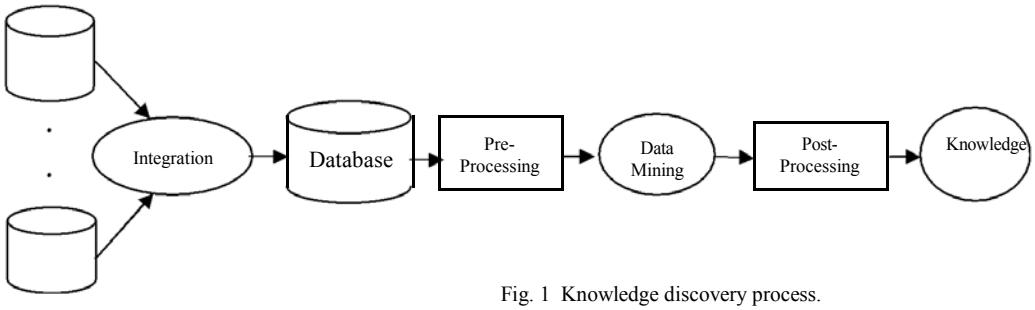

Fig. 1 Knowledge discovery process.

According to Figure 1, the first steps are to apply pre-established selection rules in the integration of primitive data, to decide whether to keep or discard data, and to decide the subset to which the data belongs. The next steps are to clean up and reorganize the data by discarding unnecessary or redundant information, to establish record-keeping formats and contents, and to ensure the integrity and consistency of the data in order to construct a d ata platform. Thereafter, the organized data is grouped into related subjects through data transformation and data mining processing methods are used to determine data models and to further define relationships among various data for reference in storage and query computation. After analysis and interpretation, the resulting models and correlations can become useful knowledge and tools for decision support.

2.1 Data Collection

Data can be collected from the following sources:
(i) The Electric News System
(ii) The customer service hotline
(iii) Word documents

The following is a description of these types of data:
i. Electric News System – Internet marketing content which includes featured reports and industry analysis.
ii. Customer Service Hotline (Customer Service Data) – The customer service hotline is established to provide appropriate and timely responses to customer demands and complaints.
iii. Word Documents – Textual Informations.

2.2 Content Analysis

The customer data can be imported from e-mail, the contents of which are diversified not manifestly structured. Therefore, content analysis for quantitative analysis is by following steps:

Step 1: User requirements
Main goals are: (1) Goal seeking or discovering core customers; and (2) evaluation of service strategies and service efficacy.
A detailed evaluation proceeded from the following three fronts:
i. Formulation of Customer Portfolios – Customer attributes and contents of inquiries are classified and their percentage shares calculated.
ii. Using Data mining technologies and methods to analyze consumer behavior and to provide the company with an ideal service model that is designed for consumers and oriented by customers.
iii. Providing the company with more effective marketing strategies that allow for timely customer retention and identification of potential customers and new customers.

Step 2: Building categories
Categories are built according to the provided company data and company needs.

Step 3: Coding and reliability analysis
Reliability is a critical factor of content analysis. If the coding process is unreliable, the analysis cannot be trusted. The reliability of content analysis is determined by the degree of inter-rater reliability - coefficient of reliability (C.R.). Before the coding officially begins, the coders must be trained to understand the subject, the goal, and the structure of the categories. Then, preliminary processing and recording may begin, and inter-judge agreement and inter-rater reliability are tested per Holsti method:

$$C.R = 2M / (N_1+N_{2)} \quad (1)$$

$$\text{reliability}=(nx (\text{average-of-C.R.})) / (1 + [(n-1) \text{ x average-of-C.R.}]) \quad (2)$$

M is the number of decisions on which the coders completely agree;

N_1 is the number of coding decisions by the first rater;
N_2 is the number of coding decisions by the second rater;
n is the total number of raters.
If reliability exceeds 0.8, then reliability meets standards and the official coding process may begin.

Step 4: Validity testing
Being a valid study based on content analysis, its findings must not be based on any specific data, method, or measured value; the findings must reach beyond any specific data, method, or measured value to reach a general conclusion.

Using content analysis methodologies to analyze text data allows inferences to be made for specific subjects or goals without having to rely on expensive and time-consuming large data warehouses or intricate information technologies.

2.3 Data Movement and Data Transformation

The following steps outline the process of data transformation:

Step 1: Environment settings
Environment settings refer to the data mining system's architecture

Step 2: Formulation of data models
After ascertaining requirements, the method of data storage should be determined. There are three levels of data modeling: (1) Conceptual: Expresses the relationships among the entities included in the data; (2) Logical: Completed without regard to which type of database is to be used. This step is only completed after the user confirms the accuracy of the model; and (3) Physical: The logical data model is actually built in an Access database.

Step 3: Data Movement and Data Transformation
The company can extract data from customer service center systems and convert the data into Excel format for study personnel. Content analysis can be used to analyze the data content of customer inquiries. After the analysis is completed, the data combined with other data in the Access database, and the necessary programs are written to transform and clean the data. Finally, the data is integrated and its accuracy verified by checking items such as customer listing and data formats.

Step 4: Data Quality Assurance
Items such as customer listing uniqueness, whether or not the data values can be null or blank, and data format are tested for accuracy. During the quality assurance process, data inaccuracies are the most likely cause of users losing confidence in database construction.

3. Artificial Neural Networks (ANN) and Fuzzy Inference Systems (FIS) Model

It is observed that the use of Neural Networks (NNs), fuzzy logic (FL) and Genetic Algorithms (GAs) has been increased during the last decade in CRM-related applications. However, the focus often has been on a single technology heuristically adapted to a problem. While this approach has been productive, it may have been suboptimal, in the sense that studies may have been constrained by the limitations of the technology and opportunities may have been missed to take advantage of the synergies between the technologies. For example, while NNs have the positive attributes of adaptation and learning, they have the negative attribute of a "black box" syndrome. By the same token, Fuzzy Logic has the advantage of approximate reasoning but the disadvantage that it lacks an effective learning capability. Merging these technologies provides an opportunity to capitalize on their strengths and compensate their shortcomings.

Fusion of Artificial Neural Networks (ANN) and Fuzzy Inference Systems (FIS) have attracted the growing interest of researchers in various scientific and engineering areas due to the growing need of adaptive intelligent systems to solve the issues related to e-commerce. ANN learns from scratch by adjusting the interconnections between layers. FIS is a popular computing framework based on the concept of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning. A key challenge for business firms is to manage customer relationships as an asset. Compared to traditional

mathematical modeling, fuzzy modeling possesses some distinctive advantages, such as the mechanism of reasoning in human understandable terms, the capacity of taking linguistic information from human experts and combining it with numerical data and the ability of approximating complex non-linear functions with simple models [5]. To create an effective toolkit for the analysis of customer relationships, a combination of relational databases and *Adaptive Neuro-Fuzzy logic* is proposed. The fuzzy Classification Query Language allows marketers to improve customer equity, launch loyalty programs, automate mass customization, and refine marketing campaigns.

3.1 Fuzzy Vs. Sharp Classification

Fuzzy logic aims to capture the imprecision of human perception and to express it with appropriate mathematical tools. The marketers are able to use linguistic variables, such as “loyalty", and linguistic terms like “high" or “low" with fuzzy classification model. There are number of advantages in using fuzzy classification for relationship management:

• Fuzzy logic enables the use of non-numerical attributes. As a result, both qualitative and quantitative attributes can be used for marketing acquisition, retention, and add-on selling.

• With the help of linguistic variables and terms, marketers may describe equivalence classes more intuitively (excellent loyalty, medium loyalty, weak loyalty). The definition of linguistic variables and terms and the naming of fuzzy classes can be derived directly from the terminology of marketing and sales departments.

• Customer databases can be queried on a linguistic level. For example, the Fuzzy Classification Query Language [2] allows marketers to classify single customers or customer groups by classification predicates such as “loyalty is high and turnover is large".

The design of fuzzy classes requires marketing specialists as well as data architects and online-shop administrators [4].

3.2 Fuzzy Classification Model

A fuzzy classification model is helpful to attract potential online customers of high quality and to retain and extend their customer value. The relational database of online customers is extended by a context model in order to obtain a classification space. To every attribute Aj defined by a domain D(Aj) there is added a co ntext C(Aj). A context C(Aj) of an attribute is a partition of D(Aj) into equivalence classes. In other words, a relational database schema with contexts R(A,C) consists of a set A=(A1,…,An) of attributes and the set C=(C1(A1)),…,Cn(An)) of associated contexts.

Customers have been evaluated by only two attributes, turnover and loyalty. In addition, these two qualifying attributes for customer equity have been partitioned into only two equivalence classes. The pertinent attributes and contexts for relationship management are:

• Turnover in Euro per month: The attribute domain is defined by [0..1000] and divided into the equivalence classes [0..499] for small and [500..1000] for large turnover.

• Loyalty: The domain {excellent, good, mediocre, bad} with its equivalence classes {excellent, good} for high and {mediocre, bad} for low loyalty behavior.

To derive fuzzy classes from sharp contexts, the qualifying attributes are considered as linguistic variables, and verbal terms are assigned to each equivalence class. The equivalence classes can be described more intuitively with linguistic variables. In addition, every term of a linguistic variable represents a fuzzy set. Membership functions defined for the domains of the equivalence classes.

As turnover is a numeric (sharp) attribute, its membership functions are continuous functions defined on the whole domain of the attribute.

For qualitative attributes like loyalty, step functions are used; the membership functions μ high and μ low define a membership grade for every term of the attribute’s domain.

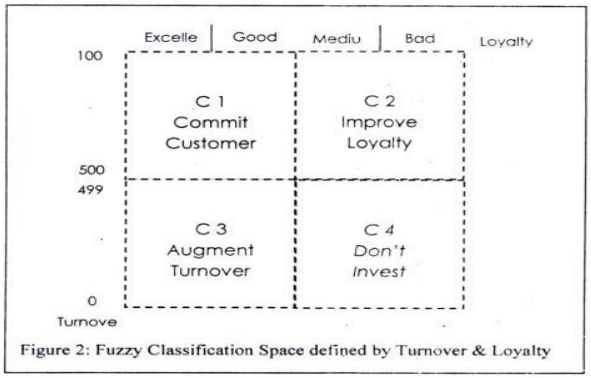

Figure 2: Fuzzy Classification Space defined by Turnover & Loyalty

The selection of the two attributes “turnover" and “loyalty" and the corresponding equivalence classes determine a two-dimensional classification space (Figure. 2). The four resulting classes C1 to C4 could be characterized by marketing strategies such as ”Commit Customer" (C1), “Improve Loyalty" (C2), “Augment Turnover" (C3), and “Don’t Invest" (C4). The selection of qualifying attributes, the introduction of equivalence classes and the choice of appropriate membership functions are important design issues [3]. Database architects and marketing specialists have to work t ogether in order to make the right decisions.

The proposed context model proved suitable to the use of linguistic variables and membership functions, the classification space becomes fuzzy. A fuzzy classification of customers has many advantages compared with common sharp classification approaches. Most importantly, with fuzzy classification a customer can belong to more than one class at the same time. This leads to differentiated marketing concepts and helps to improve customer equity.

3.3 Fuzzy Classification Query Language

The classification language fCQL is designed in the spirit of SQL. Instead of specifying the attribute list in the select clause, the name of the object column to be classified is given in the classify clause. The from clause specifies the considered relation, just as in SQL. Finally, the where clause is changed into a with clause which specifies a classification predicate.

An example in customer relationship management could be given as follows:
classify Customer
from CustomerRelation
with Turnover is large and Loyalty is high

This classification query would return the class C1 (Commit Customer) defined as the aggregation of the terms “large" turnover and “high" loyalty.

In this example, specifying linguistic variables in the “*with clause”* is straightforward. However, if customers are classified on three or more attributes, the capability of fCQL for a multi-dimensional classification space is increased. This can be seen as an extension of the classical slicing and dicing operators on a multidimensional data cube.

4. Conclusions

Content analysis is used to analyze text data. It is used to transform unstructured customer service information into structured customer service data. This paper is used to find ways to analyze text data in order to discover more latent knowledge. Content analysis is used to process text data. Data warehouse and ANFIS model are used to discover customer knowledge. In future, this model can be applicable for business with insufficient information systems.

References

[1] V.Narayani, Dr.S.P.Victor and S.Rajkumar, “Semantic investigation of unstructured datum on e vent mining analysis”, International journal of Engineering Science and Technology , Vol. 2 (7), 2010 , 2763-2769.

[2] Meier A., Werro N., Albrecht M., Sarakinos M., 2005. “Using a Fuzzy Classification Query Language for Customer Relationship Management”, Proceedings 31st International Conference on V.

[3] Werro N., Stormer H. and Meier A., 2005, “Personalized discount - A fuzzy logic apagesroach”, Proceedings of the 5th IFIP International Conference on eBusiness, eCommerce and eGovemment, Poznan, Poland, pages. 375-387.

[4] Lee J., Podlaseck M., Schonberg E. and Hoch R., 2001. “Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising”, Data Mining and Knowledge Discovery, 5, pages. 59-84.

[5] J. Bonaventura Cunha “Greenhouse climate hierarchical fuzzy modelling”, Control Engineering practice 13, pp 613-628 , 2005.

P.Isakki alias Devi receieved B.Sc and M .C.A Degree from Madurai Kamaraj University in 1997 and 2000. She is working as a Assistant Professor in MCA Department of Guru Nanak College, India. She has 10 y ears of teaching experience. She is pursuing Ph.D in Vels University, India. Her Research area is Data Mining for Customer Relationship Management.

Dr.S.P.Rajagopalan received M.Sc from IIT Madrs, M.Phil and Ph.D from Madras University, Chennai, India. He i s working as a Professor (Emeritus) in School of Computer Science & Engineering of M.G.R.University, Chennai. He has 40 y ears of teaching experience. He has published 4 Books. He has about 100 publications in International Journals and N ational Journals. His special fields of i nterest include Quantitative Techniques, Data Processing and P roject Management, Management Information System, Programming Languages, Simulation, Text generation, Cryptography and Data Mining.