Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications xxx (2008) xxx–xxx
Expert Systems with Applications
www.elsevier.com/locate/eswa

Mining the text information to optimizing the customer relationship management

Che-Wei Chang [a,*], Chin-Tsai Lin [b], Lian-Qing Wang [b]

[a] Department of Information Management, Yuanpei University, 306 Yuanpei Street, Hsin Chu 30015, Taiwan, ROC
[b] Graduate Institute of Business and Management, Yuanpei University, 306 Yuanpei Street, Hsin Chu 30015, Taiwan, ROC

Abstract

Customer data warehouse and mining are able to provide the structure of recording of the whole customers' information, the flow of detecting the important customers systematically, the change of identifying the individual and valuable customers in the whole name list of customers or discovering the royal customers. Generally speaking, it is no doubt that "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 research is to apply the data warehouse and data mining technologies to analyze the customers' behavior in order to form the right of customers' profile and it growth model under Internet and e-commerce environment. This could provide the best service model owing to the enounced of customer-orientation and making more effective marketing strategy. Consequently, a case study will be presented to verify the feasibility and effectiveness of the approach proposed in this research.

Keywords: Content analysis method; Data warehouse; Data mining; Decision tree; CRM

1. Introduction

Linoff and Berry (2002) indicate that when managing hundreds of thousands of customers, businesses will have difficulty sustaining the rising costs created by interactions among people. However, if all customer data is inputted 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 (Aha, Kibler, Albert, & Albert, 1991). 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. More precisely speaking, it replaces artificial intelligence, and for this reason the technology has become popular in recent years. Data warehousing, then, is a useful and accurate tool for assembling a business's dispersed heterogeneous data and providing unified convenient information access technique, because it can process large amounts of information with the support of its unique data storage structure and network architecture. In the business world, once the foundation of a data warehousing system is laid, data mining technology can be used to transform hidden knowledge into manifest knowledge. The results improve the independent decision-making abilities of employees and help businesses attain Gates's model: Digital activity as the kernel for building business processes and providing timely information to appropriate decision-making units (Gates, 1999).

Popularizing information automation has resulted in heavy utilization of information technologies, such as the internet and automated telephone answering systems, construction of dynamic websites, and implementation of ERP and operational CRM systems; their emphases are on process optimization and efficient, highly precise account

[*] Corresponding author. Tel.: +886 3 6102361; fax: +886 3 6102362.
E-mail addresses: tjmccw@xuite.net (C.-W. Chang), ctlin@mail.ypu.edu.tw (C.-T. Lin), Lance_Wang@e-Synergy.com.tw (L.-Q. Wang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.11.027

management. With the increased popularity of personalization management, the integrated CRM aim to seek quality service and high levels of customer satisfaction. An integrated CRM system is extremely flexible – It can adjust customer needs throughout a product's life cycle, and it has the ability 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.

Today, businesses face the challenges of using the past to predict the future and using past experiences to communicate effectively with the customer. 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) Opinions expressed through e-mail. Customer data harvested through these methods is usually unstructured; however, most data mining technologies can only handle structured data, which means that the data warehouse must have explicit field structures. Therefore, during customary data warehousing processes, unstructured data is not taken into account and much valuable customer information is lost.

This study uses content analysis to transform unstructured textual content into structured data; 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. Literature review

Content analysis is chiefly the process of establishing a framework and selecting the units of analysis based on the goal of the study. It uses the principles of "measurable" and "quantifiable" to design categories that can partition the analyzed units' data content into a series, selects representative data samples, and uses the categories to quantify (recording coding decisions and performing reliability analyses) and analyze the samples. Krippendorff (1980) praised content analysis because it is unobtrusive, accepts unstructured samples, is context-sensitive, and can handle large amounts of data, etc. McMillan (2000) considered that the ability to handle massive amounts of data is the biggest advantage of applying content analysis to web content. Empirical studies in traditional communication fields often use content analysis on various style of advertisement; in computer-mediated communications, content analysis has been used to analyze discussions in newsgroups and on electronic bulletin boards.

Linoff and Berry (2002) defined data mining as the usage of classification, association rules, machine self-learning, sequential analysis, cluster analysis, and other statistical methods to seek out implicit, unknown, yet extremely useful information from massive and diverse databases. In other words, data mining extracts accurate, previously unknown, yet significant information from large databases and uses this information to make important decisions. Romero and Ventura (2007) surveyed the application of data mining to traditional educational systems, particular web-based courses, well-known learning content management systems, and adaptive and intelligent web-based educational systems. Chau and Yeh (2004) developed concept-based cross-lingual text retrieval to discover the multilingual concept–term relationships from linguistically diverse textual data relevant to a domain. Yang and Lee (2005) developed automatic hypertext construction method is necessary for content providers to efficiently produce adequate information that can be used by web surfers. Köhler, Philippi, Specht, and Rüegg (2006) developed fully automated methods for mapping equivalent concepts of imported RDF ontologies to allow the seamless integration of domain specific ontologies for concept based information retrieval in different domains.

Among customer relationship management studies, decision trees are frequently used when studying customer portfolio management. An important function of decision trees is the construction of a branching structure by classifying known examples. The resulting decision tree has rules that can be expressed via text or data, and the decision tree model can be used in predictions beyond the existing sample. The decision tree is a figurative tree and similar to the data structure tree in that it has nodes and leaves, and an appropriate test is placed at each node. The test determines which sub-tree or condition of the node the data will be applied to for further decision-making. The goal or target of the analyzed question is arrived at through the tests at these nodes. The tree is established when all the data is distributed to appropriate tags and the results are displayed in tree form.

3. Methodology

The information flow from data collection to useable knowledge is shown in Fig. 1. According to Fig. 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 data 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 result-

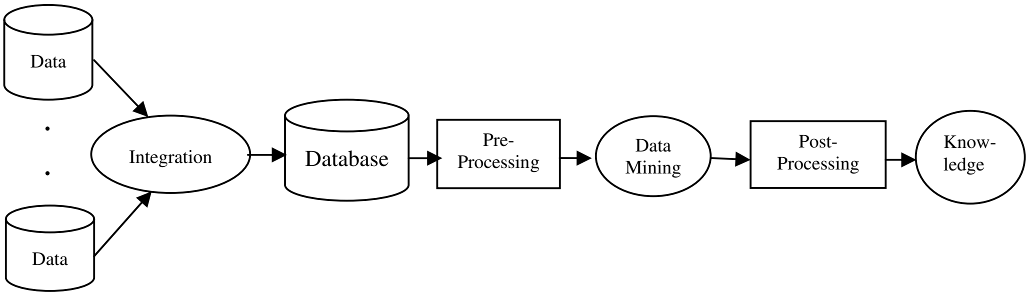

Fig. 1. Knowledge discovery process.

ing models and correlations can become useful knowledge and tools for decision support.

This section is divided into three sub-sections according to the above knowledge discovery process. First, Section 3.1 describes the process of using content analysis to analyze unstructured data. Section 3.2 explains the criteria of data movement and data transformation in data mining and the integration of related data to establish the required data source. Section 3.3 details the processes of data movement and data transformation. Section 3.4 describes the process of building business value.

3.1. Data collection

This study collected Company A's customer service center data from three sources: (1) The Electric News System; (2) The customer service hotline; and (3) Data from various conferences. The following is a description of these three types of data:

i. Electric News System – Internet marketing content which includes recent conferences, marquee information, featured reports, industry analysis, technology columns, intellectual property columns, etc.
ii. Customer Service Hotline (Customer Service Data) – The customer service hotline was established to provide appropriate and timely responses to customer demands and complaints. Currently, the team responds to anywhere between 150 and 200 e-mail inquiries per month.
iii. Data from Various Conferences – Data from 2002 to 2003 including conference type, location, fees, and participants.

3.2. Content analysis

The customer data used in this analysis was mainly imported from e-mail, the contents of which were diversified not manifestly structured. Therefore, the study used content analysis for quantitative analysis by following three steps:

Step 1: User raequirements

After discussing requirements with Company A's users, the main goals of this study 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 were classified and their percentage shares calculated. Formulation of the types of customer attributes and inquiries represented by the percentage shares.
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 were 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 of the study, the goal of the study, 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 (1969) method:

$$C.R. = \frac{2M}{N_1 + N_2}. \tag{1}$$

$$\text{reliability} = \frac{n \times (\text{average-of-}C.R.)}{1 + [(n-1) \times \text{average-of-}C.R.]}. \tag{2}$$

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. In other words, if several studies consistently reach the same conclusion, then the correspondence reflects the validity of the conclusion because it is generalized over several different studies with different circumstances.

i. Construct Validity – Refers to the degree of correlation between a certain measure within a construct and other measures within the same construct. Construct validity can be further classified as either convergent validity or discriminant validity. If a measure correlates with other measures within the construct, then the measure has convergent validity; if a measure does not correlate with other measures, then it has discriminant validity.
ii. Hypothesis Validity – Refers to the degree of inter-variable consistency and the degree to which the relationship among variables behaves as expected. If a measure has hypothesis validity, its behavior in relationship to other measures will be as predicted.
iii. Predictive Validity – Refers to the correlation between forecasted events or conditions external to the study and actual events or conditions. These predicted events include future, past, or present events.
iv. Semantic Validity – Vocabulary or other “coded units” sorted into the same class during the classification process should have similar connotations.

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.

3.3. Data movement and data transformation

Company A currently does not have a data warehousing system and cannot provide effective decisions on data integration and data analysis. The following steps outline the process of data transformation:

Step 1: Environment settings
Environment settings refer to the data mining system’s architecture. For the purposes of this study, data processing programs such as Microsoft Access, Microsoft Excel, and Visual Basic were installed in a Windows XP operating system. In addition, Analysis Services was installed as a data mining tool.

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 extracted data from customer service center systems and converted the data into Excel format for study personnel. Content analysis was then used to analyze the data content of customer inquiries. After the analysis was completed, the data combined with other data in the Access database, and the necessary Visual Basic programs were written to transform and clean the data. Finally, the data was integrated and its accuracy verified by checking items such as customer listing and data formats. The data movement and transformation process, as seen in Fig. 2, is completed when the data accuracy is confirmed and the appropriate data model is imported into the data warehousing system.

Step 4: Efficiency adjustments
Appropriate adjustments and optimizations were made to the database index file to improve query processing efficiency.

Step 5: Data quality assurance
Items such as customer listing uniqueness, whether or not the data values can be null or blank, and data format were tested for accuracy. During the quality assurance process, data inaccuracies are the most likely cause of users losing confidence in database construction. Unfortunately, this step is often omitted or forgotten under pressure.

3.4. Building business value

The decision tree generating algorithm used in this study is Quinlan’s (1993) C4.5 learning program. The C4.5 learning program is a decision tree generator that uses information theory and inductive learning methods. It involves the three steps as follows:

Step 1: Decision tree building
The first step in the C4.5 learning program involves using training data to build a decision tree. The basic concept originates from the theories proposed by Hoveland and Hunt in the late 1950s and leads to concept learning systems (Hunt, Marin, & Stone, 1966) in the following decade. To put it simply, if a set S comprised of a set of training data contains K number of classes, and $S = \{C_1, C_2, C_3, \ldots, C_k\}$, then there are three possible outcomes in the decision tree building process.

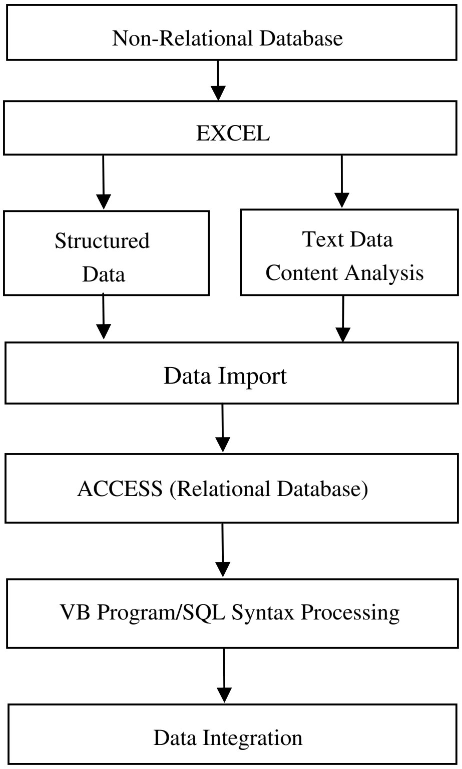
Non-Relational Database
EXCEL
Structured Data
Text Data Content Analysis
Data Import
ACCESS (Relational Database)
VB Program/SQL Syntax Processing
Data Integration

Fig. 2. Data processing.

i. If all the data in S belongs to the same class C_j, the resulting decision tree contains only one leaf that comprises all the data in C_j.
ii. If there is no training data in S, the resulting decision tree still contains only one leaf, but the class represented by this leaf is determined by training data beyond tree T.
iii. When S contains training data from many different classes, then set T is split, according to certain attributes, into many subsets S_1, $S_2, \ldots, S_n$, with one class comprising each subset if possible. The decision tree built from S includes a decision node and n branches, with each subset of training data corresponding to a branch of T.

Usually when the training data is being assigned, there are many decision trees that can accurately classify the data. In order to find the simplest decision tree that can accurately classify the data, the classifying attributes must be chosen carefully. The C4.5 program used in this study is an extension of its predecessor, the ID3 learning program. The criterion used by ID3 to select classifying attributes is called "gain", and its methods are based on information theory. It measures the amount of information in each class and calculates the average amount of information, or entropy, in the training set in order to express its level of complexity.

If the set S built from the training data has n classes, where $C_{i,j} = 1,\ 2, \ldots, n$, the quantity of data in each class is represented by $\mathrm{freq}(C_j, S)$, where $|S|$ represents the quantity of all data in S. Thus the probability of an arbitrary piece of data belonging to each class can be represented by:

$$\frac{\mathrm{freq}(C_i, S)}{|S|} \tag{3}$$

According to information theory, then, the information in each class is expressed as:

$$-\log_2\left(\frac{\mathrm{freq}(C_i, S)}{|S|}\right) \tag{4}$$

The training set contains training data of different classes, so the average amount of information (entropy) in the training set can be calculated by multiplying the amount of information in each class by the probability of an arbitrary piece of data belonging to each class and summing the products:

$$-\sum_{i=1}^{n} \frac{\mathrm{freq}(C_i, S)}{|S|} \log_2\left(\frac{\mathrm{freq}(C_i, S)}{|S|}\right) \tag{5}$$

Per the $\mathrm{info}(S)$ calculation, when set S is split into multiple subsets S_1, $S_2, \ldots, S_n$ on an attribute A, the amount of information in the set after the split is obtained by multiplying the amount of information in each subset by the proportion of each subset and summing the products:

$$\mathrm{info}A(S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \mathrm{info}(S_i) \tag{6}$$

Thus the amount of information gained by splitting set S on attribute A is the amount of information prior to the division minus the amount of information after the split, expressed as follows:

$$\mathrm{gain}(A) = \mathrm{info}(S) - \mathrm{info}A(S) \tag{7}$$

The ID3 learning system selects classifying attributes by using these calculations to find the gain values provided by each attribute and then selecting the attribute that provides the largest gain value. The decision tree is split into many training subsets based on the value of this attribute. Each sub-tree is split into more sub-trees by iterating the steps described above – selecting classifying attributes by finding the attribute that provides the largest gain value until no further sub-trees can be obtained.

The ID3 attribute selection method is sufficiently effective for common learning problems, but when the classifying conditions favor attributes that create more subsets, a unique aspect of ID3 becomes apparent. When all the subsets thus split from set S contain only one data object, the amount of information in S after the split is 0, so the largest amount of information is gained. To split the set on these types of attributes is more or less meaningless. In order to remedy this flaw, Quinlan's C4.5 proposed a way to normalize the gain and temper the effect of creating too many subsets. The normalization is obtained by dividing the original gain value by the value of split $\mathrm{info}(A)$, i.e., gain $\mathrm{ratio}(A) = \mathrm{gain}(A)/\mathrm{split\ info}(A)$, where

Please cite this article in press as: Chang, C.-W. et al., Mining the text information to optimizing the customer ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.11.027

$$\text{split info}(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \log_2\left(\frac{|S_i|}{|S|}\right) \tag{8}$$

represents the potential information generated after the set is split on attribute A. As the quantity of subsets after the split increases in number, the split info increases in value and the gain ratio decreases in value. Therefore, the split info method allows the C4.5 learning program to improve upon the flaw that emerges in ID3 when the number of subsets grows.

Step 2: Pruning the decision tree

If the data is incomplete, overly sparse, or contains noise, the decision tree built using this method usually "over-fits the data" and is excessively complex. Therefore, the decision tree needs to be pruned after it is built by C4.5.

The C4.5 pruning process uses the values of predicted error rates as judgment criteria. To prune, start from the bottom of the tree (the leaves) and work towards the top of the tree by testing the sub-trees formed by each node. If replacing the sub-tree with a leaf results in a lower predicted error rate, then the sub-tree should be pruned into a leaf; otherwise, the sub-tree should remain. The predicted error rate is the error rate in the current training data evaluated against the error rate in non-training data.

Step 3: Generating learning rules from the decision tree

After building the decision tree classification model, the natural inclination is towards obtaining accurate predictions, but the further step of building a classification model that approximates human thought processes is also desired. In other words, in addition to having high accuracy, the classification model should also consist of simple rules that are easy to understand. After building the decision tree, the C4.5 learning program also converts the decision tree into even simpler rules in order to lower the level of complexity involved in classifying the data.

The easiest way to build a set of rules from a decision tree is to write rules that describe the paths of each leaf. However, the rules obtained with this method are as complicated as the original decision tree, so there is no improvement. C4.5 adds a simplifying step to the process of rule building, and the necessity of each condition to the rules is carefully evaluated in order to keep only the simplest conditions with the lowest error rates. The criteria used by C4.5 in this evaluation are similar to the criteria used in pruning the decision trees in that they are based on predicted error rates. The basic method is as follows:

If a rule R states:

$$R : \text{If Condition} = A, \text{then Class} = C \tag{9}$$

And a rule R- that is more generalized than rule R is:

$$R\text{-} : \text{If Condition} = A\text{-}, \text{then Class} = C\text{-} \tag{10}$$

Table 1
Objects that satisfy condition X and belong in class C

	Class C	Other classes
Satisfies condition X	$Y1$	$E1$
Does not satisfy condition X	$Y2$	$E2$

The condition set A- is obtained by subtracting a condition X from condition set A, and it is expressed by $A\text{-} = A - X$. Table 1 describes the number of objects that either satisfy or do not satisfy condition X and are either in class C or other classes.

Thus according to rule R, the quantity of data that will be classified into C is $Y1 + E1$, but among this data $E1$ represents misclassified data. Therefore, the predicted error rate of R is UCF ($E1$, $Y1 + E1$). Likewise, according to rule R-, because it lacks the addition criterion of condition X, the data that will be classified into C is described by $Y1 + Y2 + E1 + E2$, but among this data $E1 + E2$ represents misclassified data. Therefore, the predicted error rate of R- is expressed as UCF ($E1 + E2$, $Y1 + Y2 + E1 + E2$). If the predicted error rate of R-is lower than that of R, it is evident that condition X does not affect predicted error rate and can be discarded. The C4.5 learning program uses this concept to eliminate the most inconsequential conditions in each rule for convenience in calculation.

4. Results

Content analysis was used to analyze text data. After discussing the results, relevant categories and their descriptions were built as per Table 2.

Three coders participated in the coding of unstructured data for this study. The reliability was 0.88, which is greater than 0.80, an indication that inter-coder consistency was up to standards, and that official coding could begin.

Content analysis was used to transform unstructured customer service information into structured customer service data. The electric news system, list of electric news subscribers, and list of conference attendees were then integrated, moved, and after data movement and data transformation, a database of the interactions between Company A and its customers was established. This database can be used to examine the different types of customer needs, evaluate electric news merits, and obtain conference attendance information. In addition, the analysis model can be used to understand possible customer needs and customer behavior patterns and to provide appropriate customer service to different customers. This study analyzed the following variables sifted from the newly built database:

i. Customer demographics variables
   (a) Represented Unit – Company of employment or school of enrollment.
   (b) Geographical Area – Location of the company or school.

Table 2
Established categories

Category	Type	Classification code	Description
1 Technological needs	Technological cooperation	1. Technology transfers 2. Investment 3. Technology and patent licensing	1. Technology transfers; cooperation in developing new technologies 2. Investment; establishing presence in industry districts 3. Licensing
	Technology services	1. Business hiring – technology 2. Business hiring – other 3. Strategy groups	1. Hiring for evaluations and testing, technology consulting, lab certifications, contracted work, and valuations 2. Hiring for market surveys, personnel training, patents licensing and auctioning, and paid studies 3. Strategy groups
2 Informational needs	Requests for technological data	Technological publications	Studies or technology reports, periodicals, theses and dissertations, technological articles, keyword definition and translation, and subscriptions
	Technology consulting	Discussion of technological issues	Technology consulting, manufacturing issues, product issues, and technological standards
	Requests for market information	1. Industry information 2. Company data 3. Product information	1. Inquiries regarding industry analyses, market surveys, trend reports, industry news, and related legislation, etc. 2. Requests for company listings and related information 3. Inquiries regarding product sales, purchasing, and product catalogues, etc.
	Conference information	1. Classes 2. Lecturers	1. Conference place, time, literature, schedules of classes, and anticipated attendance 2. List of lecturers or recommended lecturers from the company
	Administrative inquiries	1. Processes, policies 2. Administrative services	1. Procedures, policies 2. Visitations and tours; document processing
	Requests for industrial researchers	Requests for industrial researchers	Voicemails, requests for phone numbers, e-mail addresses, and physical addresses
3 Recruiting	Job searches	1. Regular employment opportunities 2. Employment opportunities for the military reserves	1. Inquiries regarding employment opportunities and employment locations 2. Inquiries regarding employment opportunities for the military reserves
4 Feedback	Personal feedback	1. Feedback from organizations 2. Feedback regarding technology or products 3. Advertisements	1. When these customers give feedback to the company, the comments are only recorded and do not need to be processed normally 2. Feedback regarding technology or products 3. Advertisements
	Regular complaints	1. Technology services 2. Administrative support 3. Employee behavior	2. Technology services (technology marketing) 3. Cases that need resolution where individuals or units are dissatisfied with the company. Examples: IT issues, community service, waste disposal, mail, parking, poor attitudes, etc. 4. Complaints regarding individual employees' behavior
	Major complaints	1. Media matters 2. Customer service 3. Employee behavior	2. Media damages (to reputation); suits; cases that result in heavy losses to customers; cases that require inter-unit handling; cases that may affect the company's promotion of technology 3. Customer service issues; industry service issues 4. Individual employees' behavior
	Website issues	1. Questions 2. Suggestions	1. Broken links, erroneous information, lack of server response, etc. 2. Suggestions about the website and individual web pages

(c) Field of Technology – Technological field of the company or school.

(d) Represented Department – Department of employment or field of study at school.

(e) Position Title – Position in company or school (faculty or student, etc.).

(f) Whether they are subscribed to newsletter or not.

ii. Customer behavior attribute variables can be divided into two main categories.
Customer service needs

(a) Class – The primary classification level of customer needs, whether it is related to research and development, industrial services, information needs, recruiting, or customer feedback, etc.

(b) Type – The secondary classification level of customer needs; established during content analysis.

(c) Classification Code – The tertiary classification level of customer needs; established during content analysis.

(d) Date – Date of contact with the customer service center.

Conferences
(a) Date – Date of the conference.
(b) Location – Location of the conference.
(c) Type – Agency or department that is hosting the conference.
(d) Monies – Cost to attend the conference.

The next steps were to use Analysis Services to generate, analyze, and calibrate the decision trees and to generate the decision rules. The decision tree analyses mainly required data as classified above; for example, in the area of customer service, analysis of class, type, and classification codes of customer service was desired. In the area of conference variables, analysis of the number of conferences and monies was desired in order to obtain new customers and to cultivate or retain current ones.

This study used the time dimension to examine the interactions between the company and its customers and attempted to discover customer behaviors at different points in time. Lastly, classification analysis was used to discover the relationship rules or behavior patterns between customer service and conference attendance: These rules or patterns can be used to anticipate customer behavior and to provide appropriate information or marketing proposals at the appropriate time.

To analyze conference variables for customers who have conference needs, customers from different geographical areas were targeted and their numbers were compiled over time using varying units of time. These statistics were compared with the predicted results in order to verify whether customers from different geographical areas can be forecasted. In the end, the predicted results can be observed by classifying levels of attendance and customer subscriptions to the newsletter, as in Figs. 3 and 4.

After analyzing all the data, the initial findings indicate that subscriptions to the electric news, company location, and conference fees have no effect on customers' inclination to attend conferences. Among all conferences, attendances at technology services, materials, and optoelectronics conferences are not affected by timing; however, other types of conferences are affected by timing.

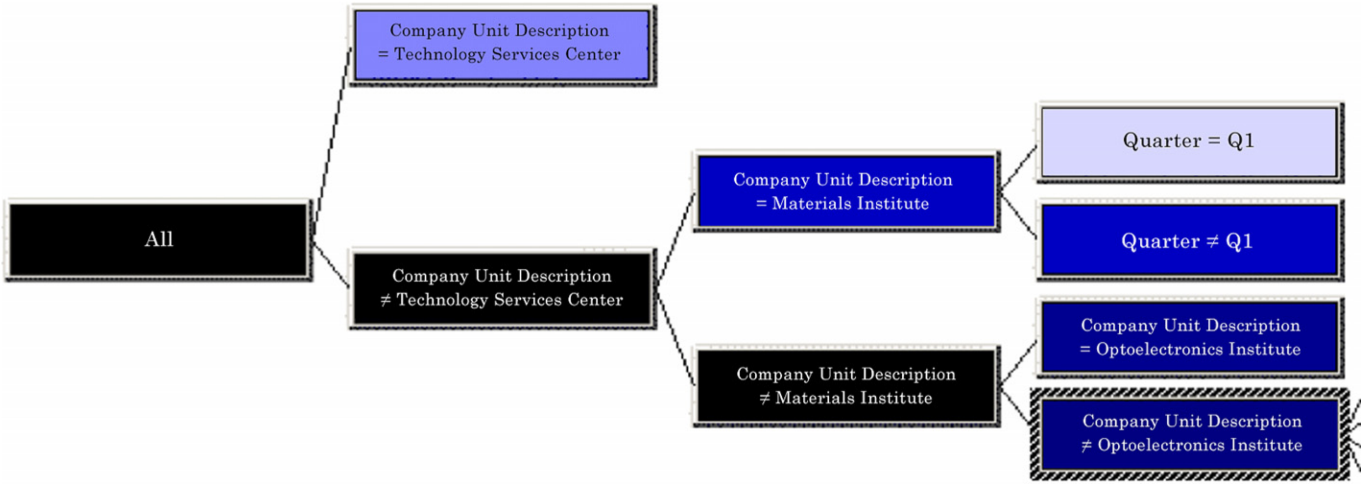
Company Unit Description = Technology Services Center
All
Company Unit Description ≠ Technology Services Center
Company Unit Description = Materials Institute
Company Unit Description ≠ Materials Institute
Quarter = Q1
Quarter ≠ Q1
Company Unit Description = Optoelectronics Institute
Company Unit Description ≠ Optoelectronics Institute

Fig. 3. Results of decision tree analysis #1.

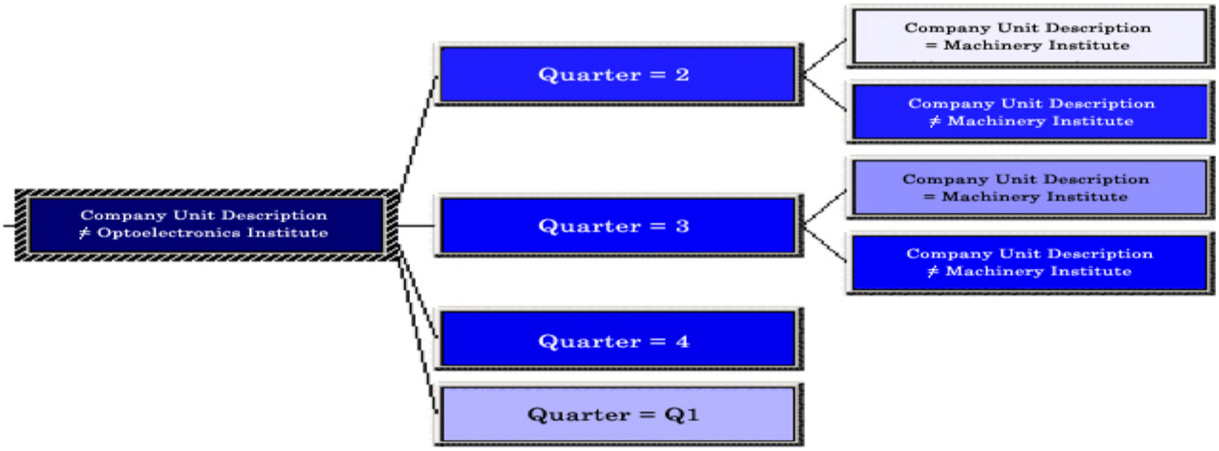
Company Unit Description ≠ Optoelectronics Institute
Quarter = 2
Quarter = 3
Quarter = 4
Quarter = Q1
Company Unit Description = Machinery Institute
Company Unit Description ≠ Machinery Institute
Company Unit Description = Machinery Institute
Company Unit Description ≠ Machinery Institute

Fig. 4. Results of decision tree analysis #2.

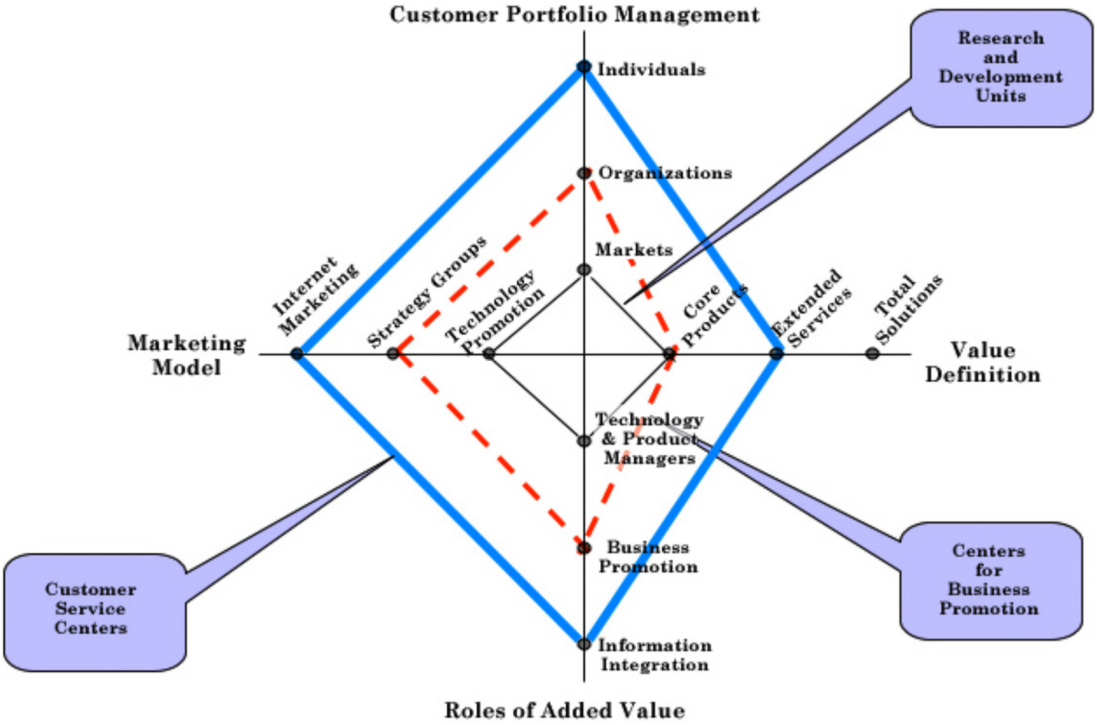

Fig. 5. Role definition compass.

Based on the analysis of customer service data and conference participation data, and borrowing from Wayland and Cole (1997) customer value compass, this study proposes a "role definition compass" (see Fig. 5) that analyzes the definition of roles played by associated units and of marketing strategies based on customer relationship management. Fig. 5 is divided into four aspects:

i. Customer Portfolio Management – The management principles of choosing valuable customer portfolios. This aspect is divided into three levels: Markets, organizations, and individuals.
ii. Value Definition – Contributions to the business value chain. This aspect is divided into three levels: Core products, extended services, and total solutions.
iii. Roles of Added Value – Business responsibilities in the business added value chain. This aspect is divided into three levels: Technology and product managers, business promotion, and information integration.
iv. Marketing Model – Refers to the marketing model that the unit can promote based on the unit's existing business foundations. This aspect is divided into three levels: One-on-one technology promotion, one-on-many strategy groups, and many-on-many internet marketing models.

The black lines in Picture 5 delineate the role of research and development units. As managers of technology and products, research and development units choose customer portfolios on a market-by-market basis and develop products or technology based on their market potential. The main marketing method is promoting the technology or products to the market as a whole.

The red lines in Fig. 5 delineate the role of centers for business promotion within research and development units. The main goal in promoting core products or technologies is to form cooperative strategy groups with appropriate companies or organizations in order to lower development costs, pioneer markets, and increase the partners' market value and competitiveness.

The blue lines in Picture 5 delineate the role of customer service centers. They integrate information and provide extended services with an emphasis on promoting internet product marketing. They serve as liaisons between research and development units and individual customers and play a role discovering potential customers.

5. Conclusion

Using content analysis and the Analysis Services decision analysis tool, this study developed an easily executed model and completed implementation of customer relationship management. The process of implementing this model not only introduces participating personnel to the concepts of customer relationship management, but also provides an actual foundation for building a customer relationship management system.

The purpose of this study is to find ways to study text data in order to discover more latent knowledge. In the example of Company A's customer service center, content analysis was used to process text data, part of the structured data was integrated into a miniature data warehouse, OLAP analysis and decision tree criteria were used to discover customer knowledge, and, finally, marketing strategies of customer relationship management models were suggested based on these criteria. However, in this example, there was not a good information system in place, and the structured data was sparse and overly dispersed. Data mining did not yield any significant discoveries, so the data analysis was indeed cursory. Therefore, the study's recommendations still focus on the execution process of complete customer relationship management and on establishing a more complete system loop in order to reinforce interactions with customers.

The contributions of this study are as follows:

i. The categories established by classifying text data can be used in follow-up studies.
ii. The process of using content analysis to transform text data into structured data can be used to supplement training of system operation personnel.
iii. An implementation model for customer relationship management was established; the model is particularly useful for smaller businesses that have incomplete information systems.

5.1. Recommendations for Company A

The recommendations for Company A are divided into three categories as follows:

i. Management of text Data – The study's empirical process revealed that not only is content analysis useful in organizing and analyzing text data, its reliability testing methods can also be used in future education and training for customer service systems. When training new personnel, after explaining the classification of current customer service data, reliability can immediately be tested. If the reliability is acceptable, it indicates that the new personnel have obtained sufficient understanding of the customer service process; if, after many tests, the reliability is still not up to standards, it indicates that there is a discrepancy between the new personnel's and the current personnel's understandings of the current classes. The latter can mean two things: The new personnel do not have sufficient understanding of the customer service process and are unsuitable for data classification tasks; or, some customer data are currently difficult to classify, and the system's current customer data classes need to be examined and modified if necessary.
ii. Analysis of Complaints – Company A currently still relies mainly on strategy groups to promote its business. If the current corporate-level customers have complaints, they usually contact the unit with which they are working. Therefore, although handling of complaints is usually the main task of customer service centers, data analysis indicates that this is not the case for Company A. Thus the role played by its customer service center is not complaint management but more along the lines of business promotion, such as providing progress updates, related data, and potential technological cooperation on the development of new technology, etc. Unless Company A plans to promote its own products, the customer service center's main responsibility should still be that of handling normal complaints.
iii. Operating Model of the Customer Service System – The data is overly dispersed on Company A's website and it is difficult to perform searches. This is made evident by the analysis of the overall business, where it is clear that its main needs are informational. This is precisely the task that the customer service center can easily handle. Therefore, the establishment of a system for searching online publications should be a priority task in digital marketing. During the beginning construction stages, the customer service center can beta-test the system internally; after beta testing is completed, the system can be unveiled to customers. This system can be used to recruit new members and systematically collect customer data by classifying according to membership and whether payment is collected, remedying the current deficiency of insufficient basic customer data. The system can also be integrated with the electric news system, allowing the customers themselves to choose whether to subscribe to electric news and the types of electric news to which they subscribe. This provides customer classification, and the electric news can actively market publications to various customer groups. Currently, the electric news carries information about conference activities, but do not contain information about conference attendance and revenue, proving that one-way communication of information is unsatisfactory. The necessity of reporting conference activities can be examined from the perspective of creating repeated interactions between the company and its customers. If the online conference registration system or related online surveys can be further integrated, creating a two-way communication channel, in addition to enhancing online reporting of conference information or content, additional customers can be attracted to the conferences. The integrated system can now actively reach out to conference attendees, and an expanded customer base will result in increased subscriptions to electric news and more effective digital marketing. This establishes the beneficial cycle of customer relationship management as seen in Fig. 6.

Follow-up studies can be considered in two areas: (1) If the main focus is on structured data, and the secondary focus is on text data, an appropriate customer relationship management system should be sought out in order to obtain more complete structured data such as detailed data

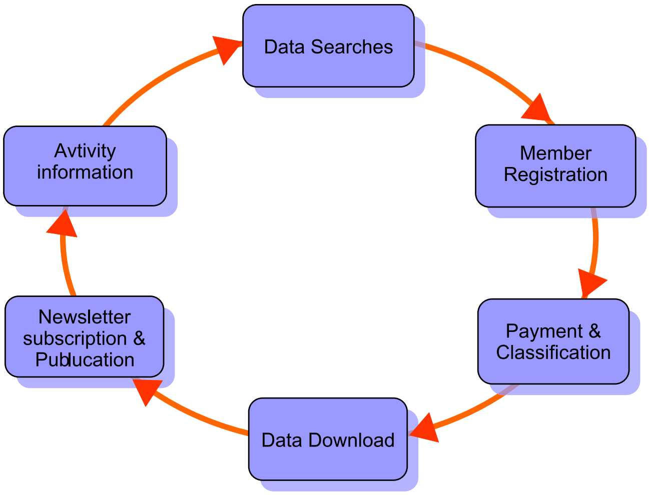

Fig. 6. Operating model of the customer service system.

on technology transfers, more complete basic data of businesses, and integrated text data such as customer feedback, etc. The results of analysis should be even more precise, the investments in customer value and customer relationship management should yield evaluation criteria that are even more valuable, and digitization and data mining can be targeted for further study. (2) If the main focus is on text data, and the secondary focus is on structured data, there should be increased quantification of text data and large amounts of text data should be studied. The resulting analysis models or customer relationship management implementation models should be even more appropriate for smaller business with insufficient information systems.

References

Aha, D. W., Kibler, D., Albert, M. K., & Albert (1991). Instance-based learning algorithms. *Machine Learning, 6*, 37–66.

Chau, R., & Yeh, C. H. (2004). A multilingual text mining approach to web cross-lingual text retrieval. *Knowledge-Based Systems, 17*(5–6), 219–227.

Gates, B. (1999). *Business @ the speed of thrust: Using a digital nervous System*. Warner Books, Inc.

Holsti, O. (1969). *Content analysis for social sciences and humanities*. Reading, MA: Addison-Wesley.

Hunt, E. B., Marin, J., & Stone, P. J. (1966). *Experiments in induction*. New York: Academic Press.

Köhler, J., Philippi, S., Specht, M., & Rüegg, Al. (2006). Ontology based text indexing and querying for the semantic web. *Knowledge-Based Systems, 19*(8), 744–754.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Newbury Park, Califormia: Sage Publications.

Linoff, G. S., & Berry, M. J. (2002). *Mining the web, transforming customer data into customer value*. New York: John Wiley.

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism and Mass Communication Quarterly*.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufman.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems With Applications, 33*(1), 135–146.

Wayland, R. E., & Cole, E. M. (1997). *Customer connections: New strategies for growth*. Havard Business School Press.

Yang, H. C., & Lee, C. H. (2005). A text mining approach for automatic construction of hypertexts. *Expert Systems With Applications, 29*(4), 723–734.

Please cite this article in press as: Chang, C.-W. et al., Mining the text information to optimizing the customer ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.11.027