Application of Text Mining for Customer Evaluations in Commercial Banking

Jing Tan
Department of Computer and Information Science, Temple University, Philadelphia, 19121, United States

Xiaojiang Du
Department of Computer and Information Science, Temple University, Philadelphia, 19121, United States

Pengpeng Hao
Department of Development Planning, China Minsheng Bank, No. 2, Fuxingmennei Avenue, Xicheng District, Beijing, 100031, China
Institute of Finance and Banking, Chinese Academy of Social Sciences, No. 5, Jianguomennei Dajie, Dongcheng District, Beijing, 100732, China

Yanbo J. Wang
Department of Development Planning, China Minsheng Bank, No. 2, Fuxingmennei Avenue, Xicheng District, Beijing, 100031, China
Institute of Finance and Banking, Chinese Academy of Social Sciences, No. 5, Jianguomennei Dajie, Dongcheng District, Beijing, 100732, China

Abstract

Nowadays customer attrition is increasingly serious in commercial banks. To combat this problem roundly, mining customer evaluation texts is as important as mining customer structured data. In order to extract hidden information from customer evaluations, Textual Feature Selection, Classification and Association Rule Mining are necessary techniques. This paper presents all three techniques by using *Chinese Word Segmentation*, *C5.0* and *Apriori*, and a set of experiments were run based on a collection of *real* textual data that includes 823 customer evaluations taken from a Chinese commercial bank. Results, consequent solutions, some advice for the commercial bank are given in this paper.

Keywords: classification; decision tree; commercial banking; association rule mining; customer evaluation.

1. Introduction

Nowadays there are increased competition between Chinese commercial banks, so if the bank can attract or maintain customers, also known as customer retention, is become a critical issue. Besides, Kandampully and Duddy attempt to clarify that attracting a new customer is about five times more costly than retaining an existing customer. Hence the retention of existing customers has become a priority for businesses to survive and prosper [1].

Most of research about this area is mainly focus on mining profits and capital, and then using formulas and predictions to analysis. But actually the comment of current customers is another direction needed to pay attention to.

Therefore, text mining of customer evaluations is a proper approach, so that customer needs can be obtained clearly (i.e. characteristics of clients and their demand, attitude about business, etc.). As a consequence, we are able find out the reasons about customer attrition, thus we can make improvements, and in some case this probably retains/returns customers (as shown in Figure 1). Last but not least, text mining is fairly sophisticated in English-speaking country, but in Chinese, there are still rooms for advancement.

The rest of this paper is organized as follows. The following section indicates related work in the previous studies. Section 3 presents the methodology of word segmentation in Chinese, classification and association rule mining by *C5.0* and *Apriori*. Results and some analysis are shown in Section 4. Finally, our discussions and direction for future work are given at the end of this paper.

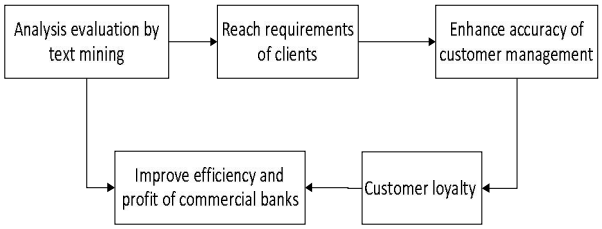

Figure 1. Cause-effect Diagram of Mining Customer Evaluation Texts

2. Related Work

2.1 Chinese Word Segmentation

Textual Feature Selection is an important step of preprocessing in text mining. It is not that difficult for *abc-based* languages, i.e. English, because there are spaces between words in general. However for other languages, i.e. Chinese, there is no space between characters or words. Hence to find appropriate positions to segment Chinese textual data, in order to obtain word-terms, is a complex problem. Before of doing text mining (for Chinese textual data), it is necessary to cut a sentence into small units, that is, *Chinese Word Segmentation.* As computer input, a Chinese sentence turns to be a single character will lose

text information. So in order to restore the information, the technique of Chinese Word Segmentation is addressed [2].

Three basic techniques for Chinese Word Segmentation are described as follows:

1. **String Matching Based Word Segmentation:** In this method, every character in a sentence will be divided into a predefined lexicon. If some continuous character matches a string in the lexicon, then it means this string can be recognized as a word.
2. **Semantic Understanding Based Word Segmentation:** In this method, computer will has ability to identify words by understanding a sentence like a human being. The basic idea is to use word segmentation and analysis grammar & semantic at the same time. There will be considerably amounts of language knowledge and information by using this approach.
3. **Statistical Probability Based Word Segmentation:** This method is relied on the probability of adjacent co-occurrence of a series of Chinese characters (usually 2 or 3) as words. If this probability is higher than a threshold, then this string will be considered to be a word [3].

2.2 Classification Approaches

In a commercial banking aspect, most customer prediction problems (i.e. identification of customer attrition risk, identification of customer credit risk, etc.) can be treated as a *binary* classification task in data mining. Wang *et. al.* [4] indicated that the common approaches of Classification in the past decades can be generally grouped into 5 models (described as follows):

1. **Artificial Based Classification (ABC) Model:** By using *Artificial Intelligence* techniques to solve problem of classification, e.g. Artificial Neural Network (ANN).
2. **Bayesian Based Classification (BBC) Model:** By using the *Bayesian* theory to solve problem of classification, e.g. *Naïve* Bayes (NB).
3. **Case Based Classification (CBC) Model:** By using training samples *directly* to solve problem of classification, e.g., *k*–Nearest Neighbors (*k*–NN).
4. **Tree Based Classification (TBC) Model:** Basic idea about this technique is greedy algorithm. The result of classifier construction using TBC is to build a *Decision Tree* (*DT*), e.g. C4.5/C5.0.
5. **Regression Based Classification (RBC) Model:** By using the statistical regression study to solve problem of classification, e.g. Logistic Regression (LR).

In the previous studies of customer attrition risk identification, Khan, Jamwal and Sepehri [5] indicated that TBC • DT, RBC • LR, and ABC • ANN can be almost equally well applied.

2.3 Association Rule Mining

Association Rule Mining (ARM) is a major data mining mechanism, which has been widely using by different business areas. It can be employed to find opportunity for customer-product cross-selling. For instance, in a retail business

aspect, in order to marketing a new brand of beer, ARM suggests us to target on male customers who also purchase diapers. This can not only lower the cost of marketing, but also improve the successful rate. In [6] ARM is applied in certain issues in commercial banks. The most widely used ARM approach is *Apriori*.

3. Methodology

3.1 Chinese Word Segmentation: N-gram

Considering about the complexity in Chinese, a combination of *string matching based word segmentation* and *statistical probability based word segmentation* is a better strategy, that is, *N-gram.* An N-gram is a sub-sequence of *n* items from a given sequence, and the items in question can be characters, words or base pairs according to the application. In our study, 2-gram or 3-gram was used to achieve the target, since for Chinese, it's enough to be meaningful and find rules.

After N-gram, there may be more than 4000 word-terms for about 800 customer evaluation texts. So in order to find specific rules in them, some meaningless words which only contain single character and such words only show up once or twice were all deleted. Consequently, there were 900 some words for further research. The algorithm flow chart of N-gram technique is attached in Figure 2.

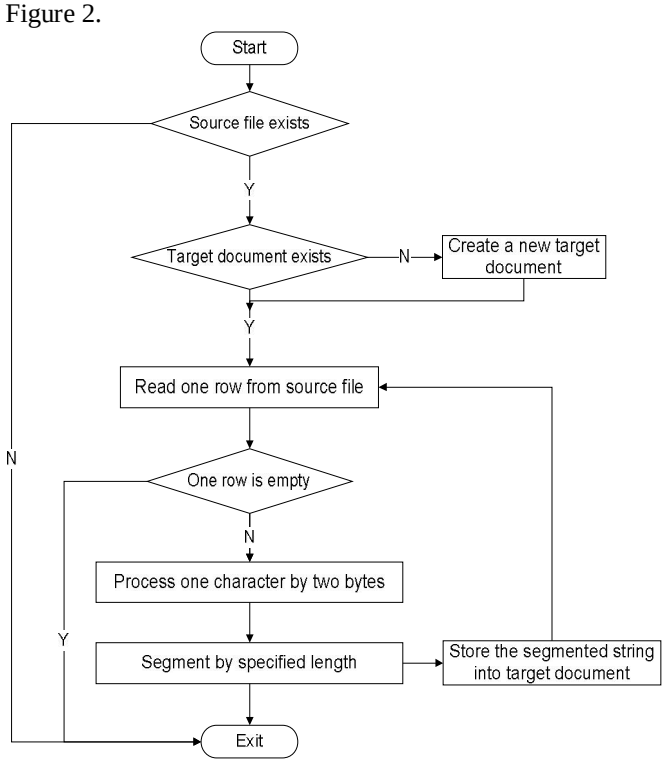

Figure 2. Algorithm Flow Chart of N-gram Technique

3.2 Decision Tree Classification: C5.0

Converting 900 some words from Chinese Word Segmentation into structured data, and turn these words into data attributes. In order to research on why customer complaints on "credit card business" (Other banking business, i.e. "counter service", "deposit card", etc. can be equally applied), the evaluation texts that contain the word "*credit card*" were labeled as *class A*, the rest of evaluation texts were labeled as *class B*. Then all data was separated into two sets randomly: 70% of the data are in training set and the rest 30% in test set. Finally, the data in training set was used to build a classification model by TBC • DT C5.0, and the test set was used for model evaluation. The process of classification analysis is shown in Figure 3.

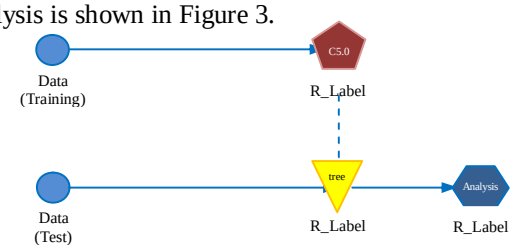

Figure 3. Process of Classification Analysis

3.3 Association Rule Mining: Apriori

Deleting the entire label in classification (i.e. shown as "*R_label*"), and still using the 933 words (after Chinese Word Segmentation), we then employ *Apriori* algorithm in ARM to find hidden business patterns. The process of association analysis is shown in Figure 4.

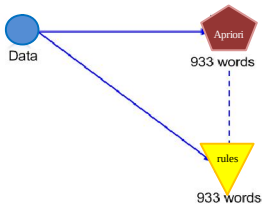

Figure 4. Process of Association Analysis

4. Results

4.1 Result of Chinese Word Segmentation

Original textual data are composed of natural language (collected from a *real* banking situation), and there are all sentences in Chinese. After *N-gram*, each sentence is separated into word-terms which are split by comma. As shown in Figure 5 and Figure 6.

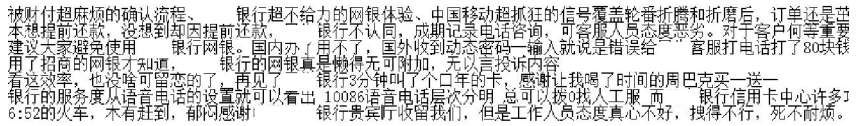

Figure 5. Original Textual Data

Figure 6. Textual Data after Chinese Word Segmentation by N-gram

4.2 Result of Classification

Figure 7 demonstrates the result of structured data and class-labels. There are more than 4,000 word-terms after N-gram, but only these words with more than two characters and also show up more than three times are further used. So when converting textual data into structured data, there are only 933 words in total, in other words, 933 data attributes/variables. Moreover, there are 212 samples that contain the word *"credit card"* (of a total of 823 evaluation texts), and the percentage for training set and test set is 70% *vs.* 30%. Part of the build decision tree is shown in Figure 8. Table 1 and Table 2 show the analysis of classification result, where the accuracy can be reached higher than 92%.

932 Illegal Use	933 Airplane	R_Label	Sets	$C-R_Label
N	N	B_Non-CreditCard	1_Training	B_Non-CreditCard
N	N	A_CreditCard	2_Test	B_Non-CreditCard
N	N	A_CreditCard	2_Test	A_CreditCard
N	N	A_CreditCard	1_Training	A_CreditCard
N	N	A_CreditCard	2_Test	B_Non-CreditCard
N	N	B_Non-CreditCard	1_Training	B_Non-CreditCard
N	N	A_CreditCard	1_Training	A_CreditCard
N	N	B_Non-CreditCard	1_Training	B_Non-CreditCard
N	N	B_Non-CreditCard	1_Training	B_Non-CreditCard
N	N	B_Non-CreditCard	1_Training	B_Non-CreditCard

Figure 7. Structured Data and Class-label

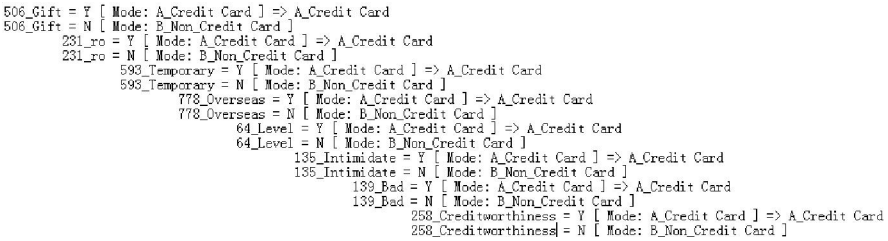

Figure 8. The Built Decision Tree

Table 1. The overall Accuracy of Classification Analysis

Accuracy	762	92.59%
Inaccuracy	61	7.41%
Total	823	100%

Table 2. The Confusion Matrix of Classification Analysis

	A_Credit Card (*predicted*)	B_Non-Credit Card (*predicted*)
A_Credit Card (*given*)	154	58
B_Non-Credit Card (*given*)	3	608

4.3 Result of Association Rule Mining

After using *Apriori*, Table 3 shows a part of result of mining association rules. Three of the most important and interesting association rules are "Menu & Telephone à Customer Service", "Credit Card à Telephone" and "Milk à Credit Card".

Table 3. Part of results of association rules

Consequent	***Antecedent***	***Support***	***Confidence***	***Lift***
Credit Card	Application Complaints	1.458%	91.667%	3.971
Credit Card	Milk	1.458%	83.333%	3.61
Customer Service	Menu Telephone	3.402%	100.0%	6.282
Customer Service	Unexpected Complaint Credit card	1.215%	97.2%	5.026
Customer Service	Consultant Credit Card	0.851%	100%	6.282
Customer Service	Process Staff	0.972%	100%	6.282
Telephone	Customer Personal	0.851%	100%	5.31
Telephone	Odious Credit	0.851%	71.429%	3.793
Telephone	Credit Card	2.673%	72.727%	3.862
Business	Counter Transaction	1.094%	88.889%	7.701
Queue	Counter Service	0.851%	71.429%	8.645
Attitude	Odious Staff	0.851%	85.714%	11.96

5. Discussions

Before of our investigation, no one would believe that one major complaint in credit card business (for our studied bank) is the "milk gift". In fact, the "milk gift" is a kind of reward for such customers who use the bank's credit card frequently. The aim of providing a proper gift (i.e. the "milk gift") is to enhance

loyalty of good customers. As an add-valued service, which actually costs nothing from customers, hence the bank would never consider about customer complaints on this issue. However, it seems that customers also quite care about service quality when accepting free gift. From our study, we discover that "free service ≠ poor service", and this point is verified, confirmed and accepted by the bank. By constructing the decision tree, there is another variable significantly related to the "credit card" complaint, which mentions a specific area in China (the "Henan province"). This tells the bank that there are more problems with credit card business in the Henan branch rather than other branches.

On the other hand, by using association rule mining, we find some other problems. "Customer service" is becoming a major issue, where the *support* value of "customer service" is higher than other variables. Herein, major types of "customer service" in complaints are "counter service" and "after-service". Moreover, "counter" and "service" imply "queue". Staffs in the bank always thought the efficiency of "ATM" is low, they intended to add more devices to enhance the efficiency. But actually, complaints about "queue-problem" are from the "counter" rather than the "ATM". Another rule seems strange at first is that "milk" and "gift" are both associated with "credit card", as we described above. After making confirmation with the bank, we know that "milk" is a "gift" for credit card customers, and this value-added service turns out to be a major concern in complaints.

Based on these classification and association rules (as mentioned above), the bank finds correct directions to make improvements. Besides, based on our evaluation system, if there is a new customer complaint coming, it will be sent automatically to appropriate departments, i.e. credit card center, Henan branch, department of retail banking, etc.

References

[1] Kandampully, J. and Duddy, R. (1999): Relationship Marketing: A Concept beyond Primary Relationship. Marketing Intelligence and Planning, 17(7), 315-323

[2] Zhuang, X. (2010): Application in Chinese Word Segmentation in Computer. Journal of Hulunbeier College, 18(3)

[3] Hu, T. and Wang, J. (2006): Research in Chinese Text Classification Techniques, Consulting Herald in Science and Technology, 9

[4] Wang, W., Wang, Y.J., Xin, Q., Bañares-Alcántara, R., Coenen, F. and Cui, Z. (2011): A Comparative Study of Associative Classifiers in Mesenchymal Stem cell Differentiation Analysis. In the book Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains, 223-243

[5] Khan, A.A., Jamwal, S. and Sepehri, M.M. (2010): Applying Data Mining to Customer Churn Prediction in an Internet Service Provider. International Journal of Computer Applications, 9(7), 8-14

[6] Wu, Q. (2008): Certain Issues applied in Commercial Banks by using Association Rules Analysis, Journal of System Simulation, 20(8)