Advance Journal of Food Science and Technology 7(1): 49-53, 2015
DOI:10.19026/ajfst.7.1265
ISSN: 2042-4868; e-ISSN: 2042-4876

Submitted: August 31, 2014 Accepted: September 20, 2014 Published: January 05, 2015

Research Article

Rice Products Feature Analyzing on the Base of Online Review Mining

Xixiang Sun, Xiaoqing Song and Xiangdong Liu
Wuhan University of Technology, Wuhan 430070, China

Abstract: This study aims to investigate the rice product features using online review mining method. The opinion mining is used to do online review data analyze. High-frequency words were extracted from non-structured online reviews text. Then the rice product features were gotten from the factor analysis on the base of high-frequency words. This provided a new method for product feature analyzing which was based on the data mining of online reviews. The proposed method can be used to compare features of rice products which were in the similar category. It also provided new views for understanding consumer brand knowledge. At the end of this study, the online review of rice products were used to verify the scientificity and rationality of this method.

Keywords: Data mining, online review, opinion mining, rice products

INTRODUCTION

Authoritative statistics from the China Internet Network Information Center shows that online customer reviews are the most important consideration in making purchase decision for online shoppers of all ages. Since Amazon first launched its online reviews system, various retail sites following launched its own review system. The reason of these activities is that online reviews can effectively promote product sales and create profits. Online customer reviews has become one of the key considerations in e-marketing, customer relationship management and brand management. Using artificial methods to do data processing was inefficient and not practical in the condition of massive online reviews. It must resort to the rapid development information technology. Opinion mining was a very suitable technology for massive online consumer review information analysis. By opinion mining, consensuses and focus can be extracted from the unstructured reviews and the product features were get form the perspective of consumers. This was good for manufacturers and markers in understanding consumer opinion in a more comprehensive and accurate way. It also provides objective basis for improving product performance and making marketing strategies.

Opinion mining, which was also known as review mining or sentiment analysis, refers to the process of analyzing, processing, inducing and deducting of the text which contains of subjective feelings. Kim and Hovy (2004) proposed that opinion contains four semantic components: Topic, holder, claim and sentiment. Wiebe *et al*. (2004) used the parts of speech, i.e., the pronouns, adjectives, cardinal, modal verbs and adverbs, punctuation and sentence position as eigenvalues to design the sentence-level classifier and then also did research on classification method on the chapter level. After that the studies of opinion mining algorithms become a hot topic in researches. Nevertheless, there were a lot of differences in semantic identification and analysis between Chinese context and English context. These research findings could not been directly applied in Chinese context.

The research of text mining at domestic began later then aboard, but there were still a lot of meaningful findings. Wang *et al*. (1999) gave a definition of text mining. And they also gave a systematic discussion about the processing, functions and methods of text mining. Wang *et al*. (2010) divided the method of discriminating sentimental polarity into two types: one method is based on the identification of semantic characteristics of emotional words and another method is based on the identification of statistical natural language processing. Xia *et al*. (2006) had studied on informal Chinese network language. In terms of the research about analysis of sentimental polarity and opinion extraction of Chinese sentences, Lou and Yao (2006) did semantic polarity analyze and opinion extraction of Chinese online reviews by using natural language processing technology. They gave an algorithm for calculating polarity of words in context. In the study, they also analyzed the matching relationship between the subject and the polar modifier components. In terms of research on the development of opinion mining system, Lou and Yao (2006) developed an opinion mining system which can be applied in various e-WOM platforms. They tested the efficiency of the system by reviews of car brands. The system could be used to extract online reviews and opinions for each car brand and judge appraise and

Corresponding Author: Xixiang Sun, Wuhan University of Technology, Wuhan 430070, China

sentimental strength of the opinion. However, all these studies had not focus on rice product features mining for the purpose of using online reviews to get product features from the consumers' perspective and making more efficient marketing strategy (Robson *et al*., 2013).

In order to investigate the rice product features mining, this study proposed a new method based on the online opinion mining. In the analysis, 2000 reviews of FuLinmen rice on JD.com and 2000 reviews of FuLinmen rice on T-mall.com were chosen to test the proposed method. The analysis results have shown that the new method was good at getting product features form consumers' perspective. Comparing with the method that based on questionnaires, the online opinion mining was more convenient, high-efficiency and accurate.

MATERIALS AND METHODS

The process of opinion mining of online reviews: The general flow of web-text mining as shown in Fig. 1 can be divided into three steps: acquisition and collection, pretreatment, text clustering and visualization. The first work was text feature extraction. Following, text representation and optimization should be done. At last text mining algorithms was used to acquire knowledge that contained in the text (Chen and Zhang, 2007). The knowledge can meet the reader's need and become useful knowledge of guiding people's practice.

There are several typical Chinese automatic text analysis software, for example, TRS CKM, ROST content mining, ROST Word Parse and so on (Zhao *et al*., 2010). In this study, we choose ROST, a free software package which was developed by Professor Shenyang, to identify emotional tendency of the review text. Specific steps applications were following:

- Select samples, including training samples and testing samples. For this study, subjective evaluation reviews and objective description reviews, subjective expression reviews and objective expression reviews, positive emotional tendencies and negative emotional tendencies reviews should be all selected.
- Building text feature classifiers with dynamic language model of Java programming.
- Training classifier with training samples that had been marked classes.
- **Classifier performance evaluation:** Use other test samples to assess the efficiency of the classifier.
- Analyze emotional features of the text. Classify emotions and calculate probability of emotional tendencies of unknown predictive text with trained classifier.

In this study, training set and test samples that needed in text feature identification were from two sources: corpus which was given by industry experts and online reviews form T-mall.com and JD.com. They were two major domestic online retail platforms. We take rice products as the object of study, the specific processes are as follow:

- For Classifying the review text into subjective comments or objective description, rice products training corpus was established in ROST. And then reviews of rice products that displayed on the first 10 pages in T-mall.com and JD.com were retrieved to be training and testing corpus. After that, ambiguous sentences were removed from the training set. At last, we get 1400 reviews which most were positive view as final training and testing corpus.
- For classifying the expression way of reviews into subjective reviews and objective reviews, 500 messages that were objectively described or subjectively expressed were randomly selected from T-mall.com as the training and testing corpus.
- For classifying emotions into positive and negative, since each review on T-mall had its emotional tendency score, it need not calculate the probability of emotional tendencies with the method of text mining. However, the mix degree of positive and negative emotion still need been gained by text mining. One thousand positive and negative reviews were selected from rice products reviews on T-mall.com to be training and testing corpus.

The working process of training and testing samples and selecting text feature with ROST were shown in Fig. 2.

Following the steps, we calculated the accuracy of the three classifiers as 92.18, 90.23 and 85.56%, respectively indicating that three classifiers were effective.

Rice product features text mining: Reviews of the high rating rice products which were sold on T-mall.com and JD.com were selected as data source. The selection principle was that the cumulative number of online reviews should more than 10,000. Reviews text form date October 1, 2013 to December 31 was retrieved by the text mining software ROST from the website.

According to these principles, 2000 reviews of FuLinmen rice on JD.com and 2000 reviews of

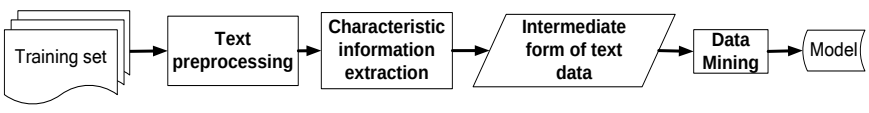

Fig. 1: The general flow of web-text mining

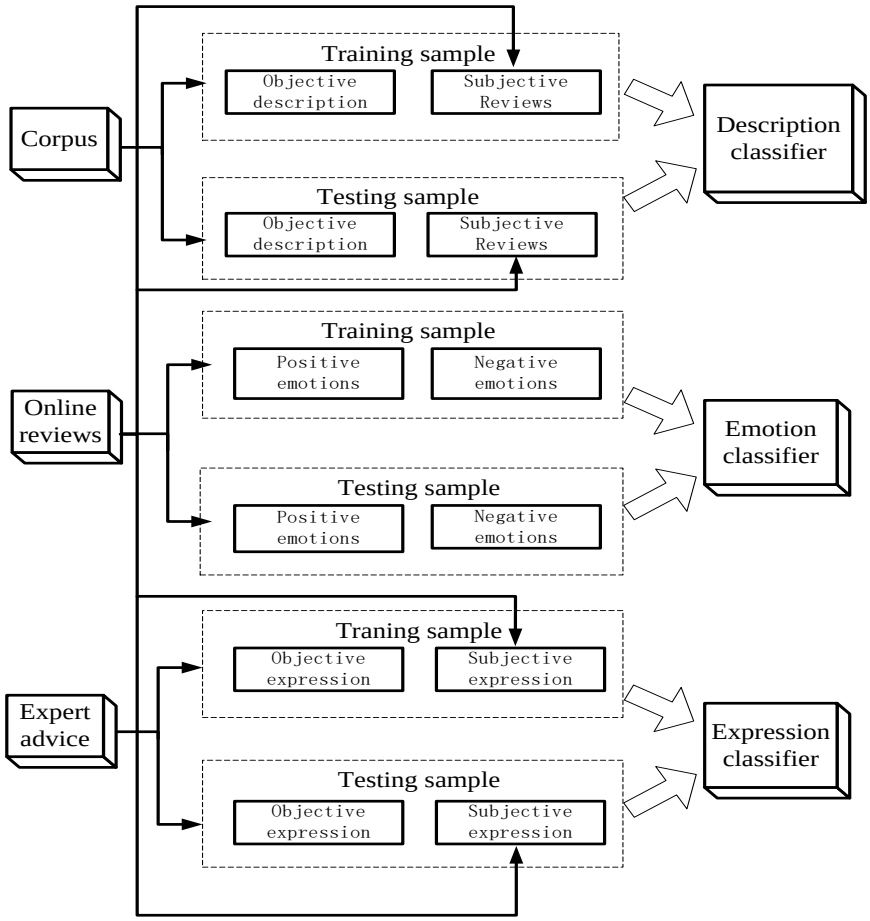

Fig. 2: The working process of text feature classifier

Table 1: Fifteen high-frequency words in two retail platform

Rank	JD.com	Frequency/thousand	T-mall. com	Frequency/thousand
1	Taste	0.232	Taste	0.328
2	Packing	0.124	Packing	0.226
3	Fresh	0.105	Fresh	0.408
4	Plump	0.008	Plump	0.053
5	Just so so	0.152	Just so so	0.255
6	Quality	0.231	Quality	0.833
7	Overall feeling	0.189	Overall feeling	0.254
8	Clean	0.110	Clean	0.758
9	Cheap	0.311	Cheap	0.225
10	Expensive	0.124	Expensive	0.456
11	Delivery	0.040	Delivery	0.422
12	Non-GM	0.103	Non-GM	0.866
13	Speed of logistics services	0.250	Speed of logistics services	0.722
14	Convenience	0.123	Convenience	0.031
15	Big grain	0.013	Big grain	0.025

Table 2: In order to make total variance explained

	Initial eigenvalue			Cumulative variance loading			Rotated cumulative variance loading		
Factor	Eigenvalue	Proportion	Cumulative	Eigenvalue	Proportion	Cumulative	Eigenvalue	Proportion	Cumulative
1	9.963	76.677	76.677	9.963	76.677	76.677	8.230	61.780	61.780
2	1.969	15.160	91.837	1.969	15.160	91.837	2.812	23.165	84.945
3	1.024	7.787	99.624	1.024	7.787	99.624	1.954	13.725	98.670
4	0.050	0.232	99.856						
5	0.024	0.130	99.986						

FuLinmen rice on T-mall.com were chosen to be dealt with ROST. All reviews were transfer into TXT format documents by the text mining software. All similar words in reviews would be expressed in a unified word. That was data preprocessing. The amended words were saved in a new TXT file which can be identified by the word-frequency analysis software. Then, ROST Content Mining and ROST Word Parse statistical software were used to do content analysis for the amended TXT document. The prepositions, conjunctions, auxiliary and other words which were common but independent with product description reviews were filtered by the software and get 15 same high-frequency feature words and their frequencies as shown in Table 1.

RESULTS AND DISCUSSION

Although opinion mining of the text have refined the original reviews, there were some thesauruses in the high-frequency words that extracted by the software. That means the product features we get from the software were not refined enough. In order to get more refined product features, factor analysis method was taken to determine the key factors.

First, it's standardization of the 15 high-frequency words which were extracted by opinion mining. High-frequency word "cheap" and "expensive" were merged as "cost", "plump" and "big grain" were merged as "grade", "overall feeling" and "just so so" were merged as "customer satisfaction". Then 12 merged high-frequency words were: taste, packaging, fresh, grade, quality, clean, cost, lustre, health, customer satisfaction, speed and convenience of logistics services.

Second, it's correlation analysis. The correlation analysis was done by SPSS software. The results showed that in the 144 data values of the C12*12 correlation matrix, 132 data values were greater than 0.3. They were accounting for 92.3% of the total correlation data. And the result of KMO test and Bartlett test of sphericity were that KMO value was 0.893 which was greater than 0.7 and Bartlett's test was under the significant level of 0.001. That means the sample data was suitable for factor analysis.

Third, it's factor extraction. Eigenvalues and eigenvectors were calculated by solving the characteristic equation, as shown in Table 2. It was seen in the table that three principal factors were met by calculating and their cumulative contribution of variance exceeded 80%. It indicated that these three factors have expressed most of the information that contained in the 13 indicators.

In order to make the factors to be more interpretive, maximum variance method was adopted to do orthogonal rotation on the factor loading matrix. After four times rotations, the rotated factor loading matrix was shown in Table 3.

Table 3: Rotated factor loadings

	Factors loadings		
Variable	F1	F2	F3
Taste (X_1)	0.933	0.059	0.161
Packaging (X_2)	0.916	0.109	0.223
Fresh (X_3)	0.906	0.129	0.251
Clean (X_4)	0.901	0.141	0.266
Quality (X_5)	0.894	0.174	0.270
Luster (X_6)	0.894	0.204	0.246
Cost (X_7)	0.890	0.182	0.976
Customer satisfaction (X_8)	-0.002	0.945	-0.002
Convenience (X_9)	0.275	0.903	0.062
Speed of logistics services (X_{10})	0.542	0.613	0.462
Quality (X_{11})	0.246	0.046	0.912
Health (X_{12})	0.628	0.086	0.699

Finally, it's calculation of factor scores. Factor score were calculated by SPSS. The factor scoring equations were:

$$F1 = 0.933X_1 + 0.916X_2 + 0.906X_3 + 0.901X_4 + 0.894X_5 + 0.894X_6 + 0.890X_7 - 0.002X_8 + 0.275X_9 + 0.542X_{10} + 0.246X_{11} + 0.628X_{12} \quad (1)$$

$$F2 = 0.059X_1 + 0.109X_2 + 0.129X_3 + 0.141X_4 + 0.174X_5 + 0.204X_6 + 0.182X_7 + 0.945X_8 + 0.903X_9 + 0.613X_{10} + 0.046X_{11} + 0.086X_{12} \quad (2)$$

$$F3 = 0.161X_1 + 0.223X_2 + 0.251X_3 + 0.266X_4 + 0.270X_5 + 0.246X_6 + 0.976X_7 - 0.002X_8 + 0.062X_9 + 0.462X_{10} + 0.912X_{11} + 0.699X_{12} \quad (3)$$

Comprehensive factor score equation was:

$$F = 0.618F_1 + 0.232F_2 + 0.137F_3 \quad (4)$$

The results of the rotated factor loading matrix indicated that 7 high-frequency online review words including "taste", "packaging", "fresh", "clean", "grade", "lustre" and "cost" were heavily loaded on common factor Fl. So, factor Fl should be treated as key factor and we named it as: the quality of goods. Meanwhile, "customer satisfaction", "convenience" and "speed of logistics services" were 3 high-frequency words that heavily loaded on common factor F2. F2 were treated as important factor and named sale service. "Quality" and "health" were greatly loaded on the F3. F3 were treated as second important factor and named: food safety.

Based on the above analysis, it verified that this research method was good at getting product features form consumers' perspective. Comparing with the method that based on questionnaires, it was more convenient, high-efficiency and accurate. It provided objective basis for manufacturing enterprise to rely on when they were willing to improve product

performance and service. It's a more effective and targeted approach. In this study, for the specific industry, manufactures and enterprises which produce or sale rice products should take their primary attention to the grade of rice products? The probable reason may be that customers were generally required relatively high on the grade of the rice. Meanwhile, their judgment mostly relied on the appearance of the rice products and packaging. Hence, improving the product quality and appearance should be known as breakpoints of the industry competition (Zhu *et al*., 2007). Meanwhile, if this method was applied to study specific product, consumer product knowledge and satisfaction of the product could be accurately grasped. In this study, for the rice product in T-mall.com, taste, packaging, fresh, clean, grade, luster and cost were the standard that customer used to evaluate the product. So, product description which was given by online seller should highlight the characteristics of products in these areas. And the "just so so" satisfaction rating reflected that most consumers who have bought the products were not satisfied with the marketing service. It should be upgraded from the terms of service attitude, service professional and logistics speed. And when this method was used to compare the selling of similar products in different platforms, it would be helpful for manufacturers and marketers in choosing marketing platform and marketing mix strategy. In this study, for the specific platforms, JD.com got a higher comprehensive factor score than T-mall.com. It indicated that, for those rice products which have higher service quality, manufacturers and marketers would better to choose JD.com as their online retail platform on which the rice products have average higher quality and service. Manufacturers and marketers would take advantage of this platform to promote sales and do brand management, too. And for those rice products which did not have higher service quality at present moment, T-mall.com would be a more suitable platform for manufacturers and marketers to promote sales.

CONCLUSION

In this study, we attempted to combine text mining technology and factor analysis to do online reviews research. This method provides new ideas and approaches for companies who would like to know their product competitiveness and marketing environment through the online consumer feedback system. However, this study also has some limitations. Such as limited data resources and different companies and different product categories also require different types of corpus. Different mining software and different mining algorithms also may come to different conclusion and so on. All these are required further study and test.

ACKNOWLEDGMENT

This study was financially supported by the National Natural Science Foundation of China (No. 71172042) and Wuhan University of Technology Innovation Fund (No. 2012-IB-092).

REFERENCES

Chen, Z. and G. Zhang, 2007. Study on the text mining and Chinese text mining framework. Inform. Sci., 25(7): 1046-1051.

Kim, S. and E. Hovy, 2004. Determining the sentiment of opinions. Proceeding of the COLING Conference. Geneva, Switzerland, pp: 1-8.

Lou, T. and T. Yao, 2006. Semantic polarity analysis and opinion mining on Chinese review sentences. Comput. Appl., 26(11): 2622-2625.

Robson, K., M. Farshid, J. Bredican and S. Humphrey, 2013. Making sense of online consumer reviews: A methodology. Int. J. Market Res., 55(4): 521-537.

Wang, J., Y. Sun and Y. Zhang, 1999. Text mining: A new research issue of data mining. J. Lanzhou Univ., 35(8): 314-318.

Wang, H., L. Xie, P. Yin and G. Liao, 2010. Literature review of sentiment classification on web text. J. China Soc. Sci. Tech. Inform., 29(5): 931-938.

Wiebe, J., T. Wilson and R. Hwa, 2004. Learning subjective language. Comput. Linguist., 30(3): 277-308.

Xia, Y., K. Wong and W. Li, 2006. Phonetic-based approach to Chinese chat text normalization. Proceeding of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, pp: 993-1000.

Zhao, Y., B. Qin and T. Liu, 2010. Sentiment analysis. J. Softw., 21(8): 1834-1848.

Zhu, J., J. Lu and J. Pei, 2007. Actors and mechanism of customer behavior loyalty. J. Wuhan Univ., Technol., 29(10): 160-163.