Research on Clothing Product Reviews Mining Based on the Maximum Entropy

Qinghong Yang
School of Software, Beihang University
Xueyuan Road, HaidianDistrict,Beijing,100191
rainbow.yang2013@gmail.com

Pengfei Feng
School of Software, Beihang University
Xueyuan Road, HaidianDistrict,Beijing,100191
fengbuaap@163.com

***Abstract*—this paper excavated the review theme of clothing products by method of association rules, and built a maximum entropy model for the reviews classification. Then this paper did experimental verification to large-scale clothing product reviews classification, which verified the practical effect that maximum entropy model had in the comment text classification problems. In the process of classification, the maximum entropy model had a good effect, of which accuracy was over 90%.**

Keywords-association rules; the maximum entropy; review classification

I. INTRODUCTION

The high-speed development of the Internet lifted the e-commerce boom [1]. Nowadays the network shopping has become a new way of shopping, which is chosen by more and more people. Clothing is the largest category in the network shopping, in 2014, the transaction scale of Chinese clothing online shopping market was 615.3 billion yuan, accounting for 22.1% of the national network shopping market [2]. People often comment on the commodities after online purchase of apparel goods. These comments often comes from personal experience, and has more user's subjective experience compared to the merchant's advertising [3]. The customers can consult relevant product reviews to understand features and credibility of commodity at the time of choosing clothing products [4, 5], and the merchants also need to know the advantages and disadvantages of goods through comments for marketing [6]. If there are too many product comments, it will be very tough to refer. However, if we are able to complete review information mining, summary and classification, it is bound to improve the efficiency of user's access to information and user experience [7].

In order to facilitate the browse of the product reviews for users, this paper excavated the review theme by method of association rules, and used the maximum entropy model for classifying clothing comments to improve accuracy of comments on the classification.

This paper excavated the reviews, researched on what subject they were classified according to, and did empirical study on reviews classification. On this basis, the paper is divided into eight parts. The first part is introduction, which introduces what work to do; The second part is literature review; The third part ,the paper research design; The fourth part, data preparation and pretreatment; The fifth part, review theme mining; The sixth part, the classification model building; The seventh part, model validation; The eighth part, conclusion and prospect.

II. LITERATURE REVIEW

For a comment, Kim and Hovy gave related definitions [8]: review consists of four basic elements, topic, holder, statement and sentiment. According to the four basic elements of comments, the reviews mining task can be divided into four sub tasks: theme identification [9， 10], holder recognition, statement screening and emotion recognition [11]. On the review theme research, the literature divided the clothing products items into 17 classifications including the material, workmanship and after-sales, et al from the products and services [12]. In addition to these comments, the customer would like to know the specific factor of special goods when buying clothing products, such as comments about "fuzz ball" of woolen sweater goods.

After determined the review theme, how to summarize and classify the comments is a text classification problem [13]. Existing classification method is mainly based on the theory of statistics and machine learning methods, the most famous text classification methods include Bayes theorem, KNN[14], LLSF, Boosting[15] SVM[16] (support vector machine), etc. Yang from Carnegie Mellon University used English standard classification corpus to compare the commonly used classification methods and drew the conclusion that the KNN and SVM had more classification accuracy and stability than the other methods [14]. The SVM method is essentially a kind of two-type classifier. For two type's classifier, whose time complexity is linear. However, if you want to use the SVM classifier to achieve more class classification, you must construct multiple SVM classifiers. This paper used a statistical model in natural language processing - the maximum entropy model for text classification.

Maximum entropy model reflects a simple principle that the human learn about the world, namely in the case of knowing nothing about an event, selecting a model to make its distribution be as even as possible. In other words, given some facts set, choosing a model consistent with the existing facts to make the distribution even as far as possible for unknown event. Adwait[17, 18] was the first to apply maximum entropy model to text classification, he compared the methods of classification based on the maximum entropy model and the decision tree using ME DEFAULT and ME IFS. Characteristics he used in

QSHINE 2015, August 19-20, Taipei, Taiwan

DOI 10.4108/eai.19-8-2015.2260919

the experiment are binary. But in text classification, we can't determine a word's contribution to the document semantic just by its existence. A more accurate method is to use word frequency.

This paper used association rules in review theme mining, making the classification of the theme more abundant, and introduced the maximum entropy classification model to classify clothing comments, and found that the maximum entropy classifier had high accuracy in text classification problem.

III. THE RESEARCH DESIGN

A. The research process

On the basis of summarizing the literature, this paper determines the following research route: Establish the research target, data preparation and pretreatment, review theme mining, classification model building and the model validation and optimization.

Figure 1. clothing reviews mining research flow chart

The research target: how to be more effective to show customers clothing comments, how to be more convenient for customers to query related comments information and how to save the customers' time of browsing comments. Collecting data: this study crawled comments data of the top 10 commodity in each of the 221 categories from taobao. Mining review theme: we excavated review theme in the comment data, and researched on classifying comments according to what theme. Building classification model: according to the mining of the review theme, the classification model of clothing comments was established based on maximum entropy model. Model validation and optimization: verify the classification accuracy and performance of the model, and find the model optimization method from the results.

B. The research methods

This article adopted the research methods of literature research and experimental validation, classifying and analyzing the related literature and method of review mining in recent years, and did experiment verification of the summed up review mining method, analyzing the experimental results and making improvement.

IV. DATA PREPARATION AND PRETREATMENT

This study took clothing products online reviews in the electronic commerce as the research object. Taobao's user scale and sales proportion is the first one in China's online shopping market, the quantity of a lot of clothing products reviews is big, and the quality of the reviews is high[19]. So this study collected online reviews data from taobao, ensuring the data quality.

On October 10, 2014, we crawled comments data of the top 10 commodity in each of the 221 categories from taobao, including a total of 3,628,637 goods comments.

Although spam comments crawled were less, but there were still a small amount. We eliminated 1357 spam comments such as pure symbols, obtaining 3,627,280 comments data finally.

V. REVIEW THEME MINING

Literature divided the clothing products items into 17 classifications including the material, workmanship and after-sales, et al from the products and services [12]. In order to excavate more review themes, we mainly used association rules.

Reference [10] obtained good effect by using the association rules in review theme mining. Firstly tagging comments corpus and extracting the noun and noun phrase from each sentence, after that, extracting the noun or noun phrase meeting minimum support as candidates for review theme from the corpus of comments by using the method of association rules mining. In actual effect validation, it is found that reviews theme often use the same or similar adjectives to modify. So we can get better coverage by analyzing the nouns and adjectives in the comments using association rules.

Firstly, withdraw reviews, extracting nouns and adjectives from each review. Secondly, find the common combinations of nouns and adjectives or the combinations of nouns and nouns. Finally, get review theme from the combinations through artificial selection. This study obtained more 28 review theme through the analysis of the association rules.

According to the literature and the review theme mining by association rules, we received 45 review themes including "fabric", "thickness", "quality" and so on.

VI. CLASSIFICATION MODEL BUILDING

A. Text features generation

The text features for training text must be generated before training model. In extracting text feature for short text, n-gram language model, one of the most used models, is based on the assumption that the appearance of a word was only associated with the previous n-1 words. This paper generated text features by using unigram and bigram language model.

After segmenting the reviews, the unigram language model took individual words as the text features, while using the frequency of the single word as its weight value; the bigram language model considered the combination of two adjacent words as the text features, while using the frequency of their occurrence as their weight value.

B. The Maximum Entropy Model

The maximum entropy model makes the distribution of the unknown events as uniform as possible in the case of known constraint. According to definition of Shannon, the entropy is calculated as follows:

$$H(p) = -\sum P(x) log_2 P(x) \qquad (1)$$

In this paper, x represents the comments and p (x) represents the probability that x belongs to the review theme. When H (p) has the maximum value, p* is the probability distribution which is consistent with the maximum entropy model:

$$P^* = argmaxH(p) \qquad (2)$$

If there is no other prior knowledge, according to the nature of entropy, the formula (1) has the maximum value when the probabilities that various events happened are equal. That is, when there is no information about a review, the probability that the review belongs to each review theme is equal in the maximum entropy model. In fact, the training text can provide the probabilities that the words belong to various review themes, and the probabilities are constraints in the maximum entropy model. Namely, the issue turns out to be seeking the maximum entropy under the constraint condition.

C. Maximum entropy model building

According to the definition of the maximum entropy model, the description of constraints is the most important issue in building maximum entropy model. In this paper, the review word 'silk' belonging to the review theme 'fabric' is a constraint. In order to describe this constraint in maximum entropy model, a characteristic function is defined as follows:

$$f(w,c)=\begin{cases} n, & (w="silk" \wedge c=" fabric") \\ 0, & otherwise \end{cases}$$

$W = \{w_1, w_2, ..., w_n\}$ is the set of text features which contain single words and two combination of adjacent words , $C = \{c_1, c_2, ..., c_3\}$ is the set of review themes and n is the number of times the text feature appears. For this characteristic function f, its expected value in the empirical probability distribution p (w,c) is as follows:

$$E_p f_i = \sum_{w,c} p(w,c) f_i(w,c) \qquad (3)$$

Its expected value in the probability distribution p (w, c) of the maximum entropy model is as follows:

$$E_p f_i = \sum_{w,c} p(c) p(w \mid c) f_i(w,c) \qquad (4)$$

So a constraint of the maximum entropy model is to making the values of the formula (3) and (4) be equal. Obviously, we can define a number (k) of such characteristic functions according to the text features generated above. Thus k groups of constraints can be made up and the issue becomes finding the optimal solution under the k constraints. The classical method to solve the optimal solution is Lagrange-Multiplier Algorithm and this paper demonstrated the conclusion directly. The probability distribution p* of the maximum entropy model has the following form:

$$p(w \mid c) = \frac{1}{a} \exp\left(\sum_i^k \lambda_i f_i(w,c)\right)$$

α is a normalization factor, λ is a parameter. After learning in the training set, we got the value of λ and the probability distribution p*, completing the construction of the maximum entropy model. The next task is to get the parameter λ of the maximum entropy model by practicing through the training set.

VII. Model Validation

Based on the theory and research above, this paper validated the model from these aspects: what is the effect of the maximum entropy classifier? We designed experiments based on this question.

We crawled 3,627,280 customer reviews from taobao, and these reviews were classified into 45 review themes. We divided them into training and test sets. The training set consisting of 2,720,460 reviews was used for learning of the classifier, and the test set consisting of 906,820 reviews was used for testing of the classifier.

A. Testing the accuracy of the maximum entropy classifier

The toolkit of maxent is used for the experiment of reviews classification and trained parameters of the maximum entropy model by iteration.

To test the effect of the maximum entropy classifier, we conduct experiments to classify the reviews using two different ways of generating text features: the first way, we generated text features using unigram language model; the second way, we generated text features using unigram combined with bigram language model. We tested the maximum entropy classifier when using two different ways of generating text features. The number of iterations was 100, and the accuracy of the classifier is shown in table 1.

TABLE I. The accuracy of the classifier when using two different ways of generating text features

ways of generating text features	*The right number*	*Accuracy*
unigram	829278	91.449%
unigram+bigram	841340	92.779%

From Table 1, it can be seen that the accuracy of the maximum entropy classifier was more than 90%. The accuracy of the second way of generating text features was higher than the first one. The results showed that text features of unigram and bigram language model was better than that of unigram language model alone, and classification accuracy had been improved by using text features of unigram and bigram language model.

VIII. The Conclusion and Prospect

In this article, we excavated special review themes that existing in some characteristic category and wanted by the users, enhancing the experience of the users when they browsing reviews. Then we used maximum entropy model to experiment the clothing reviews classification based on the large-scale reviews data. It can be seen from the experimental data that the accuracy of maximum entropy classifier reached more than 90%, which means a good practical effect. According to the results of the experiment, these conclusions are obtained:

1) The association rules had a good effect on excavating the review theme of clothing products. This paper excavated more 28 review themes using association rules, which are hidden in the reviews.

2) Maximum entropy classification results reached more than 90%, suggesting that the maximum entropy model had a good effect in the clothing reviews multiple classification issues. Besides, we combined unigram language model with

bigram language model to generate text features, making the results up to 92.779%, which improved the classification effect.

This study shows the good effect that the maximum entropy classifier has in the clothing reviews classification. The specific classification of clothing review themes does not have a unified standard, and the division of review themes also needs certain improvement and research. As a short text, the characteristics of the clothing reviews also lost some information. The next step of our work can be focused on the aspect of text feature extraction based on content.

ACKNOWLEDGMENT

This paper's work was completed during my working period in Dangdang Information Technology Company. Thanks to big data and operations Director Mr. Qiang Fu .Thanks to the support and guidance from researcher Mr. Qi Ju. And thanks to algorithm engineer Mr. Yuanshu Jiang's algorithm discussion and practical guidance.

REFERENCES

[1] Rafael Maranzato,Adriano Pereira. Fraud detection in Reputation System in e-Markets using Logistic Regression.SAC 10 March 22-26.2010.Sierre.Switzerland.

[2] CNNIC. Chinese Internet data platform[M]. http://www.cnnic.net.cn/.

[3] He Huang.Research and Application about the Sentiment Classification of Automobiles' Online Reviews[D].Harbin Institute of Technology,2013.

[4] Lian J, Lin T. Effects of consumer characteristics on their acceptance of online shopping: Comparisons among different product types[J]. Computers in Human Behavior, 2008, 24(1):48–65.

[5] Gruen T, Osmonbekov T, Czaplewski A. Ewom: The Impact Of Customer-To-Customer Online Know-How Exchange On Customer Value And Loyalty[J]. Journal of Business Research, 2006, 59(4):449-456.

[6] Litvin S W, Goldsmith R E, Pan B. Electronic word-of-mouth in hospitality and tourism management[J]. Tourism Management, 2008, 29(3):458–468.

[7] Lei Jiang.Research on Key Technologies of Opinion Mining Towards Product Reviews[D]. Harbin Institute of Technology,2010.

[8] Kim S, Hovy E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text[J]. In ACL Workshop on Sentiment and Subjectivity in Text. Alessandro Moschitti, Daniele Pighin, Roberto Basili, 2006:1-8.

[9] Hongqing Z, Yangyang W. Extracting and clustering features of evaluation object in Chinese user reviews[J]. Microcomputer & Its Applications, 2014.

[10] Hu M, Liu B. Mining opinion features in customer reviews[J]. In Proceedings of Nineteeth National Conference on Artificial Intellgience (AAAI, 2004.

[11] Zhang Z, Ye Q, Zhang Z, et al. Sentiment classification of Internet restaurant reviews written in Cantonese[J]. Expert Systems with Applications, 2011, 38(6):7674–7682.

[12] Jie Li, Xiangqian Zhang. Key Content Elements of Online Consumer Review and Effects on Customer Satisfaction for Garments in C2C E-commerce[J]. Chinese Journal of Management,2014,02:261-266.

[13] Peng W. Study on Chinese text classification based on dependency relation[J]. Computer Engineering and Applications, 2010.

[14] Y. Yang, X. Lin. A re-examinnation of text categorization methons. In The 22nd Annual Intel ACM SIGIR Conf. on Research and Develop ment in the Information Retrieval. New York: ACM Press. 1999

[15] R. E. Schapire, Y. Singer. Improved boosting algorithms using confidence-rated predications. In: proc. Of the 11th Annual Conf. on Computaional Learning Theory. New York: ACM Press,1998. 80~91

[16] T. Joachims. Text categorizstion with support vector machines:learning with many relevant feature. In Proc. Of the 10th European Conf. on Machine Learning. New York: Springer,1998. 137~142

[17] Ronglu Li, Jianhui Wang, Xiaoyun Chen, Using Maximum Entropy Model for Chinese Text Categorization[J]. Journal of Computer Research and Development,2005,01:94-101.

[18] Ratnaparkhi A. Maximum entropy models for natural language ambiguity resolution[D]. University of Pennsylvania, 1998.

[19] Korfiatis N, Garc á-Bariocanal E, S ánchez-Alonso S. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content[J]. Electronic Commerce Research and Applications, 2012, 11(3):205–217.