Clothing Product Reviews Mining Based on Machine Learning

http://dx.doi.org/10.3991/ijoe.v11i9.5069

Qinghong Yang, Pengfei Feng, Zhichao Cheng*
Beihang University, Beijing, China

***Abstract*—This paper used the method of machine learning to study clothing product reviews classification based on big enterprise data. Taking Taobao clothing reviews as the object, it firstly excavated review themes from reviews corpus by association rules, and then searched review themes related to the categories by a method of mutual information to enrich the review themes. In the process of building classification models, commonly used SVM classifiers were studied in the beginning. After training and verification of a large amount of data, the classification accuracy reached 90.597%. In order to further improve the classification accuracy, the maximum entropy model was built by adopting the maximum entropy algorithm, on the basis of the same review themes. After repeated experiments and optimization in a large-scale of clothing product reviews, the classification accuracy reached up to 93.035% finally. Compared with SVM classification algorithm, the accuracy of maximum entropy in the clothing product reviews classification is higher. This paper verified the effectiveness of maximum entropy model on comment text multi-classification problem, and the maximum entropy model has practical values in electronic business.**

***Index Terms*—mutual information, review classification, SVM, the maximum entropy.**

I. INTRODUCTION

The rapid development of the Internet has accelerated the e-commerce upsurge [1]. Nowadays, network shopping has become a convenient and relatively new way of purchasing, which is chosen by increasingly more people. Clothing is the largest category in network shopping. In 2014, the transaction scale of the 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 their commodities after the online purchase. These comments reflect more users' personal experience than the merchants' own advertising [3]. The customers can consult relevant product reviews to understand features and credibility of commodities when 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 hard to refer to them. However, if it is possible to complete review information mining, summary and classification, it is bound to improve the efficiency of users' access to the information and their experiences [7].

In order to make the classification theme more in line with the needs of users, this paper excavated the review theme by the method of mutual information, and then used SVM classification model to classify clothing reviews. Moreover, it introduced maximum entropy classification model for the reviews classification so as to improve the accuracy of the comments classification, and achieved a good effect.

II. LITERATURE REVIEW

Kim and Hovy gave related definitions for a review [8]: Review consists of four basic elements, topic, holder, statement and sentiment. According to the four basic elements of comments, the review 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 in terms of the products and services including the material, workmanship and after-sales [12]. In addition to these comments, the customer would like to know the specific factors when buying specific clothing products, such as comments about "fuzz balls" on woolen sweater goods.

How to summarize and classify the comments after the review theme is obtained is a text classification problem [13]. Existing classification methods are mainly based on the theories of statistics and machine learning methods, the most famous text classification methods include Bayes theorem, KNN [14], and Boosting [15] SVM [16] (support vector machine,). Yang from Carnegie Mellon University used an 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, whose time complexity is linear. However, if the SVM classifier is used to achieve more class classification, multiple SVM classifiers have to be constructed. This paper finally used a statistical model in natural language processing, the maximum entropy model for text classification.

Maximum entropy model reflects a simple principle that humans learn about the world. Namely, in the case of knowing nothing about an event, humans select a model to make its distribution be as even as possible. In other words, given some fact sets, choose a model consistent with the existing facts to make the distribution as even as possible for unknown events. Adwait [17, 18] was the first to apply the 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 the experiment were binary. However, in text

classification, a word's contribution to the documents semantic cannot be determined just by its existence. A more accurate method is to use word frequency.

This paper used association rules and mutual information in review theme mining, excavated the specific review theme in the category to make the classification of the theme more relevant to the user's requirements, and used SVM classifier and maximum entropy classifier to classify clothing comments. By comparison, it is found that the maximum entropy classifier had higher accuracy in text classification.

III. RESEARCH DESIGN

A. Research Procedure

On the basis of summarizing the literature, this paper determines the following research route: Establishing the research target, data preparation and pretreatment, reviewing theme mining, text features generation, and classification model filtering, building and validation.

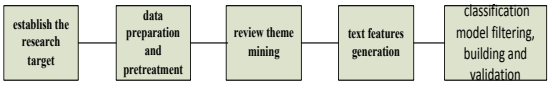

Figure 1. Clothing reviews mining research flow chart

The research target is to be more effective in showing customer clothing comments, to make customers more convenient in querying related comment information and to save the customers' time when browsing comments. In the period of data preparation and pretreatment, this study crawled comment data of the top 10 commodities in each of the 221 categories from Taobao. For the mining review theme, we excavated review themes in the comment data, and researched classifying comments according to what themes they contained. After that, the reviews were changed into the form that can be handled in the classification model. At last, establish the classification models for clothing comments according to the mining of the review theme, compare these models and optimize the model.

B. The research methods

This article adopted the research methods of literature research and experimental validation. First the related literature and method of review mining in recent years were classified and analyzed, and then the mature methods and models used to classify the reviews in enterprise. Finally, experiment verification of the summed up review mining method was made, with the aim of achieving better results.

IV. DATA PREPARATION AND PRETREATMENT

Taobao's user scale and sales proportion is the largest in China's online shopping market. Since the quantity of clothing product reviews is numerous and the quality of the reviews is high [19], this study took online review data from Taobao as the research object to ensure the data quality.

On October 10, 2014, a total of 3,628,637 comments of the top 10 commodities in each of the 221 categories from Taobao were crawled.

After 1357 spam comments were eliminated, such as pure symbols and error messages, 3,627,280 valid comments were finally obtained.

V. REVIEW THEME MINING

Literature had summarized 17 review themes from the products and services [12]. In order to excavate more review themes, the following two aspects of work were mainly performed.

A. Using the association rules in review theme mining

Reference [10] obtained good results by using the association rules in review theme mining. This was done first by tagging the comment corpus and extracting the noun and noun phrase from each sentence, and then by extracting from the corpus of comments the nouns or noun phrases which meet the minimum support as candidates for the review themes by using the method of association rules mining. For example, the words "clothes" and "fabric" often appear together. They will be picked by association rules and be the candidate for the review themes. In actual effect validation, it is found that review themes often are often modified with the same or similar adjectives. For instance, "fabric" is usually modified by "soft". As a result, better coverage can be got by analyzing the nouns and adjectives in the comments using association rules.

At first, withdraw reviews, extracting nouns and adjectives from each review. After that, find the common combinations of nouns and adjectives or the combinations of nouns and nouns. In the end, obtain review themes from the combinations through artificial selection. This study obtained a total of 17 review themes through the analysis of the association rules.

B. Using mutual information for review theme mining

When focusing on a certain category of goods, people often want to focus on specific characteristics of this category. For example, with "raincoats", people tend to focus on their waterproof properties; with "silk" commodities, people tend to be more concerned about whether or not it is easy to fad. "Waterproof," "fading" among others will also appear in the corresponding categories with a bigger probability and in other categories with a smaller probability. The mutual information of these words and categories also reflects the distribution probability in the categories.

The mutual information was used as a measure of the association degree of the two signals in information theory. Later, the mutual information was extended for statistical description of the association degree between two random variables. In this paper, the mutual information was used to measure the association degree between words and categories.

A problem needing to be considered before calculating the mutual information is that if a word W_1 rarely appears in the general comments, and only appears in a certain category C_1, then the mutual information between W_1 and C_1 is great. However, the word W_1 is not a common word and it cannot represent the category C_1. In order to solve this problem, word frequency filtering needs to be combined before the mutual information is calculated. After filtering out those uncommon words, calculate the mutual information and search for words related to

category. After word frequency filtering, the review theme model by mutual information mining method is:

$$I(w,c) = log_2\left(p(w|c)/p(w)\right)$$

Where W represents words; C represents category; P (W) represents the probability that W appears in general comments; and P(W|C) represents the probability that W appears in category C. The bigger the mutual information value is, the more relevant the word W and category C is, showing the probability that W appears in the category is higher.

Among the 221 categories of comments we collected. First withdrew words with word frequency filtering to filter out those uncommon words. Second, separate the mutual information values of each word and the 221 categories to select the top 100 words with the largest mutual information in each category. Finally, filter these words artificially and selected representative words as review themes. 11 words have been selected including "fuzz ball," and "waterproof," as our review themes.

According to the above two aspects of the review theme mining, 45 review themes were received including "fabric" "thickness" and "quality".

VI. TEXT FEATURES GENERATION

Since the classifier cannot deal with the original training text, the text features have to be generated for training text before building training model. In extracting text features 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 models.

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.

VII. CLASSIFICATION MODEL BUILDING AND VERIFICATION

In the process of classification model building, select a mature classification model, SVM model, to classify the comments in the beginning. Then calculate the classification results of SVM model. Through literature research and existing technology, use the maximum entropy algorithm to build classification model, aiming at better results.

A. SVM model building and validation

1) SVM model building

The SVM model used in this paper is a kind of classification algorithm based on statistical learning theory, and it was proposed in the 1990s. It is evolved from the optimal classification surface which is linearly separable, and its basic idea can be illustrated in the two-dimensional case. In Figure 2, the solid points and hollow points represent two categories of samples and H is the classification plane. H1 and H2 are the planes which are parallel to sorting lines and pass through the samples being closest to the sorting line in different categories respectively. The distance between them is called the classification margin. The optimal classification line is the line that separates two categories of samples while making the classification margin the largest.

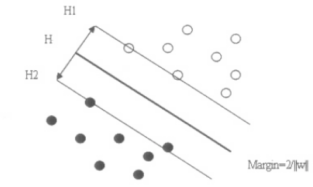

Figure 2. The Principle of SVM

Suppose the classification plane: g(x)=ωx+b, and there are n samples, $\{(X_i,Y_i)\}$,i=1,2,3... n, where n is the number of training samples; X_i is the vector after text feature generation and $Y_i \in \{-1, +1\}$ is the symbol for the category. The samples satisfy the formula:

$$y_i\left[\omega x_i + b\right] >= 1\left(i = 1,2,3...n\right) \quad (1)$$

At this time, classification margin is equal to 2/||ω||. To maximize the classification margin is equivalent to minimize || ω ||(minimizing || ω ||²/2). The optimal classification surface is to satisfy the formula (1) while minimizing ||ω||²/2.

Thus, the problem of searching the optimal classification surface is to find the minimum value of the following functions under the constraints of the formula (1):

$$\varphi(\omega) = \|\omega\|^2 / 2 = (\omega * \omega)/2 \quad (2)$$

To this end, Lagrange function is defined as follows:

$$L(\omega, b, a) = \frac{1}{2}(\omega * \omega) - \sum_{i=1}^{n} a_i \left(y_i (\omega * x_i) + b - 1\right) \quad (3)$$

Where a_i ($a_i >= 0$) is the coefficient of Lagrange function. Seek the partial differential of formula (3) for ω and b and let them equal 0, and then the problem can be transformed into the following questions:

$$\sum_{i=1}^{n} y_i a_i = 0 \quad (4)$$

$$a_i \geq 0, i = 1,2,3...n \quad (5)$$

And the problem now is to solve the maximum of the following objective function under the constraints of the formula (4) and (5):

$$\delta(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j \left(x_i x_j\right) \quad (6)$$

Let a* be the solutions that meet the maximum of the above function, and a* is the Lagrange multiplier corresponding with each constraint in the original problem. This is a quadratic optimization problem under inequality constraints, and thus there is a unique solution. Only a small part of a_is is not equal to 0, and the corresponding sample is the support vector. The optimal classification function after the solution of these problems is as follow:

$$f(x) = \mathrm{sgn}\left((\omega * x) + b\right) = \mathrm{sgn}\left(\sum_{i=1}^{n} a^* y_i (x_i * x) + b^*\right) \quad (7)$$

Where $\mathrm{sgn}(x)$ is the sign function. The category can be judged by the sign of discriminant function f (x).

For linearly inseparable issues, the input can be mapped into a high dimensional space by non-linear transformation. Then the optimal classification surface can be calculated in this new space. The non-linear transformation is achieved by defining the appropriate kernel function. The multi-classification classifier based on SVM model can be achieved by establishing n SVM classifiers.

2) SVM model validation

Based on the theory and research above, this paper validated the SVM model to test the effect of the SVM classifier. 3,627,280 customer reviews were crawled 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 the learning of the classifier, and the test set consisting of 906,820 reviews was used for the testing of the classifier.

To test the effect of the SVM classifier, we conduct experiments to classify the reviews using two different methods of generating text features: In the first method, generate text features using the unigram language model; in the second method, generate text features using the unigram combined with the bigram language model. We tested the SVM classifier when using two different methods of generating text features, 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 SVM classifier was around 90%. The accuracy of the second way of generating text features was higher than that of the first one. The results showed that text features of the unigram combined with bigram language models were better than that of unigram language model alone, and thus classification accuracy was improved up to 92.779% by using the text features of unigram combined with bigram language models.

After studying studied the literature and existing technology, the maximum entropy model was used to replace SVM model for clothing comments classification so as to improve the accuracy of the classification.

B. The maximum entropy model building and validation

1) The Maximum Entropy Model

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

$$H(p) = -\sum P(x) log_2 P(x) \quad (8)$$
(8)

In this paper, x represents the comment 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:

$$P^* = argmaxH(p) \quad (9)$$

If there is no other prior knowledge, according to the nature of entropy, the formula (8) has the maximum value when the probabilities that various events that 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.

2) Maximum entropy model building

According to the definition of the maximum entropy model, the description of constraints is the most important step in building the 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 the maximum entropy model, a characteristic function is defined as follows:

$$f(w,c) = \begin{cases} 1, & (w = \text{"}silk\text{"} \wedge c = \text{"}fabric\text{"}) \\ 0, & otherwise \end{cases}$$

$W = \{w_1, w_2, ..., w_n\}$ is the set of text features which contain single words and the combination of two 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:

$$E_p f_i = \sum_{w,c} p(w,c) f_i(w,c) \quad (10)$$

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

$$E_p f_i = \sum_{w,c} p(c) p(w \mid c) f_i(w,c) \quad (11)$$

Therefore a constraint of the maximum entropy model is to make the values of the formula (10) and (11) equal. Obviously, a number (k) of such characteristic functions can be defined 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 the Lagrange-Multiplier Algorithm and this paper demonstrated the conclusion directly. The probability distribution p* of the maximum entropy model has the following form:

$$p(w|c) = \frac{1}{a}\exp(\sum_{i}^{k} \lambda_i f_i(w,c)) \quad (12)$$

Where α is a normalization factor, and λ is a parameter. After learning in the training set, obtain 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.

3) The maximum entropy model validation

In order to investigate the accuracy and performance of the maximum entropy classifier for text classification, this study used the SVM classifier for comparison. We investigated the effect of classifiers on accuracy and classification time. In order to ensure the same experimental environment, we used the same Linux machine to conduct experiments for two different classifiers. The results of the two classifiers are shown below.

TABLE II.
COMPARISON OF THE SVM CLASSIFIER AND THE MAXIMUM ENTROPY CLASSIFIER

Classifier	***Accuracy***	***Training Time(Second)***
SVM classifier	90.597%	286.25
maximum entropy classifier with 30 iterations	90.584%	206.77
maximum entropy classifier with 100 iterations	92.779%	697.22

As can be seen from the results, when the iteration of the maximum entropy classifier was 30 times, the maximum entropy classifier had an equivalent accuracy with the SVM classifier, but the SVM classifier took 38.43% more time than the maximum entropy classifier. When the iteration of the maximum entropy classifier was 100 times, the maximum entropy classifier had a higher accuracy than the SVM classifier.

4) Maximum entropy model Optimization

In the process of classification, it is found that some words in the comments had strong correlations with the review themes. For example, in the review theme "color losing", the phrase "color losing" and "discoloration" had greater impact on classification. Therefore we gave these text features a higher weight. When these text features appeared, their weight would be their frequency multiplied by 10. In the experiment, weigh 100 text features which had the greatest impact, and then classified these reviews. Table 3 shows the results.

TABLE III.
CLASSIFICATION ACCURACY WITH NON-WEIGHTED AND WEIGHTED TEXT FEATURES

	non-weighted text features	***weighted text features***
Accuracy	92.779%	93.035%

As can be seen from the results, the accuracy of the classifiers became higher after the text features were weighted. The maximum entropy model had been optimized to a certain extent.

VIII. CONCLUSION AND PROSPECT

In this article, by using the method of calculating mutual information between words and categories, we excavated special review themes that exist in certain characteristic categories and are desired by the users, enhancing the experience of the users when they browse reviews. We then used the SVM model to test the clothing review classification based on the large-scale review data. The experimental data showed that the accuracy of the SVM classifier was about 90%. In order to improve the accuracy of the classification, the maximum entropy model was used to replace the SVM model. Finally, the accuracy of the maximum entropy classifier reached up to more than 93%, which resulted in a good practical effect. According to the results of the experiment, four conclusions were reached:

1) The maximum entropy classification results reached up to more than 93%, suggesting that the maximum entropy model had a good effect on clothing review multiple classifications. Moreover, the unigram language model was combined with bigram language model to generate text features, improving the classification effect.

2) According to the comparison of training effect and training time between the SVM classifier and the maximum entropy classifier, the maximum entropy classifier had an equivalent accuracy with the SVM classifier when the iteration of the maximum entropy classifier was 30 times. However the SVM classifier took more time than the maximum entropy classifier. When the iteration of the maximum entropy classifier was 100 times, the maximum entropy classifier had a higher accuracy than the SVM classifier, suggesting the maximum entropy model was better than the SVM in the text classification.

3) While the accuracy of the maximum entropy classifier before weighing text features was 92.779%, it was 93.035% after weighing text features. This shows that weighing text features has a positive effect.

4) This paper used SVM model to classify the reviews in the beginning, and then substituted the maximum entropy for the SVM model for higher classification accuracy. The process may have reference values for mining review data in similar enterprises.

This study expanded a way for the clothing review classification and shows the positive effect that the maximum entropy classifier has on clothing review classification. Since the review data are different among enterprises, there is still a long way in review processing. The specific classification of clothing review themes does not have a unified standard, and the division of review themes also needs further improvements and research. Moreover, as a short text, the characteristics of the

clothing reviews also lose some information. The future work will focus on the aspect of text feature extraction based on content.

ACKNOWLEDGMENT

This paper was completed during the writer's stay in Dangdang Information Technology Company. My heartfelt thanks must be given to big data and operations Director Mr. Qiang Fu. Thanks are also due to Researcher Mr. Qi Ju for the support and guidance, and to Algorithm Engineer Mr. Yuanshu Jiang for his 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. http://dx.doi.org/10.1145/1774088.1774400

[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. http://dx.doi.org/10.1016/j.chb.2007.01.002

[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. http://dx.doi.org/10.1016/j.jbusres.2005.10.004

[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. http://dx.doi.org/10.1016/j.tourman.2007.05.011

[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 Nineteenth 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. http://dx.doi.org/10.1016/j.eswa.2010.12.147

[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.

AUTHORS

QingHong Yang, a doctor of Management Science and Engineering, College of Economics and Management, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing, China, 100191. Visiting Scholar, Southern Connecticut State University, Unite State; Teacher of College of Software, BeiHang University, China. Her research fields are data mining, data analysis, and mainly analyzing data about consumer behavior, consumer review or human resource management. (e-mail: rainbow.yang2013@gmail.com)

Peifei Feng, a master of Software Engineering, College of Software, BeiHang University, No. 37 Xueyuan Road, Haidian District, Beijing, China, 100191. (e-mail: 1216862060@qq.com)

Zhichao Cheng, the corresponding author of this paper, also a professor of Management Science and Engineering, College of Economics and Management, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing, China, 100191. His research field is Human resources Management. (email: 1216862060@qq.com)

Submitted 21 September 2015. Published as resubmitted by the authors 20 October 2015.