The research for effect of aspects extraction of Chinese commodity comments on supervised learning methods

Yan Zhao[1], Suyu Dong[1], Hua Yang[2], Jing Yang [3,*]

[1] *College of Management, Inner Mongolia University of Technology, Huhhot, China;*

[2] *College of Information Engineering, Inner Mongolia University of Technology, Huhhot, China;*

[3] *College of Mechanical Engineering, Inner Mongolia University of Technology, Huhhot, China;*

Received 1 March 2014, www.cmnt.lv

Abstract

With the advent of Web 2.0, there are more and more websites for shopping. These websites often allow customers make comments of the commodity which they have purchased. Therefore, three is an increasing number of online reviews. More importantly, these reviews contain a mass of sentiment. The sentiment is meaningful for merchants and customers. This paper focuses on the extraction of aspects of online review of products.

We will use Supervised Learning methods to extract aspects of online review of products. Through the experiment of this paper, we found that Machine Learning can be used for aspects extraction of Chinese online review of products. Using ME and presence character representation can achieve 85.6% accuracy.

Keywords: Aspects extraction; Product reviews; Machine Learning

1 Introduction

With the advent of Web 2.0, the development of e-commerce is rapid. These promote more and more customers through web to shopping. Almost retailers allow customers to make comments on the websites and a large number reviews follow. In addition, these reviews are important for customers and merchants. For merchants, they can know more requirements which customers want to have for the different aspects of products; for customers, they can know more about products through the reviews from other purchasers and make better decision.

For most reviews which come from customers not only express the viewpoint of the whole product but also contain the perspective of different aspects of product. There are many aspects of every product, but for customers, they often pay close attention to some aspects of the product. We hold that aspects which customers care for are more important than others. Therefore, extracting important aspects of products is more meaningful for merchants and customers. Merchants can improve the efficiency through caring about important aspects. Customers can know more about product and to make better decisions more conveniently and rapidly. However, confronting with vast product reviews which also update quickly, it is difficult for us to extract aspects of products through manual works. So, automatically extracting aspects of product from product reviews is demanded urgently.

A main method of extracting aspects of products is to extract frequent features in the reviews. For the presentation of Chinese reviews, people usually frequently describe aspects to indicate these aspects are important. Thus, if some aspects of products appear in different reviews simultaneous, we think that these aspects are important for the products than other aspects.

This paper will use surprised learning to extract important aspects of products which come from online reviews and select different methods to extract aspects which various products contain. This paper aims to test the effecttiveness of surprised learning which is used for extracting aspects of Chinese online reviews.

2 Previous Work

Extracting aspects, this also can be regarded as information extraction [1]. For most of reviews, there are two methods to express aspects, which are explicit aspects and implicit aspects.

Explicit aspects, it means the aspects are expressed through nouns and noun phrases. For example, the “battery” in the “The capacity of battery of this phone is big.” is explicit aspect. Now, there are four main methods for the extraction of explicit aspects, extraction based on frequent nouns and noun phrase, extraction by exploiting review and target relations, extraction using supervised learning and extraction using topic modelling [1]. In the researches of aspects extraction based on frequent nouns and noun phrase, Hu and Liu [2] proposed the method through association mining to mine all frequent nouns and noun phrase. They also used compactness pruning and redundancy pruning to get rid of the nouns and noun phrase which are not the aspects of products. At last, they use threshold to select frequent nouns and noun phrase as aspects. Zhu et al. [3] proposed the method which is based on Cvalue [4] to extract aspects, these aspects were consisted of multiply nouns. Long, Zhang and Zhu [5] used the

* ***Corresponding author’s*** e-mail: yangj81@126.com

extraction based on frequent nouns and noun phrase, they also used information distance to mine the related aspects, combining the features coming from these methods to select the reviews which contain the features. Exploiting review and target relations are used widely in many researches [6]. There is tight connection between sentiment and the targets which sentiment focused on. At the same time, sentiment words are easier to be recognized. Therefore, some researchers exploit the relations to extract aspects. Hu and Liu [2] used this method to extract the infrequent features, they made use of sentiment words to find the nouns and noun phrases which had the shortest distance of sentiment words, and these nouns and noun phrases were regarded as features of products. Their research demonstrated this method can extract aspects commendably. In the research of Zhuang, Jing and Zhu [7], they use dependency parser to extract aspects. Many researchers adopted surprised learning to information extraction [8-10]. Aspects extraction as one of information extractions, some paper also used surprised learning to finish it. Kobayashi et al [11] used dependency tree to select candidate features and sentiment word, based on this, they use tree-structured classification method to extract aspects from candidate aspects. Ghani et al. [12] made use of traditional surprised learning and semi-surprised learning to extract aspects. In the paper of Yu et al. [13], one-class SVM was be used to extract aspects, this method only need to label aspect, but not non-aspect. Extraction using topic modeling is an unsupervised learning method for aspects extraction. This method was first used by Brody and Elhadad [14] to extract aspects. They also use adjectives to recognize opinion word which describes the aspects. Titov and McDonald [15] proposed a model to extract aspects, this model also can be used for information extraction which served text ranking. Lu, Zhai and Sundaresan [16] adopted a model which is based on structured pLSA (probabilistic Latent Semantic Analysis) to aspects extraction.

There are many researchers made use of other methods for the work of extract aspects. Such as, Yi [17] utilized opinion lexicon and sentiment pattern database to aspect extraction; Ma and Wan [18] applied center theory to extract targets from news comments which the sentiment words focused on. Kim and Hovy [19] wield the annotation of semantic role to aspects extraction from online news media texts.

Compared with the explicit aspects extraction, the researches about implicit aspects extraction are less. Apart from the nouns and noun phrases which described explicit aspects, the other expressions about aspects are implicit aspects. Generally, people represent implicit aspects through adjectives and verbs. For example, “expensive” is the implicit aspect in the review “This phone is expensive.”, and “using” in “Using this computer is difficult.” Hu and Liu [2] only mentioned implicit aspects extraction but not completed the work. Clustering method was used for implicit aspects extraction in Su et al[20], the main idea was constant repetition of form phrases which were consisted of sentiment word and explicit aspects simultaneously, and then adopted these sentiment words to extract implicit aspects. They considered that the sentiment words from the phrases were used for modifying aspects, so, the relation between sentiment words and explicit aspects can be used for the implicit aspects extraction.

The exciting researches about aspects extraction almost focused on English reviews. This paper will centres on experimental research about Chinese product reviews. Because of there is no obvious grammar role for Chinese reviews, it is difficult to use the related methods to extract aspects from Chinese reviews. We will verify that supervised learning can be used for aspects extraction of Chinese product reviews.

3 Methodologies

In this paper, the supervised learning methods are adopted to extract the aspects of online comments of the commodity. Meanwhile, the effects of different feature representtation methods are discussed. The method of feature selection based on word frequency (DF) is adopted. The top 150 words are selected to constitute feature vector for classification.

3.1 FEATURE WEIGHTING

Feature weighting denotes the importance of a feature in text, which can be calculated through the statistical information of text. In addition, the feature weighting play an important role for the distinguish ability of a feature of the text. In this paper three different calculation methods of feature weight for classification (Boolean, frequency and TF-IDF) will be compared.

3.1.1 Boolean

Boolean weighting represents the text according to the feature whether or not existing in the text. When the feature appears in the text, the value of weight is 1, otherwise the value is 0. Because the Boolean value cannot response the significance of feature of text, this method is often replaced by other more accurate methods. However, this method can also obtain a good effect under some circumstances. Hence, it is necessary to experiment the effect of this feature weighting method for aspect extraction of online commodity comments.

3.1.2 TF

The times of feature appearance in the text is used to represent the text for the method of TF, which is generally considered as a good weighting method. However, the weakness of this weighting method is that the distinguish ability of low-frequency features is ignored. In some cases, some low-frequency features may have a greater ability to distinguish text than high-frequency features. Hence, the effect of this weighting method is worth to explore for the application of this paper.

3.1.3TF-IDF

TF-IDF is the most widely used feature weight calculation method for the text classification. It can not only highlight weight of the high frequency features, but also care the weight of lower frequency features, which have good distinguish ability. The $TF-IDF$ value of a certain feature is calculated by the following equation:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i} \quad (1)$$

Thereinto, w_{ij} indicates the weight of feature t_i in document d_j. tf_{ij} indicates the frequency of feature t_i in document d_j. n_i indicates the number of document which contains feature t_i. N is the number of all documents. TD-IDF is based on the idea: if one feature has high-frequency, and rarely appears in other text, then the feature has a good ability to distinguish. Although it's ideas and structure of statistics are very simple, but its performance is very good.

3.2 CLASSIFIERS

3.2.1 Naive Bayes classifier

Naive Bayes classifier is a simple and popular classifier, which is based on the Bayes probabilistic model. In practice, the probability of document belonging to C_i is calculated by the following equation:

$$P(C_i|d) = \frac{P(d|C_i) * P(C_i)}{P(d)} \quad (2)$$

Thereinto, $P(C_i)$ indicates the probability of a document belonging to C_i. In this paper, we used Naive Bayes classifier with weight. Naive Bayes classifier is based on the assumption of independence conditions of different features, using the joint probability between features and categories to estimate the probability of categories given a document. Hence, this model can be detailed as following formula:

$$P_{NB}(C_i|d) = \frac{P(C_i)(\prod_{t_i \in V} P(t_i|C_i)^{W(t_i,d)})}{\sum_j [P(C_j)\prod_{t_i \in V} P(t_i|C_j)^{W(t_i,d)}]} \quad (3)$$

There into, feature t_i is independent of document d, $W(t_i,d)$ indicates the weights of feature t_i in document d. $P(t_i|C_i)$ indicates the Laplacean probability estimation value of conditional probability of documents belonging to C_i if it contains feature t_i. $P(t_i|C_i)$ is calculated by the following equation:

$$P(t_i|C_i) = \frac{1 + W(t_i, C_i)}{|V| + \sum_j W(t_j, C_i)} \quad (4)$$

$W(t_i,C_i)$ indicates the number of documents containing features t_i and belonging to C_i. $|V|$ is the size of $\{t_1, t_2, \cdots, t_m\}$, which are all features coming from all documents.

Although it's assumption conditions is very restrictive and difficult to meet in real-world, it still performed well in text classification.

3.2.2Maximum entropy classifier

Maximum entropy classifier (ME) is based on maximum entropy model, Berger et al. [21] was the first researcher who applied maximum entropy models in the natural language processing; Chen et al. [22] improved maximum entropy model. Nigam et al. [23] found that ME is better classifier than Naive Bayes classifier on text classification. Its basic idea is that it does not make any hypothesis and remain maximum entropy for the unknown information, which is an advantage for maximum entropy compared with Naive Bayes. Maximum entropy model must satisfy the constraint of known information and the principle of maximum entropy. Hence maximum entropy model is got through solving a optimization problem with constraints. The classical algorithm to solve this problem is lagrange multiplier method. In this paper, we give the conclusion directly. The result is following:

$$p^*(C_i|t_i) = \frac{1}{\sum_{C_i} \exp\left(\sum_i \lambda_i f(t_i, C_i)\right)} \exp\left(\sum_i \lambda_i f(t_i, C_i)\right), \quad (5)$$

$p*$ indicates a predictive model for classification; V indicates the feature vectors; C_i indicates the type which the document belongs to. λ_i indicates the feature weight of feature vectors containing many feature t_i. $f(t_i, C_i)$ is a indicator function.

3.2.3 SVM

Support vector machine (SVM) is generally considered as the best classifier for traditional text classification [24], it is usually better than naive Bayes and maximum entropy. Naive Bayes and maximum entropy are based on probability model, support vector machine (SVM) classifier is got by solving the optimal hyperplane represented by vector W which is used to accomplish classification which can ensure maximum separation between a certain amount of data from the training set and hyperplane. Solving the maximum margin hyperplane eventually is converted into solving a convex quadratic programming problem.

Generally, it translates the above problem into the constrained optimization problem of dual variables through Lagrange Duality. The solution can be written as:

$$\overrightarrow{W} = \sum_{i=1}^{n} \alpha_i C_i \overrightarrow{d_i} \quad (6)$$

C_i is the correct category for document $\overrightarrow{d_i}$. α_i are support vector and greater than zero.

What's more, for linear inseparable problems, kernel function can be used for SVM to convert low dimensional space nonlinear problem to high dimension space linear problem. There are many kernel functions: linear kernel, Gaussian kernel function, radial basis function and so on. In this paper, we used linear kernel function and optimize the parameter of SVM model, which will be used for following experiments.

4 Experiments

4.1 DATA COLLECTION

This paper researches aspects extraction about Chinese product reviews. We use supervised learning for our experiment. So, we created a corpus by retrieving reviews of electronic products from each big shopping website (URL: http://www.taobao.com/; http://www.jd.com/). We got 3000 product reviews through a crawler.

4.2 THE DEFINITION OF EXTRACTING ASPECT

Before we labelled reviews, we talked with specialists who are engaged in research about electronic products and concluded the aspects which people pay close attention. At last, we defined seven aspects which we will experiment. They are price, price performance ratio, screen, pixel, appearance, battery, and reaction rate. Apart from pixel, other aspects all have synonymy or implicit expression.

4.3 DATA ANNOTATION

Because we used supervised learning to complete experiment, so, we asked three students to label the reviews. They adopted same rule to label different aspects of product, 1-7 were corresponding to price, price performance ratio, screen, pixel, appearance, battery, and reaction rate respectively. For every aspect in different reviews, 0 represents this review does not contain this aspect, and 1 represents this review contains the aspect. For example, if “price” is not contained by one review, the review would be labeled by 10, if “battery” is contained by the same review, the review will be labeled by 61 simultaneously. Every review would be labeled by seven tags to indicate the appearance or not of every aspect.

4.4 PERFORMANCE MEASURES

To evaluate the experiment result, this paper uses four indexes. They are Accuracy, Precision, Recall and F1 measure. The accuracy is used to estimate the whole effect of aspect extraction. F1 measure was used to evaluate the comprehensive effect of Precision and Recall. These performance measures can be calculated through table 1. Y means the review contains the aspect. N means the review does not contain the aspect.

TABLE 1 Results of experiments

	Actual Y	**Actual N**
Labelled Y	a	b
Labelled N	c	d

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \tag{7}$$

$$\text{Precision}(pos) = \frac{a}{a+b} \tag{8}$$

$$\text{Precision}(neg) = \frac{d}{c+d} \tag{9}$$

$$\text{Recall}(pos) = \frac{a}{a+c} \tag{10}$$

$$\text{Recall}(neg) = \frac{d}{b+d} \tag{11}$$

$$F1(Y) = \frac{2*\text{Precision}(Y)*\text{Recall}(Y)}{\text{Precision}(Y)+\text{Recall}(Y)} \tag{12}$$

$$F1(\text{N}) = \frac{2*\text{Precision}(N)*\text{Recall}(N)}{\text{Precision}(N)+\text{Recall}(N)} \tag{13}$$

5 Experiment result and discussion

To complete experiment, we adopt our own implementtation for text preprocessing. On the basis of text preprocessing, McCallum’s Mallet toolkit [25] implementation of naive Bayes classifier, maximum entropy classifier, and Chang’s LIBSVM [26] implementation of a Support Vector Machine classifier are used for aspects extraction. This paper adopts different supervised learning and feature weighting methods to extract aspects. The experiment results are showed by follow tables.

5.1 THE EXTRACTION OF DIFFERENT ASPECTS

TABLE 2 The experiment result of battery extraction

		Acc	**Pre (N)**	**Pre (Y)**	**Re (N)**	**Re (Y)**	**F1 (N)**	**F1 (Y)**
ME	Boolean	0.904	0.962	0.754	0.911	0.881	0.936	0.812
	TF	0.892	0.962	0.710	0.897	0.875	0.928	0.784
	TF-IDF	0.861	0.918	0.710	0.893	0.766	0.905	0.737
NB	Boolean	0.841	0.898	0.677	0.888	0.698	0.893	0.688
	TF	0.821	0.843	0.75	0.915	0.6	0.877	0.667
	TF-IDF	0.757	0.794	0.645	0.872	0.506	0.831	0.567
SVM	Boolean	0.752	0.81	0.694	0.729	0.782	0.767	0.735
	TF	0.76	0.841	0.677	0.726	0.808	0.779	0.737
	TF-IDF	0.704	0.778	0.629	0.681	0.736	0.726	0.678
Average		0.810						

TABLE 3 The experiment result of reaction rate extraction

		Acc	Pre (N)	Pre (Y)	Re (N)	Re (Y)	F1 (N)	F1 (Y)
ME	Boolean	0.825	0.810	0.842	0.85	0.8	0.829	0.821
	TF	0.825	0.833	0.816	0.833	0.816	0.833	0.816
	TF-IDF	0.763	0.814	0.703	0.761	0.765	0.787	0.732
NB	Boolean	0.838	0.786	0.895	0.892	0.791	0.835	0.84
	TF	0.8	0.727	0.889	0.889	0.727	0.8	0.8
	TF-IDF	0.763	0.758	0.767	0.737	0.786	0.747	0.776
SVM	Boolean	0.747	0.795	0.702	0.714	0.786	0.753	0.742
	TF	0.725	0.841	0.617	0.673	0.806	0.747	0.699
	TF-IDF	0.703	0.773	0.638	0.667	0.75	0.716	0.690
Average		0.776						

The Table 2-8 shows the experiment result of seven aspects extraction. Table 2 shows the result of battery extraction. The statistical data indicate that the top accuracy 90.4% achieved by ME with presence, the accuracy of SVM with TF-IDF is the minimum value, which is about 70.4%. Generally speaking, the mean value of accuracy of battery extraction is 81%. Table 3 demonstrates the experiment result of reaction rate. The highest accuracy 83.6% and lowest accuracy 70.3% are reached by NB with presence and SVM with TF-IDF.

TABLE 4 The experiment result of price extraction

		Acc	Pre (N)	Pre (Y)	Re (N)	Re (Y)	F1 (N)	F1 (Y)
ME	Boolean	0.89	0.878	0.902	0.896	0.885	0.887	0.893
	TF	0.78	0.788	0.771	0.788	0.771	0.788	0.771
	TF-IDF	0.82	0.82	0.82	0.82	0.82	0.82	0.82
NB	Boolean	0.76	0.75	0.769	0.75	0.769	0.75	0.769
	TF	0.64	0.585	0.702	0.689	0.6	0.633	0.647
	TF-IDF	0.71	0.6	0.82	0.769	0.672	0.674	0.739
SVM	Boolean	0.818	0.824	0.813	0.824	0.813	0.824	0.813
	TF	0.798	0.804	0.792	0.804	0.792	0.804	0.792
	TF-IDF	0.697	0.706	0.688	0.706	0.688	0.706	0.688
Average		0.768						

TABLE 5 The experiment result of screen extraction

		Acc	Pre (N)	Pre (Y)	Re (N)	Re (Y)	F1 (N)	F1 (Y)
ME	Boolean	0.891	0.875	0.908	0.913	0.868	0.894	0.888
	TF	0.874	0.920	0.833	0.833	0.920	0.874	0.874
	TF-IDF	0.885	0.934	0.837	0.85	0.928	0.890	0.88
NB	Boolean	0.874	0.814	0.928	0.909	0.849	0.859	0.888
	TF	0.831	0.693	0.958	0.938	0.771	0.797	0.854
	TF-IDF	0.814	0.75	0.869	0.829	0.804	0.788	0.835
SVM	Boolean	0.895	0.944	0.846	0.859	0.939	0.899	0.890
	TF	0.901	0.944	0.857	0.867	0.940	0.904	0.897
	TF-IDF	0.862	0.9	0.824	0.835	0.893	0.866	0.857
Average		0.870						

TABLE 6 The experiment result of appearance extraction

		Acc	Pre (N)	Pre (Y)	Re (N)	Re (Y)	F1 (N)	F1 (Y)
ME	Boolean	0.738	0.842	0.643	0.681	0.818	0.753	0.72
	TF	0.788	0.787	0.788	0.841	0.722	0.813	0.754
	TF-IDF	0.75	0.622	0.914	0.903	0.653	0.737	0.762
NB	Boolean	0.838	0.787	0.909	0.925	0.75	0.851	0.822
	TF	0.75	0.9	0.6	0.692	0.857	0.783	0.706
	TF-IDF	0.688	0.703	0.674	0.65	0.725	0.675	0.699
SVM	Boolean	0.848	0.784	0.927	0.930	0.776	0.851	0.844
	TF	0.826	0.765	0.902	0.907	0.755	0.830	0.822
	TF-IDF	0.826	0.843	0.805	0.843	0.805	0.843	0.805
Average		0.783						

TABLE 7 The experiment result of pixel extraction

		Acc	Pre (N)	Pre (Y)	Re (N)	Re (Y)	F1 (N)	F1 (Y)
ME	Boolean	0.966	0.953	0.977	0.976	0.956	0.965	0.966
	TF	0.954	0.976	0.935	0.930	0.977	0.952	0.956
	TF-IDF	0.885	0.884	0.886	0.884	0.886	0.884	0.886
NB	Boolean	0.851	0.818	0.884	0.878	0.826	0.847	0.854
	TF	0.954	0.923	1	1	0.897	0.96	0.946
	TF-IDF	0.862	0.8	0.915	0.889	0.843	0.842	0.878
SVM	Boolean	0.922	0.878	0.972	0.973	0.875	0.923	0.921
	TF	0.909	0.854	0.972	0.972	0.854	0.909	0.909
	TF-IDF	0.844	0.854	0.833	0.854	0.833	0.854	0.833
Average		0.905						

TABLE 8 The experiment result of price performance ratio extraction

		Acc	Pre (N)	Pre (Y)	Re (N)	Re (Y)	F1 (N)	F1 (Y)
ME	Boolean	0.93	0.958	0.904	0.902	0.959	0.929	0.931
	TF	0.89	0.917	0.865	0.863	0.918	0.889	0.891
	TF-IDF	0.87	0.857	0.886	0.906	0.830	0.881	0.857
NB	Boolean	0.91	0.898	0.927	0.946	0.864	0.922	0.894
	TF	0.89	0.843	0.939	0.935	0.852	0.887	0.893
	TF-IDF	0.76	0.661	0.921	0.932	0.625	0.774	0.745
SVM	Boolean	0.911	0.902	0.92	0.92	0.902	0.911	0.911
	TF	0.871	0.824	0.92	0.913	0.836	0.866	0.876
	TF-IDF	0.832	0.784	0.88	0.870	0.8	0.825	0.838
Average		0.874						

The mean value of accuracy achieved by different supervised learning and feature weighting method is about 70.3%. The result of extracting price by supervised learning are showed by table 4, when using ME with presence and NB with TF to extract aspect, they attain the top accuracy (89%) and lest accuracy (64%). In the statistical data (Table 5) denote that the effect of screen extraction is better, the mean value of accuracy is 87%, the highest accuracy is about 90.1%, the lowest accuracy is about 81.4%, they are achieved by SVM with TF and NB with TF-IDF. Compared with other aspects extraction, appearance extraction is worse. NB with presence and TF-IDF reach the highest accuracy 83.8% and the lowest accuracy 68.8% respectively. The experiment result of Pixel is best, the mean value of accuracy is about 90.5%, ME with presence achieve the top accuracy 96.6%, the minimum of accuracy is 84.4% and reached by SVM with TF-IDF. The result of price performance ratio is better; whose mean value is about 87.4%.

5.2 EXPERIMENT RESULT AND DISCUSSION

5.2.1 Feature extraction of product review

Observing table 2-8, we found that the gap between experiment result of every aspect extraction which use same classifier and feature weighting method has bigger difference. When extract different features, the mean value of accuracy of pixel is the highest, moreover, the result of three classifiers with all feature representation methods are better. Comparing with pixel, battery, the experiment results of price performance ratio and screen are worse. The statistical data indicate that the effect of reaction rate, price and appearance are worst, but the mean value of accuracy of these features is all above 76%. The difference between these results of every feature extraction is derived from features themselves.

For aspects, with the growth of the number of different expression ways in reviews, the accuracy of every classifier is reducing. The basic reason is that the classifiers will not be easy to converge to a better result due to the high disrupt features. For example, the pixel only has one way to express the feature. Conversely, in reviews, more ways to express reaction rate, price and appearance. For example, “value” is the synonymy of price, and “expensive”, “cheap” are the implicit expression of price. Considering all expression of every aspect will increase the difficult of aspect extraction. So, the accuracy of features which have numerous expressions will reduce.

5.2.2 Analysis of classifiers

TABLE 9 The mean value of accuracy of every classifiers for different features

	battery	reaction rate	price	screen	appearance	pixel	price performance ratio
ME	0.886	0.804	0.83	0.883	0.758	0.935	0.897
NB	0.806	0.8	0.703	0.840	0.758	0.889	0.853
SVM	0.739	0.725	0.771	0.886	0.833	0.892	0.871

(1) Table 2-8 showed that, the experiment results of various classifiers are different. Table 9 displays the mean value of accuracy of every classifier with different feature weighting method. The statistical data indicate that different classifiers are suitable for extraction of different aspects. The best mean value of accuracy of battery, reaction rate, price, pixel and price performance ratio are achieved by ME. SVM attains the best mean value of accuracy of screen and appearance. Compared with ME and SVM, the result of NB is worse, but sometime the gap between NB and other classifiers is small, such as the result of reaction rate and pixel.

(2) Table 10 displays the mean value of accuracy of all features which products have. The accuracy is all above 80%. It explains that every classifier is good for aspect extraction. The order of integral effect of aspect extraction is ME>SVM>NB.

TABLE 10 The mean value of accuracy of extraction for all features

ME	NB	SVM
0.856	0.807	0.817

(3) The standard of compatibility can be measured as following steps. Firstly, the mean values a_{ij} of average accuracy coming from seven aspects are calculated for every classifier together with every feature representation method. Secondly, the mean values a_i of one classifier for three feature representation methods are calculated. Thirdly, the difference values $\|a_i - a_{ij}\|$ of every classifier together with every feature representation method is calculated, which are shown in the Table 11, which shows the compatibility of every classifier, and the order is ME>SVM>NB.

TABLE 11 The compatibility of classifier

	Boolean	TF	TF-IDF	SUM
ME	0.021	0.001	0.023	0.045
NB	0.037	0.005	0.042	0.084
SVM	0.025	0.010	0.036	0.071
SUM	0.083	0.016	0.101	

Based on the analysis above, the ME is the most suitable supervised learning method for aspect extraction of product reviews.

5.2.3 Analysis of feature weighting method

(1) Table 12 displays the experiment result of highest accuracy of aspect extraction with different feature representation methods. Apart from screen extraction using TF, the results of other aspect extraction with presence attach the best accuracy. Moreover, sometime the gap between presence and other feature represent-tation is about 7%.

(2) Table 13 shows the mean value of accuracy of all aspects which is achieved by different classifiers with different feature representation methods. Through observing the statistical data, we found that ME, NB and SVM all achieve the highest accuracy when they combine with presence. Comparing with presence, for all classifiers, the accuracy of aspects extraction using TF will be less than 1.5%, and the accuracy of aspects extraction using TF-IDF will be less than 4.5%.

TABLE 12 The highest accuracy for every aspect with different feature weighting method

	Presence	TF	TF-IDF
battery	0.904	0.892	0.861
reaction rate	0.838	0.825	0.763
price	0.89	0.798	0.82
screen	0.895	0.901	0.885
appearance	0.848	0.826	0.826
pixel	0.966	0.954	0.885
price performance ratio	0.93	0.89	0.87

Table 13 The mean value of accuracy for all aspects achieved by different classifiers with feature representation methods

	Presence	TF	TF-IDF
ME	0.878	0.858	0.833
NB	0.844	0.812	0.765
SVM	0.842	0.827	0.781

The data from table 11 indicates that the compatibility of presence with different classifiers is higher than TF-TDF and lower than TF. Based on the analysis of feature representation methods, presence is more suitable for the experiment of this paper. The reason is that feature extraction is different with sentiment classification, it does not need to consider the times of feature appearance. We only need to know if the aspects appear or not. If an aspect appears in one review, this aspect should be extract.

6 Conclusions

With more and more product reviews on Internet, the attention from merchants and customers improved increasingly. They aim to know the detail information about the product. Being owing to no distinct grammar rule in Chinese reviews, this paper uses supervised learning to complete the experiment about features extraction from product reviews.

The experiment result indicates that ME, NB and SVM all can be used for features extraction of Chinese product reviews. For the features extraction of Chinese product reviews, The ME is more suitable than other supervised learning methods. In this paper, comparing with other feature weighting method, presence has the better compatibility and higher accuracy with different super-vised learning. Moreover, the accuracy of using ME and presence is higher. From the analysis of experiment result, we found that supervised learning can be used for aspect extraction of other products to improve the efficiency of merchants and customers.

In practice, aspects extraction can not only be used on commodity comments analysis, but also can be used in other areas, such as news reviews analysis, classification of text, and so on. Though the current machine learning methods can have good results on aspects extraction, the new deep learning methods should be explored for getting result in the future research.

Acknowledge

This work was mainly supported by National Natural Science Foundation of China (71363038), Natural Science Foundation of Inner Mongolia, China(2012MS1008, 2013MS1009), Scientific Research Project of Colleges and universities in Inner Mongolia, China (NJSZ12047), Promotion Plan of Research Center on Modern Management of Inner Mongolia (2013-2015), Scientific Research Project of Inner Mongolia University of Technology, Inner Mongolia, China (X201302).

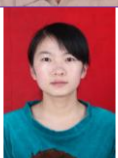
Reference

[1] Liu B. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1-167.
[2] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.
[3] Zhu J, Wang H, Tsou B K, et al. Multi-aspect opinion polling from textual reviews[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 1799-1802.
[4] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms:. the c-value/nc-value method[J]. International Journal on Digital Libraries, 2000, 3(2): 115-130.
[5] Long C, Zhang J, Zhut X. A review selection approach for accurate feature rating estimation[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 766-774.
[6] Kessler J S, Nicolov N. Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations[C]//ICWSM. 2009.
[7] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization[C]//Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006: 43-50.
[8] Sarawagi S. Information extraction [J]. Foundations and trends in databases, 2008, 1(3): 261-377.
[9] Jin W, Ho H H, Srihari R K. OpinionMiner: a novel machine learning system for web opinion mining and extraction[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 1195-1204.
[10] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 1035-1045.
[11]Kobayashi N, Inui K, Matsumoto Y. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining[C]//EMNLP-CoNLL. 2007: 1065-1074.
[12]Ghani R, Probst K, Liu Y, et al. Text mining for product attribute extraction[J]. ACM SIGKDD Explorations Newsletter,2006,8(1):41-48.
[13]Yu J, Zha Z J, Wang M, et al. Aspect ranking: identifying important product aspects from online consumer reviews[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 1496-1505.
[14]Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 804-812.
[15]Titov I, McDonald R. Modeling online reviews with multi-grain topic models[C]//Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 111-120.
[16]Lu Y, Zhai C X, Sundaresan N. Rated aspect summarization of short comments[C]//Proceedings of the 18th international conference on World wide web. ACM, 2009: 131-140.
[17]Yi J, Nasukawa T, Bunescu R, et al. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques[C]//Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003: 427-434.qz
[18]Ma T, Wan X. Opinion target extraction in Chinese news comments[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 782-790.
[19]Kim S M, Hovy E. Extracting opinions, opinion holders, and topics expressed in online news media text[C]//Proceedings of the Workshop on Sentiment and Subjectivity in Text. Association for Computational Linguistics, 2006: 1-8.
[20]Su Q, Xu X, Guo H, et al. Hidden sentiment association in chinese web opinion mining[C]//Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 959-968.
[21]Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing[J]. Computational linguistics, 1996, 22(1): 39-71.
[22] Chen S F, Rosenfeld R. A survey of smoothing techniques for ME models[J]. Speech and Audio Processing, IEEE Transactions on, 2000, 8(1): 37-50.
[23]Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification[C]//IJCAI-99 workshop on machine learning for information filtering. 1999, 1: 61-67.
[24]Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.
[25]McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit."http://mallet.cs.umass.edu. 2002.
[26]Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27,2011.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Authors

< Yan Zhao >, <July 19, 1982>,<Tong Liao, Inner Mongolia, China>

University studies: Teacher in College of Management, Inner Mongolia University of Technology, Associate Professor
Scientific interest: Internet Public Opinion; Information Dissemination; Management Science and Engineering; Emotion Recognition
Publications: more than 12 papers in the core journals
Experience: Doctor's degree in Economics from Beijing Central University of Finance and Economics in June 2011. Since September 2006, he taught in Department of Business Administration of Management College in Inner Mongolia University of Technology after he got master's degree in the field of Management Science and Engineering.



<Suyu Dong >, <October 13,1989 >,<Liang Cheng, Inner Mongolia, China >

University studies: Student in College of Management, Inner Mongolia University of Technology, Student
Scientific interest: Management Science and Engineering; Sentiment analysis
Experience: Bachelor degree in Software Engineering from East China Jiaotong University in June 2012. Since September 2012, she majored in Management science and Engineering of Management College in Inner Mongolia University of Technology.

< Hua Yang >, <June 12, 1979>,<Hohhot, Inner Mongolia, China>

Current position: Engaging in the work of educational administration and basic computer experimental teaching work at the Information engineering college, Inner Mongolia University of Technology.
Scientific interest: Image transmission, Information management
Publications: has published 2 papers in the general journals.
Experience: Master's degree in Computer Technology from Inner Mongolia University of Technology in January 2014.Since 2011, she tanght in Introduction of Computer Technology and the experimental course of language. The main research areas include Image transmission and Information management.

< Jing Yang >, <May 09, 1981>,<Feng Zhen, Inner Mongolia, China>

University studies: Teacher in College of Mechanics, Inner Mongolia University of Technology, Associate Professor
Scientific interest: Internet Public Opinion; Information Dissemination; Emotion Recognition
Publications: more than 8 papers in the core journals
Experience: Doctor's degree in Economics from Beijing Central University of Finance and Economics in June 2012. Since September 2007, she taught in Department of Industrial Engineering of Mechanics College in Inner Mongolia University of Technology after she got master's degree in the field of Technology Economy and Management. The main research areas include Internet public opinion, Information dissemination, and Emotion recognition.