ATLANTIS PRESS

Advances in Computer Science Research, volume 79

2nd International Conference on Sensor Network and Computer Engineering (ICSNCE 2018)

The Extraction of Comment Information and Sentiment Analysis in Chinese Reviews

Li Danyang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
e-mail: 821563942@qq.com

Zhao Yingze
School of Marxism
Xi'an Jiaotong University
Xi'an, China
e-mail: yingze1013@163.com

Fan Huimin
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
e-mail: 492896361@ qq.com

***Abstract*—Sentiment analysis, also known as opinion mining, refers to the emotional tendencies expressed by the critics through the analysis of the content of the text. The task of text sentiment analysis mainly includes the classification of sentiment, the extraction of sentiment information and the retrieval and induction of sentiment information. Based on CRF, this paper will extract several pairs of theme words and sentiment words exist in the e-commerce review, and judge the sentiment inclination of the extracted sentiment words. The experimental results show that CRF has a good effect on the extraction of emotional information.**

Keywords-CRF; Extract Theme Words; Extract Sentiment Words; Sentiment Analysis

I. INTRODUCTION

With the rapid development of web 2.0 technology, there have been network reviews on the platform with exponential growth, such as micro-blog reviews, news commentaries and e-commerce reviews, etc. E-commerce is a business activity based on information network technology and centered on commodity exchange. With the diversification of consumer information in twenty-first Century, the trading volume of e-commerce has increased rapidly. It has become an important part of the national economy and plays an extremely important role. For the e-commerce platform, the comment information greatly affects the consumer's purchase decision[1]. By extracting the comment information in the Chinese comment text, it can not only guide consumers to make rational consumption, but also help the merchants to improve the quality of the products. The comment information includes the theme words and sentiment words that appear in the commentary, the theme word refers to the evaluation object in the comment, which is the modification object of the sentiment word in the sentence, which is usually expressed as some attribute of the product. The extraction of comment information is one of the key tasks of text sentiment analysis, the existing methods for extracting information from reviews are mainly divided into rules/template and statistical methods.

II. EXTRACTION METHOD OF EVALUATION INFORMATION

The rule/template method is mainly based on the characteristics of the text itself, making the corresponding rules or templates to identify the specific field of evaluation objects. Liu Bing first proposed the problem of evaluation object extraction, he used the noun with high frequency as the evaluation object, and used the nearest adjective from the evaluation object as an sentiment word[2]. According to the characteristics of Chinese language, Qiu Yunfei and Chen Yifang put forward the method of extracting commodity evaluation objects by using word characteristics and syntactic analysis[3]. However, the rule/template method requires domain experts to define the evaluation objects and rules in the corresponding field, so it cannot satisfy the emerging neologisms, and has no cross domain and portability, therefore, the most effective extraction method is based on the statistical method.

The statistical extraction method uses a trained statistical model to extract comment information. Niklas Jako et al. proposed the use of conditional random field model to extract the evaluation object, and model the extraction problem of the evaluation object into a sequence marking task[4]. Jin Lijun and others have studied the method of automatic recognition based on SVM[5]. This paper mainly studies the application of CRF statistical model in the extraction of comment information.

III. COMMENT INFORMATION EXTRACTION BASED ON CRF STATISTICAL MODEL

A. Review information extraction process based on CRF

Based on the statistical method, this paper uses the CRF model as the main model and combines the constructed emotional dictionary to extract the comment information, as

shown in Figure 1, it mainly includes building emotional lexicon, data preprocessing, part of speech tagging, training CRF model, using CRF model to extract theme words and sentiment words, judging the sentiment inclination of the extracted sentiment words and exporting final results.

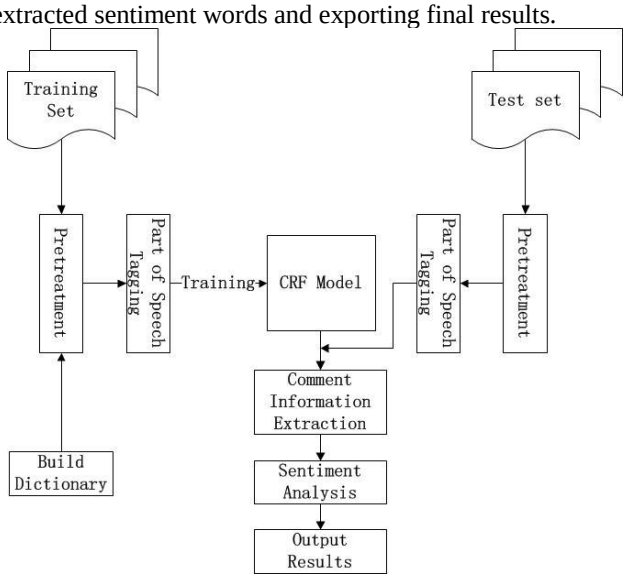

Figure 1. Review Information Extraction Process Based on CRF

B. *Data pretreatment*

Data pretreatment is an essential part of text data mining. In this paper, it mainly includes the following parts:

1) Building a sentiment word dictionary: This paper will extend the sentiment dictionary for the use of the corpus. Firstly, the new dictionary is applied to Chinese word segmentation, which makes segmentation more accurate and largely avoids the destruction of the theme words and emotional words when they are segmenting. Secondly, the new emotional dictionary is applied to the sentiment tendencies of sentiment words.

2) Chinese word segmentation: Chinese word segmentation is the basis of text mining, but Chinese is not as natural as English word, so Chinese word segmentation is much more complicated than English word segmentation. In this paper, the more mature Jieba segmentation algorithm combined with the new sentiment dictionary can be used to carry out Chinese word segmentation, which is achieve a good segmentation effect.

3) Removing the stop words: The stop words are words that are completely useless or meaningless, such as auxiliary words, mood words, punctuation marks and so on. The removal of stop words can improve the efficiency of retrieval, save storage space, and exclude interference words.

4) Sequence labeling: In order to extract more accurate theme words and sentiment words, this paper divides the elements in the text into 3 categories by using sequence labeling: the theme words is marked as T (Theme), the sentiment word is marked as S (Sentiment), and the rest of the words are labeled as O (Other).

The results of some of the data after the above pretreatment are shown in Figure 2.

Figure 2. An example of data pretreatment results

C. *Part of Speech Tagging*

CRF model is actually transforming information extraction problem into sequence labeling problem. Therefore, in order to train CRF model, we need to process part of speech tagging of corpus in addition to three kinds of customized tags. Part of speech is used to describe the function of a word in context. Part of speech tagging is also called part-of-speech tagging, which refers to the process of marking a correct part of speech for every word in the word segmentation result. Different languages have different set of part of speech tagging. In this paper, the annotation set of parsing tree is used. Part of the data after the tagging is shown in Table 2.

Figure 3. An example of part of speech tagging

D. *The Introduction of CRF*

Conditional Random Fields[6], CRF or CRFs for short, was first proposed by John Lafferty in 2001. It combines the characteristics of maximum entropy model and hidden Markov model, and it is a probabilistic undirected graph model. It is often used in sequence segmentation and tagging,

and the conditional probability of the output node can be calculated under the given input node. CRF model can better capture the context information[7], accurately identify the key information, has been widely applied to many fields of natural language processing, and has a good performance in Chinese natural language processing tasks, such as the part of speech tagging, machine translation, prosodic structure prediction and speech recognition.

CRF is an undirected graph model, and Lefferty and others define CRF as: order G=(V,E), where G represents an undirected graph, and V and E belong to a set in an undirected graph. In this expression, V represents the set of nodes, and E represents the set of edges. In the tag sequence, the elements and the nodes in the graph correspond to one by one. Under the condition of known observation sequence X, if the distribution of the random variable satisfies Markov property, that is, the node is adjacent to the node in graph G, then it is called a conditional random field. The formalized description of CRF is as follows:

Set G= (V, E) is an undirected graph, the V represents the set of vertices, and the E represents the set of the edges. Y={Yv|v∈V} represents the index of vertices in figure G, that is, each vertex corresponds to the composition Yv of the marked sequence represented by a random variable. Therefore, on the condition of X, the form of joint distribution related to G is p(y1,y2,…,yn|X), in which y is a marker sequence and X represents the observation sequence. If the random variable r satisfies the Markov property about G, that is

$$p(Y_v \mid X, Y_u, u \neq v) = p(Y_v \mid X, Y_u, u \sim v) \qquad (1)$$

In the above formula, u~v indicates that u is adjacent to v in graph G, and (X, Y) constitutes a conditional random field.

In theory, if graph G represents the conditional dependence between the labeled sequences to be modeled, then its structure can be arbitrary. But when modeling the sequence annotation task, the most simple and general graph structure is: a simple first order chain corresponding to the elements of Y is formed. This CRF is generally called linear - chain CRF, the model of linear - chain CRF is shown in Figure 2, X= (x1, x2,... xn) is the observation sequence, y= (y1, y2,..., yn) is the output sequence.

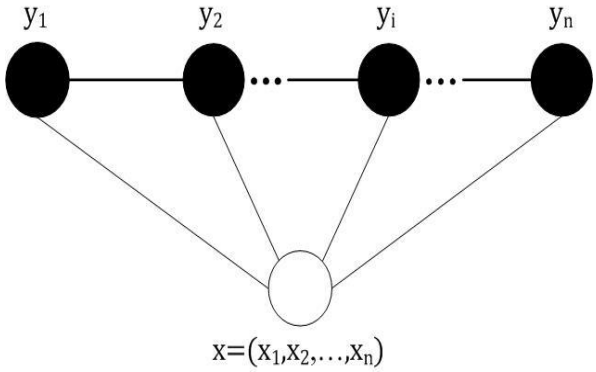

Figure 4. Structural Representation of Linear - chain CRF

Given the observation sequence, the sequence conditional probability of the output is as follows:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i \sum_k \lambda_k t_k(y_{i-1}, y_i, x, i) \right) + \sum_i \sum_k \mu_k s_k(y_i, x, i) \qquad (2)$$

tk(yi-1,yi,x,i) is a state transfer function; sk(yi,x,i) is a state feature function; tk, sk are all characteristic functions; λk, μk is the weight of the characteristic function, which is learned by training; Z(x) is a normalization factor:

$$Z(x) = \sum_y \left(\sum_i \sum_k \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_i \sum_k \mu_k s_k(y_i, x, i) \right) \qquad (3)$$

As one of the most important undirected graph structures, linear - chain CRF has been applied to the practical research, and most of the Natural Language Processing research tasks all use linear - chain CRF.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In the experiment, the data set of this paper is divided into two parts: training set and test set based on the data set provided by the Big Data & Computing Intelligence Contest in 2017. The CRF model was trained with the training set, and then the theme words and Sentiment words were extracted from the test set, and judge the sentiment inclination of the extracted sentiment words. In this paper, F1 is used as the index of evaluation model.

A. Experimental evaluation index

There are three main evaluation indexes commonly used in data mining and natural language processing, including accuracy, recall rate and F1. Among them, the accuracy rate for the prediction results is to indicate how many positive samples are true in the predicted sample. The recall rate is aimed at our original sample, which indicates that the number of examples in the sample is predicted correctly. And F1 is the harmonic average of the accuracy and recall. In a word, this paper uses the F1 value as an evaluation standard. In this paper, the calculation formula is as follows:

$$P = \frac{\text{Extracting the correct theme words/Number of sentiment words}}{\text{Extracting theme words / Number of sentiment words}} \qquad (4)$$

$$P = \frac{\text{Extracting the correct theme words/Number of sentiment words}}{\text{Theme words in data set / Total number of sentiment words}} \qquad (5)$$

$$F = \frac{2 * P * R}{(P + R)} \qquad (6)$$

B. Training CRF model

The CRF model is trained after the data set is preprocessed by word segmentation, annotation and so on. This paper uses the open source tool CRF++ to train the training model, the feature template needs to be prepared

before the training model, and the feature template file of this article is shown in Figure 3.

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

U23:%x[0,1]

# Bigram
B
```

Figure 5. The Feature Template

The T in "T**:%x[#,#]" represents the template type, where there are two templates altogether. The first is Unigram template, the first character is U, which is a template for describing unigram feature; The second is Bigram template, the first character is B. Two "#" respectively represent relative row offsets and column offsets, each line of "%x[#, #]" generates a CRF point (state) function: f(s,o), where "s " is the t time label (output), "o" is t times the context. The trained CRF model contains feature template and feature dimension, data set number, characteristic function and the weight of information, a series of information is output in the training process, and some of the information is shown in Figure 4. The meaning of the parameter information is as follows:

Iter: The number of iterations. When the number of iterations reaches the max, the iteration is terminated.

Terr: Mark error rate.

Serr: Sentence error rate.

Obj: The value of the current object. When the value converges to a definite value, the training is completed.

Diff: The relative difference between the value of the last object. When this value is lower than eta, the training is completed.

```
reading training data: 100.. 200.. 300.. 400.. 500..……700.. 800.. 20000..
Done!6.79 s

Number of sentences: 20000
Number of features:  1460484
Number of thread(s): 4
Freq:                1
eta:                 0.00010
C:                   4.00000
shrinking size:      20
iter=0 terr=0.16280 serr=0.87220 act=1460484 obj=453431.34851 diff=1.00000
iter=1 terr=0.16280 serr=0.87220 act=1460484 obj=231041.34805 diff=0.49046
iter=2 terr=0.16280 serr=0.87220 act=1460484 obj=212661.60141 diff=0.07955
……
iter=513 terr=0.00153 serr=0.02890 act=1460484 obj=13278.86569 diff=0.00005
iter=514 terr=0.00153 serr=0.02885 act=1460484 obj=13278.19406 diff=0.00005
iter=515 terr=0.00153 serr=0.02880 act=1460484 obj=13277.70131 diff=0.00004

Done!202.86 s
```

Figure 6. Output File

C. *Experimental Results and Analysis*

After extracting the theme words and sentiment words, the next step is the judgment of the sentiment inclination of the extracted sentiment words, and this step is much simpler. If the sentiment word belongs to the positive affective dictionary, the sentiment is positive. If it belongs to the negative sentiment dictionary, the sentiment is negative, otherwise it is neutral. The experimental results before and after the optimization dictionary are compared as shown in the table below. Table 3 is the experimental result of no emotional dictionary. Table 4 is the experimental result after optimization.

content	theme	sentiment_wo	sentiment_a
跟我实体店买的宝贝完全质量不一样，感觉质量好差感觉特别	质量;NULL;	差;假;	-1;-1;
酒质很一般。。。。。。。。也就这价格，期望过高。	酒质;期望;	一般;过高;	0;-1;
差的原因是没看到赠品，卖家如果没有就不要写上做虚假宣传	NULL;NULL;	差;虚假;	-1;-1;
不错，是我想要的，快递小哥服务一级的棒，	NULL;服务;	不错;棒;	1;1;
真**差，我都无语了。眼瞎才买了	NULL;	差;	-1;
给奶奶买的，效果不错，厂家服务也好，给好评[追评]	效果;服务;N	不错;也好;好评	1;1;1;
粉特别干，送的唇彩不能用，颜色很难看	颜色;	难看;	-1;
用了好几个了，很好	NULL;	很好;	1;
感觉还可以	感觉;	还可以;	1;
死机n次也是醉了			
试用中，良心产品	产品;	良心;	1;
做工很一般不知道穿了怎么样	做工;	一般;	0;

Figure 7. Examples of no optimized results

content	theme	sentiment_wo	sentiment_a
跟我实体店买的宝贝完全质量不一样，感觉质量好差感觉特别	质量;感觉;	差;假;	-1;-1;
酒质很一般。。。。。。。。也就这价格，期望过高。	酒质;期望;	一般;过高;	0;-1;
差的原因是没看到赠品，卖家如果没有就不要写上做虚假宣传	NULL;卖家;	差;虚假;	-1;-1;
不错，是我想要的，快递小哥服务一级的棒，	NULL;服务;	不错;棒;	1;1;
真**差，我都无语了。眼瞎才买了	NULL;NULL;	差;无语;	-1;-1;
给奶奶买的，效果不错，厂家服务也好，给好评[追评]	效果;服务;N	不错;也好;好评	1;1;1;
粉特别干，送的唇彩不能用，颜色很难看	粉;颜色;	干;难看;	-1;-1;
用了好几个了，很好	NULL;	很好;	1;
感觉还可以	感觉;	还可以;	1;
死机n次也是醉了	NULL;	次;	-1;
试用中，良心产品	产品;	良心;	1;
做工很一般不知道穿了怎么样	做工;	一般;	0;

Figure 8. Examples of optimized results

In the above table, there are no definite theme words in some comments, but there are corresponding sentiment words. In this case, the theme word is marked as NULL. After comparison, it is found that after optimization, the recognition of the theme words is more accurate than before, and the accuracy rate is further improved. The next table 5 is the comparison of the F1 values before and after the optimization. The comparison results from the table show that the F1 value of the optimized dictionary is about 3% higher than that before the optimization.

TABLE I. COMPARISON OF F1 BEFORE AND AFTER OPTIMIZATION

Data Set	F1
20,000 comments	0.58498
20,000 comments	0.61827

V. CONCLUSION

The accurate recognition of the theme words and sentiment words in the commentary is the key to the extraction of comment information and the basis for further analysis of the text[8]. The extraction of comment information in Chinese review is of great significance to both the merchant and the consumer. The merchant can adjust the goods according to the comment information or improve the quality of the product. The consumer can also make auxiliary decisions according to the comment information.

In this paper, the CRF statistical model is used to extract the theme words and sentiment words in the comment statement, and it is proved that the CRF model is effective in identifying subject words and emotional words. In addition, this paper optimizes the emotional dictionary, which combines data sets and CNKI's emotional lexicon to build an new emotional dictionary that is more suitable for this paper. The experimental results show that the optimization dictionary makes the recognition of the theme words and sentiment words more accurate, and the F1 value is further improved.

Of course, in addition to the CRF model used in this paper, there are other methods that can also extract comment information, such as LDA theme model, dependency parsing method and so on.

Besides, there are still some shortcomings in this paper. Chinese expression is much richer than English, In Chinese, there are some irony and even the use of network language statements can not accurately identify theme words and sentiment words, so we need further study of Chinese semantics and so on.

REFERENCES

[1] Li Piji, Ma Jun, Zhang Dongmei, etc.. “Label extraction and sorting in user reviews,” J. Journal of Chinese Information Processing, 2012, vol. 26(5), pp. 14-19,45.

[2] Hu MQ, Liu B. “Mining and summarizing customer reviews,” C.Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA. 2004, pp.168–177.

[3] Qiu Yunfei, Chen Yifang, Wang Wei, etc.. “Product evaluation object extraction based on word character and syntactic analysis,” J. Computer Engineering, 2016, vol. 42 (7), pp. 173-180.

[4] Jakob N, Gurevych I. “Extracting opinion targets in a single-and cross-domain setting with conditional random fields,” Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts. 2010.

[5] Jin Lijun. “Research on the automatic recognition of the usefulness of SVM based search commodity reviews,” D. Harbin Institute of Technology, 2013.

[6] John D Lafferty, Andrew McCallum, Fernando C N Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” Proceedings of the 18th International Conference on Machine Learning. Williamstown, MA, USA, 2001, pp. 282-289.

[7] Wang Rongyang, Ju Jiupeng, Li Shoushan, etc.. “Research on feature extraction feature of evaluation object based on CRFs,” J. Journal of Chinese Information Processing, 2012, vol. 26 (2), pp. 56-61.

[8] Xia yuan, Zhang Zheng . “Evaluation of object extraction based on CRF,” J. Computer Systemsand Applications.2017, vol. 26 (11), pp. 254-259.