Journal of Computers Vol. 28, No. 6, 2017, pp. 139-150
doi:10.3966/199115992017122806012

A Bipartite Graph Model for Implicit Feature Extraction in Opinion Mining

Li-zhen Liu[1], Wan-di Du[1], Wei Song[1*] and Han-shi Wang[1]

[1] Information and Engineering College, Capital Normal University,
Beijing 100048, China
msongwei@gmail.com

Received 23 July 2016; Revised 07 February 2017; Accepted 12 February 2017

Abstract. Along with the rapid development of online shopping, it's hard for buyers to find useful information in a short time so it makes sense to do research on opinion mining which fundamental work is focused on product reviews mining. Previous studies mainly focus on explicit features extraction whereas often ignore implicit features which haven't been stated clearly but containing necessary information for analyzing comments. So how to accurately mine features from web reviews has important significance for summarization technology. In this paper, explicit features and "feature-opinion" pairs in the explicit sentences are extracted by Conditional Random Field and implicit product features are recognized by a bipartite graph model based on random walk algorithm. The experiment results demonstrate that these two models we proposed can improve the accuracy of feature and collocation extraction are preferred over baselines.

Keywords: bipartite graph, conditional random fields, implicit feature, opinion mining

1 Introduction

Nowadays, the degree of activity in Chinese online market is still high and it's time-consuming for customers to read a flood of comments. According to China Internet Network Information Center (CNNIC) [1] survey data shows the most important factor affecting consumer purchase decisions is product reviews which can provide valuable information for buyers to do decision-making. Therefore, it has great significance and value to analyze product reviews, then opinion mining as an emerging research area of unstructured information mining will be born.

The purpose of opinion mining is to make decisions by analyzing comments on the network. In general, appropriate decisions cannot be made through small amount of comments. On the other hand, it's strenuous for artificial methods to extract classification from massive reviews. Opinion mining technology is not only based on data mining and text mining technology, but also needs some ability to understand and analyze text sentiment. Unlike traditional information extraction, the comments to be mined are usually not expressed explicitly and independently [2].

Recently, researches on English processing have achieved significant success whereas still in initial stage on Chinese. Due to the cultural background and language differences between Chinese and English, the research achievements in English comments field cannot be directly applied to Chinese. Thus it's necessary to directly discuss on Chinese network comments. Currently, a few typical Chinese e-commerce websites have done several inductive statistics, for example Tmall.com gives phrases and its quantity to other users for giving reference, Amazon.cn gives star ratings to goods based on user reviews, but all of these are coarse-grained extraction, resulting in interpreting out of context which are limited to objectively understand reviews for users, for example some extracted labels can only represent the experience of a certain people, and some phases express incompletely [3]. When the number of users is large, the problem will be more prominent.

* Corresponding Author

Therefore, this paper considers Chinese comments on the network as the research object. The problem of mining web reviews can be divided into two parts: one is mining product features and opinion words, another is extracting opinion target association. The first part can be defined as keyword extraction in which product features are divided into explicit and implicit features. Explicit features appear in the comments explicitly and implicit features don't appear explicitly in the sentences but are alluded to opinion words and often implied in the context. For example, in a cellphone comment sentence like "The quality of this phone is high, but a little expensive" contains two opinion targets, namely quality and price, however quality is an explicit feature and price is an implicit feature. We will turn the collocation extraction problem into the sequence labeling problem, using CRFs model and semi-supervised learning method to extract and associate features and opinion words in the comment text. For instance, in the above example, pixel is a feature of the phone, high is the opinion word corresponding this feature and <pixel, high> is a group correlated pairs. For implicit comment sentences, we will mine keywords and collocation pairs through building a bipartite graph and calculating probability to extract the collocation pairs which can extract implicit features effectively.

In general, our contributions are the following:

- We provide a model for extracting explicit features from large Chinese comments datasets.
- A novel model is used to solve the problem of implicit features extraction, and verify the feasibility of this model under some test.
- We experimentally evaluate our methods against some existed [14, 15, 20] on feature extraction for both precision and recall.

The remaining parts of this paper are organized as follows: Section 2 introduces some related works; Section 3 proposes related knowledge and our approach; The experimental results are presented, evaluated and discussed in Section 4; Section 5 presents our conclusions and future work detailedly.

2 Related Work

2.1 Opinion Mining

Opinion mining is widely studied in the field of information extraction and contains product features mining and opinion words mining in Chinese reviews [4]. Ding, Liu and Yu [5] define the opinion words that an opinion can be expressed on any node or attribute of the node. Hu and Liu [18] present two kind of features in product features mining, namely explicit and implicit features, but their approach maybe not effective in extracting implicit features. Up to now, explicit feature extraction methods can be divided into two types, one is the artificial defined methods, for example Kobayashi, Inui and Matsumoto [6] define the product features of cars by artificial defined method which using a triple(<Attribute, Subject, Value>) to represent each feature, but this approach may lose some important features; another is the automatic extraction methods, Stone and Choi [7] propose a visualization tool which leverages machine learning algorithms to interpret user demand from user-generated content, and no manual labeling is required by the designer, but the accuracy of labeling is not ideal. There are other models such us Conditional Random Field (CRF) Sequence Model, Latent Dirichlet Allocation (LDA) probability model [8]. Currently, more and more studies of evaluation object based on CRFs have appeared. Compared to LDA topic model, CRFs have processed better on product reviews. In 2010, CRFs model is firstly mentioned in Jakob and Gurevych's paper [9] for product features recognition, but they only deal with opinion words. After that, a lot of scholars have made some achievements in product features extraction based on CRFs model, Huang, Liu and Peng [10] use CRF Sequence Model to mine product features from comment sentences sequence. However, the applications in extracting opinion target association based on CRFs model, especially in Chinese field, is still relatively a little.

Although the above work has achieved good results in explicit features extraction, the accuracy of these methods used in the implicit features extraction cannot achieve the desired and often lead to erroneous or inaccurate extraction results [11]. Many people are aware of the existence of implicit features in [12-13], whereas the existing methods for mining implicit features are not very mature. Su, Xiang and Wang [14] mainly use Point-wise Mutual Information (PMI) to associate semantic analysis with product features and opinion words which match probability in training set, but the quantitative results are not provided. Hai, Chang and Kim [15] propose a co-occurrence association rule mining (CoAR) algorithm to select implicit product features, but the facts have been neglected. Above all sorts

of implicit product features extraction methods can be evaluated only for special words, it is not ideal for general words, such as “good”, “well”, “poor”.

2.2 Product “Feature-Opinion” Pairs Mining

Finding and establishing the relationship of opinion words and feature words accurately, which also called “feature-opinion” pair, is a hot topic in domestic and abroad academe. Extracting “feature-opinion” pairs has been studied by many domestic and international scholars. Some foreign scholars have used rule-based approach, for instance Bloom, Garg and Argamon [16] build 31 syntactic rules manually to present the relationship between object and opinion word, but they don’t consider the sentiment of pairs; Using the co-occurrence of opinion and feature, a method is proposed by Popescu and Etzioni [17] aimed at finding possible views, but it needs too much training data and seed words. In this method, the syntactic dependency and part-of-speech tags calculated by the MINIPAR framework are combining with rules to improve the accuracy. And Hu and Liu [18] find frequent product features as candidate feature indicators according to word frequency method, then using artificial methods to filter noise, but the accuracy and recall rate is not ideal enough. Through studying Chinese reviews on the Internet, we find that these reviews on the website are hardly standard in grammar, and even having many syntax errors due to the accuracy of existing dependency analysis techniques, needing large-scale training corpus and the higher quality of the corpus. This paper uses opinion to identify the opinion target association.

3 Implicit Feature Identification

A number of studies shown that it’s essential to use a special text processing technology for web produce comments with brief text, diverse language, sparse data and high in noise, which is different from traditional documents [19-21]. Therefore, we propose the approach in this chapter mostly considering opinion mining. The main content of this chapter are “feature-opinion” pairs collocation and implicit features extraction. System flowchart is shown in Fig. 1.

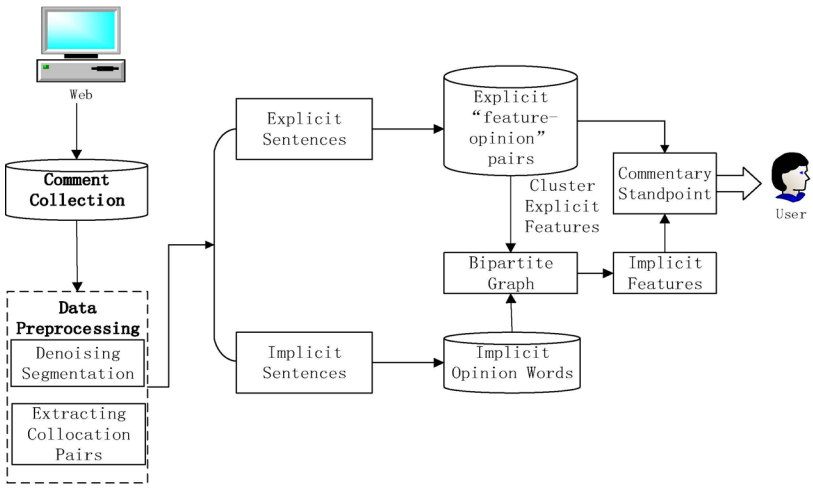

Fig. 1. System flowchart

Product reviews are climbed from e-commerce sites and all reviews can be seen as a document in which each sentence is a comment. In order to obtain the high-quality and reliable experimental data, we firstly proceed review datasets preprocessing, including segmentation, denoising which covers comments emotions, special characters, or off-topic sentences (for example “I am very happy to receive the goods”, “This style is what I want”) and so on. Because of the particularity of Chinese grammar, we need to do word segmentation using CAS ICTCLA[1] segmentation system. Training data is labeled by HowNet [22]

[1] ICTCLAS is the abbreviation for “Institute of Computing Technology, Chinese Lexical Analysis System”. We published ICTCLAS as free software. The full source code and document of ICTCLAS is available at no cost for non-commercial use. Welcome researchers and technical users download ICTCLAS from Open Platform of Chinese NLP (http://www.nlp.org.cn/).

and trained by models in order to extract product features and opinions, and comment sentences can be divided into explicit sentences and implicit sentences according to the extraction results. Then we cluster "feature-opinion" pairs in explicit sentences to construct a bipartite graph, using random walk algorithm to calculate the probability of implicit features and achieving the extraction of "feature-opinion" pairs.

3.1 Collocation Mining Based on CRFs

The main content of this section is to extract "feature-opinion" pairs based on the Conditional Random Fields. We firstly introduce the relevant research of collocation extraction based on CRFs, then propose the structure of feature template and the feature function of extraction model based on CRFs. Finally, the labeling method of CRFs model and the corresponding relationship of input and output sequences have been presented.

As CRF applied in Chinese word segmentation, sentiment analysis and part-of-speech tagging, we transform the problem of collocation extraction into the sequence annotation task. CRFs [23] is a kind of conditional probability distribution model used to construct and mark serialized data, under given a set of input random variables to output another random variables, which is characterized by the assumption of random variables constitute Markov Random Fields. The training target of CRFs is maximize the conditional probability and CRFs model is usually defined as undirected graph model or Markov Random Fields, avoiding the shortcomings of Hidden Markov Models (HMM) and Maximum Entropy Markov Models (MEMMS), calculating sequence joint probability on the basis of the given observed data sequence [24].

Let $X=(X_1,X_2,\cdots,X_n)$ and $Y=(Y_1,Y_2,\cdots,Y_n)$ represent the same length sequence of random variables, among them X is the input variable indicating the observation sequence required labeling, like a word sequence of text, and Y is the output variable indicating the symbol sequence. Let $P(Y|X)$ is the linear CRFs, when the value of the random variable X is x , the conditional probability of the random variable Y is y can be calculated as:

$$P(\mathrm{y}|x)=\frac{1}{Z(x)}\exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}w_k f_k(y_{i-1},y_i,x,i)\right\}. \quad \textbf{(1)}$$

$$Z(x)=\exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}w_k f_k(y_{i-1},y_i,x,i)\right\}. \quad \textbf{(2)}$$

Where $w=(w_1,w_2,\cdots,w_k)^T$ represents weight vector, $Z(x)$ is a normalizing factor and K represents transferring characteristics and status feature as well as their weights. $f_k(y_{i-1},y_i,x,i)$ is the feature function of the collocation extraction task marking feature vector located between T and $T-1$ for the entire observation sequence X, which considers not only the context of current term, but also previous state. The eigenvalues can be 0, 1, or maybe any real. The prediction problem of conditional random fields is given the input sequence (observation sequence) X and conditional random field $P(Y|X)$, solving the maximum conditional probability of output sequence (tag sequence) $Y*$, namely marking observation sequence. This paper uses the famous Viterbi algorithm to identify the global optimal sequence, then on the basis of the sequence, extracting "feature-opinion" pairs. With high efficiency and simple computation, this algorithm is accepted at present as the most effective training algorithm for dealing with product reviews.

Collocation Extraction is defined as extracting commodity feature and opinion word which is expressed as <product feature, opinion word> in the comment text, like the <shape, beautiful> in the comment "The shape of iPhone is beautiful." The process of identifying product feature and opinion word can be seen as under the condition that input a string of words $w_1\ w_2\ w_3\cdots w_n$, the maximum probability labeled sequence $L_1\ L_2\ L_3\cdots L_n$ is outputted. Here we introduce seven mark symbols $\{B-F,I-F,E-F,B-O,I-O,E-O,O\}$, in which '$B-F$' represents the initial word describing property features, '$I-F$' represents the intermediate term describing property features, '$E-F$' represents the end term describing property features, '$B-O$' represents the opinion word which adjacent

to the feature, '$I-O$' represents the intermediate term of opinion word, '$E-O$' represents the end term of opinion word, OFF represents the unrelated word. In this way, we turn the property feature recognition problem into the process of generating the maximum probability labeled sequence.

Choosing a good feature can greatly improve extraction performance, thus it's important to construct the feature template for labeling sequence based on CRFs. Features we used including word feature, part of speech feature, position feature, interdependent syntactic relation feature, whether is an explicit comment sentence, which are as follows:

Word feature. The words in the comments which are dealt with word segmentation contain word itself and context words. Word feature can directly affect extraction performance and especially effective for high-frequency words, due to a specific opinion word can only modify some product features which can be constructed semantic features.

Part of speech feature. Wiebe, Wilson and Bruce [25] has shown that adjectives can be used to determine the subjective and objective of sentences. In addition, through observing massive corpus of reviews, we find that idioms and phases often appear in the comments and nouns or noun phases are used to describe product objects. Therefore, this paper selects adjectives and idioms as opinion words and nouns or noun phases as candidate feature objects.

Position feature. Position Feature is the location of product feature or opinion word in the comment and the context positional relationship between product feature and opinion word.

Whether is an explicit sentence. This feature considers whether this comment is an explicit sentence or not. In general, an explicit sentence has a pair of <product feature, opinion word> at least, whereas an implicit sentence only has opinion words.

Interdependent syntactic relation feature. The dependence relationship between words can be stated as interdependent syntactic relation which is useful to extract information, improve model performance and extraction performance.

Above all, after building feature templates and training model using training corpus, we get CRFs extraction model and can mine the collocation pairs through entering new corpus. For example, the results of labeling and training the sentence "This phone is beautiful with durable battery and good performance, but the price is a little expensive." are shown in Table 1. The example has five columns, which represent "word feature", "part of speech feature", "position feature", "interdependent syntactic relation feature", "whether is an explicit comment sentence". All elements of the model in which we can obtain four groups of collocation as <phone, beautiful>, <battery, durable>, <performance, good>, <price, expensive>.

Table 1. Processing results of example sentence

This	r	0	1	OFF
phone	n	1	1	B-F
is	vn	2	1	OFF
beautiful	a	3	1	B-O
with	p	4	1	OFF
durable	a	5	1	B-O
battery	n	6	1	B-F
and	c	7	1	OFF
good	a	8	1	B-O
performance	n	9	1	E-O
,	x	10	1	OFF
but	c	11	1	OFF
the	r	12	1	OFF
price	n	13	1	B-F
is	vn	14	1	OFF
a	q	15	1	OFF
little	d	16	1	OFF
expensive	a	17	1	B-O
.	x	18	1	OFF

Different words or phrases are often used to describe the same feature of an item by customers, like "facade", "external", "aspect" and so on are used to describe the appearance of the mobile phone. In order to extract implicit feature easily, making similar features have the same description, we cluster n features of "feature-opinion" pairs $\{< f_1, o >, < f_1, o >, \cdots, < f_n, o >\}$ for matching each opinion word o. Our method of clustering is based on this paper [26] which automatically identify some labeled examples by semi-supervised method, then unlabeled features are assigned to a cluster using naïve Bayesian based on EM formulation [27]. When EM converges, the classification labels of all the attributive words give us the final grouping. Thus the implicit feature extraction problem is turned to a classification problem.

3.2 Implicit Feature Extraction

It's not hard to find that explicit features in the reviews are obvious and the number of it is limited, therefore this paper can extract explicit features accurately based on CRFs model. Whereas extracting implicit features using rule-based methods with full coverage is difficult, mining implicit features via the calculation results of random walk algorithm and calculating the probability of candidate features, this paper will identify implicit features according to the confidence to match the corresponding implicit features.

We utilize features and opinion words previously collected to build a graph $G = (V, E)$. A bipartite graph $G = (V, E) = (V_1, V_2, E)$ consists of two sets of vertices, denoted by V_1 and V_2 respectively, the weight of each edge is define by W and a set of edges denoted by $E \subseteq V_1 \times V_2$ in which the vertex of each edge (i, j) belongs to different vertex sets, namely $i \subseteq V_1$ and $j \subseteq V_2$. Our graph can be seen as a bipartite graph which contains candidate feature sets (including seed feature set) and opinion word sets, meanwhile these two sets can be seen as two vertex sets respectively. When a candidate feature appear with some opinion words at the same time, we connect them with the edge and the weight of the edge is decided by the appearance of candidate features and opinion words.

Taking some cellphone reviews for example, the more co-occurrence of product features and opinion words, the greater relevance between them. As shown in Fig. 2, the opinion word "very big" has contacted with the feature "screen" and "memory", whereas the connection with "memory" is more than "screen" and the edge weight will be higher, the feature described by "very" is more likely to "memory". A small amount of artificial features can be seen as the seed set based on our graph model, we can obtain implicit features from corresponding candidate features using random walk model with opinion words which in the implicit sentences.

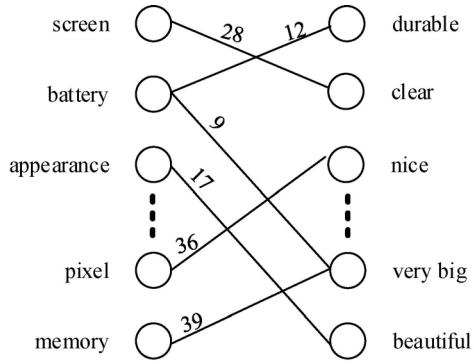

Fig. 2. The bipartite graph of some cellphone reviews

In this section, our main task is to extract implicit features. According to the definition of bipartite graph described in the section 3.2.1, a bipartite graph $G = (F \cup O, E)$ composes of candidate features and opinion words, here $F = \{f_1, f_2, \cdots, f_3\}, i = 1, 2, \cdots, m$ represents candidate features and seed features and $O = \{o_1, o_2, \cdots, o_j\}, j = 1, 2, \cdots, n$ represents opinion words. The edge of E connects the vertex F and O, w_{ij} is the edge weight of connecting the vertex f_i and o_j in the weight matrix $W = \{w_{ij}\}_{m \times n}$, implicit features are represented by b_j and the seed set of F is denoted by F_s where the feature belonging to the

extraction feature b_j is signed as a positive example and the others are signed as negative examples. According to graph G and seed feature set F_s, our algorithm calculate the probability of implicit feature b_j assigned to the candidate feature set $\{F - F_s\}$.

The size of state matrix $X(t)$ is $m \times 2$, therein the number of matrix iterative is defined by t. When $t = 0$, the initial state of the candidate feature set is denoted by $X(0)$, and when iterative stop condition $t = c$ is reached, $X(c)$ represents the final state of all candidate features via random walk algorithm. The probability of feature f_i belonging to cluster category b_j is expressed by each entry $X_{i,j}$ in the matrix X which is non-negative and can be calculated as shown in (3).

$$X(t) = \lambda HX(t-1) + (1-\lambda)X(0),\ \lambda \in (0,1). \tag{3}$$

Here, $H = D^{-\frac{1}{2}}RD^{-\frac{1}{2}}$ is a normal matrix in which diagonal matrix D is to normalize relational matrix R and each diagonal entry d_{ii} in matrix D is the sum of each element in matrix R, whereas others in matrix D are zero. The function of λ is to adjust the degree of depending on the initial state or bipartite graph when distribute candidate features. We define $j = 0,1$ when $j = 0$ is the first column of X corresponding the category b_j (positive example) and $j = 1$ is the second column of X corresponding the non-implicit features (negative examples). When reached the final state, the probability of each feature f_i is the category b_j calculated by $P(b_j \mid f_i)$ as shown in (4):

$$P(b_j \mid f_i) = \frac{X_{i,j}}{X_{i,0} + X_{i,1}}. \tag{4}$$

For the above definition, this paper uses random walk algorithm to extract implicit features is shown in Table 2.

Table 2. Random walk algorithm based on bipartite graph

Algorithm
Input: weight matrix W, category b_j, seed word set F_s, candidate set $\{F - F_s\}$
Output: $P(b_j \mid f_i)$
1. $R = WW^T$, $H = D^{-\frac{1}{2}}RD^{-\frac{1}{2}}$
2. Initialize X by $X(0)$
3. repeat
4. $X(t) = \lambda HX(t-1) + (1-\lambda)X(0),\ \lambda \in (0,1).$
5. until $X(t)$ converges to $X(c)$.

Our algorithm firstly build the relational matrix $R = WW^T$ between the candidate features according to weight matrix W to obtain diagonal matrix D and structure normal matrix $H = D^{-\frac{1}{2}}RD^{-\frac{1}{2}}$. For the state matrix initialization, there are three cases: we set $X_{i,1} = 1$ and $X_{i,0} = 0$ if f_i is the positive example of F_s; $X_{i,1} = 0$ and $X_{i,0} = 1$ if f_i is the negative example of F_s; $X_{i,1} = 0$, $X_{i,0} = 0$ if f_i belongs to $\{F - F_s\}$. Until $X(t)$ convergence to the state $X(c)$ after iterative calculation, and finally the probability of each f_i is the category b_j is calculated by $X_{i,j} \big/ (X_{i,0} + X_{i,1})$, we believe that the highest probability word is the implicit feature corresponding with the opinion word according to the probability is arranged from high to low.

4 Experiments and Results

We conduct the experiments based on the approach we proposed. The experiment results and analyses are as follows. The semi-supervised learning methods based on the results of the existing segmentation are used to extract features and opinion words in this paper to solve the congruent modified relationship between the commercial features and the opinion words. Besides, the paper combines the features of the merchandising function, capability and components which are gained from the user comments to construct a bipartite graph, then the highest probability implicit features would be computed by random walk algorithm. Thus, the corresponding relationship between product features and opinion words can be ensured. The experiment results and analyses are as follows.

4.1 Experimental Data

In this paper, the 121790 pieces of comments which are crawled from three Chinese e-commerce sites are adopted as experimental data in two areas including 79855 pieces of comments from mobile phones and 41935 pieces of comments from computers. Via observing the corpus of information, it can be concluded that most of the syntactic structure in the experimental data are short texts, then the comments are segmented by the Chinese punctuation, which leads to 368963 pieces of comment clauses. And we call the sentences which both have product features and opinion words as explicit sentences, on the contrary, the sentences which only have opinion words is called implicit sentences. After eliminating some irrelevant comment sentences which describe the facts, we deal with the remaining 311870 clauses. To alleviate the problem of candidate features low coverage in comments, using the self-heuristic, we may extract the candidate feature sets from CRFs Model, and the feature words in the candidate opinion feature is extended by using HowNet2000 conceptual dictionary. Using CRF++ toolkit to extract features and opinion words, which is a relatively matured open source toolkit of CRF. Refer to these examples provided in CFR++ for named entity and noun-phrase block to define templates, we denote the template of attribute features, in which words, part-of-speech, whether the current certain word is an emotional one, adverbs and word frequency in domain [-2, +2] will be used as characters.

When constructing the bipartite graph, a series of seeds feature sets are prepared at first and 100 features are chosen as the seeds from the “mobile phone” and “computer” these two kinds of comments. In order to decide on the final seed feature set, we collect the features which have been manually marked respectively.

In this paper, 200 pieces of comments from mobile phones and computers respectively are manually selected as the set of comments which contain 100 pieces of explicit comments and 100 pieces of implicit comments, and the explicit features and opinion words are marked by the way of manual annotation, the implicit features are added manually.

4.2 Experimental Data

Extracting explicit features and opinion words are the first step of text extraction whose results will affect the overall mining results directly, and establish the foundation for implicit features extraction as well. This paper uses the precision and recall rate as the evaluation criteria, in which precision is the ratio of retrieved correct information numbers and retrieved total information numbers and recall rate is the ratio of extracted correct information numbers and total correct information numbers. This article extracts explicit features and opinion words as well as their collocation, comparing the results with the researches in [8, 18], it’s shown in the Table 3. Hu’s approach [18] is defined as Method One.

Table 3. Explicit extraction results comparison

	Phone			Computer		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Method One	81.5%	67.5%	73.8%	63.4%	70.3%	66.7%
LDA	76.3%	65.1%	71.2%	61.2%	71.3%	65.3%
CRFs	90.3%	75.4%	82.2%	83.1%	71.2%	76.7%

Ma, Zhang and Yan [8] have presented that the LDA model tends to extract explicit features and that it

maybe unsuitable for identifying fine-grained features. Hu and Liu [18] represent that the more important commodity features are, the more often it will appear. Thus, the association rules are used to extract the high-frequency terms and noun phrases to mine commercial features, and according to setting the text window and extracting non-frequent features depending on the adjective collocations around the frequent features. This method is easy and efficient, but the effect partly lies on the selective correlation of frequent item sets support degree and both of the extraction results of F-value from the two areas of mobile phones and computers are lower than ours. Because our method can arrive at that the features and the adjectives are associated in the comments where the speech tags has been completed based on CRFs model and the higher F-value can be gotten when we deal with the sentences which are short sentences and strong regularity comment corpus.

Convergence probability has been calculated after several iterations based on random walk algorithm, in our experiment, the iterative time $t = 5$ and the value of λ is 0.63. This experiment has investigated the accuracy of "mobile phone" and "computer" these two kinds of goods when λ within the scope of different values from 0.1 to 1.0 as shown in Fig. 3 which represents the accuracy of the first 100 results from two types of commodity evaluation set. With the increase of λ, the accuracy of two types of comments changes gradually from high to low, reaching a peak at a certain point. From the Fig. 3 we can see that when λ=0.63, relatively high accuracy of extracting implicit features on these two types of goods has obtained.

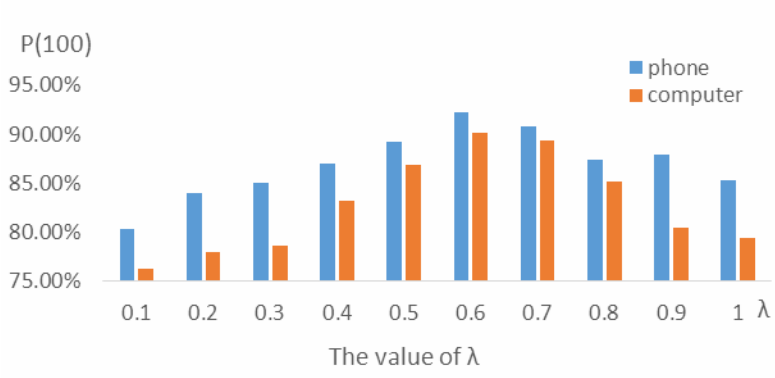

Fig. 3. The accuracy comparison of extracting implicit features using random walk algorithm based on a bipartite graph at different values of λ

The mean absolute error (MAE) is used to measure implicit feature extraction accuracy in our experiment, equals to the difference of implicit features extracted by the machine identification and human annotation, which is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|M_i - A_i\right|. \tag{5}$$

Where M and A respectively represent the implicit features extracted by machine recognition and by human annotation, the number of implicit features is denoted by n. The higher value of MAE represents the lower extraction quality, vice versa. This paper calculates the MAE of extracting implicit features using random walk algorithm based on a bipartite graph, comparing the result with PMI algorithm [14] and CoAR algorithm [15], it's shown in Fig. 4 and the values of precision, recall and F-measure in three approaches are shown in table 4. Extracting implicit features based on CoAR algorithm is described as follows: firstly, clustering explicit features from existing product reviews and selecting corresponding representative words as a representation of different product features, then implicit features are defined by the representative of product features category. PMI algorithm mainly uses the probability of opinion words and product features in the training set to analyze the semantic correlation of them. And we select the highest PMI value as the implicit feature.

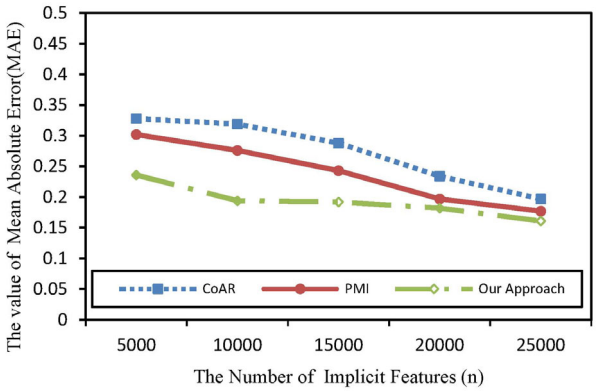

Fig. 4. The comparison of three methods on implicit features extraction

Several conclusions can be drawn from the experimental results. First, the precision, recall and F-measure of our approach are much higher than baseline approach. The precision and recall of PMI are better than CoAR, indicating that neglecting the semantic of reviews cannot significantly improve the extraction results. In the experiment, we find that product features modified by PMI algorithm and CoAR algorithm with fixed category, like through "cheap", "benefit" can get the appropriate product feature "price", but for some strong generality words, such as "nice", "general" etc., are treated unsatisfactory in effect, because these general opinion words can be used to modify almost all features. The proposed method in dealing with these opinion words has achieved good results, the mean absolute error of our method is lower than the other two methods. The more number of implicit features, the lower value of MAE. The numerical results of three methods in implicit feature extraction are shown in Table 4, it's obvious that the accuracy, recall and F values obtained by our method after processing two different areas of comments is higher than the other two methods.

Table 4. The method comparison of implicit feature extraction

Methods	Phone			Computer		
	Precision	Recall	F-measure	Precision	Recall	F-measure
CoAR	76.29%	72.71%	74.46%	70.59%	69.11%	69.84%
PMI	81.34%	79.51%	80.41%	79.16%	77.31%	78.22%
Our Approach	90.33%	85.62%	87.91%	86.49%	80.42%	83.34%

Moreover, we also find that the precision and recall blended in implicit features are higher than extraction results that only considering explicit features. Through studying Chinese reviews on the Internet, we realize that these reviews on the website are hardly standard in grammar, and even having many syntax errors which maybe effect the preprocessing performance, especially for non-standard grammatical structure or long sentences.

5 Conclusion

In this work, we present extraction models respectively for explicit and implicit features according to their characteristics. Using CRFs model to mine explicit features and "feature-opinion" pairs in the explicit sentences, then we propose a bipartite graph based on random walk algorithm to extract implicit features, thereby combining features and corresponding opinion words into binary collocation that is turning the unstructured or semi-structured text into structured text. Experimental results show that our method is reasonable and effective, the two models proposed in the mining results have achieved good results.

Opinion mining based on Chinese product reviews is a difficult subject reflecting the flexibility, complexity and uncertainty of natural language processing and can also provide useful information for sentiment analysis with great research value. Although we focus on extracting product features from Chinese online consumer reviews in this study, the proposed methodology (i.e., extraction of implicit features and collocation pairs) is generic and should be applicable to feature extraction from online

consumer reviews in other languages. On the basis of our work, an automatic Chinese comments summarization system will be generated.

Acknowledgments

This work was supported in part by National Science Foundation of China under Grants No. 61402304, the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046, the Beijing Natural Science Foundation under Grants No. 4154065, the Beijing Educational Committee Science and Technology Development Planned under Grants No. KM201610028015, Science and technology innovation platform, Teaching teacher, and Connotation Development of Colleges and Universities.

References

[1] Meng, Chinese online shopping market research report , Internet World 2013(006) 85-90.

[2] Shi, Research on opinion mining of product reviews based on syntactic tree pattern, [dissertation] Shanghai, China: Donghua University, 2013.

[3] S. Tuarob, C. Tucker, Automated discovery of lead users and latent product features by mining large scale social media networks, Journal of Mechanical Design 137(7)(2015) 071402:1-11.

[4] C.Y. Chung, M. Gertz, N. Sundaresan, Reverse engineering for Web data: from visual to semantic structures, in Proc. 18th International Conference on IEEE, 2002.

[5] X. Ding, B. Liu, P.S. Yu, A holistic lexicon-based approach to opinion mining, in: Proc. the 2008 International Conference on Web Search and Data Mining, 2008.

[6] N. Kobayashi, K. Inui, Y. Matsumoto, Collecting Evaluative Expressions for Opinion Extraction, Springer, Berlin, 2005.

[7] T. Stone, S.K. Choi, Visualization tool for interpreting user needs from user-generated content via text mining and classification, in: Proc. ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2014.

[8] B. Ma, D. Zhang, Z. Yan, An lda and synonym lexicon based approach to product feature extraction from online consumer product reviews, Journal of Electronic Commerce Research 14(4)(2013) 304-314.

[9] N. Jakob, I. Gurevych, Extracting opinion targets in a single- and cross-domain setting with conditional random fields, in: Proc. the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.

[10] S. Huang, X. Liu, X. Peng, Fine-grained product features extraction and categorization in reviews opinion mining, in: Proc. 2012 IEEE 12th International Conference on Data Mining Workshops, 2012.

[11] S. Tuarob, C.S. Tucker, A product feature inference model for mining implicit customer preferences within large scale social media networks, in: Proc. ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2015.

[12] R. González-Ibáñez, S. Muresan, N. Wacholder, Identifying sarcasm in twitter: a closer look, in: Proc. the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

[13] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in Twitter, Language Resources & Evaluation 47(1)(2013) 239-268.

[14] Q. Su, K. Xiang, H. Wang, Using pointwise mutual information to identify implicit features in customer reviews, Y.

Matsumoto, R.W. Sproat, K.-F. Wong, M. Zhang (Eds.), Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead, Springer, Berlin, 2006, pp. 22-30.

[15] Z. Hai, K. Chang, J.J. Kim, Implicit feature identification via co-occurrence association rule mining, Lecture Notes in Computer Science 6608(2011) 393-404.

[16] K. Bloom, N. Garg, S. Argamon, Extracting Appraisal Expressions, in: Proc. Human Language Technologies/North American Association of Computational Linguists, 2007.

[17] A.M. Popescu, O. Etzioni, Extracting Product Features and Opinions from Reviews, Natural Language Processing and Text Mining, Springer, London, 2007.

[18] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proc. the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2004.

[19] W.X. Zhao, J. Jiang, J. Weng, Comparing twitter and traditional media using topic models, Lecture Notes in Computer Science 6611(2011) 338-349.

[20] Y. Duan, Z. Chen, F. Wei, M. Zhou, H.Y. Shum, Twitter topic summarization by ranking tweets using social influence and content quality, in: Proc. the 24th International Conference on Computational Linguistics, 2012.

[21] Y. Wang, H. Wu, H. Fang, An exploration of tie-breaking for microblog retrieval, in: M. de Rijke, T. Kenter, A. De Vries, C. Zhai, F. de Jong, K. Radinsky, K. Hofmann (Eds.), Advances in Information Retrieval, Springer, Switzerland, 2014, pp. 713-719.

[22] HowNet, HowNet's Home Page. <http://www.keenage.com>.

[23] J.D. Lafferty, A. Mccallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data , in: Proc. 18th International Conf. on Machine Learning, 2001

[24] S.C. Ding, T. Jiang, N. Wen, Research on sentiment orientation of product reviews in Chinese based on cascaded CRFs models, in: Proc. Machine Learning and Cybernetics (ICMLC), 2012 International Conference on IEEE, 2012.

[25] J. Wiebe, T. Wilson, R. Bruce, Learning Subjective Language, Computational Linguistics 30(3)(2006) 277-308.

[26] Z. Zhai, B. Liu, H. Xu, Clustering Product Features for Opinion Mining, in: Proc. Fourth Acm International Conference on Web Search & Data Mining, 2011.

[27] K. Nigam, A.K. Mccallum, S. Thrun, Text classification from labeled and unlabeled documents using EM, Machine Learning 39(2-3)(2000) 103-134.