Opinion Mining Based on Feature-Level

Lizhen Liu, Zhixin Lv, Hanshi Wang
College of Information Engineering,
Capital Normal University, Beijing, China
Email:a839606145@126.com

Abstract—**An important task of opinion mining is to extract people's opinions on features of an entity. However, for the same feature, people can express it with many different words or phrases. To produce a useful summary, these words and phrases, which are domain synonyms, need to be grouped into the same feature group. Moreover, the sentiment relatedness between the features and opinions is usually complicated. For many cases, product feature words are implied by the opinion words in reviews. A novel method is proposed to deal with the feature-level opinion mining problems. More Specially, 1) the proposed method considers the explicit features and the implicit features. 2) the opinion words are divided into two categories, vague opinion words and clear opinion words, to identify the implicit features and cluster the features. The feature clustering depends on three aspects: the corresponding opinion words, the similarity of the features and the structures of the features. Moreover, the context information is used to enhance the clustering in the procedure, which is proved to be useful in clustering. The experimental results demonstrate that our method performs well.**

Keywords- feature-level; implicit features; opinion mining

I. INTRODUCTION

With the rapid development of the Internet, a large amount of subjective reviews is available in online forums, blogs, and shopping websites. Some researches [1, 2, 3] primarily focus on mining objective information from the text segments. While some researches [4, 5] primarily deal with classifying sentiment orientations expressed in text. Document-level sentiment analysis, or opinion mining can classify the overall subjectivity or sentiment orientation expressed in the review content, but it fails to get the sentiment associated with individual features. In recent years, many researchers focus on finer-grained opinion mining, which detects opinions on different review features as opposed to the document level .The researches on feature-level opinion mining rely on identifying the features and the corresponding opinions. However, the Chinese reviews on the Internet lack of standardization. People express their opinions using omission and free structure, which lead to the relationship between opinions and features becomes more complex, which is hard to use the syntax analysis to deal with. Particularly, the features are implicit in many cases. For example, the sentence "好贵啊，买不起" (*Too expensive to afford*), "贵"(*expensive*) imply the feature "价格"(*price*). Moreover, different words may be used to express the same product feature. For example, the words "外观"(*facade*) and "外表"(*appearance*) express the same features. The method proposed in the paper identifies the implicit features for the opinions and summarizes the features with the same meaning into one cluster. The summary is useful for people to scan the products quickly.

For feature-level opinion mining, the most important task is to identify the features and the corresponding opinions. Liu and Hu (2004), Popescu and Etzioni (2005), Kobayashi et al. (2007), Wong et al. (2008), Qiu et al. (2009), Liu and Zhang (2010) and Zhen et al. (2011) studied the problem. The proposed method uses the opinions to extract the corresponding features, and remove the noise by the support and conference of the opinions and the corresponding features. However, this problem is far from being solved. The detail information will be introduced in part three.

Recently, few studies focus on recognizing the implicit features. For example, the work in [7] completely ignored the problem of recognizing implicit features. Hu and Liu [8] partially addressed the implicit feature identification problem by applying the same method used for explicit feature extraction. This is unreasonable to ignore the implicit features in product review contents, because people often express their opinions by simple structures and brachylogies. Su et al. [9] tried to infer the implicit features for such single-word opinions, e.g. "重" (*heavy*), by using Point-wise Mutual Information (PMI) based on semantic association analysis. Zhen Hai et al. [10] used CoAR algorithm to identify the implicit features. This algorithm clusters explicit features at first, then use the clustering results to choose the feature with high support as the implicit feature. However, the proposed method uses the part-of-speech dictionary and the corresponding opinions to identify the implicit features. Experimental results in part four show that our method performs well. The detail information will be introduced in part four.

II. RELATED WORK

Hu and Liu (2004) [10] proposed a technique based on association rule mining to extract product features. Their main idea is that people often use the same words when they comment on the same product features. So the frequent item sets of nouns in reviews are likely to be product features while the infrequent ones are less likely to be product features. But, in fact, the infrequent item may also be features, which may offer more information. Their work only find the features that many people focus on, which not what we want to do. The infrequent features also are very import for people to make choice. The method proposed in this paper uses opinion words to extract the corresponding features. The relationship between opinions and features is used to remove the noise to improve the precision.

Popescu and Etzioni (2005) [11] investigated the same problem. Their algorithm requires that the product class is

978-1-4673-0964-6/12/$31.00 ©2012 IEEE

known. The algorithm only reckon noun/noun phrase as the candidate features. It determines whether a noun/noun phrase is a feature by computing the Point-wise Mutual Information (PMI) score between the phrase and class discriminators, e.g., "of xx", "xx has", "xx comes with", etc., where xx is a product class. But it calculates the PMI by searching the Web. Querying the Web is time-consuming.

Qi et al. (2008) [12] proposed a novel mutual reinforcement approach to deal with the feature-level opinion mining problem. This approach clusters product features and opinion words simultaneously and iteratively by fusing both their content information and sentiment link information. This algorithm uses the relationship between the opinions and the features. But, the relationship between the opinions and features is so complex that the errors will increase with the number of the iteration in a certain range. This paper also uses the relationship between opinions and features. But, instead of iteration, the method only uses it to remove the noise by check the mutual conference. Empirical evaluation show the good perform.

Qiu et al. (2009) [13] proposed a novel algorithm called Double Propagation. It is a state-of-the-art unsupervised technique for solving the problem. Their primary idea is that opinion words are usually associated with features in some ways. Thus, opinion words can be recognized by identified features, and features can be identified by known opinion words. So the extracted opinion words and features are utilized to identify new opinion words and new features, which are used again to extract more opinion words and features. This propagation or bootstrapping process ends when no more opinion words or features can be found. The biggest advantage of the method is that it requires no additional resources except an initial seed opinion lexicon, which is readily available. It mainly extracts noun features, and works well for medium-size corpora. But for large corpora, this method can introduce a great deal of noise (low precision), and for small corpora, it can miss important features.

Zhang and Liu (2010) [14] improved the Double Propagation. This approach used two patterns which based on part-whole patterns and "no" patterns to increase the recall and precision. As for the low precision problem, a feature ranking approach is present to tackle it. Ranking feature candidates based on the importance consists of two factors: feature relevance and feature frequency. This algorithm models the problem as a bipartite graph and uses the well-known web page ranking algorithm HITS to find important features and rank them high. But, the patterns in Chinese corpus are very litter, which maybe because many people use concise statement even the wrong grammar to write the reviews which do not have more pattern. So the role of the model will be restricted in this corpus.

The proposed approach takes adjectives as the opinion words and uses the opinion words to extract the corresponding features. The main idea is that an adjective must be used to modify something. So an adjective will correspond to a feature, weather it is the whole entity or a feature of the entity. If there is no, it must have one implicit feature for it. Hai et al. (2011) [10] used two-phase co-occurrence association rule mining approach to identify implicit features. In the first phase of rule generation, for each opinion word occurring in an explicit sentence in the corpus, they mine a significant set of association rules of the form [opinion-word, explicit-feature] from a co-occurrence matrix. In the second phase of rule application, they first cluster the rule consequents (explicit features) to generate more robust rules for each opinion word mentioned above. Given a new opinion word with no explicit feature, they then search a matched list of robust rules, among which the rule having the feature cluster with the highest frequency weight is fired, and they assign the representative word of the cluster as the final identified implicit feature. But they do not consider the opinion words which can modify the all features, e.g. "很好" (*very good*), "不错" (*not bad)*, "还可以" (*fairish*), which are common in Chinese reviews. This kind of opinions can not be used for distinguishing the features, however, that may lead to low precision and recall. So the proposed method divides the opinions into two categories and deal with separately to solve this problem.

III. THE PROPOSED METHOD

Feature-level opinion mining including three steps:

(1) Extract the features and the corresponding opinions

(2) Cluster the features

(3) Orient the opinions of the features

This paper focuses on the former two tasks. Step three can use a sentiment dictionary to orient the opinions, which is not our emphasis. The formers are the foundation, which is important for feature-level opinion mining.

A. Extract The Opinions and Features

The proposed method uses the opinions which are adjectives to find the corresponding features. The mainly idea is that every adjective is used to modify a feature, no matter what it is the whole entity or a feature of the entity. The method not only considers the noun /noun phrase but also the verb/verb phrase as the features. Studying the characteristics of the Chinese comments, we priority consider the left relation, (the feature is on the left of the opinions) because it is common in Chinese reviews. For example, in Chinese reviews, people used to using the pattern "价格有点贵"(*the price is a litter expensive*) rather than the pattern "很高的价格"(*high price*). Syntax analysis can also solve this problem, why we do not use it is because of the complexity of the algorithm. And the normative sentence which leads to the errors of grammatical structures, is hard to use syntax analysis to find the opinions and the corresponding features.

Obviously, using the opinions as the feature indicator is ambiguous. This means that it is not a hard rule. We will inevitably get the wrong features. So remove the noise is an important task. The relationship between the opinions and features is used to solve this problem. The main idea is that the words with low frequency maybe the noise. Remove the noise not by filter the low frequency of the groups which consist of the features and the corresponding opinions but mutual filter the noise. By selecting high confidence of opinions, our

approach filters the opinions with low confidence and the corresponding feature also has low confidence. First the proposed approach selects the opinions with low confidence, then check the corresponding feature weather it is with low confidence, if so, delete it and recalculate the Co-occurrence matrix. Executing this procedure again based on reversing the roles of opinions and features.

The confidence formula is represented as follows:

$$Con(x_i) = P(x_i) / N \quad \{x_i \in F\} \bigcup \{x_j \in O\} . \tag{1}$$

Here, *F* present the features, *O* present the opinions.

B. Identify The Implicit Features

According to the human way of thinking, when we all know something, we will use omission in the dialogue process. This phenomenon also appears in the comments. In the corpus, some adjectives can not find the corresponding features, there must have implicit feature for the adjectives. There are two kinds of implicit feature, one is the entity, e.g. the sentence "不错，可以选择购买" (*not bad, can choose it*). The word "不错" (*not bad)* can not find the explicit feature to be modified, but we know it is modify the entity what we are concerned about. This kind of opinions can not specified the feature without the context, so we call this kind of opinions "vague opinion"; the other is the features, for example, the sentence "便宜，买的值了"(*cheap, it is worth*), the word "便宜"(*cheap*) hinder the feature"价格"(*price*), This kind of opinions hinder the special features which even when we do not know the context, we can also know. This is called "clear opinion". This paper deals with the later. The former used the entity to replace.

People use different description to express the same features. For example, "价格"(*price*), "价值"(*value*) and so on, so witch one will be selected? It is do not matter, because everyone express the correct feature, the difference will disappear in the cluster process. This is why we do not use two-phase co-occurrence association rule mining approach which is proposed in [10]. Besides, the opinions of people about the same features are different event are opposite. For example, "有点贵，勉强可以接受"(*a litter expensive, force to accept*),"便宜，没有这么好的东西了"(*cheat, there is no better than it*). The former sentence express the negative opinion on the price, while the later is positive. The part-of-speech dictionary can be used to group the opinions. The part-of-speech dictionary not only includes the synonyms but also the antonym. The proposed method group the opinion firstly based on the part-of-speech dictionary. The features, which correspond to the same opinion group with the implicit features, are the candidate set for the implicit features. We select the feature with the highest importance as the implicit feature. The importance formula is expressed as follow:

$$imp(x_i) = weight(x_i)(sup(x_i) + con(x_i)) , \tag{2}$$

$$sup(x_i) = p(x_i) / N , \tag{3}$$

$$weight(x_i) = \sum_{f_i \in F(x_i)} con(f_i) , \tag{4}$$

$$\mathrm{Con}(x_i) . \tag{1}$$

C. Cluster The Features

People use different words express the same features. We want to cluster the same features into groups to form a summary. We use K-means to deal with it. The proposed method considers three aspects of the features:

1) the similarity of the corresponding opinions

As we have mentioned above, the "clear opinion" can identify the features. So the similarity of the opinions can be used to guide the features. The similarity of opinions considers the kinds and the associated features. The similarity of opinions is given in Equation (5).

$$simo(x_i, x_j) = index(x_i, x_j) \cdot dis(x_i, x_j) \tag{5}$$

$$index(x_i, x_j) = \begin{cases} 1 & \{x_i, x_j \in co\} \\ 0.5 & \{x_i \in co, x_j \in vo \text{ or } x_i \in vo, x_j \in co\} \\ 0 & \{x_i, x_j \in vo\} \end{cases} \tag{6}$$

Here, "*co*" represent the "clear opinion", the "*vo*" represent the "vague opinion". The $dis(x_i, x_j)$ represent the Cosine distance based on the Co-occurrence matrix which consists of features and the opinion groups.

2) The similarity of features in text

This aspect considers the features which have the same words, e.g. "运行速度" (*running speed*) and "速度" (*speed*) represent the same feature "速度" (*speed*). The similarity uses the Set Theory.

$$simt = 2 \cdot p(x_i \bigcap x_j) / p(x_i \bigcup x_j) \tag{7}$$

Here, $p(x_i \bigcap x_j)$ represent the number of words what x_i and x_j common contain. $p(x_i \bigcup x_j)$ represent the total number of words what x_i and x_j contain.

3) The structure of the features in comment

This aspect considers two indexes, one is the kinds of the features, as we introduce former, the noun/noun phrase and verb/verb phrase may be the features. The number of kinds is five: N (noun), NV(noun + verb), V(verb),VN(verb + noun), NN(noun + noun). The other is the location of features and the

corresponding opinions. The similarity uses the Cosine distance to express.

Therefore, the similarity of features is represented as follows, which based on the former:

$$sim(x_i, x_j) = \alpha simo(x_i, x_j) + \beta simt(x_i, x_j) + \gamma sims(x_i, x_j) \quad (8)$$

Here, $\alpha + \beta + \gamma = 1$ .In our experiments, they in turn are 0.7, 0.2, 0.1. These are the results of the experiment many times.

D. Clustering enhancement

The algorithm utilizes the constructed instance representation to conduct the process of clustering. Our basic idea of clustering enhancement by background knowledge comes from COP-KMeans [15]. COP-KMeans is a semi-supervised variant of K-Means. Background knowledge, provided in the form of constraints between data objects, is used to generate the partition in the clustering process. One type of constraints are used in COP-KMeans, it is the incompatibility.

Incompatibility: two data objects must not to be in the same cluster.

The context-dependent information is also useful to construct constraints. In general, review is a collection of related sentences with a single different focus. In one review, there are not two same features appear. People do their best to express their opinions use simple sentence. Our observation of product review corpus largely meets the point. For example, for an editor review on computers, reviewers may usually present their opinions on the power of the computer in a sentence, followed by their opinions on the other feature in another sentence and they do not repeat the features what they have described. That's a common case in reviews on all kinds of products. So our approach uses the incompatibility that same feature cans not be in one cluster to enhance the clustering.

IV. EXPERIMENT

This section evaluates the proposed method. We first describe the data sets, evaluation metrics and then the experimental results and analysis.

A. Data Sets

We used four diverse data sets to evaluate our techniques. They were obtained from a commercial web (360buy.com) that provides opinion mining services. Table I shows the domains (based on their names), the number of reviews in each data set ("Revi" means the review) and the number of sentence in each data set ("Sent" means the sentence which is identify by the punctuations).

TABLE I. THE DATA SETS

Date sets	Computer1	Computer 2	Phone	Camera
Revi	500	1000	1000	1000
Sent	1459	2798	3067	2674

B. Evaluation Metrics

(1) For the extracting and identifying the implicit features, we use precision and recall as the evaluation metrics.

(2) For the clustering, the proposed approach uses the VI (Variation of Information). It is an information- theoretic measure that regards the system output C and the gold standard tags T as two separate clusters, and evaluates the amount of information lost in going from C to T and the amount of information gained, i.e., the sum of the conditional entropy of each clustering conditioned on the other. More formally,

$$\begin{aligned} VI(C,T) &= H(T \mid C) + H(C \mid T) \\ &= H(C) + H(T) - 2I(C,T) \end{aligned} \quad (9)$$

Here, $H(\cdot)$ is the entropy function and $I(\cdot)$ is the mutual information.

VI and other entropy-based measures have been argued to be superior to accuracy-based measures, because they not only consider the majority tag in each cluster, but also whether the remainder of the cluster is more or less homogeneous.

C. Experimental Results and analysis

We first compare our results of extraction the features with double propagation on recall and precision for different corpus sizes and different domain. The results are presented in Table II. We did not try more sentences because manually checking is a very large job. "*Ours*" represent our method, and "*DP*" means Double Propagation.

From the Table II, we can see that for corpora in all domains, our method outperform or equivalent to double propagation on recall. On date sets for "computer2", the precision even better. This maybe because of the removing noise based on mutual information, which gets the correct features. However, for the precision of the small scale date sets, the recall and the precisions both low.

For the identifying the implicit features, we calculate the precision and recall of our results based on Manual annotation results. Table III shows the results.

For the cluster the features, figure 1 shows the value of VI on different K for the four corpuses.

TABLE II. THE RESULTS OF THE EXTRACTION

Date sets		Computer1	Computer2	Phone	Camera
Precision	***Ours***	0.57	**0.69**	**0.71**	0.64
	DP	**0.59**	0.65	**0.71**	0.65
Recall	***Ours***	0.52	**0.60**	**0.62**	**0.58**
	DP	**0.55**	0.58	**0.62**	0.57

From the Fig.1, The best numbers of the cluster for the four corpuses respectively are 23, 33, 25, 35, which is close to the

gold standard, which respectively are 29, 35, 30, 30. We also compare the K-Means algorithm with the enhancement based on the knowledge. Table 4 shows the precision and recall for the two algorithms on these corpuses. "K-Means" represent the K-Means algorithm and "Enhance" represent the K-Means based on the knowledge what is used in the proposed method. From the Table IV, the K-Means based on the knowledge outperform well. It shows the context-dependent information is a good indicator for the cluster of features.

TABLE III. THE IDENTIFYING OF THE IMPLICIT FEATURES

Date sets	Computer1	Computer2	Phone	Camera
Precision	0.65	0.72	0.79	0.74
Recall	0.56	0.67	0.70	0.65

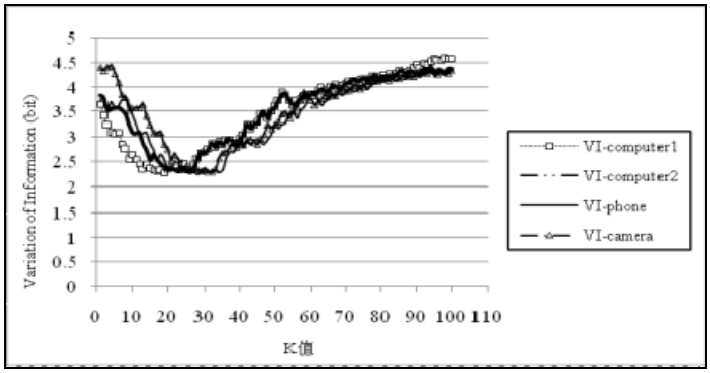

Figure 1. The Result of Cluster

TABLE IV. THE RESULTS OF COMPARE

Date sets		Computer1	Computer2	Phone	Camera
K		**23**	**33**	**25**	**35**
Precision	K-Means	0.54	0.64	0.63	0.72
	Enhance	**0.65**	**0.70**	**0.67**	**0.79**
Recall	K-Means	0.43	0.52	0.58	0.55
	Enhance	**0.53**	**0.62**	**0.65**	**0.64**

V. CONCLUSION

Feature extraction for entities is an important task for opinion mining. This paper proposed a new method to deal with this problem. The new method uses the corresponding opinion words extracting the features, and according to mutual support and confidence to filter the noise. It also identifies the implicit features and clusters the features based on the knowledge of the background which strengthen cluster results. Empirical evaluation show the proposed method outperforms. However, this method has some shortcomings. Small scale corpus cans not perform well. And the structure of the vague opinions dictionary and part-of-speech dictionary increases the cost of the method. Next, we will study the establishment of two dictionaries by automaticly and improve the precision and recall for the small scale corpus.

ACKNOWLEDGMENT

This work was supported by the Beijing Key Disciplines of Computer Application Technology, China.

REFERENCES

[1] Blair-Goldensohn, Sasha., Kerry, Hannan., Ryan, McDonald., Tyler, Neylon., George A. Reis, Jeff, Reyna., "Building Sentiment Summarizer for Local Service Reviews," In Proceedings of the Workshop of NLPIX , WWW, 2008 .

[2] Ding, Xiaowen., Bing Liu, Philip S. Yu., "A Holistic Lexicon-Based Approach to Opinion Mining," In Proceedings of WSDM, 2008.

[3] Chieu, Hai Leong and Hwee-Tou Ng, "Name Entity Recognition: a Maximum Entropy Approach Using Global Information," In Proceedings of the 6th Workshop on Very Large Corpora, 2002.

[4] Hu, Mingqin and Bing Liu., "Mining and Summarizing Customer Reviews.," In Proceedings of KDD, 2004.

[5] Kobayashi, Nozomi., Kentaro Inui and Yuji Matsumoto, "Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining," In Proceedings of EMNLP, 2007.

[6] Bing Liu, "Sentiment Analysis and Subjectivity," in Handbook of Natural Language Processing, Second Edition, pp.627-665, CRC, 2010.

[7] Hu, M., Liu, B., "Mining and summarizing customer reviews," In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, pp. 342–351, 2004.

[8] Hu, M., Liu, B., "Opinion feature extraction using class sequential rules," AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, Palo Alto, USA , 2006.

[9] Su, Q., Xiang, K., Wang, H., Sun, B., Yu, S., "Using pointwise mutual information to identify implicit features in customer reviews," In Matsumoto, Y., Sproat, R.W., Wong, K.-F., Zhang, M. (eds.) ICCPOL 2006 [LNCS (LNAI), vol. 4285, pp. 22–30.Springer, Heidelberg , 2006]

[10] Zhen Hai, Kuiyu Chang, and Jung-jae Kim A. Gelbukh (Ed.), "Implicit Feature Identification via Co-occurrence Association Rule Mining," CICLing , Part I, LNCS 6608, pp. 39 3–404, 2011.

[11] Popescu, Ana-Maria and Oren, Etzioni. 2005, "Extracting product features and opinions from reviews," In Proceedings of EMNLP, 2005.

[12] Su, Qi., Xinying Xu., Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang et al., "Hidden Sentiment Association in Chinese Web Opinion Mining," In Proceedings of WWW , 2008.

[13] Qiu, Guang., Bing, Liu., Jiajun Bu and Chun Chen, "Expanding Domain Sentiment Lexicon through Double Propagation," In Proceedings of IJCAI, 2009.

[14] Lei Zhang, Bing Liu, Suk Hwan, Lim Eamonn, O'Brien-Strain , "Extracting and Ranking Product Features in Opinion Documents," Coling, Poster Volume, pages 1462–1470, Beijing, August , 2010.

[15] K. Wagsta, C. Cardie, S. Rogers, and S. Schroedl., "Constrained k-means clustering with background knowledge," In Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577–584, 2001.