International Journal of Database and Theory and Application
Vol.9, No.2 (2016), pp. 129-136
http://dx.doi.org/10.14257/ijdta.2016.9.2.14

Mining Opinion Word from Customer Review

Jiang Tengjiao[1, 2], Zhong Minjuan[1, 2], Liao Shumei[1] and Luo Siwen[1, 2]

[1]School of Information Technology, Jiangxi University of Finance and Economics
[2]Jiangxi Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics
Nanchang, China
lucyzmj@sina.com

Abstract

Online customer review is considered as a significant informative resource which is useful for both potential customer and product manufacturers. As a result, it is one of the most challenging tasks to mine customer reviews automatically and to provide users with opinion summary. Product features and opinion word play the most important roles in the customers' opinions mining. In this paper, we dedicate our work to opinion word mining. We proposed an approach for opinion word identification based on the association rule mining algorithm. The method makes full use of co-occurrence syntactic characteristic between product features and opinion word. Firstly, the product feature is identified by two-stage filtering scheme, and secondly the opinion word is extracted through association rule mining. The final experiment results show that the proposed method could not only obtain the product features related to domain characteristics, but identify the opinion word effectively. Meanwhile, our approach possesses much higher precision and recall than Hu's work.

Keywords: *Product Feature; Opinion Word; Frequent Filtering Scheme; Semantic Filtering Scheme; Association Rule Mining*

1. Introduction

Due to proliferation of Web 2.0, a number of online shopping customers have dramatically increased and the increase of online merchants. To enhance the customer satisfaction, merchants and product manufactures allow customers to review or express their opinions on the products or services. These online customer review, therefore, become a significant informative resource which is useful for both potential customer and product manufacturers. However, the number of reviews can be in hundreds or even thousands for a popular product. This makes it difficult for potential customer to read them to make an informed decision on whether to purchase the product. How to extract the people's views and opinions automatically from a mass of data is the faced problem of text tendency analysis [1]. As the fundamental work, product features and opinion word play the most important roles on the opinions mining of customers. The task of opinion word recognition has been of great interested since the last decade.

Really, opinion word mining includes following three task [2]:

-to determine subjectivity of words in a document (*i.e.*, whether the word is subjective or objective)

–to determine orientation or polarity of words (*i.e.*, whether the word is positively subjective or negatively subjective)

–to determine strength of orientation (*i.e.*, how much positive or negative a word is).

In this paper, we dedicate our work to identify opinion word in the customer review. We proposed an approach for opinion word identification based on the association rule mining algorithm. We makes full use of co-occurrence characteristic between the product feature and opinion word. Different from the literature of [3] work, our method first presented two-stage pre-processing technique to the product features selection. The advantage of it is to avoid noise to be brought into and meanwhile to make the extracted features possess both high precision and domain correlation. After that, association rule mining algorithm is performed to opinion word identification. The experimental results showed that the proposed method could identify opinion word effectively and produce better performance.

This paper is sketched as follows: Section 2 illustrates the previous work; Section 3 describes the approach in detail; Experiment in Section 4 indicates the effectiveness of our approach. Section 5 concludes and presents our further work.

2. Related Work

Opinion word also called sentiment orientations, semantic orientations, or polarities and their recognition usually related to the product feature extraction. Actually, most product features are modified by the surrounding opinion words in customer reviews, thus they are highly context dependent on each other, which is referred to as context-dependency property henceforth. There is much work on feature extraction and opinion word identification. The existing work could be broadly classified into two categories, those based on machine learning and the ones syntactic rules.

(1)Machine learning based. Wu [4] focused on extracting relations between product features and opinion expressions by phrase dependency parsing, and classified opinion word and product features by using Tree-kernel SVM. Jim and ho [5] proposed a novel machine learning approach under the framework of lexicalized HMMs. Their approach naturally integrates multiple important linguistic features into automatic learning and is effective in determining the feature terms and opinion words. Lakkaraju [6] proposed a series of probabilistic models that jointly discover latent facets and sentiment topics, which is achieved by simultaneously capturing both short-range syntactic structure and long range semantic dependencies between the sentiment and facet words. Wangzhongqing [7] extract feature and opinion word based on combination conditional random fields model and rule. They presented one system named Suda-SAM-OMS, which is designed for joining COAE2011 and achieve a very high performance on the task of emotional evaluation unit extraction.

2) Syntactic rule based. Hu and Liu [3] firstly proposed a technique of product feature extraction to assist opinion word recognition task in production review. The key to the approach is to find product feature which is frequent itemsets of nouns based on association rule mining, and the surrounding adjectives of any extracted feature are considered as opinion words. However, it is obvious that not all frequent nouns belong to the product features and hence much more non-feature nouns are generated, leading to precision decreasing of opinion word recognition. Popescu and Etzioni [8] has utilized statistic-based point-wise mutual information (PMI) to extract product features. Based on the association of opinion words with product features, they take the advantage of the syntactic dependencies computed by the MINIPAR parser [9] to identify opinion words. Qiu *et al*. [10] proposed a double propagation method, which exploits certain syntactic relations of opinion words and features, and propagates through both opinion words and features iteratively. The extraction rules are designed based on different relations between opinion words and features, and among opinion words and features themselves. Kim *et al* [11] presents a method for identifying an opinion with its holder and topic, given a sentence from online news media texts. This method uses semantic role labeling as an intermediate step to label an opinion holder and topic using data from FrameNet. Bloom *et al* [12] describes a system for extraction target word and opinion word, based on hand-built lexicon, a combination of heuristic shallow parsing and dependency parsing. Kamal *et al*. [13] propose a text mining approach to mine product features, opinions and their

reliability scores from web opinion sources. A rule-based system is implemented, which applies linguistic and semantic analysis of texts to mine feature-opinion pairs that have sentence-level co-occurrence in review documents.

The method based on Machine learning takes product features and opinion word as sequence labeling task and hence training data is needed. Meanwhile, the components of the sentence are mutual dominated and dependent, and should satisfy their grammatical requirements. Therefore, we collected music production reviews from amazon.com as dataset, and proposed opinion word identification approach based on association rule mining.

3. Opinion Words Identification Based on Association Rule Mining

Extraction of opinion word and target word in domain is a basic and important work in text sentiment orientation analysis. In product review, target word is usually talked about as the product features. Customers express their opinion on them in review sentence. So, with the co-occurrence property, identifying opinion words could be implemented.

3.1. Frequent Product Features Extraction

Our purpose aims to find what people like and dislike about a given product, therefore, how to find the product features that people talk about is the crucial step. From the view of literature [14], product feature appears in three forms, that is, overall product; components of product; properties and its extension of product, which is corresponding to the name and attributes of product respectively. Moreover, we have observed that product features are usually nouns or noun phrases in review sentence. On basic of it, some pre-processing works, such as word segmentation, part-of-speech tagging, and dependency parsing analysis, are performed to the dataset, which is from the website of amazom.com. Subsequently, the candidate product feature set (named *T*) is formed by nouns and noun phrases extraction. Followed by it, two-stage filtering scheme is employed to the *T* in order to improve the precision of product features identification as well as domain correlation. The detailed filter scheme is as follows:

(1) Term frequency filtering

Term frequency filtering refers to filter those nouns or noun phrases with less frequency in corpus. We think that when people comment on different features of a product, the expressed vocabulary usually converges. Thus, those nouns that are frequently talked about are usually genuine and important features. On the contrary, irrelevant contents in reviews are often diverse. Hence, those infrequent nouns are likely to be non-features or less important features. On the other hand, although some low frequency words may be filtered, resulting in the decreasing recall, our ultimate purpose aims to opinion word identification. Only based on the correct product features recognition, subsequent performance of opinion word identification can be guaranteed. Furthermore, compared to recall, people prefer to precision. In a word, we think those features with low frequency is secondary property and could be ignored.

(2)PMI Semantic filtering

PMI (pointwise mutual information)score could quantify the relationships between words in the text corpus and the value between *word₁* and *word₂* is defined as follows:

$$PMI(word_1, word_2) = \log_2 \left[\frac{P(word_1 \,\&\, word_2)}{P(word_1)P(word_2)} \right] \tag{1}$$

Where $p(word_1 \& word_2)$ is the co-occurrence probability with $word_1$ and $word_2$ pairs. To ensure the domain correlation of target word, combing terminology the saurus(called *F*)by artificial construction on music field, semantic similarity value, PMI(w_F^i , w_T^j), is computed, in which w_F^i is the *ith* feature words in *F*, w_T^j refers to the *jth* candidate feature in *T*. The idea of this approach is clear. If the PMI score of a candidate feature is too low, it may not be a component of the product because w_F^i and w_T^j do not co-occur frequently. Finally, the final feature set *T* is achieved by setting a threshold. Subsequent experiments show that the filter scheme is effective.

3.2. Opinion Word Extraction

Opinion word primarily used to express subjective opinions. Based on above product features extraction, opinion word in the music field is identified by fully exploiting the semantic relationships between product features and opinion word. Here, we use associate rule mining to find opinion word. The detailed steps are as follows:

(1) Construction transaction set of web product review. The sentence is regard as the transaction unit and the contained adjective or verb that may be expressed emotional characteristic are extracted as item. Meanwhile, on basic of the above product feature set, the occurred product features are also extracted to constitute transaction set.

(2) Association mining the transaction set of web product review. The strong association frequent group is obtained and the semantic association relationship between product features and opinion words are also got.

(3) Formation of Association rules. Rules are generated to the all strong association frequent group based on minimum confidence. The rules that only contain one front piece and one consequent are extracted by filtering all rules. We get like X->Y rules according to part-of-speech, where X is certain product feature and Y refers to opinion word. All opinion words constitute opinion dictionary D_1.

(4) Expansion of D_1. To obtain a broader opinion words, by taking D_1 as a seed list of opinion words, we employ synonym expansion based on WordNet, and the final opinion lexicon *D* is formed.

4. Experiment and Analysis

4.1 Preparing for the Experiment

We perform the following steps to preparing for the experiment.

Step1: Collecting data.

In this work, we use reviews provided by Hu and Liu, who collected production review from amazom.com in June 2006 and extracted 5.8 million reviews, 2.14 reviewers and 6.7 million products [15]. Each amazon.com's review consists of 8 parts, including following information:<Production ID>, <Reviewer ID>, <Rating>, <Date>, <Review Title>, <Review Body>,<Number of Helpful Feedbacks>,<Number of Feedback>.The whole collection include 4 main categories of products, *i.e.*, *Books*, *Music*, *DVD* and *mProducts* (industry manufactured products like electronics, computers, etc).

Step2: Pre-processing data

To facilitate the implement of following task, we make the following pre-processing of the dataset. Firstly, we make document segmentation based on Production ID because the large amount of review has been collected in a single large text file, and then get all the information of each product review. Secondly, classification is performed. For a given review, we don't know it evaluate what kind of products. So, we should obtain the production categories corresponding to the Production ID. We take the following method. For example, we submit the following address to the web, http://www.amazon.com/dp/0000000868, 0000000868 is the Production ID, and the category is immediately shown in the webpage. Finally, we select part of music reviews as dataset in our study and there are 7705 products and 78521 reviews in total.

Step3: Marking data

To better evaluate our proposed approach, a human tagger manually read all the reviews and produced a manual opinion word list for each product. We select the volunteers to manually mark the above music dataset. The marking process follows the principle of the minority subordinate to the majority. For each opinion word, if the judgment of two volunteers is consistent with each other, the judgment results are used to final results. For example, two volunteers both regard it as opinion word, the word is marked as opinion word. On the contrary, we should add a new volunteer and select the side that would show more agreement as the final results.

4.2 Experiment Results

Opinion word identification is the first step for opinion lexicon construction. In our work, association rule mining algorithm is used to mine the relationship between the product feature and opinion word and on basic of it, opinion word is identified.

Because the correct identification of opinion word relies on the product features, the performance on product features extraction is tested firstly and two common measures, precision and recall are adopted. The purpose of this experiment is to examine the field correlation of extracted product features, that is to say, whether the identified product features is relevant to the music field. The results are shown on Table 1.

Table I. Product Features Identification Results in Music Field

unique words in collection	7705
candidate product features	3223
product features after term frequency filtering	1975
product features after PMI filtering	449
precision	0.90
recall	0.79

It can be seen from the Table 1 and Table 2 that the identified product features have not only higher precision, but field correlation. We also observed that the recall is slightly lower. We think two following aspects may lead to the results. On the one hand, term frequency filtering may lead some features with low frequency not to be extracted, for example, bassist, although the word is relevant to the music field, it is less appearance in the comments due to non-popularity. On the other hand, the performance of PMI filtering relies heavily on large data scale. Theoretically speaking, the more the number of comments, the more obvious of the statistical effect is, and more accurate the PMI score is. In our work, the data collection is not too large, resulting in the deviation of PMI value. Thus, the number of features is affected and recall is low. Table II gives part of extracted features results.

Table 2. Part of Extracted Features Results

Feature	Feature	Feature	Feature	Feature
musician	songwriter	tongue	rock	channels
chord	classic	Tapes	musicals	player
composition	synthesizer	artistry	tunes	albums
cd	keyboards	percussion	tone	melodies
masterpieces	artist	Glory	vocalists	choirs
mode	harmonica	market	popularity	recordings
opera	voice	Drums	concerts	disc
composers	listener	concert	chords	instruments
piano	solo	voices	orchestra	bands
orchestra	vocals	Violin	styles	rhythm

singers	audience	Jazz	performer	bass
disk	techno	episode	hazy	winners
pianist	pop	Ballad	producer	

The opinion word extraction is performed by using the mined relationships between opinion word and target and the performance results are shown on Table III.

Table 3. Performance Results on Opinion Word Identification

Performance	Data results
the total number of tagged opinion words by manual	1112
the number of extracted opinion words by algorithm	907
the number of tagged opinion words by manual from above algorithm results	834
Precision	0.92
Recall	0.75

The data of Table 3 shows that the precision of opinion word identification is high and produces better performance. It mainly depends on the effective implementation of product feature extraction and association rule mining algorithm. Meanwhile, we also noticed that some opinion words could not be identified due to the complexity and context of natural language. We can see that some features are implicit in the sentence and hard to find. In this case, it is not suitable for the association rule mining algorithm to extract opinion words.

Besides, we compared the performance with the Hu's work. Precision and recall are still used to evaluate the experiment. Table V. gives precision and recall results.

Table 4. Performance Results on Opinion Word Identification

method	precision	recall
Our approach	0.92	0.75
Hu's work	0.85	0.74

The results indicate that our proposed approach possess much higher precision and recall than Hu's work. In Hu's work, the extracted frequent features contain a lot of errors and much more non-feature noun is generated, leading to precision decreasing of opinion word recognition. On the contrary, in our approach, with two-stage filter scheme, product features are identified correctly and have got more domain correlation, resulting in high precision in subsequent opinion word recognition.

5. Conclusion

This article focuses on fundamental task in opinion mining, namely, opinion word identification. We propose an approach to extract opinion words based on association rule mining. In order to improve the precision of extraction, two-stage filtering scheme for the product feature is firstly employed, which is beneficial to the field correlation. Subsequently, opinion words are extracted with the help of the obtained association rules between opinion words and targets.

In the future, we plan to expand the experiment data scale and verify the effectiveness of the proposed method. We will fully mine the syntactic rule relationship between the target and opinion word and realize the identification task for more fine-grained target and opinion words. Furthermore, based on the above opinion word identification, we will also focus on the evaluation the polarity of the opinion word ((*i.e.*, whether given opinion word express positive or negative opinion), and further calculate its degree of positivity or negativity.

Acknowledgement

This paper is a revised and expanded version of a paper entitled Opinion Word Identification based on Association Rule Mining Algorithm, which presented at 2014 International Symposium on information Technology Convergence, Korea on October 30-31. Meanwhile, this material is based upon work supported by the Social Science Foundation of China under Grant Numbers 12CTQ042, and National Science Foundation of China under Grant Numbers 61363039 and 71361012, and Jiangxi Department of Education Science and Technology Project (No. GJJ12734), and Transformation Project of Scientific and Technological Achievements from Colleges and Universities in Jiangxi Province under Grant Numbers KJLD14035.

References

[1] S. Na, Y. Lee and S. Nam, “Improving opinion retrieval based on query-specific sentiment lexicon”, Proceedings of the 31st European Conference on Information Retrieval, Toulouse, France, **(2009)**.

[2] M. Muhammad, S. Missen, M. Boughanem and G. Cabanac, “Opinion mining: reviewed from word to document level”, Social Network Analysis and mining, vol. 3, no. 107, **(2013)**.

[3] M. Q. Hu and B. Liu, “Mining and summarizing customer reviews”, Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining, Seattle, American, **(2004)**.

[4] Y. B. Wu, Q. Zhang, X. J. Huang and L. D. Wu, “Phrase dependency parsing for opinion mining”, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, Singapore, **(2009)**.

[5] W. Jin, H. H. Ho and R. K. Srihari, “Opinion Miner: a novel machine learning system for Web opinion mining and extraction”, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, **(2009)**.

[6] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya and S. Merugu, “Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments”, Proceedings of 2011 SIAM International Conference on Data Mining, Mesa, Arizona, USA, **(2011)**.

[7] W. Zhongqing, W. Rongyang and Panglei, “Technical Report on Suda-SAM-OMS Sentiment Analysis System”, Proceedings of third Chinese Opinion Analysis Evaluation, Jinan, China, **(2011)**.

[8] A. M. Popescu and O. Etzioni, “Extracting Product Features and Opinions from Reviews”, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, (2005) October 6-8; Vancouver, B.C. Canada, **(2005)**.

[9] D. K. Lin, “Dependency-Based Evaluation of MINIPAR”, Proceedings of the Workshop on the Evaluation of Parsing Systems, Granada, Spain, **(1998)**.

[10] G. Qiu, B. Liu and J. Bu, “Opinion Word Expansion and Target Extraction through Double Propagation”, Computational Linguistics, vol. 1, no. 37, **(2011)**.

[11] S. M. Kim and E. Hovy, “Extracting opinions, opinion holders, and topics expressed in online news media text”, Proceedings of the Workshop on Sentiment and Subjectivity in Text at the joint COLING-ACL, Sydney, Australia, **(2006)**.

[12] K. Bloom, N. Garg and S. Argamon, “Extracting appraisal expressions”, Proceedings of the Human Language Technology: The Conference of North American Chapter of the Association for Computational Linguistics, New York, USA, **(2007)**.

[13] A. Kamal, M. Abulaish and T. Anwar, “Mining feature-opinion pairs and their reliability scores from Web opinion sources”, Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, Craiova, Romania, **(2012)**.

[14] S. Xiaolei, W. Suge and L. Hongxia, “Research on Comment Target Recognition for Specific Domain Products”, Journal of Chinese Information Processing, vol. 1, no. 24, **(2010)**.

[15] N. Jindal and B. Liu, “Opinion Spam and Analysis”. Proceedings of First ACM International Conference Search and Data Mining, Stanford, California, USA, **(2008)**.

Authors

Jiang Tengjiao, female, born in 1976, PHD candidate, Lecturer, her main research interests center on sentiment analysis and Web data management.

Zhong Minjuan, female, born in 1976. PHD, vice-professor, She is a member of China Computer Federation. Her main research interests center on opinion mining, sentiment analysis and information retrieval.

Liao Shumei, female, born in 1976. PhD, vice-professor. Her main research interests center on data mining.

Luo Siwen, male, born in 1971 Master vice-professor. His main research interests center on data mining.