Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 2, February 2018, pp. 357-362
DOI: 10.23940/ijpe.18.02.p17.357362

Extracting Emotional Units based on POS Templates

Zhenggao Pan*, Lili Chen

School of Information Engineering, Suzhou University, Suzhou, 234000, China

Abstract

With the increasingly popularity of electronic commerce, a large number of product reviews appeared in electronic commerce websites, which implicated a lot of valuable business information. Sentiment analysis is the core issue in disposing of business information, and the product feature words and sentiment words extraction are key technology that affect the quality of sentiment analysis. This paper proposes a simultaneous extraction algorithm of product feature words and sentiment words based on part-of-speech(POS) relation templates. Firstly, we extract possible POS dependency templates in a training set by using the supervised sequence rules mining algorithm. Secondly, we use the templates in the test samples to extract possible two tuple of product feature words and sentiment words. Finally, we test this method in a hotel review corpus. The experimental results show that this proposed method has a good application effect.

Keywords: natural language processing; product feature and sentiment extraction; part of speech

(Submitted on November 13, 2017; Revised on December 28, 2017; Accepted on January 28, 2018)

1. Introduction

With the rapid development of electronic commerce, the Internet has formed a huge amount of product review text information, which provides a wealth of raw materials for the study of text sentiment classification and provides an important way for business research institutions to obtain business information conveniently. The unstructured features of text bring a lot of inconvenience to text analysis. How to construct a ruled knowledge base system is currently a very significant research topic in the field of document analysis.

Sentiment analysis of product reviews requires a strong domain ontology knowledge base as support [2], and the domain ontology knowledge changes greatly over time [3]. The establishment of the domain ontology knowledge base is not only needs to spend a lot of manpower, but also needs the participation of domain experts, which is very difficult. Therefore, it is very important to study the domain ontology knowledge in the automatic or semi-automatic mining of product reviews.

By analyzing the commodity reviews corpus, there are a few certain language patterns between commodity entities and commodity sub-feature. One of the most easily discovered and most commonly used language patterns is (< (noun), commodity entity >, < (noun), commodity sub feature >). For example, the sentences "Hilton's bed", "Wanda's bathroom" and so on, as shown in Figure 1. These patterns are beneficial to extract comment objects from commodity reviews. Some scholars automatically constructed a domain ontology base of commodity properties words and comment words by graph-based ranking algorithm [1]. However, the graph-based ranking algorithm is based on set theory, and is ineffective to deal with one to one relationships. Hu and Liu proposed a product feature words extracting algorithm, applied class sequential rules algorithm to extracting the templates of product feature words, and the algorithm improved the generalization ability of the template [4]. By adopting bootstrap learning algorithm and starting from a small amount of product features, Zhao and Zhou found a part of speech template between product features words and sentiment words in a large number of unlabeled corpus [6]. In some previous studies, extracting product feature words is separated from extracting sentiment words [7]; in fact, the two aspects can enhance each other because a product feature word corresponds to several specific sentiment words and vice versa. For example, the sentiment word "favourable" can modify the product feature word "price", not "service".

* Corresponding author.
E-mail address: szxypzg@163.com

Therefore, by mining the collocation rules of product feature words and emotional evaluation to be used, we can improve the accuracy of extracting product feature words and sentiment words.

This paper proposes a product feature words and sentiment words extraction algorithm based on POS (part of speech) dependency templates. Meanwhile, we annotate a Hotel-Comments corpus based on our own annotation specifications to test the validity of the algorithm proposed in this paper.

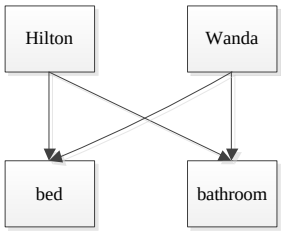

Figure 1. Hierarchical knowledge relationship

2. Product review analysis

After word segmentation and POS tagging in the comment texts, we found certain collocation rules between comment objects and evaluation words. The rules are advantageous to extracting emotion information of review texts. A few examples to introduce these part-of-speech collocation rules follows.

Step 1: Word segmentation and part of speech tagging

There are five selected sentences from an original comment corpus in Table 1.

Table 1. Sentences selected from a comment corpus

NO.	Part of speech analysised	Product feature words	Sentiment words
1	The/*r* area/*n* of/*p* the/*r* bathroom/*n* is/*v* very/*d* large/*a* ./*wp*	area	large
2	The/*r* room/*n* still/*d* is/*v* clean/*a* ./*wp*	room	clean
3	The/*r* pillow/*n* also/*d* is/*v* very/*d* large/*a* ./*wp*	pillow	large
4	The/*r* scenery/*n* is/*v* really/*d* very/*d* beautiful/*a* ./*wp*	scenery	beautiful
5	This/*r* position/*n* is/*v* very/*d* good/*a* ./*wp*	position	good

We use the 863 part-of-speech tagging set to label the words in Table 1, and product feature words and sentiment words are marked artificially.

Step 2: Label processing

In Table 1, the product feature words, sentiment words and pronouns or punctuation marks are replaced by tag <feature>, <sentiment> and <*> respectively. And we get Table 2.

Table 2. The result of label processing

NO.	The result of label processing
1	*<feature>/n<.*>/d<sentiment>/a*
2	*<feature>/n<.*>/d<.*>/v<sentiment>/a*
3	*<feature>/n<.*>/d<.*>/v<.*>/d <sentiment>/a*
4	*<feature>/n<.*>/d<.*>/d <sentiment>/a*
5	*<feature>/n <.*>/d <sentiment>/a*

Step 3: Label standardization processing

In order to find a common POS collocation relationship in several examples, we replaced all words and part of speech tagging by tagging <*> in Table 2. And we get Table 3.

Table 3. The result of label standardization processing

NO.	Templates	Explanations
1	*<feature>/n<.*>/d<.*><sentiment>/a*	*/d<.*>* replaced with */d*
2	*<feature>/n<.*>/d<.*><sentiment>/a*	*/d<.*>* replaced with */d<.*>/v*
3	*<feature>/n<.*>/d<.*><sentiment>/a*	*/d<.*>* replaced with */d<.*>/v<.*>/d*
4	*<feature>/n<.*>/d<.*> <sentiment>/a*	*/d<.*>* replaced with */d<.*>/d*
5	*<feature>/n <.*>/d<.*><sentiment>/a*	*/d<.*>* replaced with */d*

It can be seen from Table 3 that there is the same template in commodity review example sentences in Table 1; that is, comment object (product feature) is a noun, and sentiment words are adjectives. Using the <feature>/n<.*>/d<.*><sentiment>/a template, to match the sentences in Table 1, we can get the words in tag <feature> or <sentiment> position, as in Table 4.

Table 4. Results of extraction

NO.	Part of speech analysised	<feature>	<sentiment>
1	The/*r* area/*n* of/*p* the/*r* bathroom/*n* is/*v* very/*d* large/*a* ./*wp*	bathroom	large
2	The/*r* room/*n* still/*d* is/*v* clean/*a* ./*wp*	room	clean
3	The/*r* pillow/*n* also/*d* is/*v* very/*d* large/*a* ./*wp*	pillow	large
4	The/*r* scenery/*n* is/*v* really/*d* very/*d* beautiful/*a* ./*wp*	scenery	beautiful
5	This/*r* position/*n* is/*v* very/*d* good/*a* ./wp	position	good

Comparing Table 1, we found only the first sentence had an error in the product feature words' extraction in Table 4, so the patterns of product feature words extraction accuracy were 80.0%, sentiment words extraction accuracy was 100%, and the templates of the overall confidence degree is high.

From the above analysis, we can see that there are some POS dependency templates between the product features words and sentiment words in the domain corpus. Using these templates, we can extract the product features and sentiment words in certain domain.

3. Sequence mining algorithm

Using sequence mining algorithm [5], high frequency sequences supporting minimum threshold can be found. The sequence mining algorithm is very effective in mining tag sequence templates [7]. In the algorithm, support degree described tag co-occurrence frequency and confidence measured the reliability of the model. Sequence mining algorithm is described as follows.

For the given items set $I=\{i_1,i_2,...,i_n\}$, classes set $C=\{c_1,c_2,...,c_m\}$ and $C\subset I$, $s=<a_1a_2...a_r>$, represents the sequences which related the task, $a_i=(x_1x_2...x_k\}(x_j\subseteq I)$ represents database transaction. x_j can only occur once in the same sequence elements, but several times in different sequence elements. The size of the sequence is the number of elements, and the length of the sequence is the number of data items. A sequence of length k is called *k-sequence*. If there is $1\le j_1\le j_2\le...\le j_{r-1}\le j_r\le v$, then $a_1\subseteq b_{j1},a_2\subseteq b_{j2},...,a_r\subseteq b_{jr}$, the sequence $S_1=<a_1a_2...a_r>$ is a sub sequence of the sequence $S_2=<b_1b_2...b_v>$.

A series of tuple $<S_{id}, S>$ constitute the sequence database S, S_{id} represents the appeared order of S sequence. If S_j is a sub sequence of S, then the tuple $<S_{id}, S>$ included S_j sequence. Sequence mining rules are the implication like $X\rightarrow Y$, $X=<S_1x_1S_2x_2...x_lS_{r+1}>(S_i=<>$ or $S_i=<i_1i_2...i_k>$, and $i_m\notin C$, x_i represents a impossible class, $x_i\notin I$, $1\le i\le l$), $Y=<S_1c_{k1}S_2c_{k2...}c_{kr}S_{r+1}>(c_{ki}\in C,1\le i\le l)$.

In summary, S support for Y is the number of times that S contains Y, and the $X\rightarrow Y$ confidence is the ratio of Y and X's support to the degree of support of Y.

In this paper, we use the classical GSP (Generalized Sequential Pattern) algorithm. Through iterative searching of layer by layer, we find all high frequency sequences that meet the user's threshold. GSP algorithm process is as follows. The training corpus is processed by serialization, which only retains the POS information of the vocabulary and the collection of the class C($C=\{<feature>,<sentiment>\}$, as shown in Table 5.

Algorithm Generalized Sequential Pattern Mining Algorithm(GSP)

```
1: C1←init -pass(S)
2: F1←{<{f}>|f∈C1,f.count/n≥minsup}
3: For(k=2;Fk-1≠null;k++) do
4:  Ck←candidate-gen-SPM(Fk-1);
5:     For each data sequence s∈S do
6:       For each candidate c∈Ck do
7:         If c is contained in s then
8:           c.count++;
9:         End if
10:      End for
11:    End for
12:   Fk←{c∈Ck|c.count/n≥minsup}
13: End for
14: Return ∪kFk
```

Function candidate-gen-SPM(F_{k-1})
Join step: Add the last item of s_2 to s_1 when the first item removed from s_1 sequence pattern and the last item removed from s_2 are the same. Prune step: If a sub sequence of candidate sequential patterns is not sequence pattern, this pattern is not possible candidate sequence pattern, will remove it from the candidate sequence.

Table 5. Label standardization

Position number S_{id}	Sequences S
1	*<feature>/n/d<sentiment >*
2	*<feature>/d/v<sentiment >*
3	*<feature>/d/v/d <sentiment >*
4	*<feature>/d/d <sentiment >*
5	*<feature>/d <sentiment >*

4. POS templates mining algorithm

Input: The corpus S which is preprocessed, including word segmentation and part of speech tagging, $S_i(i=1,2,...,n)$represents the *i-th* sentence in S, (a_{ik},s_{ik})represents the *k-th* pair of product feature and sentiment words in S_i and $k=1,2,...,m$.

Output: High confidence product features words and sentiment words collocation templates set P.

Step 1: $t_{ik}=replace(s_i,a_{ik},s_{ik})$;

That is to say, replaced <feature> with $a_{ik}$and its POS tags, <sentiment> with S_{ik} and its POS tags, then the result t_{ik} is added to the patterns set. For example, if s_i= "The/r room/n still/d is/v clean/a ./wp", when a_{ik} represents "room", s_{ik} represents "clean", then t_{ik}="The/r <feature> still/d is/v <sentiment>".

Step 2: $c_{ik}=path(t_{ik})$;

This step extracted a dependent path from t_{ik}, and the path is from <feature> to <sentiment>. Then it is obtained the lexical pattern c_{ik}="<feature> still/d is/v <sentiment>".

Step 3: $O_{ik}=remove(c_{ik})$;

This step generalized the templates, removing the words included in c_{ik}, then the POS template O_{ik} is obtained, and O_{ik}="<feature>/d/d<sentiment>".

Step 4: $h_j=data\_mining(O_{ik})$;

In this step, sequence rules mining is executed in POS template set O, constructed high frequency generalization template set H, $h_j$represents *j-th* element, requiring both <feature> and <sentiment> are included.

Step 5: Construction extracting template p_j from h_j;

According to prior knowledge, supplement the <feature> and <sentiment> POS information, in order to improve the accuracy of the model.

Step 6: if($confidence(p_j)<beta$)then delete p_j from P;

Filtering the template p_j according to the threshold *beta*, eliminated p_j from P when $confidence(p_j)<beta$.

Step 7: return P

The template p_j confidence definition in Step 6 is as follows.

$$confidence(p_j)=\frac{\sum_{i=1}^{n_j} Ec_i}{\sum_{i=1}^{n_j} M_i}\times 100\%,\quad (j=1,...,m) \tag{1}$$

Where n_j represents the total number of the sentences that can match the template p_j in training samples, EC_i represents the times of product feature words and sentiment words two-tuples that are extracted from the *i-th* sentence matching the template p_j, and M_i represents the total number of times that the *i-th* sentence matching the template p_j.

5. POS templates mining algorithm

The adopted data set came from Hotel-Comments data in an open source data sharing website Data tang (http://www.datatang.com/data/11970), which includes 2,000 evaluation texts. We selected 5,000 reviews sentences from the texts as experimental corpus. The experiment randomly selected 4,000 reviews sentences as a training text set, and the rest as a testing set. The preprocession includes word segmentation, POS tagging and so on, which used the ICTCLAS (http://ictclas.nlpir.org/) developed by the Institute of Computing Technology Chinese Academy of Science.

In order to verify the performance of the proposed algorithm, precision rate and recall rate are used to evaluate its effectiveness in this paper, as shown in the following formula (2) and (3).

$$Precision = \frac{\sum_{i=1}^{n} Ec_i}{\sum_{i=1}^{n} E_i} \tag{2}$$

$$Recall = \frac{\sum_{i=1}^{n} Ec_i}{\sum_{i=1}^{n} C_i} \tag{3}$$

Where n represents the total number of test sample sentences in the product field, EC_i represents the number of the entries that are correctly extracted from the *i-th* sentence by using the templates, C_i represents the number of the entries in the *i-th* sentence, E_i represents the times number of the *i-th* sentence matching with the templates.

There are three kinds of evaluation criteria, which are different from the study object of the "entry", as shown in Table 6.

Table 6. Experimental evaluation criteria

The meaning of "entry"	Precision	Recall
product feature words	Pf	Rf
sentiment words	Ps	Rs
<product feature word, sentiment word>	P	R

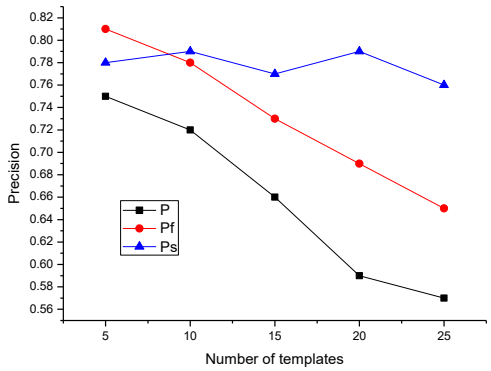

Figure 2. The precision of template extraction in the hotel reviews

Figure 2 shows that the precision of the algorithm is descending with the increase in the number of templates. That shows that the confidence level of the model in training text is basically consistent with the test text. Although the average accuracy rate is only 66% when the number of templates is 25, it is acceptable to reduce the labor cost.

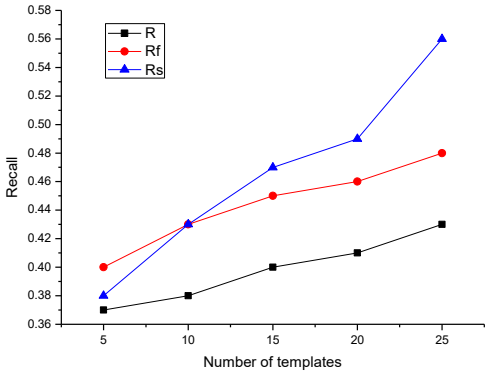

Figure 3. The recall of pattern extraction in the hotel reviews

Figure 3 shows that the recall of the algorithm is ascending with the increase in the number of templates. That shows that the templates extracted from the class sequence mining algorithm is effective for the test text.

6. Conclusions and future work

This paper proposed a new domain knowledge model, which comprehensively considers the POS of product feature words and sentiment words and the collocation relationships. In order to verify the validity of the proposed method, this paper has carried out experiments on hotel reviews corpus provided by Data tang. Experimental results show that the proposed method can substantially reduce the labor cost in constructing domain knowledge base. Actually, there is a certain relevance between product feature words and sentiments words in the semantic. So, for the next step we will research the semantic relationship and try to design an effective algorithm to extract more knowledge from product reviews corpus.

Acknowledgments

This work was supported by Outstanding Young Talent Support Program of Anhui Province of China (No. gxfxZD2016256) and Software Engineering Leader Program of Suzhou University of China (No. 2014XJZY43).

References

1. S. Cen, Y. Mao, R. Li, “Credit distribution: A graph-based approach to extract product description words,” *Knowledge Acquisition and Modeling, KAM’08. International Symposium on*, pp.398-402, 2008.
2. L. L. Dong, F. R. Zhao, X. Zhang, “Analysing Propensity of Product Reviews Based on Domain Ontology and Sentiment Lexicon,” *Computer Applications and Software*, vol. 31, no. 12, pp.104-108, 2014.
3. J. Z. Du, J. Xu, Y. Liu, “Research on Construction of Feature-Sentiment Ontology and Sentiment Analysis,” *New Technology of Library and Information Service*, vol. 30, no. 5, pp.74-82, 2014.
4. M. Hu, B. Liu, “Opinion feature extraction using class sequential rules,” *Proc. of the Spring Symposia on Computational Approaches to Analyzing Web blogs*, pp. 61-66, 2006.
5. R. Srikant, R. Agrawal, “Mining sequential patterns: generalizations and performance improvements”, *Proc. of 5th International Conference on Extending Database Technology(EDBT)*, pp. 3-17, 1996.
6. W. J. Zhao, Y. Zhou, “A template-based approach to extract product features and sentiment words”, *IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE’09)*, pp. 1-5, 2009.
7. Y. Zhang, Y. W. Liu, “Sequential Pattern Algorithm of Association Rules based on Constraint”, *Journal of Taiyuan Normal University (Natural Science Edition)*, vol. 14, no. 1, pp. 44-48, 2015.