Extracting Implicit Features in Online Customer Reviews for Opinion Mining

Yu Zhang
Zhejiang Sci-Tech University
School of Information Science and Technology
Hangzhou, 310018 China
yzh@zstu.edu.cn

Weixiang Zhu
Zhejiang Sci-Tech University
School of Information Science and Technology
Hangzhou, 310018 China
zhwx-2008@163.com

ABSTRACT

As the number of customer reviews grows very rapidly, it is essential to summarize useful opinions for buyers, sellers and producers. One key step of opinion mining is feature extraction. Most existing research focus on finding explicit features, only a few attempts have been made to extract implicit features. Nearly all existing research only concentrate on product features, few has paid attention to other features that relate to sellers, services and logistics. Therefore in this paper, we propose a novel co-occurrence association-based method, which aims to extract implicit features in customer reviews and provide more comprehensive and fine-grained mining results.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural language processing – text analysis

General Terms

Algorithms

Keywords

Opinion mining, implicit feature, co-occurrence, association

1. INTRODUCTION

With the fast development of e-commerce, the number of customer reviews increase very rapidly. On one hand, it is very difficult for users to read all these reviews and obtain useful information. On the other hand, some customers only focus on some rather than all the features of the transactions. Therefore, it is essential to provide feature-based opinion mining results to help customers make purchase decisions and help sellers and producers promote their business. One key step of opinion mining is feature extraction.

Most existing research focus on finding explicit features [3, 2], only a few attempts have been made to extract implicit features. Liu utilizes rule mining technique to map opinion words to features, and then determines implicit features according to the mapping [1]. Su develops a mutual reinforcement approach that exploits the hidden sentiment association between product feature category and opinion word group to determine implicit features [6]. Qiu proposes a regularized topic modeling framework which extracts implicit features according to the similarity of opinion word usage patterns in different reviews [4]. Our work differs from existing research and the contribution is two-fold: (1) we not only focus on product features, but also concentrate on many other features that relate to online transactions, such as sellers, services and logistics; (2) we not only concentrate on the associations between feature words and opinion words, but also utilize the associations between feature words and the rest of the notional words in the clause.

WWW2013 Rio de Janeiro, Brazil
ACM 978-1-4503-2038-2/13/05.

2. METHOD AND EXPERIMENT

Our method includes 4 steps, as shown below: 1) determine the co-occurrence matrix C; 2) determine the word modification matrix M; 3) obtain candidate feature word set F_c;

Step 1: Determine the co-occurrence matrix C

First, we identify all the notional words in text corpus D. Then for each pair of notional words, we record their co-occurrence frequency at clause level in a square matrix C.

Step 2: Determine the modification matrix M

We borrow the idea from Qiu's double propagation approach and propose a bilateral iterative method to determine the modification matrix M which records the modification relationship between opinion words and their corresponding feature words within the same clause [5]. The algorithm is shown below:

1. Select several opinion words from D and constitute the seed words set O_0, $O = O_0$. We can use different criteria to select seed words, such as word frequency and Tf-idf.

2. For each opinion word $o_j \in O$, we scan all the reviews in D and extract its corresponding feature words according to grammar rules. When a feature word f_i is found, then the modification frequency between o_j and f_i increases by 1 and is recorded in M. If $f_i \notin F$(F denotes the feature word set), then we add f_i to F.

3. For each feature word $f_i \in F$, we scan all the reviews in D and extract its corresponding opinion words according to grammar rules. When an opinion word o_j is found, then the modification frequency between o_j and f_i increases by 1 and is recorded in M. If $o_j \notin O$, then we add o_j to O.

4. When one of the two conditions is satisfied, then the iteration stops: (1) the number of newly found feature words or opinion words is less than a threshold θ; (2) the number of modification relations is less than a threshold β; Or else, return to 2 and continue;

5. After the iteration stops, if a feature word $f_x \in F$ has few relations, then we delete f_x and the x_{th} row of M. Similarly, if an opinion word $o_y \in O$ has few modification relations, then we delete o_y and the y_{th} column of M. Finally, we get matrix M.

Step 3: Obtain candidate feature word set F_c

For a review sentence $\mathcal{R}$ which has no explicit feature, we identify all the opinion words in $\mathcal{R}$ and form the set O_r. Then according to M, we select all the feature words that can be modified by the opinion words in O_r and constitute candidate feature word set F_c.

Step 4: Extract implicit features

In a review sentence, the commented features are not only related to opinion words, but also correlate with the rest notional words in the sentence. For example, "No electricity after a few phone calls." There is no explicit feature in this review. When we humans read it, we can tell that it comments on the battery feature of a mobile phone. In fact, the word "battery" often co-occurs with the words "no electricity" and "telephone call" in reviews. Figure 1 shows several example reviews both in Chinese and English. In sentence 1 to 3, we observe that "battery" and "no electricity" coexist in one clause. Meanwhile in sentence 4 to 6, "battery" and "telephone call" also coappear in the same clause. Therefore, we can infer the implicit feature according to the associations between candidate feature words and the rest of the notional words in the clause.

1. 就是带的两个电池都没电呢!
   Just the two brought batteries have no electricity!
2. 刚开始没开机是因为电池没电。
   Hadn't turned on the mobile phone at first because the battery had no electricity.
3. 电池用一天就没电了。
   The battery has no electricity after one day's usage.

4. 正常接打电话电池能用两天。
   The battery lasts for 2 days if making and answering telephone calls normally.
5. 电池接电话用用还行。
   The battery is fine if only for making telephone calls.
6. 电话还没打电池就没电了。
   The battery already has no electricity before making a telephone call.

Figure 1: Example review sentences

Given a review sentence $\mathcal{R}$ which has no explicit features. Suppose there are v notional words in $\mathcal{R}$, denoted as set $W = \{w_1, w_2, \ldots, w_v\}$. If W and candidate feature word set F_c have identical elements, we remove these elements from W and let F_c still have them. That is, $W_- = W - (W \cap F_c)$. Then, we obtain a $|F_c| \times |W_-|$ submatrix S from matrix C. The rows in S denote the candidate feature words in F_c while the columns denote the notional words in $\mathcal{R}$.

According to S, we can obtain pairwise correlations between feature words in F_c and notional words in W_-, namely the co-occurrence probability. Suppose there are n_n words in D, among which word w_a appears n_a times and word w_b appears n_b times. Word w_a and w_b co-appear n_c times in the same clause. Therefore, the probability when word w_a appears in a clause given word w_b also appears in the same clause can be calculated as follows:

$$P(w_a|w_b) = \frac{P(w_a w_b)}{P(w_b)} = \frac{n_c/n_n}{n_b/n_n} = \frac{n_c}{n_b} \quad (1)$$

Then for each candidate feature words f_i in F_c, we calculate its average correlations with each word in W_-. After calculation, we choose the one which has the highest $T(f_i)$ value as the implicit feature of review $\mathcal{R}$.

$$T(f_i) = \sum_{j=1}^{v} P(f_i|w_j)/v \quad (2)$$

We use the example review "No electricity after a few telephone calls." to illustrate our method. The text corpus D used in this paper includes 30 thousand reviews on mobile phone and clothes. These reviews were obtained from Taobao. First, we start Step 1 to determine the co-occurrence matrix C. Second, we determine the modification matrix M by using our bilateral iterative method. Then we identify the opinion word set: $O_r = \{$"no electricity"$\}$. According to the opinion word in O_r, we manage to determine candidate feature word set $F_c = \{$"battery", "screen", "mobile phone"$\}$. There are 2 notional words in review $\mathcal{R}$, : $W_- = W = \{$"no electricity", ''telephone call"$\}$. During the bilateral iteration process, we can obtain the co-occurrence frequency of candidate feature words in F_c and notional words in W_- as shown in matrix S below:

$$S = \begin{array}{c} \\ \text{battery} \\ \text{screen} \\ \text{mobile phone} \end{array} \begin{array}{c} \begin{array}{cc} \text{telephone call} & \text{no electricity} \end{array} \\ \begin{bmatrix} 1170 & 3980 \\ 79 & 63 \\ 412 & 2876 \end{bmatrix} \end{array}$$

The occurrence numbers of "telephone call" and "no electricity" are 10142 and 9874 respectively. For each $f_i \in F_c$, we have:

$$T(\text{battery}) = (\tfrac{1170}{10142} + \tfrac{3980}{9874})/2 = 0.2592$$
$$T(\text{screen}) = (\tfrac{79}{10142} + \tfrac{63}{9874})/2 = 0.0071$$
$$T(\text{mobile phone}) = (\tfrac{412}{10142} + \tfrac{2876}{9874})/2 = 0.1659$$

The result is that $T(\text{battery}) > T(\text{mobile phone}) > T(\text{screen})$, therefore "battery" is selected as the implicit feature.

We randomly select 2000 reviews from Taobao which contain 8074 clauses. Among these clauses, there are 714 clauses which have no explicit feature words. Then, we utilize them for experiment. The reviews are from two categories: mobile phone and clothes. We ask three postgraduate students who are not our team members to determine the implicit features and use their results as evaluation standards. Table 1 shows the results. We can see that the precisions of our proposed method in both categories exceed 80%. Our method has better performance in mobile phone category than in clothes category both for precision and recall. The reason is that the features of mobile phone are fixed and standardized, however, the features of clothes are more complicated and variable.

Table 1: Precision and Recall

	mobile phone	clothes
Precision	81.34%	80.17%
Recall	79.51%	77.38%

3. CONCLUSION

In this paper, we propose a novel co-occurrence association-based implicit feature extraction method. No prior knowledge nor manual tagging is needed in employing our technique. Preliminary experimental results on real-life data indicate that our proposed method is effective in performing its tasks. Although we choose Chinese web reviews as examples, the proposed technique is also applicable to reviews written in other languages. We do believe that this method presents a promising path for future research on extracting implicit features from online customer reviews.

4. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China under Grant No.61100183, and Zhejiang Provincial Natural Science Foundation under Grant No.Y1110477.

5. REFERENCES

[1] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW2005*, 2005.

[2] Q. Mei and et al. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *WWW2007*, 2007.

[3] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT/EMNLP 2005*, 2005.

[4] G. Qiu and et al. Implicit product feature extraction through regularized topic modeling. *In JZUS*, 2011.

[5] G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding domain sentiment lexicon through double propagation. In *IJCAI*, 2009.

[6] Q. Su and et al. Hidden sentiment association in chinese web opinion mining. In *WWW 2008*, 2008.