Mining and Summarizing Customer Reviews

An Exploration of Data Mining

Xuecong Tan, Yi Chen, Chengyue Lin
Simon Fraser University , School of Computing Science
xuecongt@sfu.ca yca267@sfu.ca cla325@sfu.ca

SFU SCHOOL OF COMPUTING SCIENCE

Abstract

As merchants selling is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. If customers and companies want to review some useful comments, it is going to be difficult for them to filter a huge amount of comments. Hence, we need a model to mine and summarize comments of a product. Our task has the following three steps: (1) obtain product features that have been commented on by customers. (2) identifying opinion sentences according to the product features we obtained from each review and deciding whether each opinion sentence is positive or negative. (3) summarizing the results by each features with their corresponding positive/negative sentences. This poster proposes the language model to perform these tasks along with sample output and evaluation results.

Introduction

As the rapid development of Internet, more and more products are sold online, at the same time people are willing to shop online. Products' reviews become much more important for both customers and manufacturers. For potential customers, it can help them to make decision whether to buy it or not. Furthermore, it is necessary for manufacturers to read reviews for getting suggestions which can help them to improve their products. Since some parts of the review are non-sense, we aim to fetch useful sentence from each review.

Main Objectives

1. Part-of-Speech tagging each word in all reviews. If any noun phase appears many times, appending it to feature list
2. Pruning the feature list in order to remove duplicated features
3. According to the feature list to allocate target sentences and identify sentence orientation
4. Creating two lists contain all adjective words and effective adjective words in target sentence
5. Identifying sentence by all adjective words list
6. Using effective words list to distinguish the sentence that cannot be identify by all adjective words list
7. The sentence is neutral if it cannot be identified by either all adjective words list or effective words list

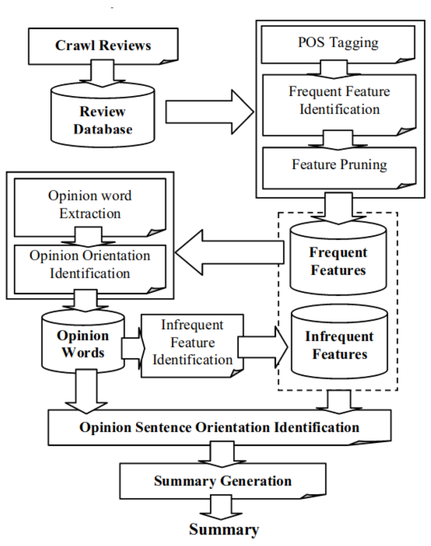

Figure 1: Figure-based opinion summarizing [1]

Methods' detail

We identify the feature which presents in each review by frequent noun phrases in each review sentences(NLTK processor / pruning). Then we identify opinion words from adjectives and predicting the semantic of opinion words(give positive/negative tag).
For tagging the opinion word,We create two list to hold all adjective words and effective adjective words(frontal and back 5 words of the) in the sentence first.
Then using NLTK.wordnet to check whether the target word, it's similarity or it's antonyms are in the list. If none of them are in the list, we apply the same method but using effective adjective words list. The sentence will be marked as neutral if these words are not in both lists.

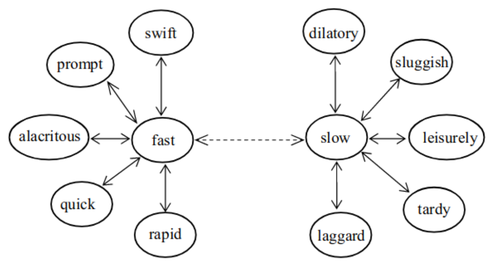

Figure 2: Bipolar adjective structure, ($\rightarrow = similarity\ OR \rightarrow = antonyms$) [1]

1. Procedure **OrientationPrediction**(*adjective_list*, *seed_list*)
2. begin
3. do {
4. $size_1$ = # of words in *seed_list*;
5. **OrientationSearch**(*adjective_list, seed_list*);
6. $size_2$ = # of words in *seed_list*;
7. } while ($size_1 \neq size_2$);
8. end

1. Procedure **OrientationSearch**(*adjective_list*, *seed_list*)
2. begin
3. for each adjective w_i in *adjective_list*
4. begin
5. if (w_i has synonym *s* in *seed_list*)
6. { w_i's orientation= *s*'s orientation;
7. add w_i with orientation to *seed_list*; }
8. else if (w_i has antonym *a* in *seed_list*)
9. { w_i's orientation = opposite orientation of *a*'s orientation;
10. add w_i with orientation to *seed_list*; }
11. endfor;
12. end

Figure 3: Predicting the semantic orientations of opinion words [1]

Data Set

We conduct our experiments by randomly picking five electronics products which are 2 digital cameras, 1 DVD player, 1 mp3 player, and 1 cellular phone from Amazon and their customer reviews from Amazon.com and C | *net.com*.

Sample Result

picture
Positive: i use this with a home theater system and it 's amazing how it sounds , the picture clarity is unmatched !
...
Negative:
maybe it 's good for tvs with not so good color settings , but not good for my tv that already has very vibrant pictures .
...
Neutral:
picture on web site clearly has the play display pictured , yet false silver plate .
...

Evaluation

Product	OS extraction (original)	OS extraction	SO accuracy(original)	SO accuracy
Digital camera1	0.643	0.739	0.927	0.529
Digital camera2	0.554	0.778	0.946	0.626
Cellular phone	0.815	0.709	0.764	0.506
Mp3 player	0.589	0.825	0.842	0.706
DVD player	0.607	0.770	0.730	0.589
average	0.642	0.764	0.842	0.591

Table 1: Results of opinion sentence extraction and sentence orientation prediction

Analysis

We improve the opinion sentence extraction precision by effectively extracting more opinion sentences from data files according to adjective and noun phrase existence. Therefore, the hitting rate of our algorithm is higher. Sentence orientation accuracy is lower because we have more features in each comment so there are more error in our algorithm. The place we use pruning is different from the algorithm of paper. This also cause some differneces in the final results.

Future Work

We are going to improve our algorithm and increase the accuracy by better pruning algorithm. Then, we need to deal with sentence that contains implicit features, such as "It cannot fit in my pockets" which is about size. Finally, we will try to use machine learning in order to figure out how to use verbs and nouns for purpose.

References

[1] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. 13:168–177, August 2004.