Application for Product Features Extraction and Sentiment Analysis from Online User Reviews

Xue Li[1, a], Lei Sun[2, b]

[1]School of Economics and Management, Xidian University, Xi'an 710000, China;

[2]School of Economics and Management, Xidian University, Xi'an 710000, China.

[a]18700937202@163.com, [b]leisun68@qq.com

Keywords: User reviews, product features, data mining, sentiment analysis.

Abstract. Based on the existing research results, the product reviews mining technology is applied in the analysis of competitive advantages of mobile phones in this paper. Taking iPhone SE and Galaxy S7 edge as the research objects, first, extract product features from the online reviews by FP-growth algorithm and rank them according to users attention. Subsequently, calculate the sentiment polarity of words. This paper ends with the advantages and disadvantages analysis of the competitive products, as well as the direction to improve. Some enlightenment from this paper to domestic consumer electronics business is expected.

1. Introduction

Network user reviews is a new source of information derived from the Web2.0 and e-commerce environment. Compared to the homepage of competitors, industry websites and other traditional sources, it has the attributes of authenticity, interactivity and relevance, etc. Research has shown that analysis and mining on product features in massive online reviews is not only convenient for customers to learn all aspects of the products performance, but also helping enterprises identify the features users concern and their attitudes. Thus, they can respond to the demand of users quickly and promote the competitiveness of enterprises at the same time [1]. In recent years, unstructured data analysis technology, opinion mining, aiming at obtaining useful information, has become a hot research topic. A large number of scholars have done research on the extraction of product features and sentiment analysis, and made certain achievements. In the field of reviews mining, unsupervised classification method based on association rules was put forward by Hu and Liu to extract English reviews on the product features, which has achieved good results [2]. Li shi introduced the method to the Chinese network user reviews, and combined the characteristics of the Chinese language to filter results, which also has obtained a good effect [3]. Zhai and Liu proposed a method inserting two constraints to Naive Bayesian classification based on EM algorithm to merge the same meaning features [4]. In sentiment analysis, Turney proposed an unsupervised point mutual information method (PMI) to determine the sentiment polarity by calculating the value of PMI between the sentiment words and "excellent" as well as "poor" separately [5]. Wang Zhenning put forward the calculation method of the sentiment polarity of words based on HowNet and PMI, improving the accuracy by 5% [6].

Based on the existing research results, the product reviews mining technique is applied in the analysis of mobile phones in this paper. The main work can be divided into the following three aspects:

(1) Extract product features from the online reviews and rank them based on users attention.

(2) Calculate the sentiment polarity of words.

(3) Analyse the advantages and disadvantages of the competitive products by visualization technology, as well as the direction to improve.

2. Method Design

In this paper, the process of research can be roughly divided into four parts, data acquisition, data

preprocessing, extracting the product features and calculating the sentiment polarity of words. It is shown in figure 1.

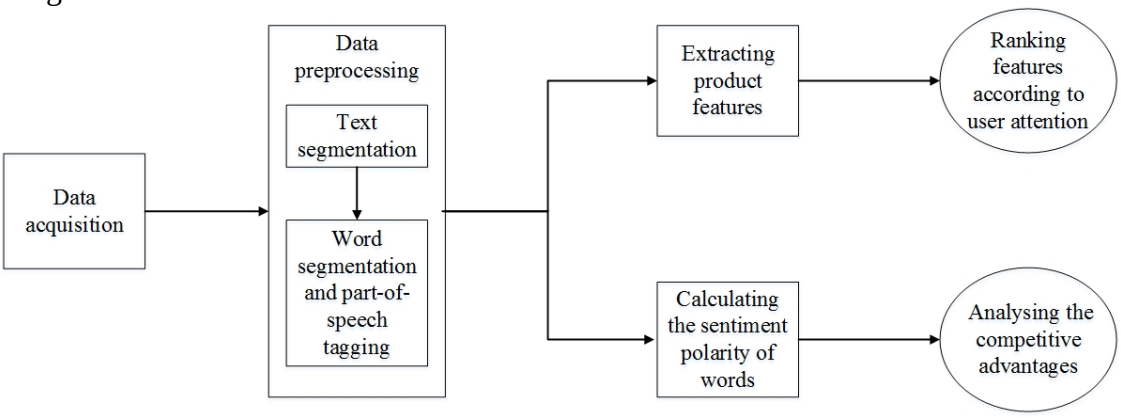

Fig. 1 The process of the research

2.1 Data Acquisition.

Considering that Meta Seeker can collect real-time comments efficiently, meanwhile, transform unstructured data into structured data identified by computer, it is selected as the web information crawler in the research.

2.2 Data Preprocessing.

1. Data cleaning. In order to improve the efficiency of mining process, data cleaning needs to delete a lot of useless and repetitive information while only reserve valuable reviews as the sources of data mining.

2. Text segmentation. The object of mining is a review in this research, thus, short sentence is chosen as the granularity to determine the product features accurately.

3. Word segmentation and part-of-speech tagging. In this paper, Chinese word segmentation tool, ICTCLAS, written by Chinese Academy of Science, is applied to do the word segmentation and part-of-speech tagging of reviews. The nouns or noun phrases are extracted by the comment corpus with part-of-speech tagging and then the transaction files are created for association rules.

4. Deleting the stop word. In the collection of reviews, many words appear with high frequency but little practical significance. For the purpose of improving the efficiency and effect of analysis, they are deleted in this paper.

2.3 Extracting Product Features.

For most of the product features are explicit, described by words directly, this paper mainly studies on extracting the explicit feature of the product. Specific steps are as follows:

1. To begin with, the Frequent Pattern growth algorithm is adopted to scan the transaction files and use the obtained frequent item sets to build a FP-tree. Then divide the FP-tree into several condition bases which are relevant to frequent item sets with the length of 1. Subsequently, do the feature recognition to frequent item sets in each condition base, separately and get frequent item sets. It is regarded as a candidate set of product features I_0.

2. The independent support is applied to filter and revise the noun and noun phrases of I_0, after that a new candidate product features set, I_1, is generated. In this paper, the number of sentences which contain noun or noun phrases of frequent features (ftr) but not contain their supersets is called the independent support of ftr. The minimum support is 1% in the research.

3. Formulate the rules of common Chinese frequent item nouns without product features and establish the corresponding noun set. Besides, filter I_1 to I_2 according to semantics and grammar. In this paper, generalization noun, colloquial reviews and address are regarded as noun common Chinese frequent item nouns irrelevant to product features.

4. The information retrieval method, TF-IDF [8], is adopted to calculate the weight of product features in user reviews. The weight of each candidate feature in reviews is shown in the following formula.

$$\mu(c_i, \mathrm{r}) = \frac{tf(c_i)}{\max tf(r)} \log\left(\frac{N}{N_{c_i}} + 1\right) \tag{1}$$

Among them, $\mathrm{tf}(c_i)$ is defined as the number of candidate feature (c_i), appearing in the review (r). $\max tf(r)$ is denoted as the maximum word frequency. Furthermore, N represents to the total number of the reviews for the product p. $\mathrm{N}c_i$ refers to the amount of reviews including c_i.In order to ensure the weight in between 0 and 1, it is normalized as follows.

$$\mu_{normal} = \frac{\mu-\mu_{min}}{\mu_{max}-\mu_{min}} \epsilon[0,1] \tag{2}$$

Thus, the formula of the weight can be set :

$$\omega(c_i, \mathrm{p}) \approx \frac{\sum_{r\in R_P} \mu(c_i,r)}{N} \tag{3}$$

From the above, the synthetical weight of each candidate feature can be calculated. According to the weight, we can get the sequence of candidate product features. Among them, higher weight implies that the feature is more significant in specific reviews. Thus, the top 10 features are chosen as the product features users concern.

2.4 Calculating the Sentiment Polarity of Words.

Sentiment polarity filtering methods can be used to remove the word with little obvious polarity view. In related literatures, HowNet thesaurus and the calculation methods of word similarity are generally applied to determine the sentiment polarity of words. However, the scope of HowNet thesaurus is limited without including many new words in network. In this paper, the method of statistics is used to judge the sentiment polarity of words, shown in the following equation.

$$\mathrm{Polarity}(0) = max \sum_{i=1}^{pos-seed} sim(o_k, w_i) - max \sum_{j=1}^{neg-seed} sim(o_k, w_i) \tag{4}$$

Among them, the kth sememe of word “o” is denoted as o_k. The positive paradigm words, i, is defined as w_i, while the negative is defined as w_j. In addition, $\mathrm{sim}(o_k, w_i)$ represents the similarity of the sememe and the positive paradigm word and $\mathrm{sim}(o_k, w_j)$ represents the similarity of the sememe and the negative paradigm word. Moreover, pos-seed and neg-seed are denoted as the number of the paradigm words. Whenever it encounters a transforming conjunction, such as, *but*, *however*, only calculate the sentiment polarity after that. What’s more, the sentiment polarity of words is on the opposite with negative adverb in the reviews. After the calculation, if the value is significantly greater than zero, it shows that the sentiment polarity word is positive, or vice versa.

3. Experiment and the Results Analysis

3.1 Corpus Data

As the representative of the consumer electronics field, Apple and Samsung are superior to other relevant enterprises, whether in the aspect of technological innovation or the brand influence. Therefore, their latest products, iPhone SE and Galaxy S7 edge, are taken as the research objects in this paper. From Zhongguancun online website, 253 items about iPhone SE and 276 items about Galaxy S7 edge are extracted. The number of effective reviews is 1482 when they are separated by sentences.

3.2 Extraction Result of Product Features

According to the feature extraction algorithm, product feature sets of iPhone SE and Galaxy S7 edge, represented by $F_{iPhone\ SE}$ and $F_{Galaxy\ S7\ edge}$, can be calculated. Then the ultimate extraction set is $\mathrm{F} = F_{iPhone\ SE} \cup F_{Galaxy\ S7\ edge}$. After combining and expanding synonymous words and normalizing the users attention, the top ten product features can be concluded, showed in table 1.

Table 1 Extraction result of product features

Ranking	Product Features	User Attention
1	Outlook	1
2	Price	0.907
3	Photograph	0.769
4	Screen	0.718
5	Battery	0.58
6	Application	0.369
7	Audio	0.243
8	Network	0.096
9	Processor	0.09
10	Memory	0

As a conclusion according to the list above, when it refers to iPhone SE and Galaxy S7 edge, product features set customers mainly concerning about is F={ outlook, price, photograph, screen, battery, application, audio, network, processor, memory }, in which five properties of outlook, price, photograph, screen and battery is most significant.

3.3 Calculation of Reviews Polarity and Strength and Competitive Analysis of Products

After quantizing the sentiment polarity of the words, the attitude of customers in each product feature, as well as the strengths and weaknesses of iPhone SE and Galaxy S7 edge, are shown clearly in the following figures.

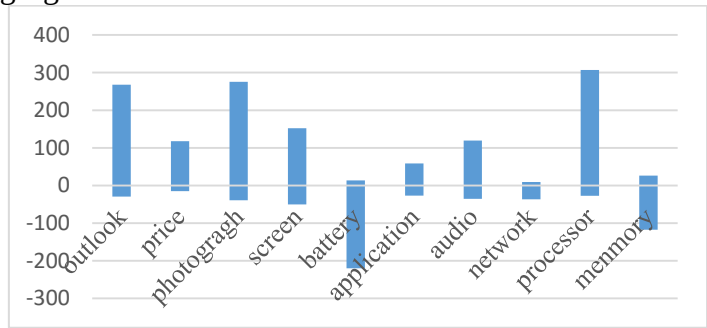

Fig. 2 Calculation of Reviews Polarity and Strength on iPhone SE

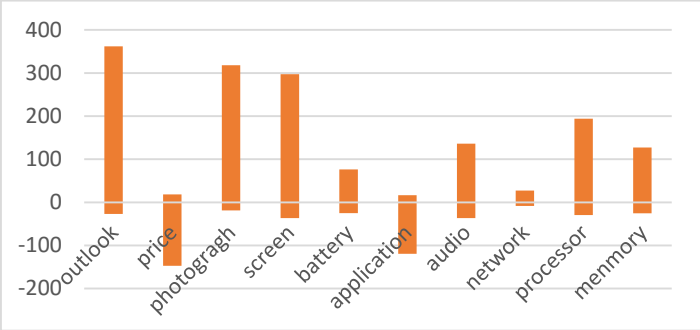

Fig. 3 Calculation of Reviews Polarity and Strength on Galaxy S7 edge

It is obvious from figure 2 and figure 3 that iPhone SE possess more positive comments in outlook, price, photograph, screen, processor and so on. Taking the parameters of iPhone SE into consideration, this result is highly correlated with the reservation of iPhone 6s' configuration, the return of classic 4 inch screen and high cost performance, and it's a satisfying choice to the group of customers in favor of small scale screen designing. Nevertheless, there exists a great many negative comments in battery and memory. It is proved by the truth that only 1600mAh battery and 2GB memory this mobile phone is built in.

Galaxy S7 edge gains lots of positive comments in outlook, photograph, screen, audio, processor, memory and so on, while a lot of negative comments on price and applications come. This is because Galaxy S7 edge adopts curved surface screen, making the whole body of the cellphone particularly delicate. Meanwhile, the memory reaching to 4G, and the adoption of Qualcomm Snapdragon 820 in processor also contributes a lot. But the optimization of its ROM is imperfect, and particularly for domestic users, it fails to localize. Moreover, it is more expensive than iPhone

SE by about 2000 RMB. These reveal that Galaxy S7 edge is more appropriate for consumers who possess high-income and pursue novelty and fashion.

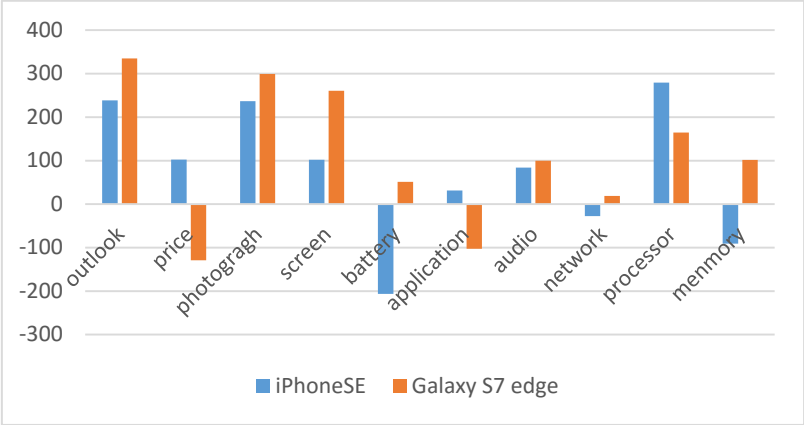

Fig. 4 The cancelled out result of positive and negative reviews of iPhone SE and Galaxy S7 edge

Combining the positive and negative customer perspectives of iPhone SE and Galaxy S7 edge, it is obvious that both kinds of products gain more positive comments on outlook, photograph, screen, processor, according to figure 4. Contrasting with Galaxy S7 edge, iPhone SE has the advantage of price and processor, and it has the disadvantage of battery and memory. Instead, Galaxy S7 edge has outstanding performance lying in outlook, photograph and screen, while the price and application of it should be improved.

4. Summary

From the above experimental analysis, obviously, for Apple, the core competitive advantage is making good use of its own technological innovation ability and creating new products by mining and stimulating market demands. Relying on the leading fingerprint identification technology, great IOS system and high software security, it has acquired a good reputation in terms of user experience and appearance design. As a mobile phone manufacturer with the most complete vertical integration, Samsung firstly unveils smartphone with curved surface screen resulting from advanced design and production capacity of critical components, which actually restricts Apple in this respect.

As the representative of the consumer electronics field, if the apple is called the pioneer of innovation, achieving success by its core technical advantages and the product strategy of “fewer, better”; then, Samsung, as the attacker, which shorts the gap with leading enterprises by imitation innovation and gradually forms the competitive advantages, eventually, realizes the process of surpass, is also admirable. Compared to them, the relevant domestic enterprises still have a certain gap in the field of user experience, technology innovation and so on. We should learn from the successful experience of Apple and Samsung and improve the advantage of technological innovation to cultivate more world-class innovation enterprises.

This research mainly focuses on extracting the explicit feature of product, and the implicit still remains to be further studied.

Acknowledgement

Thanks for the support of the National Natural Science Foundation (project number: 71573199).

Reference

[1].Zhai Dongsheng, XuYing, Huang Lucheng, et al. The Advantage Analysis of Competitive Product Based on Product Reviews Mining. Journal of Intelligence. Vol. 32 (2013) No. 2, p. 45-51.

[2].Hu M，Liu B. Mining opining features in customer reviews. The 19th National Conference on

Artificial Intelligence. California, 2004, p. 755-760.

[3].Li Shi, Ye Qiang, Li Yijun, et al. Mining Features of Products from Chinese Customer Online Reviews. Journal of Management Sciences in China. Vol. 12 (2009) No. 2, p. 142-152.

[4].Zhai Z, Liu B, Xu H, Jia P. Grouping Product Features Using Semi-supervised Learning with Soft-constraints. The 23rd International Conference on Computational Linguistics. Beijing. 2010, p. 1272-1280.

[5].Turney PD. Thumbs up or thumbs down? : Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia , 2002, p. 417-424.

[6].Wang Zhenyu，Wu Zeheng，Hu Fangtao. Words Sentiment Polarity Calculation Based on HowNet and PMI. Computer Engineering. Vol. 38 (2012) No. 15, p. 187-193.

[7].Salton G. Full Text Information Processing Using the Smart System. Database Engineering Bulletin. Vol. 13 (1990) No .1, p. 2-9.

[8].Wang Yong, Zhang Qin, Yang Xiaojie. Research on the Method of Extracting Features from Chinese Product Reviews on the Internet. New Technology of Library and Information Service. (2013) No. 12, p. 70-73.

[9].Yin Pei, Wang Hongwei. Sentiment Classification for Chinese Online Reviews at Product Feature Level through Domain Ontology Method. Journal of Systems & Management. Vol. 25 (2016) No. 1, p. 103-114.

[10].Ji Shunquan, Zhou Yi. Application of Product User Review in Research on Enterprise Competitive Intelligence. Journal of Modern Information. Vol. 35 (2015) No.6, p. 114-121.