Sentiment topic mining based on comment tags

To cite this article: Daohai Zhang et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 322 052048

View the article online for updates and enhancements.

Related content
- Rotation, Reflection, and Frame Changes: Closing remarks
  R M Brannon
- sMath for Physics: Trajectory of a projectile
  B V Liengme
- An implementation of support vector machine on sentiment classification of movie reviews
  I M Yulethia, S A Faraby, Adiwijaya et al.
Sentiment topic mining based on comment tags

Daohai Zhang*, Xue Liu, Juan Li, Mingyue Fan
School of Management, Jiangsu University, Zhenjiang, China

*Corresponding author e-mail: zdh@ujs.edu.cn

Abstract. With the development of e-commerce, various comments based on tags are generated, how to extract valuable information from these comment tags has become an important content of business management decisions. This study takes HUAWEI mobile phone tags as an example using the sentiment analysis and topic LDA mining method. The first step is data preprocessing and classification of comment tag topic mining. And then make the sentiment classification for comment tags. Finally, mine the comments again and analyze the emotional theme distribution under different sentiment classification. The results show that HUAWEI mobile phone has a good user experience in terms of fluency, cost performance, appearance, etc. Meanwhile, it should pay more attention to independent research and development, product design and development. In addition, battery and speed performance should be enhanced.

1. Introduction

Nowadays, the e-commerce platform is not only a platform for the user to buy goods, but also become a carrier for users to comment, businesses to analyze information. With the development of the e-commerce platform, a large number of user comment tags are generated based on the goods (services). Extracting effective information from these massive unstructured comment tags has become the main means and methods to get user’s feelings. The research and application of online comment tags is generated. Comment tags mining is to extract hot content the user concerned from unstructured related content that many users evaluate the views expressed by a product (service) and its characteristics, which is used in user tracking, service feedback and opinion improvement, etc. Therefore, it has a strong practical value for the platform and business, such as business can understand the consumer demand, adjust product structure and marketing strategies timely to improve product competitiveness according to the hot content mined.

In the current research, the study is mainly based on sentiment analysis and subject extraction. In sentiment analysis, Wang Gen et al. (2007) proposed a CRF method with multiple redundant markers to analyze the emotional polarity of the sentence, unified subjective and objective classification, polarity classification and appraise grading task in a model, restricting the spread of the fractional completion error[1]. Zhang Hongbin et al. (2012) introduced attitude word, and used ROSTEA text sentiment analysis tools to design sentimental tendency analysis algorithm, and described the sentimental inclination of the text from the perspective of qualitative and quantitative[2]. Chen Guoliang et al. (2016) carried out the sentiment analysis by using the hierarchical multi-strategy method of support vector machine (SVM) to extract the features of the meta word, the binaray word and the position of the word[3]. Li Hanyu et al. (2017) established sentiment analysis knowledge base and the correspondence between attribute and emotional words by extracting commodity attribute to calculate the degree of association between emotional words and positive and negative categories, thus...
determined the emotion tendency[4]. In the subject extraction research, it is mainly based on the mining of the LDA theme model, and the corpus is formed by lexical annotation and semantic analysis, then the semantic similarity calculation is carried out in the corpus, and the semantic reintegration is completed and mapped to the subject[5]. Li Fang et al.(2012) used heuristic rules and identify co-occurrence probability statistical method of combining text in the set phrase, and removed the smoke candidate words co-occurrence probability calculation and the threshold filtering noise word combination, then used LDA model to dig potential evaluation subject, and calculated the tendency the sentence using the method of multi- feature fusion[6]. Zhang Jianhua et al. (2014) used the artificial emotion dictionary after pretreatment of text extraction emotional words, through the LDA model for feature classification to construct SVM vector, to pass judgment on binary classification[7].

The existing research provides the basic theory and method for the practical application. This study combines sentiment analysis with LDA theme mining, and carries out specific application research with HUAWEI mobile phone comment label as the research object, the purpose is to provide an example for the promotion of this method in practical application. The following includes data preparation, experimental process, experimental results and analysis. During the experiment, we first perform topic mining on unclassified comment tags, grasp the user’s focus, and then make the sentiment classification for comment tags. Finally, mine the comments again and analyze the emotional theme distribution under different sentiment classification.

2. Data preparation

2.1. Topic mining process

The general steps of topic mining based on comment tags are as follows: the first step is to pretreat the comment tags data source, such as word segmentation, word dry and so on. The second step is to make the sentiment classification for comment information, finally the LDA is modeled after the emotion classified to dig out user’s comments topic word distribution. The process is shown in Figure 1.

![Fig.1 Topic Ming Process](image)

2.2. Data acquisition and pretreatment

Jingdong, Tmall, Suning, etc. are the most commercial value of the mobile phone products’ business websites, we use them as platforms for HUAWEI mobile phone sales. We collected 45000 related comment data, and retained 39690 data for the experiment through screening and duplicate removal. (Customer comment data is from [https://item.jd.com/3893493.html](https://item.jd.com/3893493.html)).

Topic mining of the emotions of comment tags is the processing of the subjective information extract in text. The data source contains the original comment information, which must be preprocessed before use: (1) De-duplication. Remove some duplicate invalid content. (2) Segmentation. All the comment data are worded and divided into phrases by using the Chinese word segmentation tool jieba. (3) Remove the stop word.
3. Mining procedure

3.1. LDA model building process

LDA (Latent Dirichlet Allocation) is a thematic model, generating theme based on the word frequency in a set of documents. LDA is a very effective method for the probability distribution of mixed words that can be found accurately and reasonably in a given document set. The general steps of LDA topic model construction are as follows: (1) After all the comment data is cleaned and segmented, the corpus doc_clean is obtained. (2) Create a dictionary of words in a corpus “dictionary”. (3) Convert corpus into document matrix “doc_term_matrix”. (4) Put the preceding dictionary and document word matrix into the function called gensim.models.LdaModel(). (5) Output the topic model.

3.2. Comment topic extraction

All the comment labels are used as input to the LDA model, and analyzed by the LDA model, the top n words with larger probability distribution in each topic model are extracted to generate the topic word matrix, that is the tag topic model. For the data collected, this study sets up 5 topics, each topic prints 5 topic words and probability distribution, getting the related thesaurus as shown in Table 1.

| Topic1 | Topic2 | Topic3 | Topic4 | Topic5 |
|--------|--------|--------|--------|--------|
| Good   | Apple  | Good   | Good   | Jingdong|
| 0.033  | 0.012  | 0.052  | 0.028  | 0.036  |
| Photo  | Android| Support | Like   | Delivery|
| 0.025  | 0.009  | 0.026  | 0.019  | 0.031  |
| Satisfy| Good-looking | Domestic goods | Very | Easy to use |
| 0.025  | 0.009  | 0.021  | 0.018  | 0.015  |
| Appearance| Come with | Made in China | Deserve | Order |
| 0.023  | 0.008  | 0.017  | 0.014  | 0.014  |
| Touch  | 4g     | Receive | Fluent | Good |
| 0.021  | 0.007  | 0.013  | 0.013  | 0.014  |

Topic1: including “take photos”, “appearance”, “touch” and other key words, indicating that users are more satisfied with camera function and appearance. It shows that HUAWEI’s design in line with human habits and can bring comfort to the user and focus on consumer experience.

Topic2: The topic includes “Apple” and “Android” that indicates users compare system and characteristics between APPLE and HUAWEI, showing that HUAWEI mobile phone has exceeded the international brand APPLE in some extent. “4G” shows that the HUAWEI mobile phone and 4G network fit high, and the market effect is good.

Topic3: The topic includes “good”, “support”, “made in China” and other key words, indicating that many users support domestic mobile phones and reflects the increasingly strong domestic brands.

Topic4: The topic includes “good”, “like”, “very”, “smooth” and other key words, indicating that users are very satisfied with the system and have good experience.

Topic5: “Jingdong”, “express delivery”, “easy to use” and other key words indicate that Jingdong is a trusted platform for users to buy mobile phones in electricity business website. It has been affirmed by consumers because its fast delivery and good service.

Draw the related word cloud based on the comment tags. We can get the main content that users concerned while buying HUAWEI mobile phones, including “cost performance”, “smooth”, “appearance”, “speed”, “domestic goods”, etc.

3.3. Sentiment classification

In this study, sentiment analysis was used to divide the seizure data into sentiment classification. There were 35,320 comments about positive sentiment accounted for 88.99%, 1770 neutral sentiment accounted for 4.46%, and 2600 of negative sentiment accounted for 6.55%. The distribution is shown in Table 2.
Table 2 Sentiment Classification Statistics

| Name  | Positive sentiment | Neutral sentiment | Negative sentiment | Totality |
|-------|--------------------|-------------------|-------------------|----------|
| Result | 88.99%             | 4.46%             | 6.55%             | 39690    |

Table 2 shows that the proportion of positive emotions is as high as 88.99%, indicating that most users have a positive attitude towards HUAWEI mobile phones.

Positive sentiment segmentation statistics are shown in Table 3.

Table 3 Positive Sentiment Segmentation Statistics

| Grade | General | Moderate | High | Total |
|-------|---------|----------|------|-------|
| Result | 23.34%  | 27.49%   | 38.16% | 88.99% |

Table 3 shows that the proportion of users’ affirmative comments on HUAWEI mobile phones from general affirmation to high affirmation increased from low to high, indicating that the user’s recognition of HUAWEI mobile phones is very high.

Negative sentiment segmentation statistics in Table 4.

Table 4 Negative Sentiment Segmentation Statistics

| Grade | General | Moderate | High | Total |
|-------|---------|----------|------|-------|
| Result | 4.64%   | 1.31%    | 0.15% | 6.55% |

Table 4 shows that the proportion of users’ negative comments on HUAWEI mobile phones from general negation to high negation decreased from high to low, indicating that although some users are negative about the HUAWEI mobile phones, their negativeness is not tall. Meanwhile, notwithstanding some problems with HUAWEI mobile phones, it is still within the range accepted by most users.

3.4. Topic extraction under different sentiments

Topic words are mined under different sentiment classification by LDA after making sentiment classification for comments. The results are shown in Table 5 and Table 6.

Table 5 Positive Sentiment Related Thesaurus

| Topic1 | Topic2 | Topic3 | Topic4 | Topic5 |
|--------|--------|--------|--------|--------|
| Good   | 0.040  | Good   | 0.025  | Still  | 0.014  | Like   | 0.021  | Use    | 0.009  |
| Very   | 0.031  | Photo  | 0.014  | Haven’t | 0.013 | Good   | 0.009  | Speed  | 0.008  |
| Jingdong | 0.025 | Fluent | 0.013  | Screen | 0.012 | Exactly | 0.009  | Still   | 0.008  |
| Like   | 0.022  | Feel   | 0.013  | Good   | 0.008  | Made in China | 0.009 | This    | 0.007  |
| Support | 0.021 | OK     | 0.010  | Feel   | 0.007  | Support | 0.008  | System  | 0.007  |

The Table 5 shows that users made positive emotional evaluation concern system-fluency, cost performance, appearance, etc.

Table 6 Negative Sentiment Related Thesaurus

| Topic1 | Topic2 | Topic3 | Topic4 | Topic5 |
|--------|--------|--------|--------|--------|
| Feel   | 0.010  | Feel   | 0.010  | Still  | 0.012  | Can’t  | 0.016  | Screen | 0.029  |
| Battery | 0.008 | Little | 0.010  | Work   | 0.009  | System | 0.011  | Haven’t | 0.010 |
| Little | 0.007  | Haven’t | 0.009 | Always | 0.007  | Hot    | 0.011  | Still   | 0.010  |
| Durable | 0.007 | Jingdong | 0.009 | Know   | 0.006  | Little | 0.009  | Touch   | 0.009  |
| Camera | 0.005  | Quality | 0.009 | Haven’t | 0.006 | Exactly | 0.008  | Feel    | 0.009  |

The Table 6 shows that users made negative emotional evaluation concern battery, generating heat, speed, etc.
4. Result and analysis
The comment tags reflect the user’s attitude toward the product. The processing of the comment tags can perform text analysis and semantic description of the comment content. Extracting valuable information from the unstructured data that can accurately describe the entity helps to provide decision-making support. We can draw the following business strategies according to the topic mining.

(1) The key point that the user focus when buying HUAWEI mobile phones is its position in the domestic mobile phone market. Similarly, good user experience, cost performance, system fluency, appearance and others are also the focus of the user. HUAWEI should pay attention to independent research and development, improve the quality of domestic mobile phones and focus on the user experience.

(2) HUAWEI mobile phones exist generating heat, running and charging slowly and other issues that are also the focus of users, so that businesses should pay attention to product design and development for further optimization.

(3) The e-commerce sales platform can adjust the product structure according to the results, apply the extracted comment hotspots to the commodity search, customize the user preference model individually, and recommend the good products to them.

5. Conclusion
With the gradual deepening of network socialization and the popularity of e-commerce, the user’s comment information has gradually become an important data source for enterprise competition and intelligence analysis. This paper studies the mining of emotional theme from user comment tags in e-commerce, so as to realize the mining analysis of automated commodity comment tags. Through word segmentation and sentiment classification of comment tags, the LDA model is used to generate the related thesaurus, and the valuable information in the user comments is excavated according to the thesaurus, which provides powerful support and help for the business management decision. This study also provides a practical method for Internet unstructured text data mining.

Acknowledgement
This paper has obtained Chinese National Natural Science Foundation(71471076) funded, Chinese Postdoctoral Science Foundation(2015M581749) funded, Jiangsu Province Social Science Foundation(13JDB031) funded and Jiangsu University senior talent foundation(15JDG007) funded.

References
[1] Wang Gen, Zhao Jun. Sentence Sentiment Analysis Based on Multi-redundant-labeled CRFs[J] National Conference on Computational Linguistics, 2007, 21: 51-55.
[2] Zhang Hongbin, Li Guangli. Research on Sentiment Orientation Analysis of Product Online Reviews[J]. New Technology of Library and Information Service, 2012(10): 61-66.
[3] Chen Guoliang, Tang Wanmei. Text Theme Ming Based on Sentiment Classification[J]. Journal of Chongqing Normal University (Natural Science Edition) , 2016(1): 92-96.
[4] Li Hanyu, Qian Fei, Zhou Pengfei. Emotional Analysis and Mining for Product Comment Texts[J]. Information Science, 2017, 35(1): 51-55.
[5] Ruan Guangce. Topic Extraction Research of Net Reviews Based on Latent Dirichlet Allocation[J]. Intelligence magazine, 2014(3): 161-164.
[6] Li Fang, He Tingting, Song Le. Opinion Topic Ming and Orientation Identification[J]. Computer science, 2012,39(6): 159-162.
[7] Zhang Jianhua, Liang Zhengyou. A Sentiment Analysis Method Based on Sentiment Words Extractin and LDA Feature Representation[J]. Computer and Modernization, 2014(5): 79-83.