A Comment Analysis Method Based on Support Vector Machine and Topic Model

Wang Peng*

School of Economics and Management, Dalian University, No.10, Xuefu Avenue, Economic & Technical Development Zone, Dalian, Liaoning, China

Abstract: In mobile applications, user reviews are an important user feedback path. Users may mention some problems in the use of mobile applications, such as system compatibility issues, application crashes, etc. With the popularity of mobile applications, users provide a large number of unstructured feedback comments. In order to extract effective information from users' complaints, a feedback analysis method based on SVM and LDA (Research) based on support vector machine and topic model is proposed to help developers better. Learn user feedback faster. First extract features from the middle and bad reviews of the mobile app, then use the support vector machine to classify the comments into multiple tags. Then use the LDA theme model (latent dirichlet allocation) to topic the comments under each question type. Extract and delegate sentence extraction. Crawl 5 141 user original comments from two mobile applications, and process these comments with the RASL method and the ASUM method respectively to get two new texts. Compared with the classic method ASUM, RASL The method is less confusing, understandable, contains more complete original comment information, and has less redundant information.

Keywords: User review, Classification, Subject analysis

1. Introduction

With the rise of the mobile Internet, a large number of online commentary information for mobile applications has been generated. The mobile application has a wide user base, user feedback is abundant, and it is updated rapidly with version iteration. Especially the user's evaluation and difference for mobile applications. Comment (referred to as the bad review), is to collect user questions

Important data sources. Existing application distribution platforms support users to comment on applications. Compared with 360 Mobile Assistant, it is a mobile application platform with a large domestic market share, providing a series of services such as application uninstallation, installation, upgrade and comment. Some empirical studies have shown that user reviews contain valuable information such as bug reports, functional requirements, and user experience [1]. For developers, user reviews in the app market can help them better understand user feedback. Improve the quality of software. With the popularity of mobile applications, the number of user reviews is huge, and it is unstructured. Manual inspection is time-consuming and inefficient. Therefore, information mining is required to deal with the user's bad reviews, so that users complain. The core content of the information is visually presented to the developer, allowing developers to update the software more quickly and more interactively. In order to analyze user comments, existing research mainly uses classification or main extraction methods [1].

The first is to categorize user reviews. Panichella et al. found that the best combination of text analysis, natural language processing, and sentiment analysis yielded the best classification results; Maalej et al. tried a variety of techniques to comment on users. Through processing and classification, it is found through experiments that multiple binary classifiers are superior to single multivariate classifiers; McIlroy et al. compared several machine learning classifiers, and finally used support vector machines for classification.

Second, some studies use topic extraction methods to analyze comments. Galvis et al. used the ASUM model in the opinion mining domain for user comments in software applications to automatically extract the topics contained in the comments; Jiang Yan proposed a problem-oriented mining problem. The domain's associated LDA model is applied to users' online comments. The above research work simply considers the classification or topic extraction methods, and does not combine the two methods to analyze the comments. We use two comments as an example to illustrate the difference between classification and topic analysis: (1) “The button in the page does not respond”; (2) “The button should be added to the page”. These two comments are handled in a classified manner and will be divided into two categories: (1) software error; (2) Request to add functions. The theme analysis method will be used to extract the theme of “page” and “button”. It can be found that the classification method can understand the types of problems encountered by users, but it is difficult to get the pieces in the comments. Feature; subject analysis method can get feature information, but it is difficult to distinguish the user's intention. If you combine the classification with the topic analysis, then you can get the comment by classification. Type problem, but the software feature obtained by the comments relating specifically to mining. For a review of the previously mentioned two examples.

2. Comment Analysis Method Based on Support Vector Machine and Topic Model Rasl

2.1 Category Type of User Comments

In order to classify the comments, it is necessary to determine the type of user comments. The extracted user comments are manually annotated according to the method proposed by Seaman et al., and the types of questions included in the comments are analyzed. The analysis process is as follows.

First, select the set of problem types defined by McIlroy et al. as the starting set. For each comment, manually check and mark the type of question indicated by the comment.

If the question in the comment is not included in the question type set, set a new question type and add it to the question type set. Then, restart the labeling process based on the new problem type set. This process is performed by 3 people in parallel. After the completion of this process, the results of the three people are compared. The reliability of the labeling result is measured by the intra-class correlation coefficient (ICC). ICC is an inferred statistic, which describes the degree of similarity of elements in the same group can be used to assess the consistency or repeatability of different quantitative measurements performed by different observers. If the ICC is less than 0.4, the similarity is poor; if the ICC is between 0.40 and 0.59, indicating similarity; if the ICC is between 0.60 and 0.74, the similarity is better; if the ICC is between 0.75 and 1.00, the similarity is good. We calculate the ICC between the labeled results. To measure the reliability of manual labeling. For each question type, ICC is better or better, and found that the results of independent labeling are not particularly different. Then discuss and eliminate the difference The ICC of all problem types is 1, that is, there is no difference [2].

2.2 User Review Classification Method

In order to automatically mark the types of questions that new user comments belong to, this paper uses machine learning methods to classify user comments. This classification method includes feature extraction part and model construction part: the goal of feature extraction part is to extract the characteristics of comment text. It is converted into a form available for the classification model; then, the support vector machine is used to construct a multi-label classification model for user comments. At the same time, in order to mitigate the impact of unbalanced data, a cost-sensitive learning method is used

3. Supervised Entity Relationship Extraction Method Based on Deep Learning

3.1 Supervised Entity Relationship Extraction Framework Evolution Process

In the feature extraction phase, the goal is to extract the features of the comment text. Since the text is unstructured data, it must first be converted into a computer parsable form. The Vector Space Model (VSM) is a text suitable for large-scale literature. Represents a model in which the text space is recognized a vector space consisting of a set of orthogonal feature vectors. Each dimension of the vector corresponds to a feature in the text, and each dimension itself represents the weight of the corresponding feature in the text. Using the vector space model to describe the text data requires determining the text. Characteristics and weights. For English text, a word is a feature. Chinese text needs to be segmented first. The Jieba participle is a Python word segmentation tool. This article uses it to segment words and delete numbers and non-Chinese characters, but stop words. Need to be retained, because some of them can help determine the type of problem, such as “don't.” Refer to existing work, filter out words that appear less than 3 times, remove misspelled or unimportant words, and reduce the complexity of the classification. For the weight of features, the tf-idf algorithm is a common method for calculating weights. Its main idea is: if a word or phrase appears multiple times in a document and is rare in other documents, then the word or Phrases are considered to have good classification capabilities. For example, the word “installation” will appear under the category of content issues, but it is rarely used in resources, etc. He appeared under the classification, so we can “install” the word as one classification. In order to build the feature vectors, as used herein, String to Word Vector filter, which is an implementation of WEKA tf-idf algorithm [3].

3.2 The Generation of Subject Words and Representative Sentences

In this section, the topic of statistical extraction is performed on the classification result, and the topic and representative sentence are further generated through the obtained topic. After classification using the classification model based on support vector machine, after obtaining the classification result, it is necessary to determine the total number of topics to be extracted (for example Expect to extract Y topics per X comments, then calculate the total number of topics extracted based on the number of user comments.) Since the “Other” type contains useless comments, the topic is not extracted for the type. The remaining question types are based on the results of the classification. The proportion of each of them (excluding the “other” type) is calculated by the number of topics corresponding to each question type, and then the topic analysis of the comments under each question type is performed. Here, the determination of the number of topics under the multi-category is illustrated, for example, the total number of topics is M, then the subject of each category is the proportion of the number of comments under the category to the total number. Specifically, if the “content problem” proportion is Ratio, then the number of topics classified as content issues is M×Ratio.

In machine learning and natural language processing, the topic model is a statistical model used to discover abstract “themes” in a collection of documents. It is a text mining tool that is often used to find hidden semantic structures in text bodies. Comments are related to certain topics, so specific words will also appear in the comments of different topics. This paper uses LDA (latent dirichletallocation) model to generate the keywords and representative words. This model is a typical word bag model, that is, a comment is composed of a group of words, does not consider the order of the words, thus simplifying the complexity of the semantic association problem. The LDA model contains the generation of the theme, the selection of keywords according to the threshold, and the generation of the sentence.

4. Experimental Verification

4.1 Test Subject

In this paper, the RASL method uses the LDA topic model to extract the subject words and representative sentences based on the support vector machine classification algorithm. This paper uses the Jieba word segmentation tool to input the Chinese word segmentation into the classical method ASUM [4-5]. Unlike the RASL method, the ASUM method is a combination. This method treats sentences as documents. Each word in the sentence is the distribution of implicit topics, and then the topic mining. On this basis, the topic features and emotional information are combined to analyze the user's views on these topics. Preference, and the subject word, representative sentence as the output. In this section we compare the method RASL and ASUM method proposed in this paper. 360 mobile assistant is a domestic market share application platform, providing mobile applications a series of services such as uninstall, installation, upgrade and evaluation. This article randomly selects a high-scoring application (score 9 or above) and a low-rated application (score 6 or less) from 360 mobile assistants. These two applications are today. Headlines and 360 cloud discs, all of them are collected in the difference. 360 cloud discs have a total of 3950, and today's headlines have a total of 1191 Articles. These 5141 pieces of data are used as the original user comment information data, and the results obtained by text preprocessing are processed by the ASUM method and the RASL method respectively to obtain the data required for the experiment. According to the statistical calculation, the 5141 we crawled. Among the data, only 0.027% of the comments consisted of consecutive paragraphs, and 99.973% of the comments consisted of a single paragraph, so this article did not consider the comment segmentation problem, and divided the comments into multiple comments as a single comment.

4.2 Test Subject

In the questionnaire survey of the topic analysis method RASL, two possible methods were compared: one is to separate the two groups and evaluate the ASUM method and the RASL method separately, but in the case that there are not enough people, it is difficult to eliminate different judgment standards of different people. Problem; therefore, this article randomly divides subjects into in the two groups, each subject reads the two methods. Although there are two methods that can be scored under the same standard, it will bring about the reading order. Therefore, this paper uses the order of change to reduce the difference between the two methods. Interactions. In addition, ASUM and RASL produce results for 200 topics. The number of topics is large, and it is difficult for participants to score each topic in detail. In the future, we will try to contact more participants to participate in the questionnaire to reduce the reading order. In the end, the research in this paper did not add praise, which may lead to some omissions, because there may be some complaints from users about the application in the praise. However, even if there are some complaints in the praise, it will not affect the usability of this method. In the future, we will conduct more experiments to improve the research on all comments.

5. Conclusions

Based on the classification results of the classification model of support vector machine, the method determines the number of topics for each problem type according to the proportion of each problem type in the classification result. Then select the LDA model for topic analysis and use LDA model for each problem. The comments under the type are subject extraction and representative sentence extraction. Then, the design experiment compares the ASUM method to estimate the RASL method. Firstly, the confusion of the number of different topics in the two methods is calculated, and the result is that the RASL method has significantly reduced the confusion. Then Evaluation using a questionnaire, the experimental data is the department of the two applications

In the middle of the evaluation, the software engineering subjects were invited to evaluate the results of the original review, ASUM topic analysis and RASL topic analysis methods. The experimental results show that the RASL method is better understandable and complete than ASUM., contains less redundant information.

Acknowledgement

This article was specially funded by Dalian University's 2019 Ph.D. Startup Fund (20182QL001) and 2019 Jinpu New District Science and Technology Project.

References

[1] Zhang L. (2015) Analyzing helpfulness of online reviews for user requirements elicitation, Chinese

[2] Journal of Computers, 36, 119-131.

[3] Blei DM. (2008) Chinese entity relationship extraction based on syntactic and semantic features: IEEE Press, 8, 69-73.

[4] Ratnaweera A. (2004) Self-organizing hierarchical particle swarm optimizer with time-varying acceleration.

[5] Coefficients. IEEE Transactions on Evolutionary Computation, 6, 712-731.