Intelligent Acquisition and Analysis of Commodity Review Data

Lianwu Deng, Jiaxin Song, Zejun Zhang, Bochu Deng, Songlin Cao*
Hunan University of Finance and Economics, Changsha 410205, China
1757526023@qq.com*

Abstract. When shopping on online e-commerce platforms as they can only contact products through photos and other means, it is easy to have cognitive deviations on the products on online platforms. This paper makes intelligent acquisition and analysis of commodity review data by using web crawler tools. Through a series of operations such as comment data capture, data preprocessing, Chinese word segmentation, emotional tendency analysis, LDA model analysis, it analyzes buyers’ emotional tendency of this commodity to help consumers choose the commodities they really like.

Keywords: Onlineshopping; Commodity reviews; Web crawler; Intelligent acquisition and analysis

1. Introduction
With the development of online shopping today, users attach more and more importance to the online shopping experience. They hope to buy high-quality goods through the Internet platform. However, customers can only determine whether the target goods meet their needs through the details and comments of goods. And in the end, this action has led to a wave of malicious brushing and false reviews, making it even harder for users to judge goods on their own. Firstly, the product review data will have an impact on the potential customers of the product. Secondly, the user experience will be reflected in the review data. This kind of user experience is an important factor to decide whether potential buyers should buy this product. For those who have not experienced this product, it is a prior, which enables them to determine whether this product is suitable for them.
2. Related work

Now some large e-commerce websites have set up statistical labels. For example, Tmall mall highlights the nature of goods by analyzing the emotional orientation in user comments. For example, labels can only indicate the emotional orientation of most people, but they are only the reflection of a statistical collection. For individual users, the emotional orientation of most people does not mean that they’re the same as most people, which will reduce the personalized experience of users. This problem becomes more prominent when the user base becomes larger and larger. The traditional evaluation abstract is the emotional analysis and statistical quantification of argument statements. This generalization is just a simple set of emotional tendencies, which cannot correctly judge the collective orientation of expressing the same set of emotional argument sentences and cannot accurately determine whether it conforms to users[1].

3. Analysis method and process

This article uses the goods comments on the data from the domestic online shopping platform jingdong Ming Chen graphics review data, and comments on the graphics data are collected in the first place. Then the obtained data are data preprocessing, jieba segmentation processing, such as the last through the establishment of mathematical model, expect to get the text value of hidden.

The main process is as follows:

![Flow of data analysis](image)

**Fig.1 Flow of data analysis**

The crawler tool -- the Octopus collector with data acquisition software was used to collect the comment data of jingdong mall video card. The online octopus data collector is more suitable. After the appropriate collection logic is set, a visual window is adopted to create the task and complete the data collection task. New Data comment task of Ming Xuan Graphics Card in Octopus Data collector, input url: https://item.jd.com/11150977711.html?&none

Through the comments on the product details page, the main flow chart of data collection can be designed.
3.1. Preprocessing of comments

After collecting the commodity comment data of Ming Xuan Graphics card in Jingdong Mall, the data should be preprocessed.

The results of data preprocessing have a direct impact on the selection of appropriate text features and the representation of text models. The pre-processing flow of data is mainly shown in the figure below.

3.2. Use LDA model for analysis

3.2.1. Introduction to LDA model

Late Dirichlet Allocation (LDA for short) was first proposed by Blei et al. and consists of three levels: word, subject, and document. The reason why such a word appears in the article is these words belong to a certain topic in a certain probability, which is what the generation model means. Through the LDA model, the common themes selected by these words can be found to analyze the proportion of positive, neutral, and negative emotional attitudes in the critical data[2].

3.2.2. Implementation of LDA model

The LDA model can directly analyze the commodity review data, however, as the positive and negative data of the commodity review data are mixed together, it is not conducive to making
judgments. This article deals with this problem by dividing the review data into two documents: positive and negative. Conduct LDA analysis on "Positive" and "native" documents respectively to find the advantages and disadvantages of the products.

3.2.3. Use the third party library for word segmentation

After the analysis of the LDA model, the comment text is collected into the topic shown in the figure above, and the probability of occurrence is calculated. The following conclusions can be drawn:

(1) In terms of user experience, the video card involved in the experiment has received relatively satisfactory user evaluation. As it can be seen from the word "fluent", the video card makes the computer run smoothly and satisfies the needs of many users.

(2) As it can be seen from the word "satisfied", this video card that participated in the experiment met the expectations of most buyers, who said they were satisfied with the purchase.

(3) It can be concluded from the theme word "game" that the audience of this video card is mainly game lovers, and this video card meets the needs of game lovers. So when looking for a video card, fans can give priority to recommend this video card.

(4) From the keyword "cost performance" can be seen that the performance of this video card and buyer price expectations consistent, highlighting the cost performance of this key, the pursuit of price buyers can recommend this product.

4. Summary and prospect

This paper is about the acquisition and analysis of commodity review data on JINGdong online platform. And obtain product review statistics, so that consumers can easily and quickly determine whether the product meets their expectations. While writing this article, I learned much knowledge about the Python language and the power of its third-party libraries. So I can understand why the language is so popular. I also learned how to use the data collector to collect the data of the web page. The captured data cannot be analyzed directly, and a series of processing such as data removal, cleaning, and word segmentation should also be carried out. Then the data was analyzed for emotional orientation and the LDA model was used to analyze the data. Through the above operation, intelligent segmentation and false comment analysis, and qualitative analysis are obtained. The product review data of a video card on Jingdong Mall is processed effectively, which has certain practical significance for consumers. This method can meet the qualitative judgment of different consumers on different aspects of different commodities and has strong applicability. However, there are still some shortcomings in this experiment. For example, when analyzing the emotional orientation of the comment text with ROSTCM6, the manual detection of the tendency analysis is not very accurate. In addition, due to limited ability in undergraduate study, I have a relatively shallow understanding of LDA model and cannot apply it well.
Acknowledgments
This work was supported in part by the National Natural Science Foundation of China, grant number 72073041. Open Foundation for the University Innovation Platform in the Hunan Province, grant number 18K103; 2011 Collaborative Innovation Center for Development and Utilization of Finance and Economics Big Data Property, Universities of Hunan Province. Open project, grant number 20181901CRP03, 20181901CRP04, 20181901CRP05. 2020 Hunan Provincial Higher Education Teaching Reform Research Project under Grant HNJG-2020-1130, HNJG-2020-1124. 2020 General Project of Hunan Social Science Fund under Grant 20B16.

References
[1] Juntao S, Bingfeng C, and Aihua H 2018 An intelligent acquisition and analysis technology of commodity review data based on Python J. Modern Computer (Pro) 14 16
[2] Huan Z, Yunquan H, Jinzhong L, Zhenming D, and Mansheng L 2018 Research on Network Comment Viewpoint Mining J. Technology Wind 29 218