Mining Product Innovation from Online Reviews

Xirong Gao, Yiming Hou

School of Economics and Management, Chongqing University of Posts and Telecommunications, Nanshan streets, Chongqing, China

a1774010027@qq.com

Abstract

Currently, online reviews are the main way for consumers to interact with information and express their needs. Mining potential user needs from the massive number of online reviews to precisely identify specific product improvement and innovation opportunities helps companies to improve product and service quality in a targeted manner and enhance their competitive advantage in the market. Firstly, we crawl online cell phone reviews and use LDA topic model to perform topic clustering and product feature word mining, so as to identify user demand attributes. Secondly, user satisfaction and the importance of each attribute are calculated using sentiment analysis; finally, the innovation opportunities of each attribute of the product are quantified using the idea of opportunity algorithm. Finally, the feasibility of the method is verified by using cell phone online reviews as an example. The application results show that the product innovation opportunity identification method proposed in this paper can provide a scientific reference basis for enterprises to carry out product innovation activities precisely and efficiently for decision making.

Keywords

Online Reviews; Product Innovation; Data Mining; Opportunity Identification.

1. Introduction

With the increasing popularity of new generation information technology such as Web 2.0 and mobile Internet and the growing consumer awareness of users, the demand for products and services is becoming increasingly diverse. In order to maintain a competitive edge in the ever-changing market environment, it is important for companies to continuously innovate to enhance their core competitiveness. However, the big data can increase diversity and also create information overload problems. Therefore, extracting valuable information from the massive amount of data and turn consumers’ potential needs into product/service innovation opportunities, is the focus of many companies.

The role of consumers in product innovation was first introduced by Eric von Hippel in 1978 (von Hippel,1978). Since then, a large number of scholars have studied the important role of users in product innovation (Chatterji & Fabrizio,2012; Von Hippel & Berg Jensen,2014; Kaiserfeld, 2017). Traditional ways such as questionnaires, market research, in-depth interviews (Zhang et al.,2021) and other methods of acquiring user needs are time-consuming, costly and difficult to obtain samples. To meet these challenges in the Internet age, the company has developed an open innovation community that emphasizes the contribution that users bring to product innovation (Dahlander & Wallin,2006). For example, Huawei’s Pollen Club, Xiaomi’s MIUI Forum, Starbucks’ My Starbucks Idea, Dell’s IdeaStorm, etc. are some of the more representative open innovation communities at home and abroad. While open communities can be great for generating innovative ideas, they may only be applicable to specific products and user groups (Cooper & Edgett,2008).
Online reviews, as a new type of data source, have received a lot of attention from researchers. Online reviews can help companies understand consumer needs, are an important reference for companies to adopt product development and improvement strategies, and are of high value for product innovation. Previous studies on online reviews have focused on the impact on sales (Sonnier et al., 2011; Duan & Gu, 2008; Chevalier & Mayzlin, 2006), the usefulness of online reviews (Mudambi & Schuff, 2010; Li & Zhan, 2011), and user behavior in engaging with online reviews (Dellarocas et al., 2007). Furthermore, although many studies mention the importance of online reviews for product innovation, most of them are theoretical studies (Xu et al., 2016; Qi et al., 2016), Few scholars have used online reviews for product innovation from the perspective of user needs.

Based on above, this paper proposes to use online reviews to explore the innovation opportunities of products/services from the perspective of customers' needs. In this paper, we construct a product/service innovation opportunity mining framework to identify user needs by extracting important product/service attributes that customers care about from online reviews using latent Dirichlet allocation (LDA). Then, the satisfaction and importance of each attribute are calculated using sentiment analysis methods, and finally the innovation opportunities in product/service attributes are quantified by combining the ideas of opportunity algorithms. The combination of text mining and opportunity algorithm can effectively identify the product/service innovation opportunities and provide new ideas for enterprise product/service iterative innovation.

2. Technology-driven Framework

This paper constructed a framework for mining innovation opportunities in online reviews, as shown in Figure 1. The framework consists of three stages, i.e.

- Stage 1. data collection and data preprocessing
- Stage 2. extracting product/service attributes and estimating each attribute's performance and importance
- Stage 3. mining product/service innovation opportunities

![Figure 1. The framework of product/service innovation opportunity mining](image)

3. Methodology

3.1. Data Preparation

Preparation of the data includes data collection and data preprocessing. First, a website with high quality online reviews and reliable content is selected and the start time and end time of the data source is set. Then, a crawler is written to download the online comments. Online reviews are filled with a large number of internet terms that cannot be used directly for analysis. In order to improve the efficiency of extracting product/service attributes in the first place,
online reviews need to be pre-processed. Specifically, the original data is preprocessed by cleaning, segmentation, part of speech tagging, removing stop words, and so on.

3.2. Extracting the Important Attributes of the Product/Service from Online Reviews Using LDA

LDA (Latent Dirichlet Allocation), also known as the document topic generation model, is a three-level Bayesian probabilistic model that contains two levels of relationships and three levels of structure, with the two levels referring to the relationship between documents and topics and the relationship between topics and words. The three-layer structure refers to documents, topics and words, with both documents and topics and topics and words obeying the Dirichlet prior distribution (Blei et al., 2003). The process of using the LDA model to identify product/service quality attributes can be seen as calculating the similarity of each potential word in an online review by repeatedly using the principle that two words obey the Dirichlet distribution and the binomial distribution of the topic and document until the two words satisfy the hyperparameter definition of the fit, at which point the two words are considered to belong to the same topic. Based on the obtained distribution of topics and words, the main themes of online reviews can be inferred, and the label of each important theme can be considered as an attribute of the product/service. This paper writes a Python program based on the LDA model to calculate the similarity in the set of nouns in online reviews to find the product/service quality attributes that customers care about. The Concept of the LDA is shown in Figure 2:

![Diagram of LDA-based topic model](image)

**Figure 2.** Concept of LDA-based topic model

In the LDA model, a document is generated in the following way.

- **Step1:** Generate the topic distribution \( \theta_i \) of document \( i \) by sampling from the Dirichlet distribution \( \alpha \)
- **Step2:** Generate the topic \( Z_{i,j} \) for the \( j \)th word of document \( i \) by sampling from the polynomial distribution \( \theta_i \) of the topic
- **Step3:** Generate the word distribution \( \phi_{Z_{i,j}} \) corresponding to the topic \( Z_{i,j} \) by sampling from the Dirichlet distribution \( \beta \)
- **Step4:** Sampling from the polynomial distribution of words \( \phi_{Z_{i,j}} \) to eventually generate the word \( W_{i,j} \)

3.3. Sentiment Analysis

Different users have different perceptions of product/service quality attributes after experiencing the same service, and the content of the online reviews they post is also inconsistent. Therefore, after determining the product/service quality attributes, it is necessary to classify the importance level given by users for the product/service attributes, which includes users’ satisfaction and attention to the attributes. The determination of user importance consists of two main parts: (1) the establishment of a sentiment dictionary and the method of processing sentiment words; (2) the calculation of user satisfaction and the calculation of importance.
3.4. The Construction of Sentimental Dictionary

In this paper, the HowNet sentiment dictionary is used, and the steps are to merge the positive sentiment words in HowNet with related evaluation words to form a positive base sentiment dictionary, and to merge the negative sentiment words in HowNet with related evaluation words to form a negative base sentiment dictionary.

With regard to the degree words collected by Hownet, there are five intensity levels. The weights are as follows: 2.5 (words such as extreme), 2 (words such as very), 1.5 (words such as more), 1 (no degree words), and 0.5 (words such as insufficiently). Here, 1 indicates the absence of any degree adverbs, greater than 1 indicates enhanced sentiments, and less than 1 indicates weakening sentiments.

3.5. Product/Service Attribute Satisfaction Calculation and Importance Calculation

In online review data, user satisfaction is usually expressed as the strength of the user's sentiment towards the attributes of a product/service. The process of calculating user satisfaction is also the process of calculating the sentiment intensity of the product/service. In the process of calculating sentiment intensity, there are usually two cases, the first is when only the sentiment word and the product/service attribute are present, and the second is when both the sentiment word and the degree adverb are present. Considering these two cases, when only the sentiment word is present, the sentiment intensity score is the weight of the sentiment word; when the sentiment word and the degree adverb are present at the same time, the sentiment intensity score is equal to the score of the product of the sentiment word and the degree adverb weight. The formula for calculating user satisfaction for the attribute is shown in Equation (1):

\[ v_j = \frac{\sum_{n=1}^{N} S(w_i) \times D}{T} \]  

In Equation (1), \( v_j \) denotes the sentiment value of the attribute \( j \); \( S(w_i) \) denotes the sentiment value of the word \( w_i \) belonging to the attribute \( j \) (if it belongs to the positive sentiment, the sentiment value is 1; if it belongs to the negative sentiment, the sentiment value is -1); \( D \) denotes the weight of the degree word; \( N \) denotes the total number of reviews; and \( T \) denotes the number of reviews with sentiment values.

The importance of attribute \( j \) is a measure of how much the user cares about attribute \( j \). The number of comments containing attribute \( j \) in the valid comment data reflects the extent to which users care about the attribute, and the process of calculating the importance of attribute \( j \) is also the process of identifying and counting the feature words and opinion words that embody attribute \( j \) in the text. Therefore, the importance of the user to attribute \( j \), \( I_j \) can be calculated as shown in Equation (2):

\[ I_j = \frac{N_j}{N} \]  

In Equation (2), \( N_j \) denotes the total number of comments containing the feature word or opinion word corresponding to user requirement \( j \) in the comment data; \( N \) denotes the total number of valid comments.

3.6. Quantification of Product Innovation Opportunity

According to the concept of opportunity algorithm, the innovation opportunity of the product is measured by user attention and customer satisfaction, and the improvement degree of the mobile phone is quantified. For service defects, the defect degree is used to replace customer satisfaction, and the change of the value is consistent, that is, the higher the defect degree value is, the better the defect performance is. It is measured by the attribute importance and defect degree of the target enterprise itself. Based on the attribute importance of the target enterprise,
when the defect degree score is greater than the importance, it means that the defect is performing well and the original importance of the attribute is maintained; When the low score of defect degree is less than the importance, it means that the defect is more important, but the customer feedback is poor. Use the difference to increase the opportunity for improvement. The calculation formula is shown in Equation (3):

\[
opportunity_j = I'_j + \max[I'_j - v'_j, 0]
\]

(3)

In Equation (3), \(\opportunity_j\) denotes the innovation opportunity of service attribute \(j\); \(I'_j\) denotes the normalized importance value of service attribute \(j\), see Equation (4); and \(v'_j\) denotes the normalized satisfaction service attribute \(j\), see Equation (5):

\[
I'_j = \frac{I_j - I_{\min}}{I_{\max} - I_{\min}}
\]

(4)

\[
v'_j = \frac{v_j - v_{\min}}{v_{\max} - v_{\min}}
\]

(5)

When the score is greater than 1.5, it represents an extreme opportunity that cannot be ignored; when the score is between 1.2 and 1.5, it is described as a 'low-hanging-fruit', which means that it is time to improve and innovate; when the score is between 1 and 1.2, it is still worth considering innovation, especially in a mature market; when the score is less than 1 point, it means that in most markets, these service attributes are not attractive to customers, and the innovation benefits are less.

4. Case Study

4.1. Data Collection

We selected the Huawei mate 40 5G phone as the research object and used Octopus Collector to collect online reviews on the Jingdong platform. The data was collected from March 2021 to December 2021 and a total of 22,160 raw online reviews were collected. Using the product review collection template in the Octopus Collector, the completed data was saved in Excel file.

4.2. Data Cleaning and Preprocessing

In order to reduce the interference of irrelevant data to the experiment, some noisy data need to be removed. By observing the online reviews of the Huawei mate 40 5G mobile phone, the data to be filtered included duplicate reviews and short texts that did not contain product/service attribute words. A total of 19,460 valid reviews were eventually obtained. After the previous step of data cleaning, the next step is to pre-process the data. The first step in the pre-processing process is word separation, and the jieba library in Python was chosen for this experiment. The deactivated word list used in this experiment is a combination of the deactivated word list from HIT, the deactivated word list from the Machine Intelligence Laboratory of Sichuan University, and the deactivated word list from Baidu, with some additional deactivated words added to take into account the characteristics of online reviews of mobile phone products.

4.3. Extracting Product/Service Attributes Using LDA

Based on the results of word segmentation, nouns, verbs and gerunds are extracted from the valid reviews as candidate attribute words. Then, LDA is used to model candidate attribute words. When setting the parameters, three enterprise experts are invited to determine and adjust the number of topics. Based on experience, the number of topics is initially set to 5-20. With the increasing number of topics, the attributes of product/service that can be reflected are also increasing. Until the attributes of product/service are overlapped and the attribute words begin to disperse, the number of topics is basically determined. Finally, three experts made the
judgment and summarized the product/service attributes by domain knowledge. Seven attributes have been identified, as shown in Table 1. The ‘word frequency’ indicates the number of keywords included in each service attribute; and the ‘number of reviews’ indicates the number of reviews belonging to the service attribute.

<table>
<thead>
<tr>
<th>Code</th>
<th>Attributes</th>
<th>Keywords</th>
<th>Word Frequency</th>
<th>Number of reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fee</td>
<td>free, price...</td>
<td>6</td>
<td>583</td>
</tr>
<tr>
<td>2</td>
<td>System</td>
<td>5G, Android...</td>
<td>7</td>
<td>730</td>
</tr>
<tr>
<td>3</td>
<td>Service attitude</td>
<td>attitude, services attitude...</td>
<td>9</td>
<td>1560</td>
</tr>
<tr>
<td>4</td>
<td>Function</td>
<td>pixel, Sound effect...</td>
<td>8</td>
<td>1360</td>
</tr>
<tr>
<td>5</td>
<td>Hardware</td>
<td>Screen, fingerprint...</td>
<td>5</td>
<td>983</td>
</tr>
<tr>
<td>6</td>
<td>Appearance</td>
<td>color, size...</td>
<td>3</td>
<td>550</td>
</tr>
<tr>
<td>7</td>
<td>Additional products</td>
<td>gift, earphone...</td>
<td>4</td>
<td>420</td>
</tr>
</tbody>
</table>

5. Mining Product Innovation Opportunity

5.1. Assessing Customer Satisfaction and Importance of Attributes

The importance and satisfaction were evaluated using Equation (1) and Equation (2). The results are shown in Table 2.

From the evaluation results of importance, it can be seen that customers generally pay attention to ‘Service attitude’ (3), ‘Function’ (4) and ‘Hardware’ (5), among which ‘Service attitude’ (3) is the most important. In addition, ‘Function’ (4) and ‘Hardware’ (5) follow closely. It can be seen from the top three attributes of importance ranking that customers are more concerned about the services with which they have direct communication and interaction. Understandably, the customers are concerned about ‘Function’ (4). Because customers pay more attention to the use experience of product functions when purchasing electronic products.

From the evaluation results of satisfaction, it can be seen that the overall sentiment polarity is positive.

<table>
<thead>
<tr>
<th>Code</th>
<th>Attributes</th>
<th>Importance</th>
<th>Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fee</td>
<td>0.025</td>
<td>0.720</td>
</tr>
<tr>
<td>2</td>
<td>System</td>
<td>0.031</td>
<td>0.793</td>
</tr>
<tr>
<td>3</td>
<td>Service attitude</td>
<td>0.072</td>
<td>0.439</td>
</tr>
<tr>
<td>4</td>
<td>Function</td>
<td>0.069</td>
<td>0.380</td>
</tr>
<tr>
<td>5</td>
<td>Hardware</td>
<td>0.057</td>
<td>0.464</td>
</tr>
<tr>
<td>6</td>
<td>Appearance</td>
<td>0.023</td>
<td>0.727</td>
</tr>
<tr>
<td>7</td>
<td>Additional products</td>
<td>0.019</td>
<td>0.225</td>
</tr>
</tbody>
</table>

5.2. Quantification of Innovation Opportunities

By using Equation (4) and Equation (5), the importance and satisfaction are normalized, and then the innovation opportunities of express service attributes are calculated according to Equation (3), as shown in Table 3.

According to the opportunity interval, ‘Function’ (4) was classified as extreme innovation opportunities that could not be ignored.
The importance of this attributes is very high, but the customer satisfaction is very low. At present, the functions of Huawei Mate 40 5G mobile phone cannot meet the needs of consumers. Therefore, this is the biggest attribute of enterprise innovation opportunities.

In addition, ‘Service attitude’ (3) and ‘Hardware’ (5) are classified as worthy of consideration. All the remaining four attributes scored less than 1 point for innovation opportunities; that is, they were classified as unattractive innovations.

<table>
<thead>
<tr>
<th>Code</th>
<th>Attributes</th>
<th>$I_j$</th>
<th>$v_j$</th>
<th>Opportunity</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fee</td>
<td>0.042</td>
<td>0.652</td>
<td>0.042</td>
<td>Unattractive</td>
</tr>
<tr>
<td>2</td>
<td>System</td>
<td>0.203</td>
<td>0.852</td>
<td>0.203</td>
<td>Unattractive</td>
</tr>
<tr>
<td>3</td>
<td>Service attitude</td>
<td>0.736</td>
<td>0.467</td>
<td>1.005</td>
<td>Worth of Consideration</td>
</tr>
<tr>
<td>4</td>
<td>Function</td>
<td>1.13</td>
<td>0.679</td>
<td>1.581</td>
<td>Extreme Opportunity</td>
</tr>
<tr>
<td>5</td>
<td>Hardware</td>
<td>0.893</td>
<td>0.767</td>
<td>1.019</td>
<td>Worth of Consideration</td>
</tr>
<tr>
<td>6</td>
<td>Appearance</td>
<td>0.764</td>
<td>0.787</td>
<td>0.764</td>
<td>Unattractive</td>
</tr>
<tr>
<td>7</td>
<td>Additional products</td>
<td>0.301</td>
<td>0.091</td>
<td>0.511</td>
<td>Unattractive</td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper, we propose a method to identify product innovation opportunities based on online reviews, and empirically test the feasibility of the method. First, online reviews are collected using a data collection tool. Second, the LDA topic model is used to identify the topics of online reviews and extract the main topics and the feature words under each topic. Finally, user satisfaction and the importance of each topic are calculated using sentiment analysis techniques, and the innovation opportunities of each topic are quantified using the idea of opportunity algorithm.

The research results found that the key elements in the text data can be better extracted when the LDA topic model is used for topic identification of online reviews, and the keywords of online reviews mainly revolve around the categories of system, function, application software, hardware and accessories of cell phones. ‘Function’ attribute that belong to extreme innovation opportunities has been dug out.

The contributions of this paper are as follows: proposed a method of identifying product innovation opportunities from online reviews. Combining natural language processing, LDA topic model and other text mining technologies, we use the idea of opportunity algorithm to quantify and sort product innovation opportunities, and take Huawei mobile phone online review as the data source to identify feasible product innovation opportunities and give corresponding improvement strategies.

The limitations and future work of this paper are as follows. (1) Online reviews are used as the data source, but many social media platforms have false reviews such as malicious comments from competitors, which will cause some errors to the data analysis results. (2) A general approach is used to evaluate the importance of service attributes, but there are still some scholars exploring more complex methods. In the future, we can consider the importance evaluation of service attributes from the two directions of enterprises and customers, and explore more accurate methods for importance evaluation.

References


