Topic Extraction Based on LDA and Its Application in Tourism

Hui Peng[1,2], Jiapei Huang[1,2], Xi Li[1,2], Danyang Dong[1,2], Peiying Fan[1,2]

[1]*Tourism science school of Beijing international studies university, Beijing, China;*
[2]*Research center for Beijing tourism development, Beijing, China;*

Abstract

This paper introduces LDA, an algorithm that automatically extracts text topics from a large amount of text, and presents a case study of its application: extracting the features of recommendation information of travel microblog key opinion leaders. Using these features to construct a travel decision influence model and analyzing the influence of travel microblog key opinion leaders' recommendation information on travellers' travel decisions. The following conclusions were drawn: the information recommended by travel microblog key opinion leaders provides a certain reference role for travellers' decision-making, and among the six features of travel microblog key opinion leaders' recommended information, the degree of quantification of recommended information is the most important factor that has an impact on travel decision-making.

Keywords

LDA Model, Topics Extraction, Text Mining, Recommended Information of Microblog Opinion Leader, travellers' decision-making

1 Introduction

In the era of information explosion, in order to obtain effective information from massive texts, we need to automatically classify, cluster and extract topics from texts. LDA is a method that uses the probabilistic production model to model the implied topics of text. The basic idea is to assume that there are several independent implied topics in the corpus. According to the probability distribution of these topics, all words in each document of the corpus can be generated, so that the document can be understood as the distribution of specific implied topics. At present, LDA model is widely used in topics mining, text retrieval, text classification, citation analysis and social network analysis. This paper introduces the principle of LDA model, applies it to tourism text data processing, analyzes the recommendation information of key opinion leaders of travel microblog, extracts the features of recommendation information of key opinion leaders of travel microblog with Python of LDA model so that further analysis of relevant issues in the tourism field can be carried out.

2 The LDA model and its Python implementation

2.1 Model LDA introduction

The Latent Dirichlet Allocation (LDA) was proposed by Blei etc. in 2003[1]. It is a three-layer Bayesian probabilistic generation model which contains a three-layer structure of documents, topics, and words[2]. As an unsupervised type of machine learning, the LDA model consists of two main steps which are word generation and topic generation. After determining the number of topics K during training, running the model yields the probability of the distribution of words under each topic and the probability of the topic corresponding to the document.

The modeling process of LDA for text was shown in Figure 1, where the circle indicates the potential variables. The arrow with direction can indicate the relationship between two variables, and the

AIoTC2022@International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology
EMAIL: *Corresponding author's email: penghui@bisu.edu.cn (Hui Peng)

rectangular box means repeated sampling was performed. The specific steps of the LDA topic model are as follows:

$\vec{\alpha}$ and $\vec{\beta}$ are the prior parameters of the Dirichlet function, $\vec{\theta}_m$ is the parameter of the multinomial distribution of the topic in the document, and $\vec{\phi}_k$ is the parameter of the multinomial distribution of the word in the topic, which obey the Dirichlet distribution with hyperparameters $\vec{\alpha}$ and $\vec{\beta}$, respectively[3]:

$$\vec{\theta}_m = Dirichlet(\vec{\alpha}) \tag{1}$$

$$\vec{\phi}_k = Dirichlet(\vec{\beta}) \tag{2}$$

M represents the total number of documents, N represents the number of feature words contained in the documents. According to the topic distribution $\vec{\theta}_m$, for the nth word in any document m, the distribution of its topics $Z_{m,n}$ is obtained as follows:

$$Z_{m,n} = Multionmial(\ \vec{\theta}_m)\ ^{[4]} \tag{3}$$

By combining the topic $Z_{m,n}$and the distribution of words $\vec{\phi}_k$, the distribution of specific words $W_{m,n}$ is obtained as follows:

$$W_{m,n} = Multionmial(\ \vec{\phi}_{Z_{mn}})\ ^{[4]} \tag{4}$$

To move in cycle, a document containing N words was obtained by cycling. Finally, M documents under K topics were generated.

To extract topics using the LDA topic model for travel microblog key opinion leaders' recommendation information, it is necessary to determine the optimal number of topics to extract. The representative method is to measure topic consistency or perplexity. The consistency is used to measure the coherence of words within the same topic, the higher the value of consistency index means that the words within the same topic have strong coherence, the better the model fit; while the perplexity is the degree of uncertainty whether the topic belongs to the document or not. Perplexity is the most common evaluation metric in natural language processing[5], which is used to test the trained language model. The smaller the perplexity is, the relatively stronger the generalization ability of the topic. The specific formula for the perplexity is as follows[6]:

$$\text{Perplexity}(M) = \exp\left(-\frac{\sum \log\ p(n)}{\sum_{m=1}^{M} N_m}\right) \tag{5}$$

$$p(n) = p(k|m) * p(n|k) \tag{6}$$

$p(n)$ represents the probability of each word in the test set, and $\sum_{m=1}^{M} N_m$ denotes the sum of all feature words. $p(k|m)$ means the probability of topic k in a given document, and $p(n|k)$ means the probability of each word under a given topic.

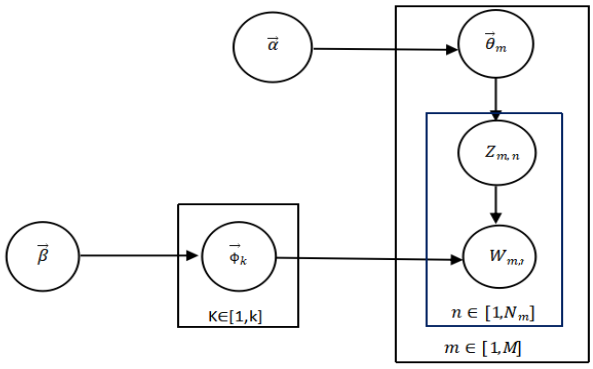

Figure 1 Model LDA structure and its workflow

2.2 Python Implementation of LDA

In the application of LDA model, we need to download and preprocess the text data, then call the modeling function in Python and obtain the relatively ideal number of topics and keywords under

each topic by adjusting the consistency parameter and confusion parameter. The Figure 2 is the main code of the modeling part.

```
#lda modeling
num_topics=7  # Number of topics
lda =  LdaModel(
                corpus=corpus,
                id2word=dictionary,
                alpha='auto',
                eta='auto',
                iterations=300,
                num_topics=num_topics,
                random_state =1
                )
# Save lda topic distribution
topic = []
for i in range(num_topics):
    topic.append(np.array(lda.top_topics(corpus,topn=100)[i][0])[:,1])
    topic.append(np.array(lda.top_topics(corpus,topn=100)[i][0])[:,0])
topic = pd.DataFrame(topic).T
topic.columns = list(itertools.chain(*[['topic{}_word'.format(i),
        \'topic{}_distribution'.format(i)] for i in range(num_topics)]))
topic.to_excel('LDA Model Results.xlsx')
# save topics-document distribution
data_lda = lda.get_document_topics(corpus,minimum_probability=0)
data_lda = pd.DataFrame([dict(data_lda[i]) for i in range(data.shape[0])])
data_lda.columns=['topic{}'.format(i) for i in range(num_topics)]
for i in range(num_topics):
    data['topic{}'.format(i)] = data_lda['topic{}'.format(i)].values
data.to_excel('datalda.xlsx',index=0)
```

Figure 2 The main code of the modeling part

3 Extracting recommended information features from travel microblog key opinion leaders

3.1 Prepare sample data

This study mainly selects the microblog content and user comments published by the current microblog platforms "Top Ten Influential Travel Bloggers in 2020" and "Top Ten Popular Travel Bloggers in 2020" from June 2019 to June 2021 as samples source.

Python was used to capture the content, number of likes and comments, posting time and text of comments of the 20 travel microblog key opinion leaders in the field of tourism during the period of 2019.6-2021.6, and a total of 7,879 pieces of content and 363,779 microblog comments recommended by the 20 travel microblog key opinion leaders in the field of tourism were obtained. The top 100 microblog users' comments and interaction data were collected under each microblog content. To ensure the scientific validity of the research results, the data is screened to eliminate invalid and meaningless comments, and the following processing is carried out on the collected data. It mainly includes the construction of deactivation dictionaries, text splitting, synonym replacement.

3.2 LDA Topic Analysis

3.2.1 Determination of topics for user comment

By continuously changing the value of K, we observe the change of confusion and consistency value, when the value of K is between 3 and 5, the value of confusion is relatively low at this time, as shown in the figure. When the value of K is 4, it can better reflect and cover the meaning of the semantics of visitors' comments, and the consistency between topics is the highest, so the number of topics is set to 4. See Figure 3 and Figure 4.

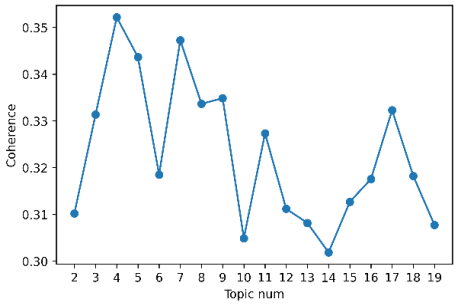

Figure 3 User Review Topic Consistency

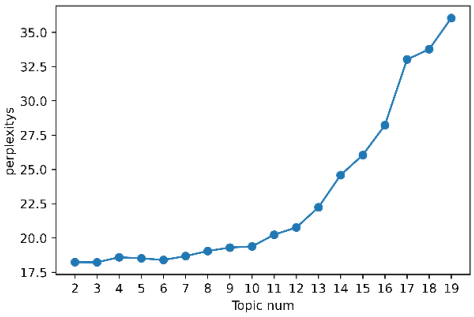

Figure 4 User Comments Topic Confusion

The high-frequency keywords in the four themes coalesced into the recommended information of travel microblog key opinion leaders are listed in Table 1. In Theme 1, words such as "comment", "retweet" and "follow" reflect the interaction between microblog users and opinion leaders, categorizing Theme 1 as: information interactivity. In Theme 2, users comment on words such as "nice", "good", "cute", etc., expressing their praise and compliments on the content recommended by the opinion leaders, with their own feelings, so Theme 2 is categorized as information expression. In Theme 3, users commented on the words "link", "web page", "video" and "image", reflecting the diversity of information presentation forms used by opinion leaders in recommending content, and users' expectation and demand for diverse information presentation forms. Therefore, Theme 3 is categorized as: Information presentation formats. In Theme 4, users comment on words such as "raffle" and "prize" as a form of interaction between opinion leaders and their followers through rewarding activities such as raffles, and words such as "envy" and "rule" as a form of interaction between travel microblog key opinion leaders and their followers. Therefore, Theme 4 is also categorized as information interactivity.

Table 1 Travel microblog key opinion leaders recommend user comments under the message theme keywords

Topics	High-frequency Keywords	User comment theme qualities
1	Comment, live, retweet, share, etc.	Information interactivity
2	Good-looking, nice, like, happy, etc.	Emotional expressions, information expressions
3	Links, web pages, accompanying images, etc.	Information presentation format
4	Sweepstakes, prizes, details, links, etc.	Incentive mechanism, interactivity

3.2.2 Identifying the qualities of travel microblog key opinion leaders ' recommended messages

Combining the perceived characteristics of the information recommended by the travel microblog key opinion leaders as reflected in the above microblog user comments, as well as the data of the likes and comments of the 20 t microblog travel key opinion leaders and the topics to which the highly interactive content belongs. The two were compiled and compared to arrive at the following characteristics of the travel microblog key opinion leaders' recommended information, and then study their influence on travelers' destination decisions.

（1）The quantitative degree of information recommended by travel microblog key opinion leaders: the comprehensive degree of the number of retweets, likes, comments, etc. of the recommended content.

（2）Information quality: including the accuracy, completeness and interest of the description of the tourist destination, tourist products or services, etc.

（3）Information timeliness: the frequency of recommended information, whether it is combined with current hotspots, leading the latest developments in the field of tourism, etc.

（4）Information interactivity: travel microblog key opinion leaders recommend information in the process of using questions, @, add topic tag, super talk and other ways to communicate and interact with potential tourists; microblog users interact with each other in the type of comments that occur after a microblog opinion leader releases a microblog.

（5）Form of information presentation: The expressions used by travel microblog key opinion leaders to disseminate information: plain text, long text, combination of pictures and text, video, live broadcast, etc.

（6）Information expression: Objective description of tourism products or services information, the post-purchase experience of tourism products or services released and recommended, adding their own attitude, with a certain emotional color.

4 Model Construction

From the above recommended information features the following tourism decision model can be constructed. It shows in Figure 5.

Based on the model, through questionnaires and hypothesis testing, it is concluded that the information recommended by tourism travel microblog key opinion leaders provides a certain reference role in travelers' decision-making behavior, and among the six characteristics of information recommended by travel microblog key opinion leaders, the quantitative degree of recommended information is the most important factor that has an impact on tourism decision-making.

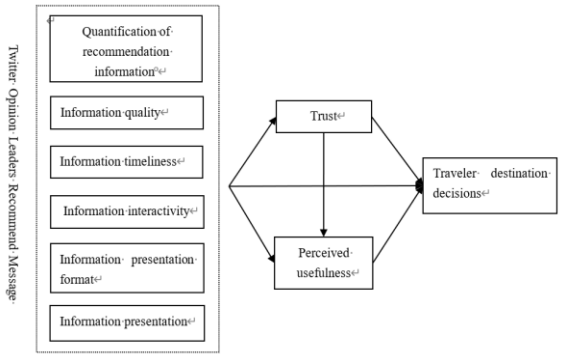

Figure 5 Theoretical model of the influence of microblog opinion leaders' recommendations on travellers' destination decisions

5 Conclusion

In this paper, the LDA model is used to mine the information recommended by travel microblog opinion leaders, and six features of the information are summarized. They are the quantitative degree of information, information quality, information timeliness, inforrmation interactivity, form of information presentation and information expression. Applying these features in the model of traveller‘ decision making, it shows the degree of quantification of recommended information is the most important factor that has an impact on travel decision-making.

6 Related works

Hoffmann[7] proposed the Probabilistic Latent Semantic Indexing (PLSI) model, which uses probabilistic generative models for topic analysis and extraction of text. Blei[1] etc. improved the PLSI model by proposing the LDA (Latent Dirichlet Allocation) model, which is currently the most widely used model in the field of topic modeling research. Xu, Ge and Wang, Houfeng[8] introduced and analyzed the important role of probabilistic implicit semantic indexing and LDA in the development of topic models, and classified and introduced various models derived from LDA. An important discussion of LDA-based text segmentation[9] and topic extraction[10] is provided by Jing Shi etc.

In the field of tourism, LDA models are widely used in research. chao Huang[11] etc. used LDA methods to refine a seasonal theme model to analyse the themes corresponding to each attraction in different seasonal contexts. Zhou Wenliang[12] used LDA to mine textual themes to obtain relatively objective tourism destination evaluation indicators, thus reconstructing the tourism destination evaluation system to evaluate tourist attractions in Jiangxi Province.

7 Acknowledgement

This research was financially supported by science research project of Beijing International Studies University (LYFZ18B003).

8 References

[1] BleiDM, NgAY JordanMI. Latent dirichleta llocation [J]. Journal of machine Learning research, 2003, (1): 993-1022.
[2] Xing Feng, Liu Xingxu. Practice and Application of Machine Learning in data analysis [J]. Telecommunication Engineering Technology & Standardization, 2021,34(12): 82-84 + 88.
[3] Li Tingyi. Study on the factors influencing the travel intention of Guangzhou residents in the context of the normalization of epidemic [D]. Guangxi Normal University.
[4] Zhou Wen-Liang. Research on AHP tourism destination evaluation based on LDA improvement [D]. Jiangxi University of Finance and Economics.
[5] Guan Peng, Wang Yuefen. Research on the Optimal Topic Number of LDA Topic Model in Scientific and technological Information Analysis [J]. Modern Library and Information Technology,2016(9):42-50.
[6] Zhao Zixuan. Emotional Evaluation and Influencing Factors Analysis of museum Tourism [D]. East China Normal University, 2022.
[7] Hofmann T. Probabilistic latent semantic indexing [J]. International ACM SIGIR conference on research and development in information retrieval, 1999, 51(2):50-57.
[8] Xu Ge, Wang Houfeng. Development of Topic Models in Natural Language Processing [J]. Journal of Computers,2011,34(08): 1423-1436.
[9] Shi Jing, Hu Ming, Shi Xin, etc. Text Segmentation Based on LDA Model [J]. Journal of Computers,2008(10): 1865-1873.
[10] Shi Jing, Fan Meng, Li Wanlong. Thematic analysis based on LDA model [J]. Acta automatica sinica,2009,35(12): 1586-1592.
[11] Chao Huang, Qing Wang, Donghui Yang, etc. Topic mining of tourist attractions based on a seasonal context aware LDA model [J]. Intelligent Data Analysis, 2018, 22: 383-405.
[12] Wenliang Zhou. Research on AHP tourism destination evaluation based on LDA improvement [D]. Jiangxi University of Finance and Economics, 2021. doi:10.27175/d.cnki.gjxcu.2021.000448.