Application of LDA Topic Model in E-Mail Subject Classification

Hechen Gong [a], Fucheng You [b], Xinxin Guan [c], Yue Cao [d] and Shuren Lai [e]

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China

[a] gonghechen123@163.com, [b] youfucheng@bigc.edu.cn, [c] 1145512971@qq.com,

[d] yuecao1993@foxmail.com, [e] 513324679@qq.com

Abstract. Text classification is an important research direction in the field of natural language processing. With the development of Internet, the use of e-mail is becoming more and more common. It is very important to quickly understand the subject and content of an e-mail in a mailbox. For example, when the police handle a case and face this demand, the intelligent processing of text information by computer has been extensively studied. This paper proposes an application of mail classification based on LDA topic model, which combines information extraction, information retrieval and natural language processing. SVM classifier is used, and TF-IDF technology, which is a technical evaluation method for the classification of this application, is also proposed. It is believed that LDA has a certain effect in mail classification, and the uncertainty and subjectivity in LDA classification are also proposed.

Keywords: LDA topic model, E-mail, TF-IDF, Text classification.

1. Introduction

Since the 1990s, the rapid development of the Internet has brought tremendous changes to the world, information technology has penetrated into all corners of the world. On the one hand, a large number of information resources have brought us great convenience; on the other hand, facing such a rich digital information resources, most of the information is not what users need. For e-mail, it is difficult to get valuable information quickly and effectively from huge information resources, so we need to classify these resources properly, so as to help us get useful information better [1].

Email is also known as EMILL. Email is a way of transmitting information. Compared with the traditional way of writing letters, email is more convenient, faster and lower cost. According to statistical reports, as of June 2018, the size of e-mail users reached 283.06 million, the user utilization rate reached 55%, an increase of about 6% over the previous year. According to another survey on the most helpful tools at work, 61 percent of respondents thought e-mail was very helpful to their work, followed by the Internet, and only 24 percent thought mobile phones were very useful in their work. With the huge scale of users and diversified application scenarios, e-mail has become one of the most widely used means of information exchange in daily life.

E-mail is a necessary tool for transmitting documents or information to each other in the workplace. At present, e-mail has become an important way of communication between superiors and subordinates, colleagues and colleagues in major companies, schools and government departments. It is very important to quickly understand the contents and themes of mail by categorization. In this paper, automatic mail classification technology is based on LDA probabilistic topic model and SVM classifier. LDA model is used to model text sets, and the latent semantic relations in text sets are mined to build corpus. Finally, the application of LDA thematic model in e-mail subject classification is realized.

Statistical topic model has been widely used in text mining since it was proposed. Based on LSA and p LSA, B lei et al. put forward LDA (Latent Dirichlet Allocation) topic generation model [2]. This model is used as a complete Generative Probabilistic Model [3], by introducing generalizations. The theory of rate statistics is used to describe the whole document generation process. Over the past few years, natural language processing has been emphasized, and LDA topic models have been constantly improved. B Lei [4] et al. have proposed CTM model, mainly to better express the potential

topic information of text sets. Logic-normal distribution is used to replace the Delichlet distribution in LDA model, which can better find the hidden topic between. Correlation information. Li Wenbo [5] proposed the Labeled-LDA model. The innovation of this model is to introduce class identification information into LDA, which reduces the training number of LDA model and achieves good classification results. Shi Jing [6] and others successfully introduced LDA into text cutting and semantic information analysis technology, and achieved good results; Yuan Boqiu [7] and others applied LDA model in feature selection technology, mainly for spam processing. LDA model is also used in this paper, which is mainly used in the mail topic classification, and combined with TF-IDF technology to determine the reliability of topic classification.

2. Related Work

2.1 Data Acquisition.

The data in this article is mainly from the e-mail gate, which contains 7,000 messages that Hillary communicates with others. The e-mail gate is a WikiLeaks (Wikileaks) e-mail that was released online after the computers of Hillary and other important people around her were hacked / spilled. The following diagram shows the communication relationship between people expressed in an email, so in the case of a large number of emails, it is very important to know the subject and content of the email.

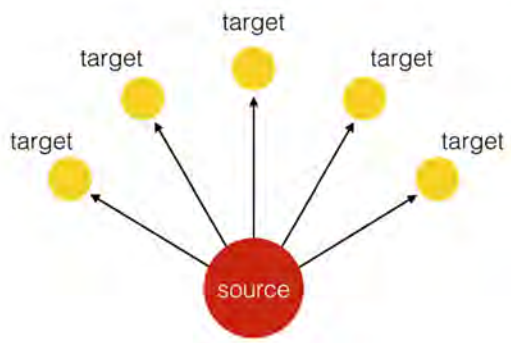

Fig 1. Communication relationship between people expressed by an email

2.2 Text Preprocessing.

In natural language, text preprocessing is very important. Through the author's observation, regular expression technology is used to filter some meaningless characters and numbers such as time, date, website, e-mail address. A regular expression is a logical formula for string manipulation, which consists of a "regular string" of predefined characters and combinations of these specific characters, which is used to express a filtering logic for strings. In this project, we use the re module in Python3 to realize regular matching.

2.3 To Stop the Word.

Stopping words refer to words that have no meaning in the text, such as the word "the," which are almost useless in determining the subject, but account for a large proportion. That is to say, the frequency should not be considered when measuring correlation. In English, stopwords also has "to", "a", "of", "are", "is" and so on. The stop word used in this article is stopwords from the nltk library.

2.4 Construction of LDA Model.

A. a function: Gamma function

B. four distributions: two item distribution, multinomial distribution, beta distribution, Dirichlet distribution

C. a concept and a rationale: conjugate priors and Bayesian frames

D. two models: PLSA, LDA
E. a sampling: Gibbs sampling

3. Related Technology Theory

3.1 LDA Topic Model [8].

3.1.1 Lda Core Function.

As the middle layer, the theme can give the probability of occurrence of word w in the file d by the current $\theta_d$and φ_t. The p(t|d) is calculate by θ_d, and the p(w|t) is calculate by φ_t.

$$P_j(\mathrm{w}_i|d_s) = P(\mathrm{w}_i|t_j) * P(t_j|d_s) \tag{1}$$

Among them: w_i stands for words; d_s for documents; t_j for topics; $\theta_d$for subjects in different documents; and φ_t for words in different topics

3.1.2 Gamma Function

Integer case:

$$\Gamma(n) = (n-1)! \tag{2}$$

Real situation:

$$\Gamma(x) = \int_0^\infty t^{x-1}\, e^{-t} dt \tag{3}$$

3.1.3 Multinomial Distribution.

The probability density function of multinomial distribution is:

$$P(x_1, x_2, \dots, x_k; n, p_1, p_2, \dots, p_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \tag{4}$$

3.1.4 Dirichlet Distribution.

$$f(x_1, x_2, \dots, x_k; \alpha_1, \alpha_2, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha^i - 1} \tag{5}$$

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha^i)}{\Gamma(\sum_{i=1}^{k} \alpha^i)}, \sum x_i = 1 \tag{6}$$

3.1.5 Text Generation Topic Model

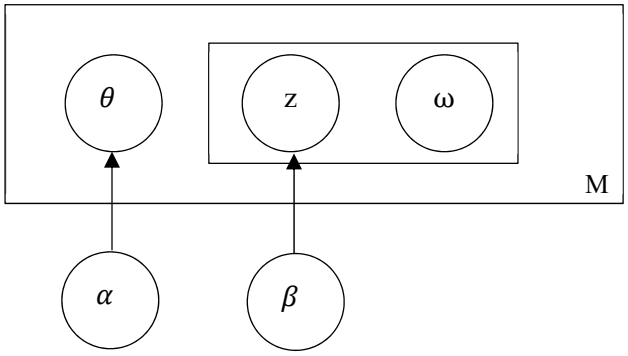

Fig 2. LDA Model diagram

In the fig 2, M represents the number of Chinese texts in corpus, L denotes the length of a text, Z denotes the topic, ω denotes the term, α and β are hyper-parameters, in which β is a $k \times V$ matrix, k is the number of topics, V is the number of words, β_{ij} denotes the probability of the j word under topic i, and θ denotes the probability distribution of the topic in the document. The basic idea of the LDA thematic model is to randomly generate a document of N words,, each term chooses a topic with a certain probability, and selects it from the topic with a certain probability. Given the α and β, the LDA model is represented by a probabilistic model, as shown in formula 7.

$$P(\theta, z, w|\alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} P(z_n|\theta) P(w_n|Z_n, \beta) \tag{7}$$

The probability of the whole corpus is shown in formula 8.

$$P(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) (\prod_{n=1}^{N_d} \sum_{Z_{nd}} p(z_{nd}|\theta_d)\, p(w_{dn}|z_{nd}, \beta)) d\theta_n \tag{8}$$

D represents a collection of documents, N_d denotes the length of the document in chapter d, θ_d denotes the topic probability distribution of the document in chapter d, w_{dn} represents the nth word of the d document, and z_{nd} represents the subject of the nth word in document d.

3.2 TF-IDF [9].

TF-IDF (term frequency–inverse document frequency), It is a commonly used weighting technique used in information retrieval and data mining.

3.2.1 TF（Term Frequency)

Word frequency: the frequency of a specified word or phrase in a given text.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{9}$$

Among them, $n_{i,j}$ denotes the number of occurrences of word w_i in text d_j, and the alphabet denotes the sum of occurrences of all characters in text d_j.

3.2.2 IDF (Inverse Document Frequency)

Inverse document frequency: measure the general importance of words or phrases in text.

$$IDF_i = \log \frac{m_i}{n_i} \tag{10}$$

Of these, m_i represents the total number of documents in the corpus, and the denominator represents the number of documents containing n_i.

That is to say, combine formula (9) and formula (10).

$$\mathrm{TF} - IDF_{i,j} = TF_{i,j} \times IDF_i \tag{11}$$

4. Experimental Process

4.1 Experimental Data.

The data in this paper is mainly from the mail scandal. There are 7000 emails that Hillary communicated with other people, which have been sorted into CSV files, making it convenient to use pandas to invoke data.

4.2 Hardware Environment and Experimental Platform.

The experimental environment is shown in Table 1.

Table 1. Experimental environment

CPU	Intel(R) Core (TM) i5-7200U
Memory	4.00GB
Programing language	Python Language
Edition	Python 3.6.5
IDE	Jupyter Notebook

4.3 Experimental Evaluation.

4.3.1 Keyword Matching.

The evaluation system of this paper is based on TF-IDF technology, because the high-frequency words in a particular file and the low-frequency words in the whole file set can produce high TF-IDF. Suppose that I topic is classified as L, L_1 as the first topic, the top 30 keywords of each topic in L are extracted, and LDA is used. The first five weighted words of each topic classified by the topic model are matched to find the matching accuracy. Then, modify the generated topic to produce L, and match the keywords generated by TF-IDF, repeat it until the matching accuracy reaches the maximum.

4.3.2 Subjectivity Counts Valuable Words.

Because the obtained keywords are not necessarily meaningful in analyzing the content of the e-mail, the proportion of the valuable words is obtained by calculating the value words of the five keywords under each topic generated by the LDA model artificially, and then the number of the generated topic L is modified to obtain the value words proportion, and the calculation is repeated until the proportion is the highest. So far.

By comparing the matching accuracy and the value word ratio obtained by the two methods, the number of topics can be used to judge the effect of LDA topic model in mail topic classification.

5. Experimental Result

The matching accuracy of LDA topic model is changing with the increasing number of topics. The matching accuracy is the highest when the number of topics is 30. The result is shown in Fig 3.

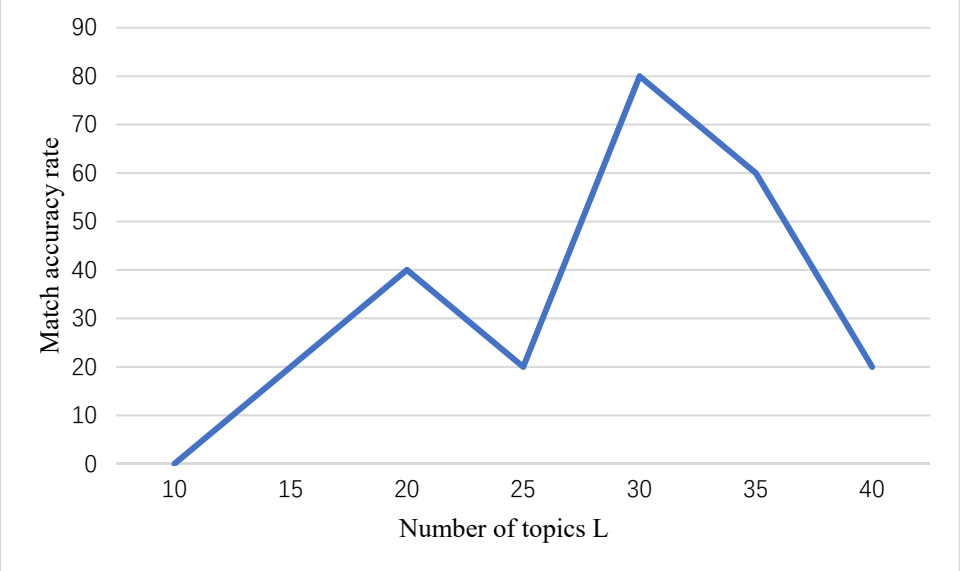

Fig. 3 matching accuracy of LDA topic model under different topic numbers

The proportion of value words in the LDA topic model is changing with the increase of the number of topics. The proportion of value words is the highest when the number of topics is 26. The result is shown in Figure 4.

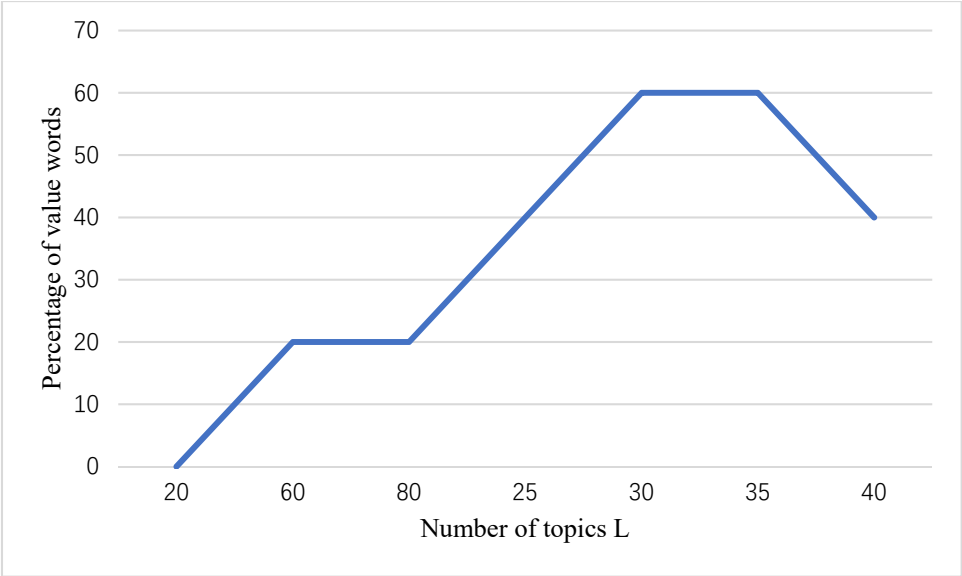

Fig. 4 the ratio of value words under different topic numbers in LDA topic model

As can be seen from the table above, when the number of method topics of LDA topic model reaches 30-35, the matching accuracy and the proportion of value words basically reach the maximum, so the classification results are more ideal.

6. Summary

In this paper, LDA topic model is used, LDA topic model is applied to email topic classification, which solves the problem of reading many emails. In view of the powerful text representation ability and dimensionality reduction effect of LDA model, the corpus is modeled and the topic classification is carried out. Then the accuracy and statistics of keyword matching are obtained by IF-IDF. Value words ratio verifies the reliability of the classification, the results show that the method has a certain effect, can be applied to mail classification. However, the problem of subjective factors in statistic value words and the trial scope of evaluation system will be the next step to be studied.

Acknowledgements

This work was partially supported by Joint Funding Project of Beijing Municipal Commission of Education and Beijing Natural Science Fund Committee (KZ201710015010), Project of National Scientific Found (No.61370188), Project of Beijing Municipal College Improvement Plan (PXM2017_014223_000063) and New Project of Green Printing and Publishing Technology by Cooperative Creating Center (PXM_014223_000025) and BIGC Project (Ec201803 Ed201802 Ea201806).

References

[1]. H. Borko, M. Bernick. Automatic Document Classification. Journal of the ACM. Vol. 11(1964) No.2, p.145-150.

[2]. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. The Journal of Machine Learning Research. Vol. 3(2003) No.1, p. 993-1022.

[3]. Yang Y, Zhang J, Kisiel B. A scalability analysis of classifiers in text categorization. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. Proc. of the 26th ACM Int´1 Conf. on Research and Development in Information Retrieval (SIGIR-03). Toronto: ACM Press, 2003, p. 96-103.

[4]. D. M. Blei, J. D. Lafferty. Correlated topic models. Advances in Neural Information Processing Systems,2006, p. l47-154.

[5]. Li W B, Sun L, Huang R H. A new algorithm for text categorization based on Labeled-LDA model. Journal of Computer Science, Vol. 31(2008), p.620-627.

[6]. Shi J, Li W L. Topic analysis based on LDA model. Automatic newspaper, Vol. 36(2009), p. 1586-1593.

[7]. Yuan Bo qiu, Zhou Yi min, Li Lin. LDA based feature selection for spam filter. Computer Engineering and Applications, Vol. 45(2009) No.25, p.121-124.

[8]. Shi J, Fan M, Li W. Subject analysis based on LDA model. Journal of automation, Vol. 45(2009) No. 12, p.1586-1592.

[9]. Chen S, Jin Z. Weibo topic detection based on improved TF-IDF algorithm. Science & Technology Review, Vol. 34(2016) No. 2, p.282-286.