Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 26 (2023), No. 4, pp. 1207–1212
DOI : 10.47974/JDMSC-1613

Information security topics extraction and classification method based on modified LDA model

Lubna Emad Kadhim *
Saif Aamer Fadhil †
Department of Computer Techniques Engineering
Imam Al-Kadhum College (IKC)
Baghdad
Iraq

Abstract

With the rapid development of social informatization, information plays a pivotal role in people's lives. The popularization of information has also brought serious information security problems, information security incidents become hot topics for the public. However, it is difficult for the researchers to quickly locate information security incidents because the information security topics are overwhelmed by massive news topics. To solve this problem, a modified LDA model using Python is proposed to improve the classification effect and accuracy of information security topics by perfecting the determination index of the number of topics. Simulation results show that the proposed method can obtain better classification performance compared with the traditional text classification method.

Subject Classification: *03C45, 13C05.*

Keywords: *Information security, LDA model, Topics determine indicators, Text classification method.*

1. Introduction

With the continuous development of Internet technology, informatization has penetrated all aspects of people's lives, and the rapid development of informatization has also led to the emergence of various information security incidents, posing a serious threat to people's property security [1]. At the same time, information security incidents are often submerged in massive news topics, making it difficult for researchers

* *E-mail:* `lubnaemad @alkadhum-col.edu.iq (Corresponding Author)`
† *E-mail:* `saifaamer@alkadhum-col.edu.iq`

to quickly It is difficult to provide targeted solutions [2]. Therefore, how effectively extracting information security topics from mass news topics is of positive and important significance for maintaining social stability and public interests [3]. In recent years, scholars at home and abroad have carried out a lot of research on the classification and extraction of information security topics [4]. In this paper, TextRank and latent Dirichlet distribution (LDA) model is used to classify and extract information security topics. However, the LDA model needs to preset parameters in advance, and the parameter selection directly impacts its classification performance.

2. Traditional LDA Model

The Latent Dirichlet Allocation (LDA) model was proposed by DM Blei et al. It analyzes each document's topic distribution in the document set through an unsupervised learning method [5]. The basic idea of the LDA model: First, generate a "topic-word" multinomial distribution, which obeys the Dirichlet prior distribution with a parameter of β, and then generate a "document-topic" multinomial distribution[6]. This multinomial distribution is a Dirichlet prior distribution with a parameter of α, and the specific process is shown in Figure 1 [7]. where α and β are the prior distributions of θ and φ, respectively, m is the topic probability distribution of the mth document, and φk is the vocabulary probability distribution, Zm,n is the topic of the nth word in the mth document, m,n is the nth word in the mth document, Nm is the number of words in the mth, M is the number of documents, K number of topics. The specific implementation of the LDA model is to determine the parameters (α, β). The specific process is as follows [8]:

1) Select the appropriate number of topics K, and initialize the parameter vector α and β;
2) Corresponding to each word of each document in the corpus, The machine is given a subject number Zm,n;
3) Re-scan the corpus, and for each word, use The Gibbs sampling formula to update its topic number and updates the number of the word in the corpus;

The specific implementation of the LDA model is to determine the parameters (α, β). Therefore, an important factor affecting the topic classification effect of the final LDA model is the determination of the optimal number of topics.

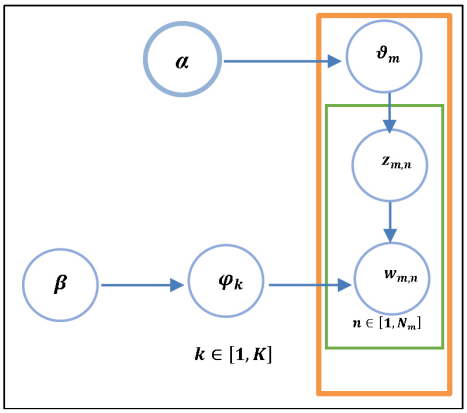

Figure 1

Traditional LDA Model Implementation Diagram

3. Improved LDA Model

At present, the indicators commonly used to measure the optimal number of topics in the LDA model include Perplexity, JS Divergence (Jensen-Shannon Divergence), etc. [9]. However, the Perplexity indicator focuses on the model's ability to predict new documents, which often results in a large number of optimal topics, while the JS divergence focuses on the differences and stability between topics. Although it makes up for the shortcomings of Perplexity, it requires the number of topics obtained is too small [10]. Therefore, this paper proposes an improved optimal topic number determination indicator Perplexity-Aver by combining the advantages of the two to obtain the number of topics more suitable for actual requirements. To measure the overall difference of the subject space, the concept of JS divergence average Aver(T) is first introduced based on JS divergence. The expression is as follows:

$$Aver(T) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} JS(k_{i_{\|}} k_j)}{C_k^2} \tag{1}$$

To sum up, the topic modeling using the improved LDA model is divided into seven steps:

1. Preprocess N documents mixed in various categories, including Include participles and stop words;
2. Input the preprocessed document set into the improved LDA topic model, and input the model parameters K, α, β;

3. Use Gibbs sampling to calculate the θ and φ distributions of the model;
4. Change the value of K, repeat steps 2) and 3), and get different Modeling results under the number of topics;
5. Determine the optimal number of topics K according to Perplexity-Aver;
6. Trim irrelevant vocabulary and output various information security topics with the following keywords;
7. Output topic modeling results.

4. Candidate Topic Set Construction

To conduct topic modeling through the improved LDA model, several categories of information security topics can be obtained, and the subject words under each category of topics can be obtained [11]. The set is called the candidate topic set. However, not all topics with trigger words are information security-related topics. For example, the word "Virus" is used for disease and computer worms.

5. Experimental Data and Preprocessing

This article uses Python programming language to improve the model and the web crawler tool GooSeeker to collect information security incidents reported by domestic mainstream information security websites [12]. The content of the original data set is the news content, title, and release time of the above websites. The news content only retains the text and removes data in formats such as videos and pictures. Since the collected raw data contains junk data, a total of 13,684 pieces of valid data are obtained after the collected data is deduplicated and blank data without text is deleted.

5.1 *Analysis of Experimental Results*

In this paper, after preprocessing the original data for word segmentation and removal of stop words, the improved LDA model is used to model the original data set, and the number of iterations is set to 100. Set to 50/k and 0.01. Table 1 shows in the experiment, the cross-validation method is used to perform 5-fold cross-validation, that is, 1/5 of the experimental data is selected as the test data each time, and the remaining 4/5 is used as the training data, 5 trials are performed, and the results of

the 5 trials are used. The average of precision, recall, and F-value is used as the final precision, recall, and F-value. To verify the effectiveness of the information security event extraction method proposed in this paper, the classification algorithms proposed in the literature [4] and literature [12] are selected for comparison. The literature [4] uses the traditional LDA.

Table 1

The performance comparison of this method and other literature methods

Classification Model	Evaluation standard (unit: %)		
	Accuracy	recall	F value
The method of this paper	84.6	81.7	83.1
Traditional LDA method	82.6	80.0	81.3
Improved BP method	83.2	80.9	82.5

6. Conclusion

This paper proposes an information security event extraction method based on the LDA model and voting mechanism. By using the LDA topic model and the improved topic number determination indicator Perplexity-Aver to model news topics, the information security candidate topic set is obtained, and finally, the voting is used. The mechanism identifies true information security incidents from the set of candidate incidents. The simulation results show that the information security event extraction method proposed in this paper is superior to the traditional text method in terms of accuracy rate, recall rate, and F value, and can obtain better text classification performance.

References

[1] "Proceedings of the 9th International Conference on Computer Engineering and Networks", Springer Science and Business Media LLC (2021).

[2] Zeng Jianbo, Li Baiyang. Exploring the problems and countermeasures of bank information security in the era of network big data [J]. *Cyberspace Security,* 11(9): 90-93 (2020).

[3] Ekin Ekinci, Sevinç İlhan Omurca. "ConceptLDA: Incorporating Babelfy into LDA for aspect extraction", *Journal of Information Science* (2019).

[4] Fan Zhang, Hui Jiang, Minghuan Wu, Jianchun Peng. "Consensus-Based Distributed Optimization of Generation Dispatch of Multi-Area AC Systems Interconnected by DC Lines", IOS Press (2021).

[5] Damir Korenčić, Strahil Ristov, Jan Šnajder. "Document-based topic coherence measures for news media text", *Expert Systems with Applications* (2018).

[6] Graovac J, Kovacevic J, Pavlovic-Lazetic G. Hierarchical vs. flat n-gram-based text categorization: Can we do better? [J]. *Computer Science and Information Systems*, 14(1):103-121 (2017).

[7] Kazuki Komatsu, Yuichi Miyaji, Hideyuki Uehara. "Iterative Nonlinear SelfInterference Cancellation for In-Band FullDuplex Wireless Communications Under Mixer Imbalance and Amplifier Nonlinearity", *IEEE Transactions on Wireless Communications (*2020).

[8] Blei DM, Ng A Y. Jordan M I. Latent Dirichlet Allocation[J]. *Journal of Machine Learning Research*, 3 : 601-608 (2003).

[9] Meyerovitch T. Gibbs and equilibrium measures for some families of subshifts [J]. *Ergodic Theory and Dynamical Systems*, 33(3) : 934-953 (2012).

[10] Maimaiti Ayifu, Wushouer Slamu, Palitan Muhetar, Yang Wenzhong. Uyghur sentiment classification based on LDA and deep neural network [J]. *Computer Simulation*, 36(10) : 194-204 (2019).

[11] Mao Ling. Research on text topic mining based on LDA [D]. Wuhan: Huazhong University of Science and Technology (2018).

[12] R. Li, S. Su, G. Wang, J. Qu, Y. Cao. "Rail transit fault text classification based on the latent dirichlet allocation", *2021 IEEE International Intelligent Transportation Systems Conference* (ITSC) (2021).

[13] Shubham Sharma & Ahmed J. Obaid, Mathematical modelling, analysis and design of fuzzy logic controller for the control of ventilation systems using MATLAB fuzzy logic toolbox, *Journal of Interdisciplinary Mathematics,* 23:4, 843-849 (2020), DOI: 10.1080/09720502.2020.1727611.

Received February, 2022

Revised April, 2022