Analysis of Food Safety Public Opinion Based on LDA Theme Model

Ting zhang[1, a], Yankun Wang[2, b], Cao Yuan[3, c *] and Kaiqiong Sun[4, d]

[1, 3, 4]School of Computer,Wuhan Polytechnic University,Wuhan Hubei, 430023,China

[2] State Key Lab for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China

[a]361554730@qq.com, [b] 15172537265@163.com[c]40176442@qq.com, [d]40473153@qq.com

* please mark the corresponding author with an asterisk

Keywords: Food safety; Theme model; LDA

Abstract. Because food is the material basis for the survival and development of mankind, and food safety is closely related to public health. But with the rapid development of the Internet age,It's very difficult to find information about food safety from huge amounts of big data.The Latent Dirichlet Allocation (LDA) topic model help to the public opinion analysis of the food safety problems we want.

Introduction

With the development of society, the food industry there have been a lot of harmful people food safety behavior.Food safety issues have become increasingly important. Food safety issues in all walks of life have set off a heated discussion[1]. We should quickly informed about food safety issues on some important news.The LDA theme model[2] is an unsupervised machine learning technique that can be used to identify topic information that is hidden in large-scale document sets or corpora, so this article will use the LDA theme model to complete the Security information exploration .

LDA Theme Model

The Idea of LDA Theme Model. The LDA theme model is a probability growth model for modeling discrete data sets. It is a three-layer Bayesian model[3,4,5], which is divided into document hierarchy, thematic layer and feature Word layer, each layer has a corresponding random variable or parameter control, its basic idea is the text from the implicit theme of random mixing, each subject corresponds to a specific feature word distribution.

If you want to generate a document, the word inside it is the probability of:

$P(\text{words}|\text{Document})=\sum_{\text{Theme}} P(\text{words}|\text{Theme}) \times P(\text{Theme}|\text{Document})$

Where the matrix is shown in Fig. 1:

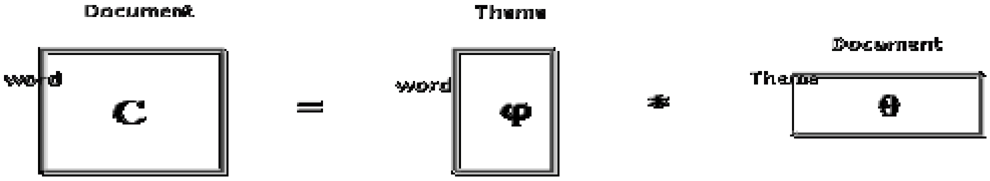

Figure 1. Finite The matrix

The Generation of LDA Theme Model. A document in the LDA theme model is generated as follows[6]:

(1)Samples are generated from the Dirichlet distribution to generate the principal distribution of document i.

(2)Sampling from the multinomial distribution of the topic generates the document i, the subject of the word j.

(3)Sampling from the Dirichlet distribution beta generates the word distribution of the topic .

(4)Sampling from the multinomial distribution of words,eventually generating words.

Among them, the similar Beta distribution is the conjugate prior probability distribution of the binomial distribution, and the Dirichlet distribution is the conjugate prior probability distribution of the polynomial distribution. Since LDA is a three-layer Bayesian model, the above generation process can be described in Fig. 2:

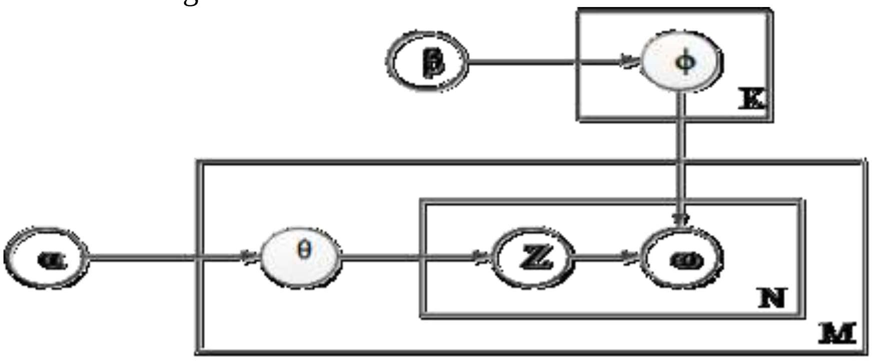

Figure 2. Finite Bayesian network diagram of the subject model

Where: M represents the number of articles in the training corpus (the total number of documents);K represents the number of subjects set; V represents the vocabulary of all the words that appear in the training corpus; θ represents a matrix of M × K, representing the subject distribution of the m-th article ;α represents the parameter of the Dirichlet distribution of the a priori distribution of the subject distribution of each document; β represents the parameter of the prior distribution of the Dirichlet distribution of the word distribution of each subject; Is an observable value.

Gibbs Sampling Estimates of LDA Parameters. In the LDA theme model, it is necessary to estimate the unknown parameters. The commonly used method is the variational -EM algorithm, and then it is found that the Gibbs Sampling[7,8] method is easier to understand and convenient for the estimation of the parameters. Therefore, the Gibbs Sampling algorithm becomes the most commonly used Parameter Estimation Algorithm for LDA Probability Theme Model.

If a set of documents is given, w is a known variable that can be observed, and a priori parameter based on experience, and the variables z, θ, and φ are unknown implicit variables, unknown implicit variables It is necessary to estimate the observed variables. Based on the LDA generation process and its model mentioned in the previous section, we can conclude that the formula for the joint probability distribution for all variables is (1):

$$p(\overrightarrow{w_m},\overrightarrow{Z_m},\Phi|\vec{\alpha},\vec{\beta}) = \prod_{n=1}^{N_m} P(W_{m,n}|\overrightarrow{\phi_{Z_{m,n}}}) * P(\,Z_{m,n}|\overrightarrow{V_m}) * P(\overrightarrow{V_m}|\vec{\alpha}) * P\left(\Phi|\vec{\beta}\right) \tag{1}$$

Wherein,$Z_{m,n}$as defined above$Z_{i,j}$, $W_{m,n}$ is equivalent to what is defined above$W_{i,j}$, $\phi_{Z_{m,n}}$as defined above$\phi_{Z_{i,j}}$, $V_m$corresponding to what is defined aboveθ_i.

Since α produces the distribution of the theme θ, and the subject distribution θ determines the specific subject, and β produces the distribution of the word o, and the word distribution Ø determines the specific word, so the above formula is equivalent to the formula described below The joint probability distribution $P(\vec{w},\vec{z})$ is the formula (2):

$$P(\vec{w},\vec{z}\,|\,\vec{\alpha},\vec{\beta}) = P(\vec{w}\,|\,\vec{z},\vec{\beta}) * P(\vec{z}\,|\,\vec{\alpha}) \tag{2}$$

Where the first factor after the equal sign indicates the process of sampling the word according to the determined subject and the a priori distribution parameter β of the word distribution. The second factor is the a priori distribution parameter α To sample the subject of the process, these two factors need to calculate the two unknown parameters.

Experiment

Acquisition of Experimental Data. Because this article is to solve the problem of food safety public opinion analysis based on LDA theme model, it is necessary to study the food safety problem on the network. so the need to use web crawler technology[9,10]to obtain data on food safety issues, so as to carry out experimental analysis.

Data Preprocessing. The first thing to do is to clean the data , because we directly from the network to crawl the page can not be directly experimental, because these pages contain not only the data we want, so need to do these irrelevant content are washed away, only to retain useful information. Followed by word processing[11], the clean text content is divided into words as a combination of units, this time used NLPIR Chinese word segmentation system.

Experimental design and evaluation methods.

This paper uses Precision (P), Recall (R) and the two comprehensive evaluation index F-Measure (F) to evaluate the results[12,13].

The formula is (5)(6) and (7):

$$p=\frac{A}{B} \quad (5) \qquad R=\frac{A}{C} \quad (6) \quad , \qquad F=\frac{2*P*R}{P+R} \quad (7) \quad .$$

Where A represents the number of correct text extracted, B represents the total number of extracted text, and C represents the number of texts of the same class.

Experimental Results and Analysis. We can get some of the main themes of food safety and the most important vocabulary that makes up the top of these topics. This experiment has four major themes and four The highest frequency of the 10 words, the results As shown in Table 1:

Table 1 4 major topics and 10 important words related to them

Topic1: healt	Topic2: Harmful to health	Topic2: additive	Topic4: Violate management
nutrition	Raw Four seasons beans	Trans fat	Pesticides
delicious	Raw soybean milk	sodium	Severe pollution
clean	Variety of vegetables	High fructose corn syrup	Overdue
healthy	Long spot of sweet potato	Acesulfame	chemicals

As a result of this paper, the LDA theme model is compared with the traditional TF-IDF statistical method. In this experiment, we selected the number of entries for the 20, the number of iterations for 1000 times, to analyze the results of the experiment to predict, and recorded the relevant F value. Therefore, the comparison of the F value of the comprehensive evaluation index of the two methods is shown in Fig. 4.

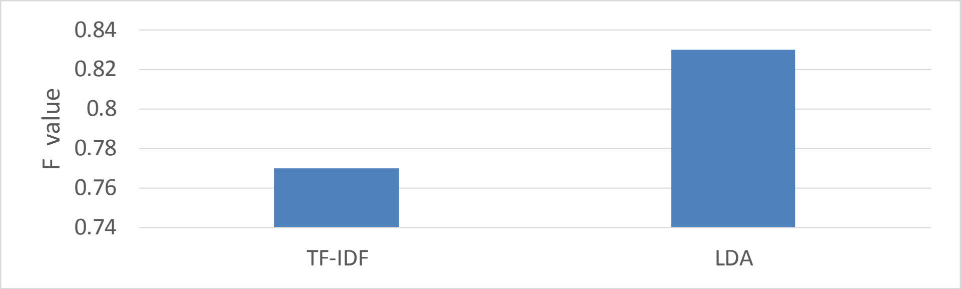

Figure 4. Finite Comparison of the F values of the two methods

It can be seen from Figure 4, based on the LDA theme model algorithm than the traditional TF-IDF statistical method is much better, In this experiment has achieved a better effect, which shows that this method is feasible.

Conclusion

In this paper, the problem of food safety analysis based on LDA theme model is put forward. The Gibbs Sampling algorithm is used to estimate the parameters of the model, and in the experiment, by comparing with the traditional TF-IDF statistical method, and the results show that the method is feasible.

Acknowledgement

This work was supported by the Natural Science Foundation of Hubei Province of China under Grant No.2016CFB470.

References

[1] Sun Baoguo ,Wang Jing, Sun Jinyuan.Perspectives on China Food Safety Problems.Beijing Laboratory for Food Quality and Safety, Beijing Technology & Business University, Beijing 100048.

[2] Blei D, Ng A, Jordan M. Latent Dirichlet allocation［J］. Journal of Machine Learning Research, 2003(3) : 993 － 1022.

[3] Zhang Peijing, Song Lei.Overview on Topic Modeling Method of Microblogs Text Based on LDA.1:Office of Chinese People’s Public Security University, Beijing 100038.2:Department of Police Technology, Beijing Police College, Beijing 102202.

[4] Zhu Ting, Qin Chunxiu, Ma Xiaoyue, Li Zuhai.Research on Text Resource Recommendation Method Based on Ontology and LDA Topic Model. School of Economics and Management, Xidian University, Xi’an 710071

[5] Zhang Liang.Research on Tagging Recommendation Method Based on LDA Topic Model.School of Management, Wuhan Institute of Technology, Wuhan Hubei 430205, China.

[6] SHI Qingwei, GUO Pengliang. Conditional random fields topic model based on LDA model. Computer Engineering and Applications, 2015, 51(7): 131-135.

[7] WANG Peng, GAO Cheng, CHEN Xiao-mei. Research on LDA Model Based on Text Clustering.1.School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130002, China; 2. School of Management, Jilin University, Changchun 130022,China.

[8] Hu Jiming ,Chen Guo. Mining and Evolution of Content Topics Based on Dynamic LDA .Center for Studies of Information Resources, Wuhan University, Wuhan 430072.

[9] WANG Shao-peng, PENG Yan , WANG Jie.Research of the text clustering based on LDA using in network public opinion analysis.1.College of Information Engineering, Capital Normal University,Beijing 100048, China;2. School of Management, Capital Normal University, Beijing 100089, China.

[10] Yu Juan,Liu Qiang.Survey on topic -focused crawlers.School of Economics and Management, Fuzhou University,Fuzhou 350108,China.

[11] Li Xiangdong,Gao Fana,Ding Cong.Study on influences of different Chinese word segmentation methods totext automatic classification based on LDA model.a:School of Information Management;b:Center for Studies of Information Resources, Wuhan University, Wuhan 430072, China.

[12] SUN Chang nian,ZHENG Cheng,XIA Qingsong.Chinese Text Similarity Computing Based on LDA.1:School of Computer Science and Technology, Anhui University, Hefei 230039, China;
2:Key Lab. of Intelligent Computing & Signal Processing, Mini. of Edu. , Anhui Univ. , Hefei 230039, China.

[13] Guan Peng,Wang Yuefen, Fu Zhu,.Effect Analysis of Scientific Literature Topic Extraction Based onLDA Topic Model with Different Corpus.1:School of Economics and Management, Nanjing University of Science ＆ Technology, Nanjing 210094;