Optimization of Topic Recognition Model for News Texts Based on LDA

Hongbin Wang, Jianxiong Wang, Yafei Zhang*, Meng Wang, Cunli Mao
Faculty of Information Engineering and Automation
Kunming University of Science and Technology, China
{whbin2007@126.com} {916864282@qq.com} {76326474@qq.com} {85175849@qq.com}
{59058012@qq.com}
* Corresponding author

Journal of Digital Information Management

ABSTRACT: *Latent Dirichlet Allocation (LDA) is the technique most commonly used in topic modeling methods, but it requires the number of topics generated by LDA to be specified for topic recognition modeling. Except the main iterative methods based on perplexity and nonparametric methods, recent research has no simple way to select the optimal number of topics in the model. Aiming at appropriately determining the number of topics and then optimizing the LDA topic model, this paper proposes a non-iterative method for automatically determining the number of topics. The clustering method is based on fast seeking and locating density peaks. This method transforms the traditional topic cluster number selection problem into clustering problem and thus can be used to optimize the topic recognition model for news texts. It does not need iterative optimization and can simplify model development. This method uses Word2Vec for word embedding on corpus text to explore the superior performance of word-related relationships and to express the implicit semantic relationship between topic corpora. Then, using a clustering algorithm that quickly searches for and finds the cluster peaks; the word vectors after word embedding are clustered to obtain the number of word vector clusters after word embedding. The number of clusters is used as the number of topics in the text. Finally, the experimental results show that the proposed method enjoys better precision and F1 value than the perplexity-based method, and is suitable for the identification of the number of topics in corpora in different sizes. This method can effectively find the appropriate number of topics from the news text dataset and improve the accuracy of the LDA theme model.*

Subject Categories and Descriptors: [H.2.4 Systems]: Textual databases; **[I.2.7 Natural Language Processing]:** Text analysis; **[H.3.3 Information Search and Retrieval]:** Clustering

General Terms: Latent Dirichlet Allocation, Topic Recognition Modelling, Text Corpus

Keywords: LDA, Topic Recognition, Word2Vec, Cluster

Received: 7 March 2019, Revised 5 June 2019, Accepted 18 June 2019

Review Metrics: Review Scale- 0/6, Review Score-4.85, Inter-reviewer Consistency- 82%

DOI: 10.6025/jdim/2019/17/5/257-269

1. Introduction

Topic modeling is actively researched in the field of machine learning, and mainly used to construct models from unstructured data that are manifested as a set of textual documents in the form of potential topics to extract the macrostructure of the document set (usually the multinomial distributions on words). In topic modeling, it is assumed that there are a certain number of potential topics in a set of given unstructured documents, and that each document contains multiple topics in different sizes. Researchers have developed a variety of subject models which are commonly used in NLP [1-3], such as Latent Dirichlet Allocation [4] (LDA) and its variants, Latent Semantic Indexing [5] (LSA) and Probabilistic Latent

Semantic Analysis (PLSA) [6, 7]. Besides, recent work has been conducted on neural topic models [8-10]. Topic modeling has been widely used in various fields including text mining, image retrieval, text retrieval, text categorization, citation analysis, network paradox analysis, and bioinformatics analysis [11-14].

LDA is the most common technique in topic modeling and is an unsupervised probabilistic method for modeling corpora. It is assumed that each document can be represented as a probability distribution of potential topics and that the topic distributions in all documents share the previous Dirichlet. Similarly, each potential topic in the LDA model is represented as a probability distribution of words, and the word distribution of the topic also shares the previous Dirichlet. In LDA, the "document-topic" vector θ_d is generated by a Dirichlet distribution with hyperparameter α, and the "topic-words" vector ϕ_d is generated by a Dirichlet distribution with hyperparameter β. Suppose that a corpus D is composed of M documents and that the document d has N_d words ($d \in \{1,..., M\}$), then an LDA model comes into being according to the following generation process [4]:

(a) Select a multinomial distribution ϕ_t of topic t ($t \in \{1,..., T\}$) from a Dirichlet distribution with parameter β.

(b) Select a multinomial distribution θ_d of document ($d \in \{1,..., M\}$) from a Dirichlet distribution with parameter α.

(c) For the word w_n ($n \in \{1,..., N_d\}$) in document d,

(i) Select a topic Z_n from θ_d.

(ii) Select a word w_n from ϕ_{Z_n}.

In the above generation process, the words in the document are the only observed variable, while the others are latent variables (ϕ and θ) and hyperparameters (α and β). The number of related topics and content in the document sets are already known. However, for unstructured document sets, the number of documents and related topics remains to be elucidated. That is to say, the optimal number of topics that the best topic model involves is unknown, and different numbers of topics may affect the accuracy and complexity of the topic model. An insufficient number of topics may make the LDA model too coarse to accurately distinguish topics. By contrast, an excessive number of topics may lead to more complex models, making topic interpretation and subjective verification more difficult [15]. At present, there is an apparent lack of a better way to automatically obtain the topic number of unstructured documents. Researchers have no alternative to solve this issue except informed guesses and time-consuming iterative methods. Therefore, this paper proposes a non-iterative method for automatically determining the number of topics to optimize the LDA-based topic recognition model for news texts.

2. Related Work

Blei et al [3] proved the reliability and effectiveness of the LDA model under the exact number of topics. However, how to select the number of topics remains to be effectively solved, which seriously affects the effectiveness of the LDA model. Currently, apart from informed guesses and time-consuming iterative methods, there is no other way to locate the appropriate number of topics in the model. Due to the importance of the number of topics, many researchers have proposed methods to solve the problem of locating the number of topics. After detailed analysis, we summarized them into the following five categories:

(1) A heuristic empirical setting method. The selection of the number of topics is equivalent to the evaluation of the model which appears complex. The method adopts a heuristic empirical setting. During continuous debugging, empirical subjective judgment is made to determine the number of topics. The method is simple and easy to operate. For example, Guan et al [16] selected the number of topics using a heuristic empirical setting method based on the LDA theme model, and touched upon other topics in the extraction effect analysis in different corpora.

(2) A method based on perplexity. The method mainly compares the quality of two models. Perplexity is a commonly used measure in information theory to assess the extent to which a statistical model describes a dataset, and lower perplexity indicates a better probability model. In the LDA theme model, the formula for calculating the perplexity is shown in equation (1), where D is the test set, w_d is the sequence of observable words in document d, and N_d is the number of words in document d.

$$perlexity\,(D) = exp\left\{ \frac{\sum_{d=1}^{M} logP(w_d)}{\sum_{d=1}^{M} N_d} \right\} \qquad (1)$$

Perplexity is used to measure the LDA model and iterations are performed on the number of topics in the dataset to get the number of topics in the corpus. Since the perplexity indicator responds to the generalization ability of the model itself, the perplexity-based method may produce meaningful results in some cases. Nonetheless, it is not stable, and it can only show the generalization ability of the model to new samples, thus leaving its judgment of the number of topics wanting in logical rigor. For example, the method proposed by Liao [17] and Liu [18] used the perplexity indicator to select the number of topics.

(3) A Bayesian statistical standard method. The method using the Log-boundary likelihood function is also commonly used. Griffiths [19] proposed the Markov chain Monte-Carlo algorithm for reasoning in the LDA topic model. The extracted subject captures the meaningful structure in the data, and this method is completed by Gibbs sampling algorithm. For illustration, Hajjem M and Latiri C [20] developed applications based on text

segmentation and microblog information filtering respectively, using the Bayesian statistical standard method. Nevertheless, the method is still semi-heuristic and involves highly complex calculation owing to empirical values and Gibbs sampling algorithm.

(4) A nonparametric method [21, 22]. The main idea of this method is make the number of topics non-parametric to obtain the optimal number of subjects in the process of model calculation without human intervention. For instance, Teh [23] proposed Hierarchical Dirichlet Processes (HDP), in which the subjects are generated from data rather than pre-fixing the subjects and then restoring them by reverse engineering the data. By analyzing the sample histogram of mixed components in HDP, it is found that the optimal number of mixed components is the same as the optimal number of topics in LDA, thereby solving the problem of selecting the optimal number of topics in LDA. Experiments have shown that the optimal number of topics selected by HDP is consistent with that selected based on perplexity. However, this method needs to establish an HDP model and a LDA model for the same set. Worse still, the algorithm is time-consuming and incurs large maintenance cost for the code.

(5) A method based on the similarity between topics [24]: using KL divergence to measure the similarity between topics. When the average similarity between topics is the smallest, the greater is the degree of recognition between topics; the better is the corresponding model. For instance, Guan et al. [25] put forward subject variance, namely, the average sum of the squares of the distance between each topic and its mean. It measures the overall difference and stability between the topics. When the subject variance is larger, the greater is the difference between the topics, the better is the distinction between themes, and the more stable is the subject structure. The degree of perplexity reflects the predictive ability of the model, but blindly pursuing the predictive ability of the model will inevitably lead to the problem of too large number of topics to be extracted. As a result, the full consideration of both factors can effectively solve the problem of low recognition of topics. As for the generalization ability of the model, the smaller is the subject perplexity; the better is the generalization ability of LDA. Moreover, regarding the theme extraction effect of LDA, the corresponding LDA topic model is the optimal when the average similarity of the topic structure is the smallest [21]. Hence the smaller is the average similarity of the topic structure; the greater is the difference between the topics. At this time, the variance of the subject structure is larger. Therefore, when the subject variance is larger, the LDA theme extraction is better, and the same perplexity-theme variance indicator is smaller. Based on the analysis above, it can be concluded that when the perplexity-thematic variance index is the smallest, the corresponding LDA topic model is the optimal.

The comprehensive analysis found that the above methods for determining the optimal number of LDA topics belong to iterative methods. Each step needs to solve an objective function, making the calculation and model highly complicated. The iterative algorithm needs to evaluate the performance of the LDA topic model, which incurs a lot of uncertainty, thus increasing the complexity and uncertainty of the algorithm. The number of iterations required in the control of the iterative process cannot be determined. Furthermore, the condition used to end the iterative process needs analyzing further. The following two situations may occur:

(1) The condition used to end iteration obtained by the algorithm will not converge, turning the iterative process into an infinite loop. Therefore, before using the iterative algorithm, we should first check whether the condition used to end iteration converges, and limit the number of iterations in the program;

(2) Although the condition used to end iteration converges, the selection of the iterative condition is improper, or the initial value selection of the iteration is unreasonable, which may cause the iteration to fail.

Taken together, the above situations will complicate the model, making the final result uncertain.

In this paper, a new topic recognition model method considering word correlation was proposed based on the traditional LDA model. Firstly, the method used a Word2Vec model to explore the superior performance of semantic relationship. Then, it clarified the implicit semantic relationship between topic corpora, and clustered word vectors after word embedding. The cluster peak was found by applying a fast search. The clustering algorithm obtained the number of word vector clusters after word embedding, which was used as the number of text topics. It solved the problem of different corpus structures due to the number of different topics.

The newly proposed method adopted a non-iterative algorithm, which reduced the complexity of the model. The traditional LDA model follows the bag-of-words hypothesis, and the correlation between words is ignored. By contrast, the proposed model applied the word2vec model to identifying the relationship between words and meanings, so that the model could pinpoint the number of topics with great accuracy.

3. Automatically Determining the Number of Topics

3.1. Model Construction

After a sequential preprocessing of the original text (including cleaning and word segmentation), we could use the Word2vec model for word embedding, and then cluster the word embedding based on fast search for

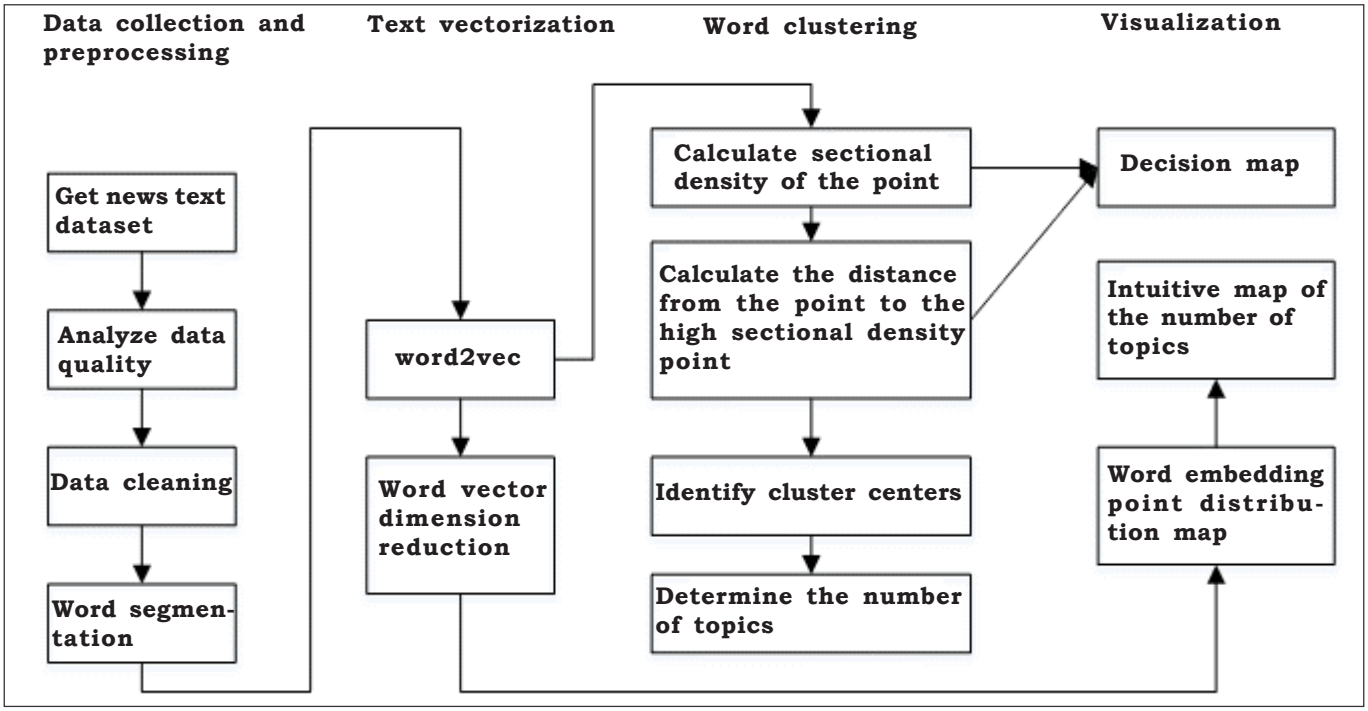

Figure 1. Flowchart for identifying the number of topics

and find of density peaks to obtain the number of word vector clusters, which was the optimal topic number of news texts. The number of optimal topics was used as a parameter of the LDA topic model for topic extraction, and different methods of identifying the number of topics were compared. The flowchart was shown in Figure 1.

3.2 Word Segmentation and Word Vector training for Word2Vec Model

As regards the text data after preprocessing, Jieba word segmentation tool was applied to word segmentation processing.

The vectorization of words provides a mathematical method of transforming symbolic information of a natural language into digital information in a vector form. Since the 21st century, people have gradually developed from the original word vector sparse representation to the dense representation in the current low-dimensional space. Sparse representation often encounters dimensional disasters when solving practical problems. Specifically speaking, the semantic information cannot be expressed and the potential connection between words cannot be revealed. The low-dimensional spatial representation not only solves the dimensionality disaster but also mines the association properties between words, thus improving the accuracy of vector semantics.

Word2Vec, as a model for learning semantic knowledge in an unsupervised way from a considerable number of text corpora, has been extensively used in NLP as an open-source algorithm. In addition, its principles have been elaborated in many studies. The architecture used in this paper was the Skip-gram. The Word2Vec was packaged into genism, which is a Python package for natural language processing and information retrieval. It trains large-sized corpora to enable topic modeling, literature indexing and similarity calculation.

3.3 Clustering

3.3.1 Clustering Algorithm

Clustering is the process of dividing a sample of a dataset into disjoint subsets, so that subsets in the same cluster have high similarity while subsets in different clusters have distinct dissimilarity. The clustering algorithm contains the following three types: clustering algorithms based on partitioning, density, and hierarchy respectively. This paper employed a new density-based clustering algorithm proposed by Rodriguez et al. [26].

3.3.2 Methodology

First, the method was based on the assumption that the center of the cluster was surrounded by some points with a low sectional density, and that these points had a relatively large distance from any other point with a higher sectional density. For each data point i, the sectional density ρ_i of the point and the distance σ_i from the point to the point with a high sectional density were calculated. These two values depended only on the distance d_{ij} between the two points. The sectional density ρ_i of the data points i was defined by equation (2):

$$\rho_i = \sum_j X(d_{ij} - d_c) \qquad (2)$$

where if $x < 0$, then $X(x) = 1$; if $x \geqq 0$, then $X(x) = 0$, and d_c was an intercept. Basically, ρ_i was equal to the number of points whose distance from point i was less than d_c. The algorithm was only sensitive to the relative size of ρ_i at different points, which meant that for large datasets, the

analysis results showed high robustness to the choice of d_c.

The δ_i of data point *i* was the minimum value of the distance from the point to any point greater than its density (3):

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \quad (3)$$

For the point with the highest density $\delta_i = max_j (d_{ij})$.

3.3.3 Clustering Process

The point with large sectional density ρ_i and large δ_i was considered as the center of the cluster. The point with small sectional density ρ_i and large δ_i was regarded as a cluster consisting of a single point, which was also known as an abnormal point. After the cluster center was determined, each remaining point was attributed to its cluster of nearest neighbors with a higher density. Moreover, the assignment was done in one step without iterative optimization of the objective function.

3.4 Visualization

The process was visualized to better denote the entire number of identified topics. After preprocessing and word embedding operations, corpus embedding was performed. In order to facilitate the display of the clustering process, the words were embedded in dimensionality reduction and mapped onto the two-dimensional vector to obtain the distribution map of embedded word points. Subsequently, the clustering process was exhibited, which was a decision map with ρ_i as the abscissa and δ_i as the ordinate. The points in Figure 3 with high δ_i and high ρ_i were considered to be cluster centers. Figure 4 distinguished cluster center points from non-cluster center points by color. After the cluster center was determined, the number of topics in the cluster center was fixed, and other points were assigned to the same cluster group as the nearest neighbors with a higher density, which represented different themes in different colors.

4. Experiment and Analysis

4.1 Experimental Data and Data Preprocessing

In this study, Tencent News' hot topic dataset was used to test and evaluate the proposed approach. By searching for popular news topics from Tencent News website, three datasets consisting of 100 news texts, 400 news texts and 1000 news texts were found respectively. The three text datasets with different numbers were preprocessed by Chinese word segmentation and removal of stop words. Then the topics and the number of related topics were manually extracted. Taking the 400 news texts as an example, the number of topics obtained was shown in Table 1:

4.2 Determination of the Optimal Number of Topics

As for the preprocessed corpus, the word embedding operation was carried out and the dimension selection of the Word2Vec ranged between 200 and 300. In order to facilitate the visual display of the clustering process, the dimension was reduced to two by operating word embedding. Figure 2 showed the two-dimensional space points where embedded words were mapped in a two-dimensional space. It could be roughly seen from Figure 2 that the point with the highest density was defined as the cluster center. Figure 3 offered the decision map with ρ_i as the abscissa and δ_i as the ordinate. In Figure 3, the point with high δ_i and high ρ_i was regarded as the cluster center, while the point with small ρ_i and large δ_i was regarded as a cluster consisting of a single point known as an abnormal point. We eliminated exception points and clusters based on the purpose of the subject extraction.

After the cluster center was determined, the remaining points were assigned to the nearest cluster with a higher density. Unlike other iteratively optimized clustering algorithms, class cluster assignments were performed in a single step. The results of allocation were shown in Figure 4. Different colors represented different class clusters. At this point, the number of clusters obtained was the subject of the text in the LDA.

Topics	Number	Topics	Number	Topics	Number
一带一路	40	红芯浏览器被质疑	3	英国脱欧	16
通海地震	3	长生生物	19	2018亚运会	51
房租涨价	16	滴滴乘客被害	23	我不是药神热映	29
安南去世	4	鸿茅药酒造假	16	韩国乐天	29
美国加征关税	23	杭州保姆纵火	13	中兴芯片	36
江歌日本被害	12	红黄蓝幼儿园虐童	33	朴槿惠	34

Table 1. Corpus topics and number of topic articles

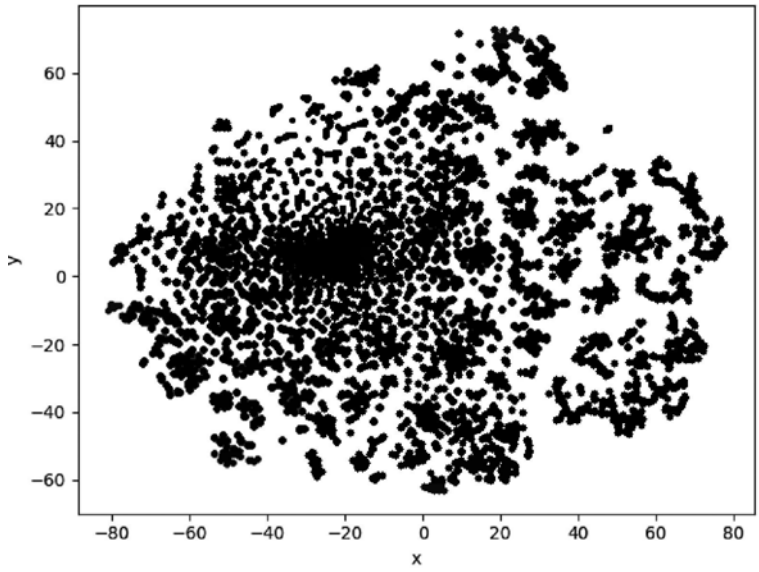

Figure 2. Distribution map of embedded word points

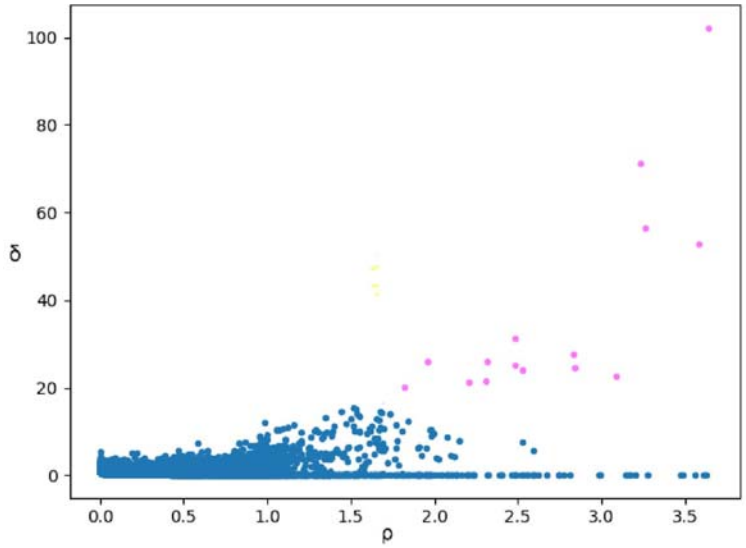

Figure 3. Decision map

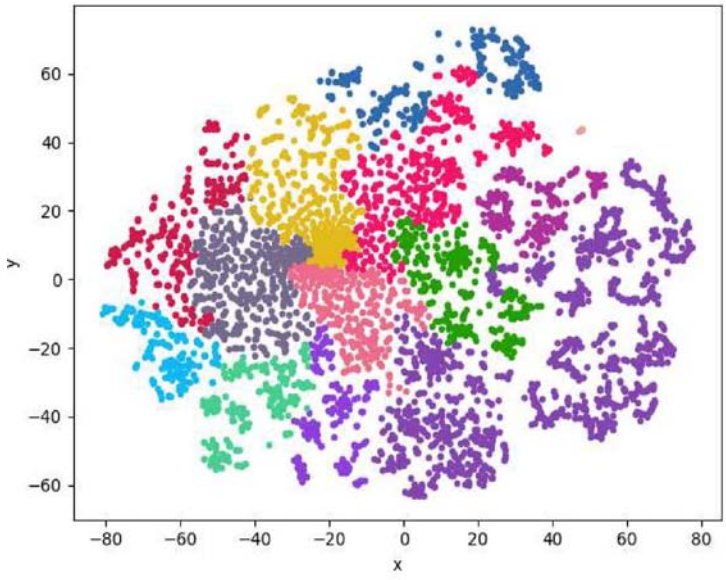

Figure 4. Distribution map of embedded word points with each point representing its cluster in a color

4.3 Methodological Evaluation

Among the methods for determining the optimal number of LDA topics, the empirical method requires plenty of time and effort while the Bayesian statistical standard method and the non-parametric method are complex. Therefore, the perplexity-based method was selected as the comparison object of the proposed method. The methods were evaluated in terms of accuracy and comprehensiveness of news text topic extraction.

4.3.1 Determination of Optimal Number of Topics Based on Perplexity Method

The experiment used a perplexity method. To be specific, after iterating over a certain number of topics and when the perplexity degree slowed down within a certain range, the number of topics at this time could best represent the optimal number of topics in the dataset. The experimental results of the 400 Tencent news texts were shown in Table 2 and Figure 5. Judging from Table 2 and Figure 5, the number of topics decreased significantly from 26 and tended to be stable. Therefore, the optimal number of topics should be 26.

n	Perplexity
1	-11.080522603744297
6	-11.528040290223517
11	-11.686841810925171
16	-11.983907486919497
21	-11.949689567973851
26	-12.04940281527981
31	-12.082360553543022
36	-12.188879226036369
41	-12.164626720613548
46	-12.171651997524351
51	-12.290086556491347

Table 2. Calculation of topics based on perplexity

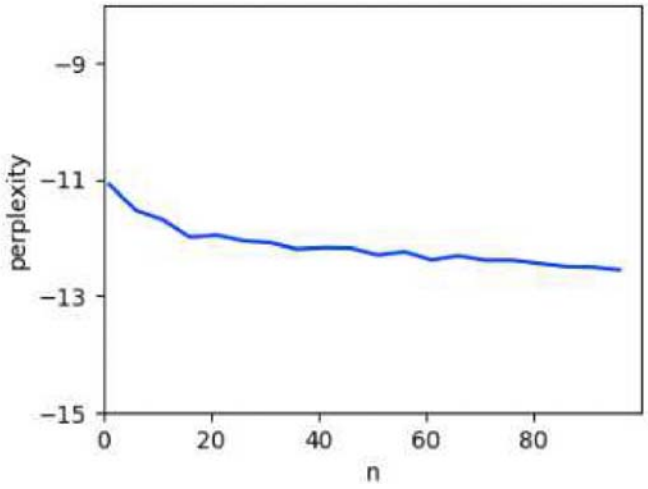

Figure 5. Calculation of topics based on perplexity

Topic	Topic words				
1	疫苗	长春	谭秦东	长生	长生生物
2	莫焕晶	放火	乐天	被告人	保姆
3	票房	米	亚运会	选手	中国队
4	幼儿园	红黄蓝	孩子	家长	幼儿
5	幼儿园	红黄蓝	家长	孩子	亚运会
6	乐天玛特	门店	乐天	韩国	比赛
7	朴槿惠	韩国	总统	检方	乐天
8	中国男篮	亚运会	幼儿园	比赛	韩国
9	疫苗	长春	谭秦东	长生	长生生物
10	莫焕晶	放火	乐天	被告人	保姆

Table 3. Extraction result of LDA topic model based on clustering

Topic	Topic words				
1	幼儿园	红黄蓝	莫焕晶	孩子	家长
2	关税	特朗普	美国	长春	钢铁
3	干政	朴槿惠	辛东彬	亲信	宣判
4	鸿茅药酒	金博洋	候选	服务区	艺术节
5	巴尼耶	中国女足	禹柄宇	朴槿惠	韩国
6	米	放火	中国男篮	中国台北	莫焕晶
7	鸿茅药酒	王霜	光伏	公司	朴槿惠
8	票房	药神	上映	西虹市	首富
9	谭秦东	鸿茅药酒	凉城县	内蒙古自治区	声誉
10	CMBS	朝阳区	朴正熙	张业遂	朴槿惠

Table 4. Extraction result of LDA topic model based on perplexity

4.3.2 Accuracy and Comprehensiveness of the Assessment Method

According to the experimental results, we obtained 26 optimal clusters based on perplexity, and 15 optimal clusters based on fast search for and find of density peaks. Firstly, the dataset underwent word segmentation and removal of stop words. Secondly, LDA was adopted to extract topics. To facilitate the display and comparison, the top 5 words of each topic were selected for evaluation, and the results were analyzed, as listed in Table 3 and Table 4.

The LDA topic model employed the integrated semantics of words from different topics to explain the extraction of topics. The results were compared with the manual judgment of topics to calculate the precision P, the recall ratio R and the F1 metric of the LDA topic extraction under different topic number optimization methods, as described in equation (4):

$$R = \frac{N_1}{N_2},\ P = \frac{N_1}{N_3},\ F_1 = \frac{2PR}{P+R} \qquad (4)$$

where N_2 was the number of valid topics extracted by LDA; N_1 was the number of correctly extracted and effective topics that were included in the domain topics based on experts' evaluation; N_3 was the number of domain topics based on literature research and experts' judgement. Both methods were compared with the number of topics based on manual judgment. There were 9 topics according to the perplexity-based approach that involved interference terms while there were 2 topics according to the fast search for and find of density peaks. The obtained data were shown in Table 5, and the two methods were compared and evaluated in Table 6.

It could be seen from the results that the perplexity-based method had more effective topics extracted and a higher recall rate. However, the similarity and cross-discipline between the themes were strong, and the probability of repetition and synonymous vocabulary was high. The interpretation of the theme was weak, and the theme recognition effect was not ideal. The relative dispersion of subject terms within a single topic made it difficult to focus on the topic. The method based on fast search for and find of density peaks had a low recall rate; but the precision was high, the discrimination between the topics was more obvious, and the cross-references of the keywords within each topic were lower. In a word, it was

Optimal subject number selection method	Based on clustering	Based on perplexity
Number of manually identified topics	18	18
Number of topics	15	26
Accurate number	13	16
Correct subject number	10	13

Table 5. Data result based on different methods of selecting the number of topics

Optimal subject number selection method	Based on clustering	Based on perplexity
Precision rate	86.67%	61.53%
Recall rate	55.56%	72.22%
F_1 value	66.50%	66.45%

Table 6. Comparison of LDA topic extraction effects based on different methods of selecting the number of topics

more cohesive than the perplexity-based method. Therefore, the method based on fast search for and find of the peak value of density not only could determine the optimal number of topics autonomously, but also had certain advantages in terms of subject recognition when compared the perplexity-based method.

4.3.3 Applicability of the Evaluation Model

In order to compare the applicability of the model, the method based on perplexity and the method based on fast search for and find of density peaks were used to select the best number of topics concerning three datasets. The procedure for the dataset consisting of 400 texts was repeated to select the optimal number of topics and to establish an LDA theme model. The applicability of the proposed method was evaluated by comparing output words of the three datasets based on the LDA topic model. The comparison was shown in Table 7.

It could be seen that the precision rate of the method based on fast searching for and finding density peak clustering was always higher than that of the method based on perplexity, and that the method based on fast searching for and finding density peak clustering showed certain practical applicability to corpora in different sizes.

In order to further verify the effectiveness of the proposed method, we used open Sogou news from Sogou laboratory, Sohu news collected from June 6, 2012 to July on automobile, finance, IT, health, sports and so on, and news from channel 18 for experimental verification. 10 mini topic datasets were chosen as experiment datasets; there were 100 documents in each topic. Word segmentation and removal of stop words were carried out for each text. Then, word embedding was executed to obtain embedded words of the corpus. Finally, clustering was carried out to obtain the number of clusters as the number of topics in the text dataset. The distribution map of embedded word points was shown in Figure 6, the decision map was shown in Figure 7, and the results of the distribution were shown in Figure 8. Different colors denoted different clusters.

Data set	Optimal subject number selection method	Precision rate (P)	Recall rate (R)	F1value
100	based on perplexity	62.50%	83.33%	71.43%
	based on clustering	**85.71%**	66.67%	75.00%
400	based on perplexity	61.53%	72.22%	66.45%
	based on clustering	**86.67%**	55.56%	66.50%
1000	based on perplexity	61.66%	77.41%	71.91%
	b**ased on clustering**	**85.00%**	61.76%	71.53%

Table 7. Comparison of LDA theme extraction based on datasets in different sizes

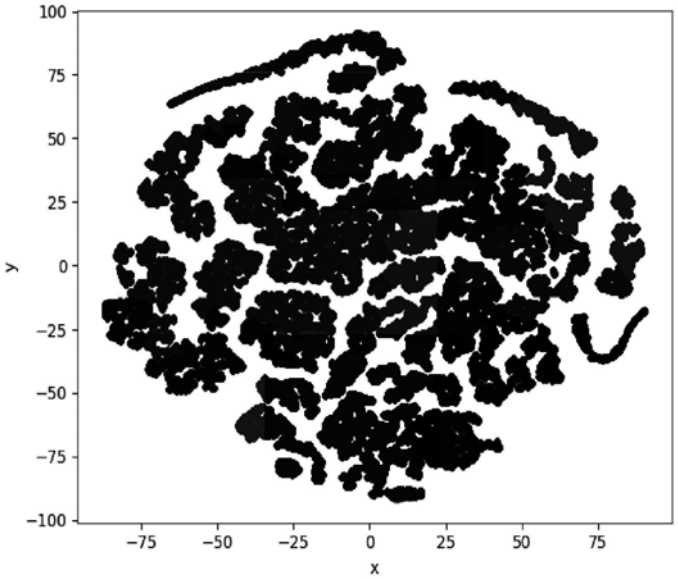

Figure 6. Distribution map of embedded word point

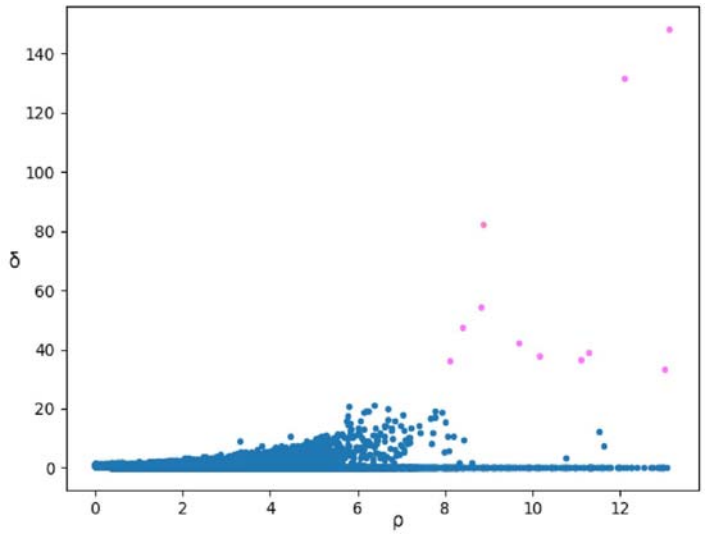

Figure 7. Decision figure

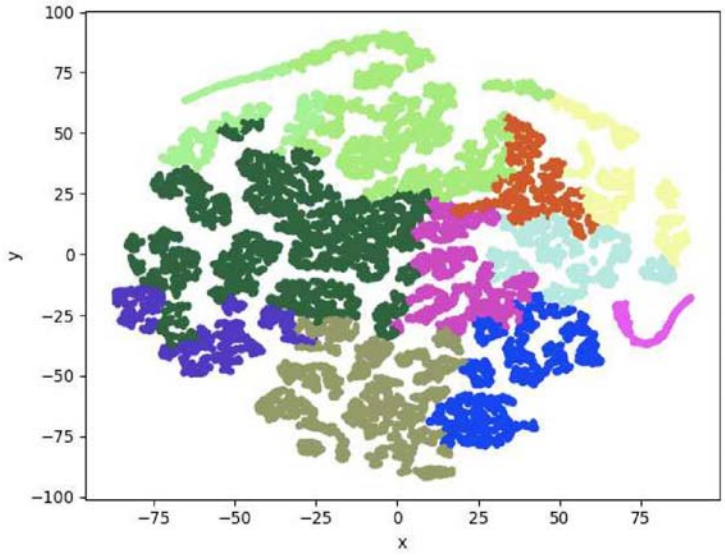

Figure 8. Distribution map of embedded word points

Judging from Figure 5 and Figure 6, and according to the experimental results, the optimal number of clusters arrived at 11 based on fast search for and find of density peaks.

The results obtained based on the perplexity method were shown in Table 8 and Figure 9. As denoted by Table 8 and Figure 9, the number of topics decreased significantly from 40 and tended to be stable. It could be known that the number of topics was significantly reduced at 40 and tended to be stable. Therefore, the number of best topics obtained was fixed at 40.

It was known in advance that the optimal number of topics was 10 in the experiment dataset. Obviously, the optimal number of topics obtained by clustering method based on fast search for and find of density peaks was better. In order to compare the effects of the two methods more vividly, we used LDA to extract topics for comparison. The topics from the experimental dataset were listed in Table 9.

n	Perplexity
10	-10.329534376805187
20	-11.037781888476468
25	-11.258965029874286
30	-11.569587921987886
35	-11.796853894277073
40	-11.960192359098103
45	-12.019705714209104
50	-12.088392856349472
60	-12.112101362304733
70	-12.241816610546948
80	-12.313608778580336

Table 8. Calculation of topics based on perplexity

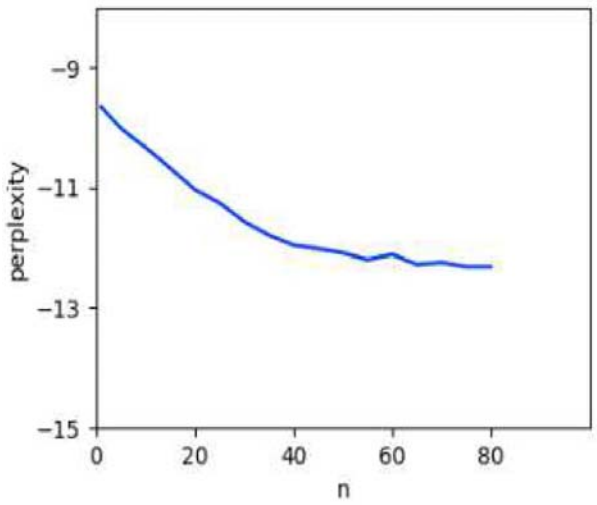

Figure 9. Calculation of topics based on perplexity

Topic	Number	Topics	Number	Topics	Number
汽车	100	财经	100	IT	100
健康	100	体育	100	旅游	100
教育	100	招聘	100	文化	100
军事	100				

Table 9. Corpus topics and number of topic articles

We used LDA to extract topics from 10 preprocessed topic datasets. The first five words of each topic were selected for evaluation, and the extraction results were shown in Table 10 and Table 11. The obtained data were shown in Table 12. The recall rate R, precision rate P and F1 values of the LDA topic extraction results of the number of topics obtained by different methods were shown in Table 13.

Topic	Topic words				
1	旅游	黄金周	游客	五一	说
2	页	增长	中海	企业	出口
3	增长	市场	通	证券	元
4	考生	专业	上海	日	美国
5	毕业生	企业	工作	市场	增长
6	新浪	美国	产品	互联网	上海
7	美国	俄罗斯	政府	空军	巴基斯坦
8	企业	汽车	增长	日	产品
9	医药	市场	健康	生活	营养
10	体育	中国	日	比赛	球队

Table 10. Extraction results of LDA topic model based on clustering

Topic	Topic words				
1	营养	丽江	浪漫	临床	评选
2	增长	企业	汽车	工作	元
3	亿美元	景区	美国	第一季度	车
4	凯旋	G	雪铁龙	车	Kashya
5	云药	天窗	上海大众	兵器	武警
6	学员	电信	央视	民航	IPTV
7	车贷险	不良贷款	备考	微软	欧盟委员会
8	通	证券	行情	万吨	KL178
9	孙子兵法	伊拉克	查尔斯	姚明	敌人
10	用人单位	面试	SUV	就业	应聘

Table 11. Extraction results of LDA topic model based on perplexity

Optimal subject number selection method	Number of manually identified topics	Number of topics	Accurate number	Correct subject number
Based on clustering	10	11	8	7
Based on perplexity	10	40	23	9

Table 12. Data result based on different methods of selecting the number of topics

Optimal subject number selection method	Precision rate(P)	Recall rate(R)	F_1 value
Based on clustering	80.00%	70%	74.67%
Based on perplexity	57.50%	90%	70.17%

Table 13. Comparison of LDA topic extraction effects based on different methods of selecting the number of topics

It could be seen from the results that the precision of the method based on density peaks in this dataset was always superior to the method based on the perplexity. The experimental code and corpus could be seen in: https://github.com/Hayden-z/Optimization-of-Topic-Recognition-Model-for-News-Texts-Based-on-LDA.

5. Conclusion

This paper has further optimized the LDA model and proposed a method based on fast search for and find of density peak clustering to determine the optimal number of topics, so that it can better serve the analysis and processing of news text data. Based on the traditional LDA model, the Word2Vec model of deep learning level has been included to enhance the similarity relationship between topic words. As a result, the problem of topic number identification has been transformed into clustering problem, or more vividly a strange problem has been changed into a statistical problem supported by relatively mature theories, thus effectively overcoming the problem of selecting the number of topics. The performance of the new and old models has been evaluated by such indicators as the recall rate, precision and F1 value. It has been found that the performance of the new model is better than that of the traditional LDA model in terms of the recall rate and F1 value. The limitation lies in the contradiction between the applicability of the clustering method and the complexity of the algorithm. Therefore, the future research direction is to combine the multi-statistical analysis theory to break through the traditional clustering method, and to apply the clustering technique that considers both methodology and efficiency to the selection of the number of topics.

The experimental data are recent hot topics taken from Tencent News. The experimental process for the corpus has reflected the superior performance of the new model. Although this paper has not coped with other types of texts, such as scientific literature and short texts from social media, the content and technical means mentioned in the above work can be applied to other large-sized text corpora. Therefore, the model proposed in this paper provides certain reference for the study of other types of text corpora. In addition, the results based on fast search for and find of density peak clustering are limited to the threshold range, only becoming the local optimal solution. But considering the actual effect and demand of cluster analysis, the threshold is much higher than the actual number of divisions. Arguably, the obtained results still enjoy strong persuasiveness.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (61462054, 61762056, 61563025, 61662041), and the Science and Technology Plan Projects of Yunnan Province (2016FB109, 2016FB101).

References

[1] Newman, David., Baldwin, Timothy., Cavedon, Lawrence., Karimi, Sarvnaz., Martinez, David., Zobel, Justin. (2010). Visualizing document collections and search results using topic mapping. *Journal of Web Semantics*, 8(2–3) 169–175.

[2] Wang, Xuerui., McCallum, Andrew. (2006). Topics over time: a non-Markov continuous-time model of topical trends. *In*: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 424–433.

[3] Hall, David., Jurafsky, Daniel., Christopher, D., Manning. (2008). Studying the history of ideas using topic models. *In*: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, USA, 363–371.

[4] David, M., Blei, Andrew, Y., Ng, Michael, I., Jordan.

(2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3. 993–1022.

[5] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science and Technology*, 41(6) 391407.10. 1002(SICI)10974571(199009)41:6<391::AIDASI1>3.0.CO; 2-9.

[6] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*. 42 (1-2) 177- 196.

[7] Hofmann, T. (1999). Probabilistic latent semantic indexing. *In*: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 50-57.

[8] Cao, Ziqiang., Li, Sujian., Liu, Yang., Li, Wenjie., Heng Ji. (2015). A novel neural topic model and its supervised extension. *In*: Proceedings of the 29th Annual Conference on Artificial Intelligence (AAAI-15). Austin, Texas, 2210–2216.

[9] Hugo Larochelle and Stanislas Lauly. (2012). A neural autoregressive topic model. *In:* Advances in Neural Information Processing Systems 25. 2708– 2716.

[10] Geoffrey, E., Hinton, Ruslan, R., Salakhutdinov. (2009). Replicated softmax: an undirected topic model. *In:* Advances in Neural Information Processing Systems 21 (NIPS - 09). Vancouver, Canada, 1607– 1614.

[11] Lee, H., Kihm, J., Choo, J. (2012). iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. Computer Graphics Forum, 31(3) 1155-1164.

[12] Kabán, A., Girolami, M. A. (2002). A Dynamic Probabilistic Model to Visualise Topic Evolution in Text Streams. *Journal of Intelligent Information Systems*, 18(2-3) 107-125.

[13] Chua, F. C. T., Lauw, H. W., Lim, E. P. (2013). Generative Models for Item Adoptions Using Social Correlation. IEEE Transactions on Knowledge and Data Engineering, 25(9) 2036-2048.

[14] Lidong, Wang., Baogang, Wei., Jie, Yuan. (2012). Document Clustering Based on Probabilistic Topic Model. Acta Electronics Sinica, 2012, 40(11) 2346-2350.

[15] Wei, Zhang., Shuo, Xu., QiaoXiaodong. (2014). Are view of the development of topic models incorporating the internal and external features of scientific literature. *Journal of the China Society for Scientific and Technical Information*, 33(10) 1108-1120.

[16] Peng, Guan., Yufen, Wang., Zhu, Fu. (2016). Effect Analysis of Scientific Literature Topic Extraction Based on LDA Topic Model with Different Corpus. *Library and Information Service*, 2016, 60(2) 112-121.

[17] Lifa, Liao., Fugang, Le., Yalan, Zhu. (2017). The Application of LDA Model in Patent Text Classification. Modern Times, 2017, 37(3) 35-39.

[18] Jianghua, Liu. (2017). A Text Retrieval Method and Validation Based on kmeans Clustering Algorithm and LDA Topic Model, *Information Science*, 35(2) 16-21.

[19] Griffiths, T. L., Steyvers, M. (2004). Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(S1) 5228-5235.

[20] Hajjem, M., Latiri, C. (2017). Combining IR and LDA Topic Modeling for Filtering Microblogs. Procedia Computer Science, 112. 761-770.

[21] Duanwu, Yan., Zhiheng, Tao., Lanbin, Li. (2016). A Method of Automatic Recommendation of Subject Documents Based on HDP Model and Its Application. Information Studies: Theory & Application, 39(1) 128-132.

[22] Haohao, Tang., Bo, Wang., Yaoyi, Xi. (2015). Unsupervised Sentiment Orientation Analysis on Micro-Blogs Based on Hierarchical Dirichlet Processes. *Journal of Information Engineering University*. 2015, 16(4) 463-469.

[23] Teh, Y., Jordan, M., Beal, M. (2007). Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, 101(476) 1566-1581.

[24] Juan, Cao., Yongdong, Zhang., Jintao, Li. (2008). A Method of Adaptively Selecting Best LDA Model Based on Density. *Chinese Journal of Computer*. 2018. 31(10) 1780-1787.

[25] Peng, Guan., Wang Qi Fei. (2016). Research on the Method of Determining the Optimal Topic Number of LDA Topic Model in Scientific and Technological Information Analysis, *New Technology of Library and Information Service*, 2016(9) 42-49.

[26] Rodriguez, Alex., Laio, Alessandro. (2014). Clustering by fast search and find of density peak. Science, 344(6191) 1492-1496.