Research on text categorization model based on LDA – KNN

Weihua Chen, Xian Zhang
School of Computer Science and Technology, Wuhan University of Technology
Hubei, China
1542248882@qq.com,1012005403@qq.com

***Abstract*—In the text classification, The similarity between the text need to be calculated, but the existing classification methods only consider the similarity between feature words and categories and does not involve the semantic similarity between feature words. In this paper, a new classification model LDA (Latent Dirichlet Allocation) – KNN（K-Nearest Neighbor） is proposed. LDA is used to solve the problem of semantic similarity measurement in traditional text categorization. The sample space is modeled and selected by this model. In the reduced feature space, KNN classifier is used to classify the sample. The experiment was based on the Matlab software platform, and the data set was obtained from the Chinese corpus of Fudan University, and the high precision classification result was obtained with the average value of 0.933. LDA-KNN model is compared with MI(Mutual Information)-KNN model and LSI(Latent Semantic Index)-KNN model. The results show that LDA-KNN model has superior classification performance in automatic text categorization.**

Keywords—Text classification; LDA theme model; KNN classification algorithm; Feature select

I. Introduction

The rapid development of the Internet makes the emergence of a large number of data resources on the network, which accounts for the vast majority in the text form. In order to extract valuable information from these textual resources, the researchers applied data mining technology to the field ,thus form the current text mining technology. As the technology has a strong practicality, it has attracted the interest of many researchers.

The reason for the study of the content in this field comes from the actual business needs of a digital communication company in Wuhan. The company has a content editing and content accurate push platform, but all the content is manual, including the sorting, classification and uploading, so it is an urgent need to integrate an automatic and efficient text classification system to reduce the costs.

Current text categorization techniques tend to be automated and intelligent, and its primary goal is to automatically map text to pre-defined class labels based on the content of each document in the document set. In classification process, it involves the text preprocessing, feature extraction, classification algorithm and evaluation [1]. At present, there are a lot of research on each step, in the part of text representation and feature extraction: Vishwanath Bijalwan[2] and so on using VSM (Vector Space Model) model for text representation; Albitar S [3] uses the improved TF-IDF（Term frequency-inverse document frequency）to improve the efficiency of feature extraction. In addition, there are many studies on classification algorithms. For example, Zhengdong Lu [4] used an improved Boosting algorithm to improve the classification accuracy. The study of text categorization has achieved a great deal of achievements, but still faces with the following challenges: reducing the dimension of the sample space, effectively extracting and selecting the text features, designing the efficient classification algorithm under the large data set. Therefore, a new classification model named LDA-KNN is proposed to improve the accuracy of text feature selection for feature similarity measurement of feature words in the feature extraction phase, so as to improve the classification efficiency.

978-1-4673-8979-2/17/$31.00 ©2017 IEEE

II. LDA Topic Model and KNN Algorithm

LDA [5-6] model, proposed by Blei in 2003, is a probabilistic topic model that can be applied to text categorization to explore potential topic information in large-scale document sample sets. Its theoretical basis is: Each article is related to topic probabilities, and each topic is related to a number of specific vocabulary probabilities. However, it can not be directly used to classification, it is an unsupervised model, and the specific classification algorithm must be used for classification.

KNN algorithm was originally proposed by Cover, which occupies a large position in the machine learning method [7-8]. The core idea of KNN algorithm is to find the K samples which are closest to each other from the training samples, and then classify the samples according to their similarity. The advantage of KNN classification is that it is simple, effective, and low training cost. It is suitable for the overlapping sample sets. The disadvantage is that the computational cost is large and the interpretability of the output is not strong.

The classification can be executed by using the traditional text classification method. However, there is a large cost of classification due to the large sample space, and because the semantic similarity between terms is not considered in the feature selection stage, the extracted features are not obvious, Which can not represent the sample text very well, thus reduce the judgment accuracy on the feature set. Therefore, a new classification model LDA-KNN is proposed in this paper. LDA introduces a topic layer between lexical items and document categories, and consider the semantic similarity between them. The LDA theme model improve the efficiency of feature selection。 Using KNN classifier in the narrow feature set on the classification, you can get a good classification results.

III. LDA - KNN Model

A. Text representation

At present, there are many representation models for text, including two-dimensional view, term probability distribution and VSM model. In this paper, the most commonly used vector space model, in which the document set can be expressed as θ , which D represents the document set, d_i on behalf of each document, l represents the number of samples of the document, each document d_i can be expressed as $d_i = (t_1, t_2, \cdots; t_n)$, t_i represents each feature item of document d_i , n is the number of feature words in the document. For a text d_i that contains n features, it is usually necessary to compute its weight w_i . Each document can then be represented as containing a characteristic word and weight vector d_i , which can be expressed as follows:

$$d_i = (t_1, w_1; t_2, w_2; \cdots; t_n, w_n) \quad (1)$$

B. Text preprocessing

As the study object of this paper is unstructured or semi-structured, the characteristics of text datasets make them unable to be directly classified. Therefore, we firstly extracted the feature and then form a structured intermediate form.

For the Chinese text, the most important part of the text preprocessing is the word segmentation processing, it tends the content of the text paragraph structure into one of the words. The purpose of the word segmentation is to prepare for later processing. There are many specific word segmentation methods, so it is needed to choose the word segmentation scheme according to the completeness and completeness of the lexicon. The words and phrases obtained after the general word segmentation also need to be purified, and the words after the word segmentation need to be filtered by using the disabled vocabularies and high frequency vocabularies. Stop words refer to stop words, including conjunctions, prepositions, modal particles.The high-frequency words in the word table is usually underrepresented, It is also need to be removed.

C. Feature extraction

After the above processing, the result of document set is generally presented in the form of text vector space and category vector space, which has a high dimension and is unfavorable to the subsequent classification calculation. So we need to select the characteristic words. As a key part of the text classification system, feature extraction directly determines the time complexity of the subsequent classification and the accuracy of the classification results. The feature extraction method based on VSM model is based on weight statistics. The weight of feature item is calculated by a certain method and then the vector space with larger weight value is selected. For the feature selection of the class vector space, the mutual information, information gain and other mature methods are generally adopted.

D. LDA-KNN classification model

Usually using traditional text categorization method, you can get the appropriate results. However, traditional text feature extraction methods generally consider only the correlation between lexical items and document categories, and the less semantic similarity between lexical items is considered. Therefore, an improved feature extraction method-LDA topic-based model has been used to improve the effect of feature extraction.

The basic idea of LDA-KNN model is that compared with the traditional feature selection method, LDA introduces the topic layer between lexical item and document category, and considers the semantic similarity between the text's topic and improves the classification accuracy, Another advantage of LDA is that feature space is reduced in feature selection of document samples, and the KNN training time is cutted down.

LDA-KNN model includes the following modules: preprocessing, LDA topic modeling, feature word weight calculation, KNN classification, experimental results analysis. The specific structure shown in Figure 1:

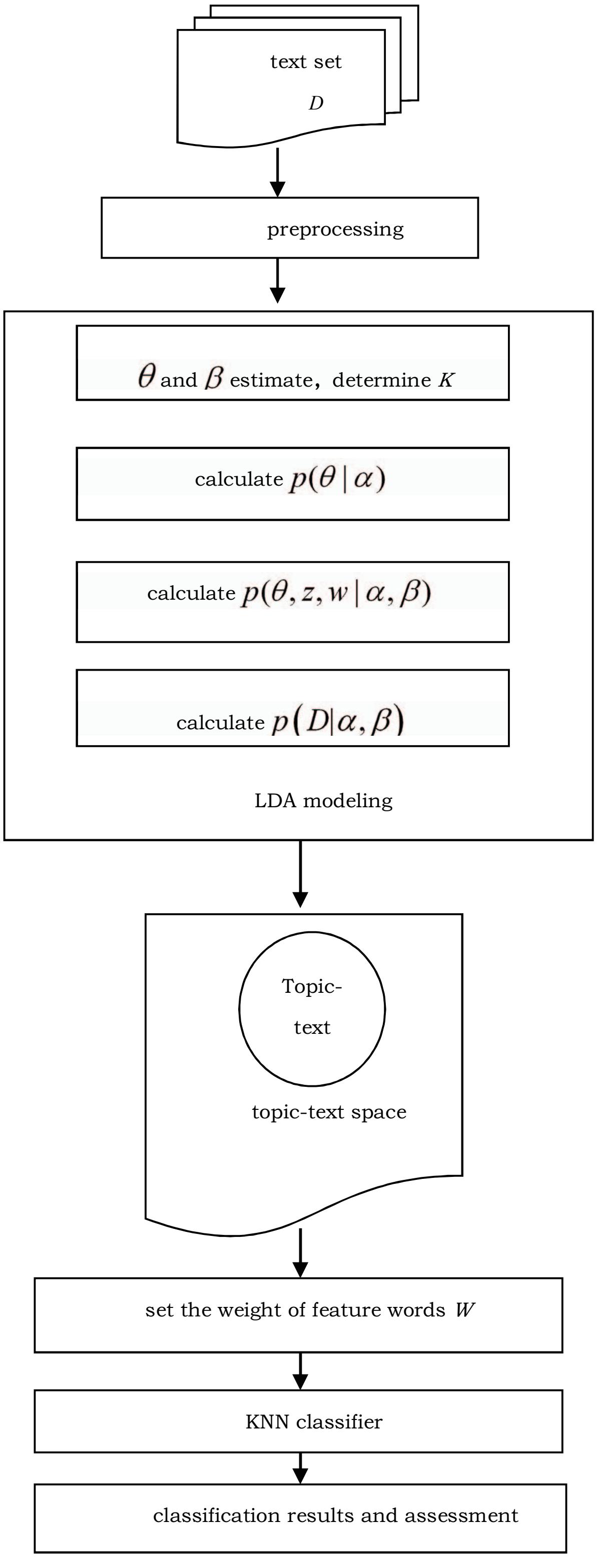

text set
D
preprocessing
θ and β estimate, determine K
calculate $p(\theta \mid \alpha)$
calculate $p(\theta, z, w \mid \alpha, \beta)$
calculate $p(D|\alpha, \beta)$
LDA modeling
Topic-text
topic-text space
set the weight of feature words W
KNN classifier
classification results and assessment

Fig. 1. Technical roadmap for the LDA-KNN model

1) LDA modeling

LDA model is used to do the feature extraction for the result after the text processing. Figure 2 shows the hierarchical structure of the LDA model. It consists of documents, topics, and terms. For each Child element in document set D, it can be thought of a distribution of topics. Each topic k can also be viewed as a distribution of several words. The word w_{dn} is the gray circle in the figure, which is the nth word of the dth document, $w_{dn} \in V$,V is the set of lexical items, z_{dn} represent the topic that generated w_{dn} , α is the parameter of the topic probability distribution of the document set, θ_d obeyes Dirichlet distribution $Dir(\theta_d \mid \alpha)$ .Topic ϕ_k is the distribution of the words in the set V; the distribution $\phi_{1:k}$ in the graph represents the distribution of the K topics in the word, and N represents the number of words contained in the document d. LDA is the topic probability model. If the number of subjects is K and the parameters α and $\phi_{1:k}$ is sure, the document generation process of the model of LDA topic is: 1) randomly select a vector θ_d from Dirichlet distribution $p(\theta \mid \alpha)$, the dimension is the number of topics, produce the probability distribution of the document d; According to $p(w_{dn} \mid \theta_d,\phi_{1:k})$ to produce each word of document d.

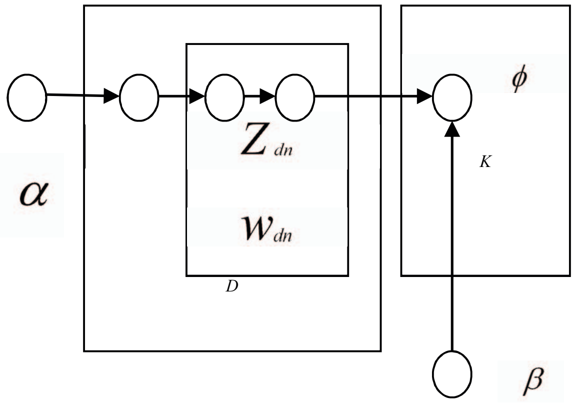


Fig. 2. Dirichlet distribution theme model

The process of generating words is the key point of LDA model. Before the generation process, parameter estimation is needed. The commonly used parameter estimation methods include Gibbs [9] sampling, variational reasoning and so on. In this paper, we take the Gibbs sampling method to estimate the number of subjects K and parameters α , and then obtain the probability distribution of keywords. The introduction of parameters β ,

$$p(d_i,c_j)=\sum_{d_j \subset KNN} Sim(d_i,c_j)\,y(d_j,C_j) \tag{2}$$

$$\phi_{mj} = \frac{C_{mj}^{VK} + \beta}{\sum_{m'} C_{m'j}^{VK} + V\beta} \tag{3}$$

$$\theta_{dj} = \frac{C_{dj}^{DK} + \alpha}{\sum_{j'} C_{dj'}^{DK} + K\alpha} \tag{4}$$

ϕ_{mj} is the probability of the current word m on topic j; θ_{dj} is the probability that the document d contains the topic j; C_{mj}^{VK} is a frequency matrix representing the number of occurrences of V characteristic words between K topics, and the value m represents the frequency of the word w_{di} in the subject j ; The frequency matrix C_{dj}^{DK} representing D documents between K topics, j' representing the number of topic j in document d; and $z_{di} = j$ indicates that the assigned topic of w_{di} is j.

Using LDA modeling, we can obtain the distributions of documents d_i on each topic in D and the distributio n of each topic under feature words., thus obtain the text-item matrix $M_{l \times s}$ of the document set D.

$$M_{l \times s} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1s} \\ m_{21} & m_{22} & \cdots & m_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ m_{l1} & m_{l2} & \cdots & m_{ls} \end{pmatrix} \tag{5}$$

Here, l is the number of samples and s is the dimension of the characteristic word.

2) Weight settings

The feature words in the text vector space are generally given a certain weight to reflect its importance. In this paper, TF-IDF is used to weight the feature words. The principle is as follows:

For Word t_i in a sample document, it’s weight can be expressed as follows：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (6)$$

Here， the word frequency $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$,

$n_{i,j}$ is the number that the word t_i occurrences in the document d_j. The denominator represents the sum of all the words in the document d_j.The reverse document frequency $idf_i = \log \frac{|D|}{|\{j : t_i \in d_i\}|}$, $|D|$ is the number of documents in the corpus, $|\{j : t_i \in d_i\}|$represents the number of documents that contain words t_i.

3) Training and classification algorithms

Text classification is not only a state of supervision but also a machine learning process. It can be divided into two stages including training and testing. The task of training is to produce a classifier with the experience and knowledge. Specifically, the class vector matrices are obtained by counting the vocabularies corresponding to the training samples, followed by normalization processing to obtain the processed vector matrices. The vector matrix can be used as a classification reference for later test phases. The testing phase needs to classify the test text set according to the knowledge learned in the training phase.

The detailed process of training phase is as follows: 1, do word segmentation to the training text set $S\{s_1, s_2, \cdots, s_n\}$ to get $(w_1, w_2, \cdots)$. 2, do the feature extraction according to the LDA model.3, determine the value of K the number of features extracted, and finally get the class vector matrix.

The detailed process of the classification phase is as follows: 1. For each document d_i in the set $D\{d_1, \cdots, d_i, \cdots, d_r\}$, calculate

$$p(d_i, c_j) = \sum_{d_j \subset KNN} Sim(d_i, c_j)\, y(d_j, C_j) \quad (7)$$

Here, $y(d_j, C_j)$ is used to determine whether the document d_j belongs to the class C_j. And $Sim(d_i, c_j)$ represents the similarity between the eigenvectors of the text d_i and c_j. The mathematical theory is usually calculated by the cosine of the angle between the eigenvectors of each other, which is:

$$Sim(d_i, c_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{\left(\sum_{k=1}^{M} W_{ik}^2\right)\left(\sum_{k=1}^{M} W_{jk}^2\right)}} \quad (8)$$

W_{ik} , W_{jk} represent the weights of the text d_i and c_j the k-th feature respectively.

2, Select the one that has the biggest similarity form the most similar K to the text as the classification of the test sample, that is, select the most similar category as the category of d_i.

4) Effect evaluation

In order to test the results of the experiment, the paper uses the common evaluation standard: Recall, Precision and F_1 Value.

Here,

$$R_i = \frac{M_i}{N_i} \quad (9)$$

$$P_i = \frac{M_i}{C_i} \quad (10)$$

$$F_1 = \frac{2 \times P_i \times R_i}{C_i} \quad (11)$$

M_i represents the number of documents classified correctly in class i, C_i represents the number of occurrences in the result, and N_i represents the number of documents actually included.

In addition, the paper also calculated the average recall avg_R of the entire document space, the average precision rate avg_P , the average avg_F_1 were:

$$avg\_R = \frac{1}{k}\sum_{i=1}^{k} R_k \quad (12)$$

$$avg\_P = \frac{1}{k}\sum_{i=1}^{k} P_k \quad (13)$$

$$avg\_F_1 = \frac{2 \times avg\_R \times avg\_P}{avg\_R + avg\_P} \qquad (14)$$

IV. Experimental design and analysis

A. Lab Environment

In order to test the classification effect of the model, the experimental data sets were collected from the Chinese corpus of Fudan University. Hardware environment: CPU for the Intel Core i5, clocked at 2.5GHZ, memory 4GB. Software environment: operating system Windows10, MATLAB software platform.

B. Data sets

In this paper, we use the Chinese corpus of Fudan University to carry out experiments. We selected Art, Education, Economy, Environment, History, Politics, Sports, Transport, and then randomly select the number of training and testing documents, as shown in Table 1:

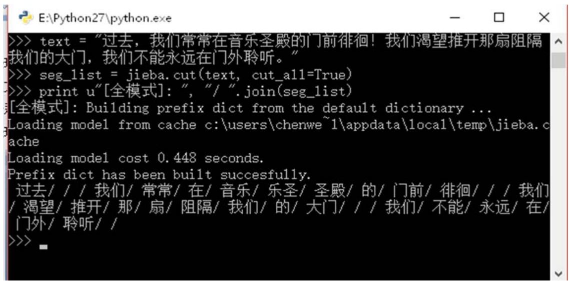
TABLE I. SAMPLE DATA

topic	training documents	testing documents
Art	100	92
Education	125	110
Economy	101	88
Environment	98	93
History	116	105
Politics	96	90
Sports	120	107
Transport	114	102
Total	870	787

C. Experimental setup

1) Text preprocessing

a) *We use the Jieba word segmentation system [10] to segment the 870 training data sets to get the word sequence. Here we select some parts of a document under the education class and input the text: "过去，我们常常在音乐圣殿的门前徘徊！我们渴望推开那扇阻隔我们的大门，我们不能永远在门外聆听。"The result of the processing after the word segmentation is shown in Figure 3:*



```
>>> text = "过去，我们常常在音乐圣殿的门前徘徊！我们渴望推开那扇阻隔
我们的大门，我们不能永远在门外聆听。"
>>> seg_list = jieba.cut(text, cut_all=True)
>>> print u"[全模式]: ", "/ ".join(seg_list)
[全模式]: Building prefix dict from the default dictionary ...
Loading model from cache c:\users\chenwe~1\appdata\local\temp\jieba.c
ache
Loading model cost 0.448 seconds.
Prefix dict has been built succesfully.
 过去/ / / 我们/ 常常/ 在/ 音乐/ 乐圣/ 圣殿/ 的/ 门前/ 徘徊/ / / 我们
/ 渴望/ 推开/ 那/ 扇/ 阻隔/ 我们/ 的/ 大门/ / / 我们/ 不能/ 永远/ 在/
 门外/ 聆听/ /
>>> 
```

Fig. 3. document word effect

b) *To do the processing of removing the Chinese stop word for the word sequence after the word segmentation(Including adverbs, conjunctions, prepositions, mood words), while removing the numbers, punctuation;*

c) *And then create documents - entry matrix for the word in second part, get the size of the matrix.*

2) LDA modeling

In this paper, the Gibbs sampling algorithm described above is used to estimate the model parameters, and 10-fold cross validation is used to evaluate the model. The topic K is set in the arithmetic sequence of the interval value $[1,30]$ to find the optimal topic number K . Judging by the subjective confusion of Gibbs sampling, the average confusion degree is the lowest when the topic K is 8.

The value of α is obtained according to the formula $50/K$, we can get $\alpha = 6.25$, β is usually 0.01. And then calculate the probability distribution of each sample under the K topic set, and get the text-term matrix $M_{l \times s}$.

3) KNN training

The TF-IDF method is used to set the weights of the feature words, and then the remaining 95% feature words are removed. The following table2 shows the characteristics of the word list, the table shows only some of the characteristics of words:

TABLE II. IS A PARTIAL FEATURE WORD SORTED BY WEIGHT

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
教育	历史	经济学	艺术	体育	环境	政治	汽车
爱国	史学家	财经	语言	体质	森林	制度	铁路
中国	现代	贫困	感官	运动	水土	民主	公路
学校	社会	流通	形象	肌肉	城市	人权	桥梁
学生	思想	市场	音乐	身体	水库	政党	机动车

文化	文物	资源	文学	动作	风景	阶级	线路
…	…	…	…	…	…	…	…

The text-term matrix $M_{l\times s}$ on the reduced feature space is classified by KNN classifier and then tested.

D. Experimental results and analysis

In this experiment, the performance of LDA-KNN model is compared and verified from two aspects. All the experiments are based on the same data set, the same experimental platform.

(a) After pre-processing, the LSI and MI are used to classify the samples, and the results are listed in Table 3:

TABLE III. COMPARISON OF FEATURE SELECTION METHODS

Classify method	*avg_R*	*avg_P*	*avg_F*
LDA-KNN	0.938	0.932	0.933
LSI-KNN	0.903	0.897	0.900
MI-KNN	0.893	0.887	0.890

TABLE IV. COMPARISON OF CLASSIFICATION METHODS

Classify method	*avg_R*	*avg_P*	*avg_F*
LDA-KNN	0.938	0.932	0.933
LDA-NB	0.890	0.910	0.901
LDA-J48	0.935	0.931	0.932

(b) Feature selection method is LDA algorithm, and classification methods were KNN, naive Bayesian, decision tree , decision tree is J48. The experimental results obtained are shown in Table4:

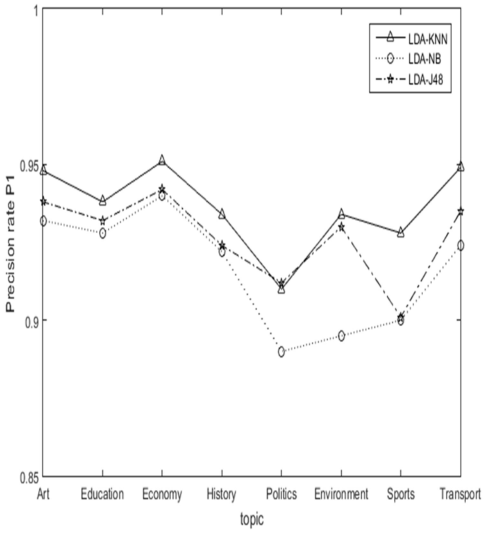


Fig. 4. Comparison of superiority of KNN

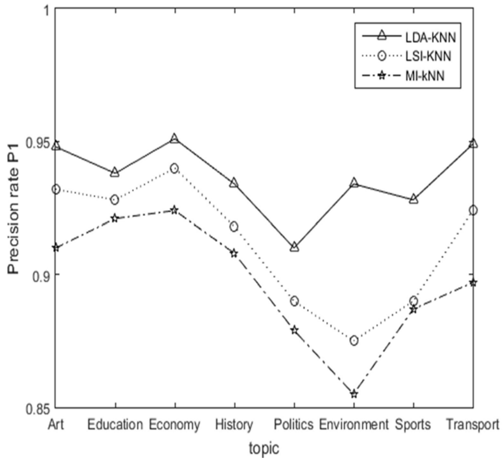


Fig. 5. LDA superiority comparison

The dashed line represents the LDA-NB, and the dotted line represents the LDA-J48. All the models in Figure 4 can be classified differently in different categories. For example, the LDA-KNN model is represented by a solid line triangle, The specific performance is related to the content of the text, the number of documents under each category, but LDA-KNN model in each category, compared with other models, the classification effect is better. In Figure 5, the solid-line triangle represents the LDA-KNN model. The dashed line represents the LSI-KNN and the dotted line represents the MI-KNN. As can be seen from the figure, the LDA-KNN model performs better in all categories. The above experiments verify the efficiency of the proposed classification model.

V. Conclusion

The LDA-KNN model proposed in this paper combine the simplicity and fastness of KNN and the similarity measure of LDA and the reduces feature space to improve the performance of text categorization. The experiments were done through building corpus, sample preprocessing, LDA modeling, and training and testing with KNN algorithm. The experimental results show that the model performs well in classification, but the model needs to be improved in terms of time efficiency. The next work will focus on controlling the time cost of the model and guaranteeing the classification accuracy.

Acknowledgements

This paper is supported in part by the National Natural Sci ence Foundation of China (NO. 61602353).

References

[1] Hui Zhang, Deqing Wang, Li Wang, et al. A semantics-based method for clustering of Chinese web search results[J]. Enterprise Information Systems, 2014, 8(1):147-165.

[2] Vishwanath Bijalwan，Vinay Kumar，Pinki Kumari，et a1．KNN based Machine Learning Approach for Text and Document Mining[J] ．International Journal of Database Theory and Application，2014，1(7)：61-70.

[3] Albitar S, Fournier S, Espinasse B. An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification[M]// Web Information Systems Engineering – WISE 2014. Springer International Publishing, 2014:105-114.

[4] Zhengdong Lu，Cane Wing-ki Leung，Qiang Yang．Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining，Chicago，August 11 – 14，2013[C]，Chicago:KDD2013，2013．

[5] Lauren R．Biggers，Cecylia Bococich，Riley Capshaw，et a1．Configuring latent Dirichlet allocation based feature location [J] ．Empirical Software Engineering，2014，3(19) ：465-500．

[6] Yuan S, Qian Z. Tibetan-Chinese Cross Language Text Similarity Calculation Based on LDA Topic Model[J]. Open Cybernetics & Systemics Journal, 2015, 9(1):2911-2919.

[7] Mathmoud Mejdoub，Chokri Ben Amar. Classification improvement of local feature vectors over the KNN algorithm[J] ．Springer Link，2013，5(64) ：197-218．

[8] Wang Meng，Lin Lanfen，Wang Jing，et a1．Improving Short Text Classification Using Public Search Engines[J]．Integrated Uncertainty in Knowledge Modelling and Decision Making，2013，vol 8032：157-166．

[9] Tang Hong，Shen Li，Qi Yinfeng，et a1．A Multiscale Latent Dirichlet Allocation Model for Object-Oriented Clustering of VHR Panchromatic Satellite Images[J]．IEEE Transactions on Geoscience and Remote Sensing，2013,51(3)：1680-1692．

[10] Ni C, Leung C C. Investigation of using different Chinese word segmentation standards and algorithms for automatic speech recognition[C]// International Symposium on Chinese Spoken Language Processing. IEEE, 2014:44-48.