International Journal of Database Theory and Application
Vol.9, No.7 (2016), pp.51-66
http://dx.doi.org/10.14257/ijdta.2016.9.7.06

Employing Latent Dirichlet Allocation Model for Topic Extraction of Chinese Text

Qihua Liu

School of Information Technology, Jiangxi University of Finance and Economics
qh_liu@163.com

Abstract

The hidden topic model of Chinese text, which possesses complicated semantics, is urgently needed, since China has occupied an increasingly significant role during the booming development of globalization over recent years. This paper details and elaborates the basic process of extracting latent Chinese topics by demonstrating a Chinese topic extraction schema based on Latent Dirichlet Allocation (LDA) model. Furthermore, the application was practiced in CCL, an authoritative Chinese corpus, to extract topics for its nine categories. With rigorous empirical analysis, extracting the LDA results has a considerably higher average precision rate as opposed to other three comparable Chinese topic extraction techniques; however the average recall rate is worse than KNN and almost the same with the PLSI model. Moreover, the recall rate and precision rate of LDA-CH is worse than LDA-EH. Therefore, the LDA model should be improved to adapt to the distinctive feature of Chinese words with the purpose of making it better for Chinese topic extraction.

Keywords*: LDA, topic model, extraction schema, Chinese text, Gibbs sampling*

1. Introduction

Topic extraction means to extract central meaning of the given texts in the subject field (a corpus or a catalog) to make them easy to identify and classify [1]. This is significantly different from "keyword extraction" in that any important words/phrases can be regarded as keywords in a number of fields, while topic terms must come from the corpus's thesaurus. Therefore, extracting keywords is a potentially large and irrelevant undertaking, while topic extraction deals with a smaller, more relevant number of topics.

Topic extraction is widely used in various fields, such as text classification, similarity calculation and collaborative filtering. Rapid search and retrieval can be achieved by classifying and filtering topic words, while the field of text can be effectively identified by calculating the similarity of topic words, in addition to a collaborative filtering of topic words, automatically excluding any text not belonging to specific areas.

In pace with the internationalization of China, Chinese resources are expanding to an enormous scale, but most of them are disordered and poorly organized despite their expansion. In order to organize, retrieve, classify and cluster the Chinese resources validly, the topic extraction method is urgently needed to discover the hidden topic model of numerous Chinese resources [2].

However, Chinese words place more emphasis on the meaning of a word rather than its patterns, so the combination and function of Chinese words embody considerable flexibility, which is different from other languages. What follows are concrete manifestations.

First, Chinese words possess varied functions, for the parts of speech are changeable. For instance, the adjectives, which aren't restricted to rigid grammar rules, are able to switch to verbs under any demanding circumstances. It also has strong capacity of word-formation, which means the Chinese language has the freedom of creation and extension.

ISSN: 2005-4270 IJDTA

According to authoritative statistics, the Chinese language contains 44300 common items, but only uses 3730 single words. Finally, the language possesses complex semantics. It is pretty obvious in terms of its volatile parts of speech and flexible word-building that any Chinese words can be put together as long as the association is reasonable, logical, and semantic. Given the above features, topic extraction of Chinese text remains a fairly difficult theme in academic research.

The initial idea of the LDA (Latent Dirichlet Allocation) model is based on the unigrams model, the mixture of unigrams, latent semantic indexing and probabilistic latent semantic indexing, which was first introduced by Blei *et al.* (2003) [3]. The model had achieved outstanding performance on the topic extraction of English texts, but it rarely applied to Chinese topics. According to this, our paper presents a schema based on LDA to extract topics in Chinese corpus-CCL. The extraction results were compared to other three topic modeling techniques. Recall rate, precision rate and F1are used to evaluate the extracting results.

The remainder of this paper is organized as follows. Section 2 reviews literature. A brief introduction of the LDA model and detailed descriptions of our schema are given in Section 3.Section 4 manifests some relevant preparations for experiments. With empirical analysis, the results and findings are exhibited in Section 5 and the last section illustrates some concluding remarks and our future work.

2. Literature Review

2.1. Research of Topic Extraction for English Text

Topic extraction is of great importance in the area of computer science, information systems and artificial intelligence. Many scholars have put forward a series of methods to extract the hidden topic model, among which the most typical one is the LDA model. Blei *et al.* (2003) has accomplished a great deal of meaningful research on the subject [3]. Griffiths and Steyvers (2004) conducted a further study using the Gibbs sampling method to extract dynamic topics varying over time and annotating semantic texts [4]. As for the case study, Maskeri and Sarkar (2008) investigated the LDA in the context of comprehending large software systems and proposed a human assisted approach for extracting domain topics from source code [5]. In the original work by Wu *et al.* (2010), a topic-level Eigen factor algorithm was created to assess the relative importance of academic entities by applying the LDA model [6]. Momtazi and Naumann (2013) propose a topic modeling approach for the task of expert finding [1]. The proposed model uses latent Dirichlet allocation (LDA) to induce probabilistic topics. Li *et al.* (2014) propose an extension of L-LDA, namely supervised labeled latent Dirichlet allocation (SL-LDA), for document categorization [7]. Lee *et al.* (2015) presented a news topics categorization method using latent Dirichlet allocation (LDA) and sparse representation classifier (SRC) [8].

Besides these, many other methods were applied to extract topics as well. Nauda and Usui (2008) presented a preliminary analysis of the neuroscience knowledge domain, and an application of cluster analysis in order to identify topics in neuroscience [9]. Topics were extracted from the abstracts of posters by clustering the documents using a bisecting k-means algorithm and then selecting the most salient terms for each cluster by rank. Nakatsuji *et al.* (2009) combined topic extraction with ontology technology, and generated user-interest ontology to measure the similarity between user interests, and finally extracting innovative topics based on user-interest ontology [10]. Ouyang *et al.* (2010) carried out a study of a sentence ranking approach, which was based on inter-topic information mining using a *K*-medoids algorithm to cluster the word tokens in all the topics in order to assign the conceptual labels to the words [11]. Srivastava *et al.* (2013) proposed a graph-based topic extraction algorithm, which can also be viewed as a soft-

clustering of words present in a given corpus [12]. Saga *et al.* (2014) proposed a new measurement called topic coverage on the basis of the assumption that the keywords extracted by a superior method can express the topic information efficiently [13]. Bastian *et al.* (2014) presented their experiences developing this large-scale topic extraction pipeline, which included constructing a folksonomy of skills and expertise and implementing an inference and recommender system for skills [14]. They also discussed a consequent set of applications, such as Endorsements, which allowed members to tag themselves with topics representing their areas of expertise and for their connections to provide social proof, via an "endorse" action, of that member's competence in that topic [14].

2.2. Research of Topic Extraction for Chinese Text

Due to the characteristic of Chinese texts, the above approaches may be not suitable for Chinese topic extraction. The history of Chinese topic extraction is divided into two periods by the usage of Chinese word segmentation. Representing the first period was a novel Chinese text subject extraction method based on character co-occurrence put forward by Ma *et al.* (2003) [15]. Neither word segmentation nor word extraction was required in this method, and it could also be used to process Multilanguage text, but the recall rates could not be guaranteed. When Chinese word segmentation was involved, many techniques were added to extract latent Chinese topics. Chen and Zhang (2005) illustrated a novel Chinese text subject extraction method based on word clustering [16]. The method used relativity calculation to create semantic relativity, which generated a word cluster, and its subject genes were then extracted by means of that word cluster. Liu (2007) applied association rules to mining thematic terms of Chinese text. Thematic terms consisted of key phrases and related terms [17]. But neither of the two methods above provided high precision rates. Xie *et al.* (2011) proposed an approach to implementing ontology-based data access in WordNet with the distinguishing feature of optimizing density-based clustering OPTICS algorithm (DBCO) to extract topics [18]. Tian *et al.* (2012) made experiment in Chinese Event Corpus CEC, and presented a method of extracting event trigger word automatically that combined extended trigger word table and machine learning [19]. Liu *et al.* (2012) presented an efficient adaptive focused crawling framework based on LDA and domain ontology, which integrates the content analysis stage, the link analysis stage and the relevance computation stage [20]. An and Liu (2013) combined domain ontology and LDA model to propose a new method of hierarchical web text classification [21]. Li and Xu (2014) proposed and implemented a novel method for identifying emotions in microblog posts [22].

Because of the excellent performance of the LDA model, more and more Chinese scholars put it into various aspects, including of text segmentation, automatic summarization, and topic evolution. Shi *et al.* (2008) demonstrated a text segmentation method based on LDA model corpora and texts with LDA [23]. Parameters are estimated using Gibbs sampling of MCMC and the word probability was represented. Yang *et al.* (2010) proposed a multi-document summarization method based on the LDA model; the number of topics was determined by model perplexity and used the Gibbs sampling method to estimate parameters, and the topic importance was determined by the sum of topic weights on all sentences [24]. Chu and Li (2010) discovered the topic's evolution over time by topic detection and relating topics in different time periods, and then applied the LDA model on temporal documents in order to extract topics [25]. Fu *et al.* (2013) proposed an unsupervised approach to automatically discover the aspects discussed in Chinese social reviews and also the sentiments expressed in different aspects [26]. They applied the LDA model to discover multi-aspect global topics of social reviews, and then extracted the local topic and associated sentiment based on a sliding window context over the review text. Cui *et al.* (2014) proposed a novel algorithm based on the co-occurrence

of the visual and annotation words. In this paper, the LDA topic model is used to cluster the images [27].

In sum, the LDA model has a wide application in English topic extraction and performs well. Chinese scholars have extended the model to various aspects, also gaining satisfactory achievement, but little concern has been given to the topic extraction of Chinese text based on the LDA Model. Regarding the metrics aspect of topic models, applying the LDA model to the topic extractions in Chinese text is inevitable. Therefore, this paper will combine Chinese segmentation with the LDA model to extract scientific topics from Chinese text.

3. Methodology

3.1. LDA Model

The LDA model is the typical representative of topic models. LDA is a generative probabilistic model for collections of discrete data such as text corpora [3]. Documents are represented as random mixtures over latent topics, and each topic is then characterized by a distribution over words, shown in Figure 1. The text generative model describes the generative process of words through documents based on latent variable and simple probabilistic sampling rules, and probabilistic topic models have been used to analyze the topic structure of given texts and the implication of each word.

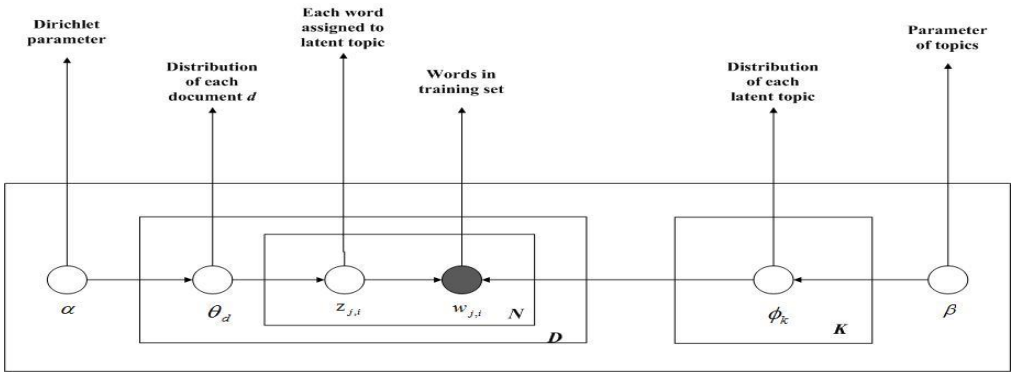

Figure 1. Basic Idea of LDA Model

The LDA model is also considered a bag-of-words model with three levels displayed in Figure 1. As the figure clearly shows, the corpus contains a collection of *D* documents. Each document *w* consists of *K* topics, and each topic *k* is characterized by a distribution over *N* words. The generative process for each document *w* in a corpus *D* is as follows [3] [20-21]:

(1) Choose $\phi \sim$ Dir (β): For each topic *k*, sampling a multinomial distribution ϕ_k subordinates a prior Dirichlet distribution parameterized by β , a distribution of *K* were taken.

(2) Choose $\theta \sim$ Dir (α): For each document *d*, sampling a multinomial distribution θ_d subordinates a prior Dirichlet distribution parameterized by α , a distribution of *D* was taken.

(3) For each of the document *d* in corpus and all words w_{d_i} in document d_i

(a) Choose topics $z_i \sim$Multinomial(θ): Extract topic z_i from θ_d

(b) Choose words $w_i \sim$Multinomial(ϕ): Extract topic w_i from ϕ_k

Therefore, the probability of a corpus based on LDA model can be presented as follows:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{D} \int p(\theta_d \mid \alpha) \left[\prod_{n=1}^{N_d} \sum_{z_{d_i}} p(z_{d_i} \mid \theta_d) p(w_{d_i} \mid z_{d_i}, \beta) \right] d\theta_d \quad (1)$$

K is the number of latent topics and D is the number of documents. The parameters α and β are hyper parameters in the corpus level, having been sampled once in the process of generating a corpus. The variable θ_d is a document-topic distribution sampled once per document, and the variable ϕ_k is a topic-word distribution sampled once per latent topic. Finally, z_n and w_n are word-level variables sampled once for each word in each document.

3.2. Topic Extraction Model for Chinese Text Based on LDA

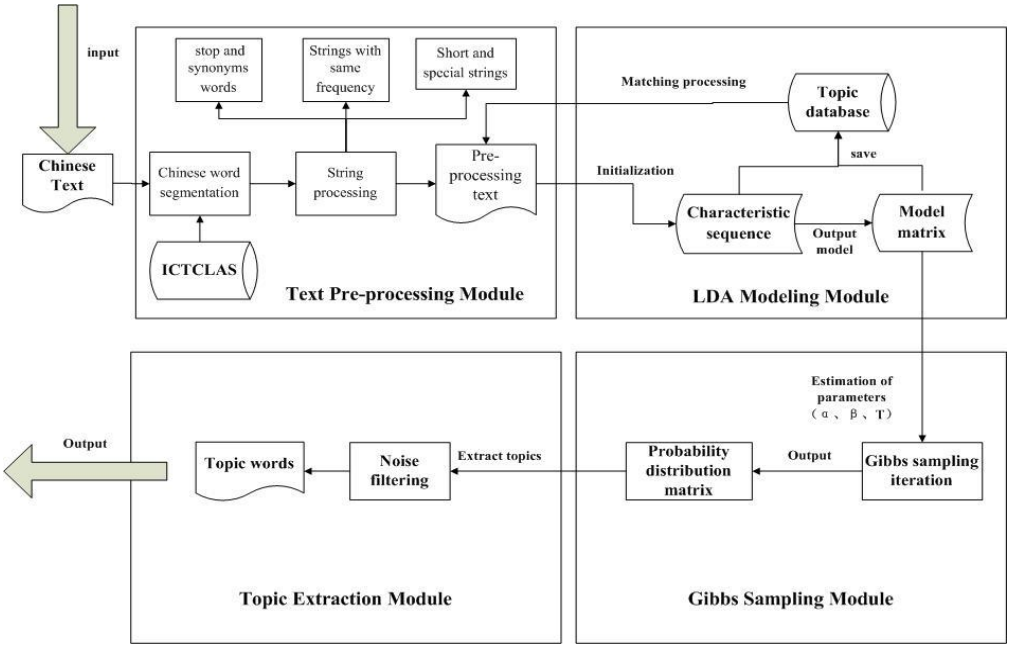

Figure 2. Flow Chart of Chinese Topic Extraction Based on LDA

Based on the LDA model, we summarized a topic extraction model for Chinese texts, which is shown in Figure 2 [2]. The model is made of the following four modules: a text pre-processing module, an LDA topic modeling module, a Gibbs sampling module and a topic extraction module [2].

3.2.1. Text Pre-Processing Module

The most distinctive trait for Chinese topic word extractions according to the LDA reflects on the text pre-processing module due to the diversity of Chinese words and their contextual semantic meanings.

Because Chinese retains its own characters of various structures, stop words and synonym processing are needed. Some meaningless modal particles and prepositions, as well as words with high frequency but that are insufficient to describe the main content of the articles will be left out, including numbers, individual characters, and some function words.

Next the segmentation of Chinese text must be implemented due to the superior capacity of Chinese word-formation and semantics. Currently, the existing two segmentation methods are based on ontology and thesaurus matching respectively; the latter is relatively mature while the former still calls for sufficient experimental practices, so we have adopted the latter, the exponent of which is representative of ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System)-a precise Chinese word segmentation system developed by the Institute of Computing, CAS (Chinese Academy of Sciences).

Finally, the processed text will be scanned three times using statistical methods in order to find all strings with a frequency greater than two, while at the same time, deleting short strings and special strings, and eventually generating the pre-processing text.

3.2.2. LDA Topic Modeling Module

The LDA topic-modeling module is the key part of the Chinese topic extraction process. The basic principle of the LDA model has already been introduced, and the specific steps are as follows:

First, initialize the pre-processing text, thus, match the text with the model matrix in the topic database in order to select the most suitable model entering the training set;

Afterward, the training set will generate characteristic sequences, which is the data source for LDA modeling.

Then, according to the basic principle of the LDA model, the characteristic sequence will be modeled to generate a topic-word model matrix.

Finally, output and save the matrix to the topic database in order to prepare for the next round.

The core codes are as follows:

```
Public class LDA implements
{
   Int K; // number of topics
   Double α;  //α is the hyperparameter; Dirichlet(α) represent the distribution of topics
   Double β; // β is the hyperparameter; Dirichlet(β) represent the topic-word distribution
   Double tAlpha;
   Double vBeta;
   InstanceList ilist;  //  Characteristic sequence
   Public void estimate (InstanceList documents, int numIterations, int showTopicsInterval,
                         int outputModelInterval, String outputModelFilename, Randoms r)
   {
      ilist = documents.shallowClone();
      topics = new int[numDocs][];
      tAlpha =α * K;
      vBeta =β* K;
      long startTime = System.currentTimeMillis(); // Initialization
      int topic, seqLen;
      FeatureSequence fs;
      seqLen = fs.getLength();
      numTokens += seqLen;
      topics[di] = new int[seqLen];
      for (int si = 0; si < seqLen; si++)
      {
          topic = r.nextInt(numTopics);
          topics[di][si] = topic;
          docTopicCounts[di][topic]++;
          typeTopicCounts[fs.getIndexAtPosition(si)][topic]++;
          tokensPerTopic[topic]++;
        }
   }//  Generating characteristic sequence
   Public void sampleTopicsForAllDocs (Randoms r)
   {
      double[] topicWeights = new double[numTopics];
      for (int di = 0; di < topics.length; di++)
```



```
            {
                sampleTopicsForOneDoc ((FeatureSequence)ilist.get(di).getData(), topics[di],
docTopicCounts[di], topicWeights, r);
            }
        }// Modeling characteristic sequence
    }
```

3.2.3 Gibbs Sampling Module

We used Gibbs sampling to carry out our parameter estimation. Gibbs sampling is a special case of the Markov chain Monte Carlo; it simulates a high-dimensional distribution by sampling lower-dimensional subsets of variables, which are conditioned on the value of all others. It is easy to implement and provide an approach that can extract a set of topics from a relatively large corpus [4]. The sampling will iterate by sequence until the Markov chains converge to the target distribution. Instead of estimating θ and ϕ explicitly like EM, the Gibbs sampling method considers the posterior distribution over the assignments of words to topics, $P(z \mid w)$, to indirectly obtain the parameters θ and ϕ. Gibbs sampling traverses all the words in the document in turn. Current word w_i is assigned to topic *j* conditioned on the topic assignments to all other word tokens, then marginalizes θ and ϕ to get their estimates. Griffiths and Steyvers (2004) showed how this can be calculated by [4]:

$$P(z_i = j \mid z_{-i}, w_{d_i}, \alpha, \beta) = \frac{\frac{N^{WK}_{w_{-i},j} + \beta_{w_i,j}}{\sum_{w=1}^{W}(N^{WK}_{w_{-i},j} + \beta_{w,j})} \times \frac{N^{DK}_{d_{-i},j} + \alpha_{d,j}}{\sum_{k=1}^{K}(N^{DK}_{d_{-i},k} + \alpha_{d,k})}}{\sum_{j=1}^{K}\frac{N^{WK}_{w_{-i},j} + \beta_{w_i,j}}{\sum_{w=1}^{W}(N^{WK}_{w_{-i},j} + \beta_{w,j})} \times \frac{N^{DK}_{d_{-i},j} + \alpha_{d,j}}{\sum_{k=1}^{K}(N^{DK}_{d_{-i},k} + \alpha_{d,k})}} \quad (2)$$

The left part is the probability of word *i* under topic *j*, where $z_i = j$ refers to the topic assignment of token *i* to topic *j*; z_{-i} refers to all other word tokens excluding token *I*; w_{d_i} represents all words in document *d*; α and β are hyperparameters with observed information.

The right part is the approximate probability that topic *j* is under current distribution for document *d*, where N^{WK} and N^{DK} are matrices of counts with dimensions $W \times K$ and $D \times K$ respectively. *W* refers to the number of words, *K* refers to the number of latent topics, $N^{WK}_{w_{-i},j}$ represents the number of times word *w* is assigned to topic *j*, excluding the current token *i*, $N^{DK}_{d_{-i},j}$ represents the number of times topic *j* is assigned to document *d*, excluding the current instance *i*.

3.2.4. Topic Extraction Module

The Gibbs sampling method directly estimates topics *z* for each word. The values of θ and ϕ can be calculated by:

$$\hat{\theta}^{(d)}_{z=j} = \frac{N^{DK}_{d_{-i},j} + \alpha_{d,j}}{\sum_{k=1}^{K}(N^{DK}_{d_{-i},k} + \alpha_{d,k})} \quad (3)$$

$$\hat{\phi}_{w_i}^{(z=j)} = \frac{N_{w_{-i},j}^{WK} + \beta_{w_i,j}}{\sum_{w=1}^{W}(N_{w_{-i},j}^{WK} + \beta_{w,j})} \quad (4)$$

$\hat{\theta}_{z=j}^{(d)}$ is a document-topic distribution sampled once per document, and $\hat{\phi}_{w_i}^{(z=j)}$ is a topic-word distribution sampled once per latent topic.

The last part comes to the topic extraction module. It will take a series of post-processing of the extracted topics following the three basic steps:

First, extract a set of topics from the generated topic-word matrix. Next, improve the precision of extracted topics via noise filtering, and finally, output the topics.

The principle of noise filtering is to count topic words, the occurrence of which has a frequency greater than the threshold value, deleting the lower ones.

4. Experiments

The corpus of our experiment is one of the most prestigious Chinese corpuses -- the CCL (Center for Chinese Linguistics Peking University) corpus (http://ccl.pku.edu.cn:8080/), which is manually written by scholars at Peking University. It consists of 1696 articles in9 categories; our experiment would implement 9 times the extracted latent topics for all the categories that were regarded as a document. The number of words in each document is displayed in Table 1.

Table 1. Number of Words in Each Category

Categories	Number of words	Categories	Number of words
Category1 (Press)	129,615,553	**Category6** (Drama)	408415
Category2 (Practical writing)	502,364	**Category7** (Biographical)	14,475,420
Category3 (Translation works)	78,162,797	**Category8** (Online text)	29,023,002
Category4 (Literature)	102,316,537	**Category9** (Spoken language)	7,026,370
Category5 (Movies)	133,909	**Total**	361,664,367

4.1. Evaluation Merits

Generally speaking, the most extensive approach used to evaluate topic extraction results is precision rate. In addition, we added recall rates and F1 to make the comparison more diverse and creditable.

After Chinese word segmentation, we manually marked some key words (no more than 6) for each category, if the current extracted topic word was similar or synonymous with one of the marked words, it was a correct assignment. Furthermore, the specific definitions of precision rate, recall rate and F1. Among the words assigned to current topic, we assumed the number of correct assignments is *a*, and assumed the number of wrong assignments is *b*. Among the words not assigned to current topic, the number of words that should be assigned to current topic is *c*, and the ones that should not be assigned is *d*, which are shown in Table 2, and then $r = \frac{a}{a+c}$, $p = \frac{a}{a+b}$, $F1 = \frac{2 \times r \times p}{r+p}$.

Table 2. Number of Words in Different Collections

Number of words	Correct assignments	Wrong assignments
Assigned to current topic	A	b
Not assigned to current topic	C	d

4.2. The Value of Observed Parameters

Hyperparameters α , β and number of topics K are either known or observed information, which must be fixed to facilitate the observation and evaluation of results. Griffiths and Steyvers (2004) [2] suggested $\beta = 0.1,\ \alpha = 50/K$. As for K, we computed the posterior probability distribution $P(w \mid K)$ for K values of 10, 20, 30, 40, 50, 60 and 100. For each values of K, we ran a Markov chain, iterated 1000 times, and took 10 samples from the last 100 iterations, because only the last 100 iterations are effective ones. The estimation results of $P(w \mid K)$ are shown in Figure 3. $P(w \mid K)$ initially increases as a function of K, reaches a peak at K=50, and then decreases thereafter, so data are best accounted at the point of K=50.Therefore, for each category in CCL, the number of topic words are K=50.

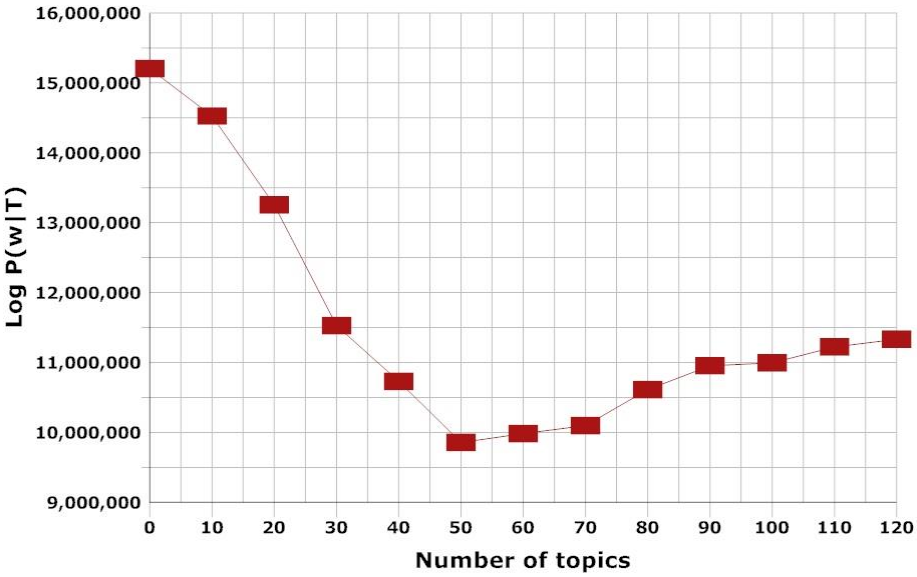

Figure 3. The Value of $P(w \mid K)$ for Different *K*

4.3. Other Compared Topic Modeling Techniques

For comparison, we also used a KNN algorithm, a TF-IDF schema and a PLSI model to extract topics for our corpus. The KNN algorithm is one of the best algorithms in topic extraction aspect by using vector space models. The TF-IDF schema has applications in information retrieval and text mining. PLSI is an aspect model of two-mode and co-occurrence data based on a mixture decomposition derived from a latent class model.

4.3.1. Basic Idea of KNN Algorithm

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

4.3.2. Basic Idea of TF-IDF Schema

The TF-IDF of a given word for a given document can be represented as follows:

$$tf_i idf_{i,j} = tf_{i,j} \times idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|d : d \ni t_i\}|} \quad (5)$$

Where $n_{i,j}$ represents the number of occurrence for word t_i in document d_j, $\sum_k n_{k,j}$ refers to the number of occurrence for all the word in d_j, $|D|$ refers to the total number of documents in corpus, and $|d : d \ni t_i\}|$ refers to the number of documents involving word t_i.

According to the principle of TF-IDF, it tends to filter common words and reserves relatively important words.

4.3.3 Basic Idea of PLSI Model

Probabilistic latent semantic indexing (PLSI) is a statistical technique for the analysis of two-mode and co-occurrence data based on a mixture decomposition derived from a latent class model [29]. PLSI has applications in information retrieval and filtering, natural language processing, machines learning from text, and other related areas. The representation of the PLSI model is exhibited in Figure 4.

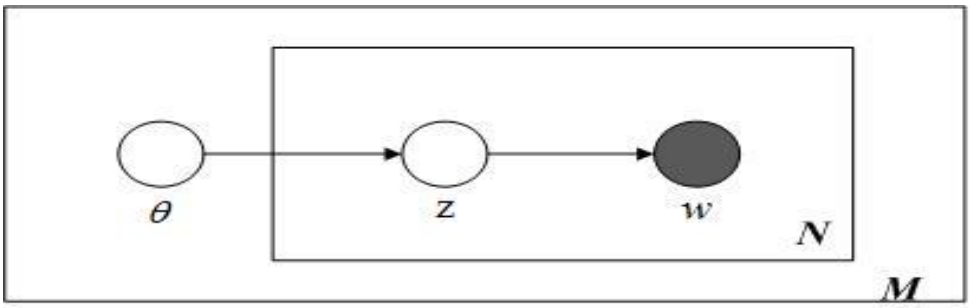

Figure 4. Representation of PLSI Model

5. Results

After pre-processing all the articles in category1, we got a vocabulary of 120,113,024 words, which were modeled by Equation (1). Taking *K*=50, the parameters were estimated by using Equation (2), (3) and (4), finally outputting the topic-word matrix. The number of correct assignments, total assignments and the precision rate for each topic above in category1 are exhibited in Table 3.

Table3. Number of Correct Assignments, Total Assignments and Precision Rate for Ten (Out of 50) Topics Extracted from Category1 in CCL Corpus

	Correct assignments	Total assignments	Precision rate(%)
Topic14	685611	2135861	32.1%
Topic22	661833	1975620	33.5%
Topic20	628543	2014562	31.2%
Topic41	723928	2241264	32.3%
Topic32	682290	2042785	33.4%
Topic02	625938	1902548	32.9%
Topic17	709603	2156851	32.9%
Topic25	734658	2310248	31.8%
Topic06	579844	1876521	30.9%
Topic39	640077	1963425	32.6%

5.1. Comparison with Other Techniques

Meanwhile, taking *K*=50 as well, we conducted the experiment for all nine categories by using the four different topic extraction techniques (including LDA) introduced above. The average precision rate, recall rate and F1 value for each were manually calculated, and the compared results are shown in Figure 5, Figure 6 and Figure7 respectively.

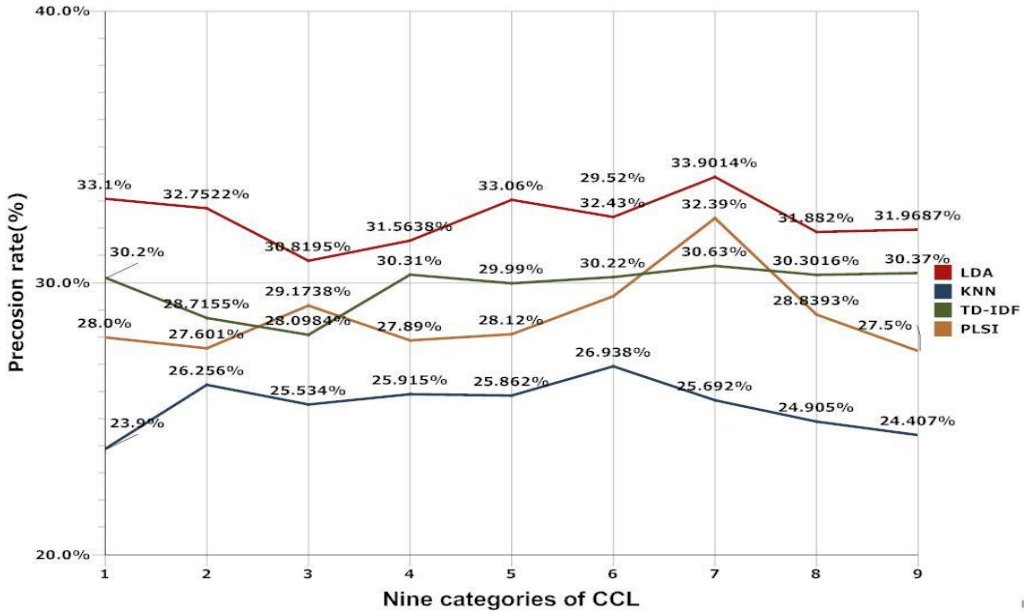

Figure 5. Average Precision Rate Performed by Four Compared Techniques on Nine Categories

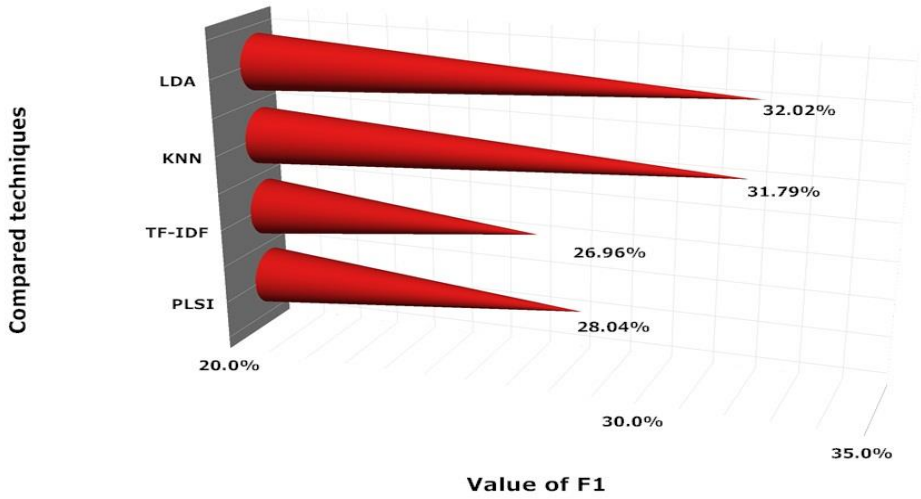

Figure 6. Average Recall Rate Performed by Four Compared Techniques on Nine Categories of CCL

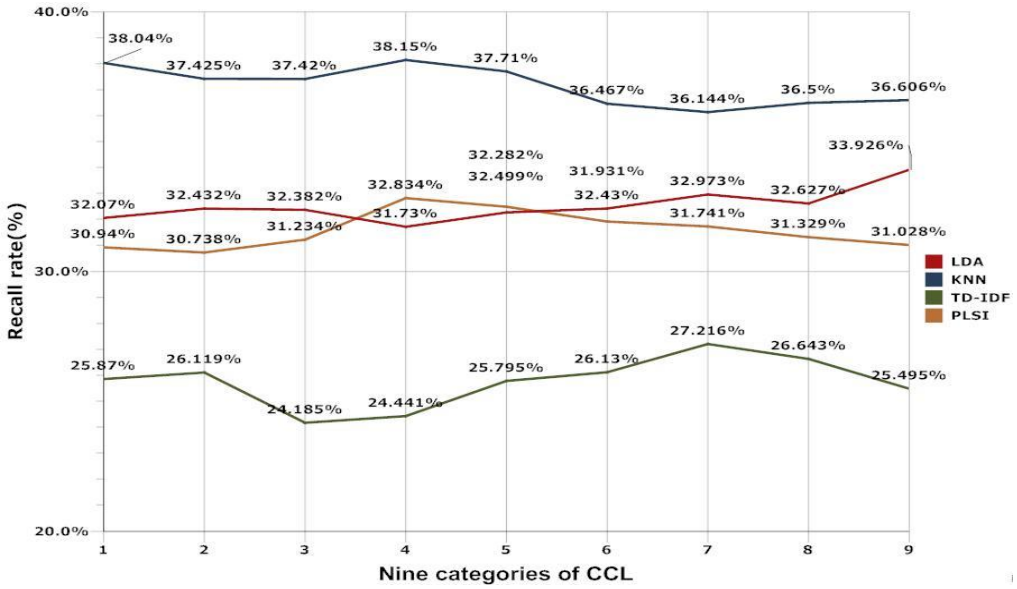

Figure 7. Average F1 Value Performed by Four Compared Techniques on Nine Categories of CCL

5.2 Comparison between LDA-CH and LDA-EH

In order to verify the suitability for LDA model to extract Chinese topics, we also compared Chinese topic extraction based on LDA (LDA-CH) with English topic extraction based on LDA (LDA-EH). This paper selects the DUC 2005 (http://www-nlpir.nist.gov/projects/duc/duc2005/) as the English corpus. The National Institute of Standards and Technology (NIST) initiated the Document Understanding Conference (DUC) series to evaluate automatic text summarization. Its goal is to further progress in summarization and enable researchers to participate in large-scale experiments.

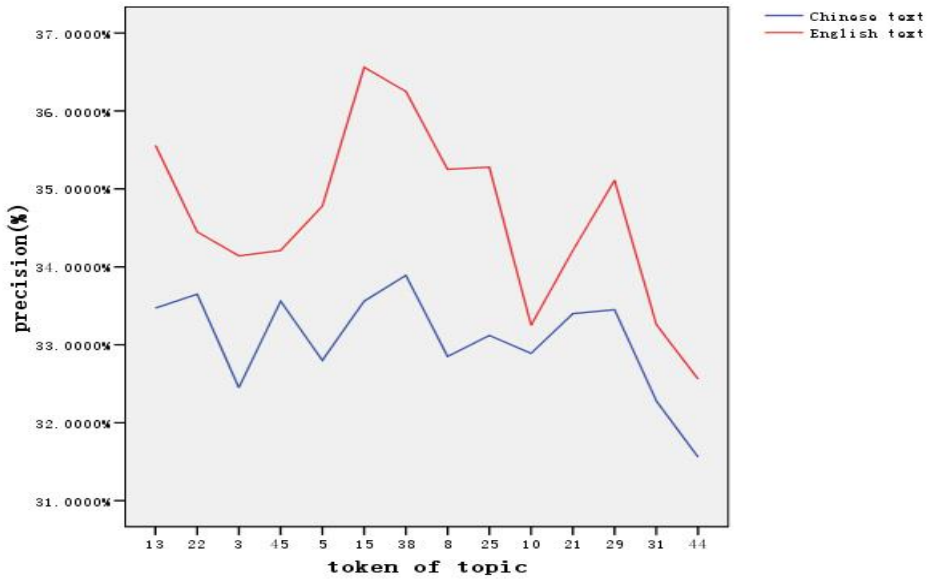

Figure 8. Comparison of Precision Rate between LDA-CH and LDA-EH

During the comparison experiment, we use LDA model to extract topics in two corpus with K=50, count each recall rate and precision rate for each topic. The comparison results are manifested in Figure 8 and Figure 9.

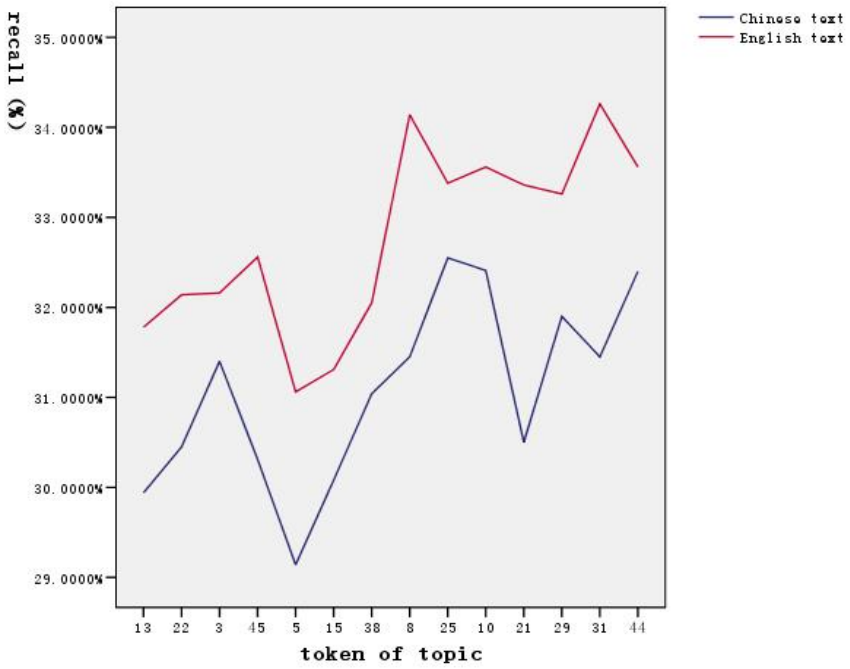

Figure 9. Comparison of Recall Rate between LDA-CH and LDA-EH

From Figure 8 and Figure 9, the recall rate and precision rate of LDA-CH is worse than LDA-EH.

6. Conclusions and Discussion

In this paper, we applied the LDA model to extract Chinese topics. Our results exhibit LDA has better performance in precision rate and F1 value than any other comparable techniques, since the Chinese words demonstrated in Table 1, Figure 5 and Figure 7 are distributed to each topic automatically. This is a perfect endorsement for our method to prove the LDA model is relatively scientific and advanced in Chinese topic extraction. Nevertheless, Figure 6 reveals that the recall rate is worse than using KNN and almost the same as with PLSI. We suggest the main reason is because KNN assumed words are independent to each other, so every word should be compared to all word samples in documents to seek out its nearest neighbors, which contributes to the recall rate. As for LDA, we assume that the order of a word's occurrence can be ignored; only taking the number of occurrences into consideration. When assigned to each topic, words strongly are associated with each other because they are in the same latent topic. Words not assigned to the same topic will not be covered.

Moreover, we find that the recall rate and precision rate of LDA-CH is worse than LDA-EH. Therefore, whether the LDA model is suitable for Chinese text topic extraction should be discussed further. The following reasons are pointed out:

First and foremost, the LDA model has a weak point of arbitrariness originating from operating latent variable distributions directly in multi-level graphical models. The meaning of Chinese words varies indifferent contexts, so the "context-level" should be taken into consideration in the LDA model.

Secondly, we all know that LDA has an excellent performance in extracting English topics. Whether the design of LDA model may be compatible only with English texts is a matter that should be taken into account by performing some comparison work.

Finally, Chinese texts have their own characteristics, which make them distinct from English. English words are separated by a space interval while Chinese words are not. If words are not separated appropriately, the result will be less accurate and reliable, whereas Chinese word segmentation is essential for Chinese topic extraction. Although ICTCLAS holds a remarkable precision, it still needs to be improved.

As for the drawbacks above, we will improve the LDA model in our future wok by adding a new level: sentence-level, to reduce arbitrariness because context is much more easily obtained in sentences than in a single word. The superior performance of LDA in English text encourages us to further our study by conducting the same algorithm on same-sized Chinese and English corpus in order to extract topics based on the LDA model so as to discover the disparity between them. Ultimately, a more precise Chinese word segmentation system should be developed as well.

Acknowledgements

This paper was supported by National Natural Science Foundation of China (No: 71363022, No: 71361012).

References

[1] S. Momtazi and F. Naumann, "Topic modeling for expert finding using latent Dirichlet allocation", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 3, **(2013)**.
[2] Q. Liu, "A Novel Chinese Text Topic Extraction Method Based on LDA", Proceedings of the 4th International Conference on Computer Science and Network Technology, Harbin, China, **(2015)**.
[3] D. M. Blei, A.Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 5, no. 3, **(2003)**.
[4] T. Griffiths and M. Steyvers, "Finding scientific topics", Proceedings of the National Academy of Science of the United States of American, vol. 1, no. 101, **(2004)**.
[5] G. Maskeri and S. Sarkar, "Mining Business Topics in Source Code using Latent Dirichlet Allocation", Proceedings of the 1st India software engineering conference, Hyderabad, India, **(2008)**.

[6] H. Wu, J. He and Y. J. Pei, "Scientific Impact at the Topic Level: A Case Study in Computational Linguistics", Journal of the American Society for Information Science and Technology, vol. 11, no. 61, **(2010)**.
[7] X. Li, J. Quyang, X. Zhou, Y. Lu and Y. Liu, "Supervised labeled latent Dirichlet allocation for document categorization", Applied Intelligence, vol. 3, no. 42, **(2014)**.
[8] Y. S. Lee, R. Lo, C. Y. Chen, P. C. Lin and J. C. Wang, "News topics categorization using latent Dirichlet allocation and sparse representation classifier", Proceedings of IEEE International Conference on Consumer Electronics, Taiwan, China, **(2015)**.
[9] A. Nauda and S. Usui, "Exploration of a collection of documents in neuroscience and extraction of topics by clustering", Neural Networks, vol. 21, **(2008)**.
[10] M. Nakatsuji, M. Yoshidab, and T. Ishidac, "Detecting innovative topics based on user-interest ontology", Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, **(2009)**.
[11] Y. Ouyang, W. J. Li, S. J. Li and Q. Lu, "Intertopic Information Mining for Query-Based Summarization", Journal of the American Society for Information Science and Technology, vol. 5, no. 61, **(2010)**.
[12] A. Srivastava, A. J. Soto and E. Milios, "A graph-based topic extraction method enabling simple interactive customization", Proceedings of the 2013 ACM symposium on Document engineering, Florence, Italy, **(2013)**.
[13] R. Saga, H. Kobayashi, T. Miyamoto and H. Tsuji, "Measurement Evaluation of Keyword Extraction Based on Topic Coverage", Proceedings of the 16th International Conference on Human-Computer Interaction, Crete, Greece, **(2014)**.
[14] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. Kim, S. Uryasev and C. Lioyd, "LinkedIn skills: large-scale topic extraction and inference", Proceedings of the 8th ACM Conference on Recommender systems, Silicon Valley, USA, **(2014)**.
[15] Y. H. Ma, Y. C. Wang, G. X. Su and Y. M. Zhang, "A Novel Chinese Text Subject Extraction Method Based on Character Co-occurrence", Journal of Computer Research and Development, vol. 6, no. 40, **(2003)**.
[16] J. Chen and Y. K. Zhang, "Novel Chinese text subject extraction method based on word clustering", Computer Applications, vol. 4, no. 25, **(2005)**.
[17] F. Liu, "Research and Application of Chinese Text Topic Extraction", Master thesis, Fudan University, Shanghai, **(2007)**.
[18] M. Xie, C. Wu and Y. Zhang, "A New Intelligent Topic Extraction Model on Web", Journal of Computers, vol. 3, no. 6, **(2011)**.
[19] L. Tian, W. Ma and W. Zhou, "Automatic Event Trigger Word Extraction in Chinese Event", Journal of Software Engineering and Applications, vol. 12, no. 5, **(2012)**.
[20] Q. Liu, "An Efficient Adaptive Focused Crawler Based on LDA and Domain Ontology", Journal of Computational Information Systems, vol. 8, no. 5, **(2012)**.
[21] W. An and Q. Liu, "Hierarchical Text Classification based on LDA and Domain Ontology", Applied Mechanics and Materials, vol. 414, no. 10, **(2013)**.
[22] W. Li and H. Xu, "Text-based emotion classification using emotion cause extraction", Expert Systems with Applications, vol. 4, no. 41, **(2014)**.
[23] J. Shi, M. Hu, X. Shi and G. Z. Dai, "Segmentation Based on Model LDA", Chinese Journal of Computers, vol. 10, no. 31, **(2008)**.
[24] X. Yang, J. Ma, T. F. Yang, Y. Q. Du and H. M. Shao, "Automatic Multi-document Summarization Based on Topic Model LDA", CAAI Transactions on Intelligent Systems, vol. 2, no. 5, **(2010)**.
[25] K. M. Chu and F. Li, "Topic Evolution Based on LDA and Topic Association", Journal of Shang Hai Jiao Tong University, vol. 11, no. 44, **(2010)**.
[26] X. Fu, G. Liu, Y. Guo and Z. Wang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon", Knowledge-Based Systems, vol. 3, no. 37, **(2013)**.
[27] J. Cui, L. Yu, and P. Li, "Image Clustering via Combined Visual and Annotation Information", Journal of Harbin University of Science and Technology, vol. 2, no. 19, **(2014)**.
[28] T. Hofmann, "Probabilistic Latent Semantic Indexing", Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, Berkeley, USA, **(1999)**.

