Topic Model Based Text Similarity Measure for Chinese Judgment Document

Yue Wang[1,2], Jidong Ge[1,2(✉)], Yemao Zhou[1,2], Yi Feng[1,2], Chuanyi Li[1,2], Zhongjin Li[1,2], Xiaoyu Zhou[1,2], and Bin Luo[1,2]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
gjdnju@163.com
[2] Software Institute, Nanjing University, Nanjing 210093, China

Abstract. In the recent informatization of Chinese courts, the huge amount of law cases and judgment documents, which were digital stored, has provided a good foundation for the research of judicial big data and machine learning. In this situation, some ideas about Chinese courts can reach automation or get better result through the research of machine learning, such as similar documents recommendation, workload evaluation based on similarity of judgement documents and prediction of possible relevant statutes. In trying to achieve all above mentioned, and also in face of the characteristics of Chinese judgement document, we propose a topic model based approach to measure the text similarity of Chinese judgement document, which is based on TF-IDF, Latent Dirichlet Allocation (LDA), Labeled Latent Dirichlet Allocation (LLDA) and other treatments. Combining with the characteristics of Chinese judgment document, we focus on the specific steps of approach, the preprocessing of corpus, the parameters choices of training and the evaluation of similarity measure result. Besides, implementing the approach for prediction of possible statutes and regarding the prediction accuracy as the evaluation metric, we designed experiments to demonstrate the reasonability of decisions in the process of design and the high performance of our approach on text similarity measure. The experiments also show the restriction of our approach which need to be focused in future work.

Keywords: Chinese judgment documents · Data science · Machine learning · Natural language processing · Text similarity · TF-IDF · Topic model · Latent Dirichlet Allocation · Labeled Latent Dirichlet Allocation

1 Introduction

In the recent informatization of Chinese courts, the huge amount of law cases and judgment documents, which were digital stored, has provided a good foundation for the research of judicial big data and machine learning. In law, a judgment is a decision of a court regarding the rights and liabilities of parties in a legal action or proceeding. Judgments also generally provide the court's explanation

B. Zou et al. (Eds.): ICPCSEE 2017, Part II, CCIS 728, pp. 42–54, 2017.
DOI: 10.1007/978-981-10-6388-6_4

of why it has chosen to make a particular court order. Judgment document is the documented judgement with relevant content. In 2013, China Judgment Online System officially opened. Up to now, it has recorded more than 26 million electronic judgment documents and became the largest judgment document sharing website around the world. The achievements of the informatization of Chinese courts not only provides the benefit of digitization, but also is a great help to judges and relevant parties.

In this situation, some ideas about Chinese courts can reach automation or get better result through research of machine learning. For example, Judge can find the similar judgement documents by the basic situation of the case to contribute to the process of judgement; court can evaluate workload of a judge by the similarity of the judgement documents it handled; Even relevant parties can input the situation of case to view the relevant statutes. In trying to achieve all above mentioned, a text similarity measure for Chinese judgment documents is being called.

As an important category of natural language processing, text similarity has developed from String-based algorithms, Corpus-based algorithms, to Knowledge-based algorithms [1], including TF-IDF, topic model, distributed representation, etc. When the target of text similarity is changed to Chinese judgement document, there are some new challenges as follows:

1. It needs to focus on the semantic layer when measured the text similarity of judgement document.
2. Judicial specific words existing in various types of judgement documents may influence the text similarity.
3. Chinese judgement document is semi-structured, which means it includes not only expression with natural language, but also a relatively fixed standard. The standard may provide a chance to improve the result of text similarity measure.
4. In Chinese judgement document, besides the process of reasoning and judgement, the claims and evidence of pleadings also need be recorded. It's a critical factor to influence text similarity that how to judge the important of similarity measure for different part of judgement document.
5. Chinese legal system, which is embroidered on legislation and assisted by administration, is obviously different to other countries with adequate legal system [2]. Referencing civil law system, Chinese legal is grounded on statutory code instead of law precedent. It means Chinese judgement document depends on relatively fixed statutes, which may help the work of text similarity.

In this paper, we propose a topic model based approach to measure the text similarity of Chinese judgement document. For the challenges mentioned above, the approach is based on TF-IDF, Latent Dirichlet Allocation (LDA), Labeled Latent Dirichlet Allocation (LLDA) and other treatments. The approach can be used to develop corresponding applications, such as similar documents recommendation, workload evaluation based on similarity of judgement document, and prediction of possible relevant statutes.

The remainder of this paper is laid out as follows. Section 2 introduces related work. Section 3 introduces our approach in detail. Section 4 shows the implementation of our approach and the experiments and Sect. 5 makes conclusion and discusses the future work.

2 Related Work

Text similarity measure play an important role in natural language processing research and applications such as information retrieval, text summarization, text classification, topic detection and so on. Developing to this day, from String-based algorithms to Corpus-based algorithms to Knowledge-based algorithms [1], many text similarity measures have been proposed for using in different scenes.

Topic model, as a Corpus-Based algorithms, originated from Latent Semantic Indexing (LSI) [3]. Although LSI is not a probability model, Hofmann, based on the main idea of LSI, proposed Probabilistic Latent Semantic Indexing (pLSI) [4]. After that, Blei et al. proposed Latent Dirichlet Allocation [5], which introduces Dirichlet distribution to further improve topic model. Based on LDA, more improved topic model, such as Supervised LDA [6], Labeled LDA (LLDA) [7], Hierarchically Supervised LDA [8], etc., are proposed to solve the specific problems. Topic models have been applied successfully in documents modeling [9], image modeling [10,11] and etc.

LDA, as one of the most important topic model, assume each document is modeled as probability distribution over an underlying set of topics, which in turn are modeled as probability distributions over words. LDA has been widely applied in various scenes, such as sentiment analysis [12], bug localization [13], image classification [14] and text segmentation [15]. In process of LDA training, finding the number of topics is a difficulty. For this problem, Arun et al. showed some observations [16] and some extensions of LDA, just like Hierarchical Dirichlet Processes, were designed [17].

LLDA is An extension of LDA with supervision [7]. This model allows corpus to be labeled by tags, and output a list of labeled topics. LLDA has been demonstrated the potentiality for fine grained topic modeling [18]. It also be applied to text classification [19] and social relation [20].

3 Approach

In this section, we describe the text similarity measure for Chinese judgment document in detail as follows. Subsection 3.1 presents an overview of the workflow of our approach. Subsection 3.2 describes the preprocessing of corpus. Subsection 3.3 introduces the choice of model, Subsect. 3.4 describes the process of topic model training. Subsection 3.5 introduces the meaning and method of corpus segmentation with different weight. Subsection 3.6 introduces the evaluation method for result.

3.1 Overview

Figure 1 presents an overview of workflow for Our approach. Because of the characteristics of Chinese natural language, the process is based on Chinese word segmentation. The steps of workflow is as follows:

1. Collect Chinese judgement documents, with structured information, including cause of action and category of cases, as the target corpus.
2. Divide the target corpus into two parts. The first part is used as the training corpus and the second part called test corpus, which needs to be labeled with practical similarity, is prepared for the evaluation of model.
3. Preprocess the judgment documents of the training corpus.
4. Choose the high reliable and important segments as the inputs.
5. Define and decide the parameters of training and model.
6. According to LDA or LLDA, Use the training corpus to complete training.
7. Use test corpus to evaluate the result of previous training model.

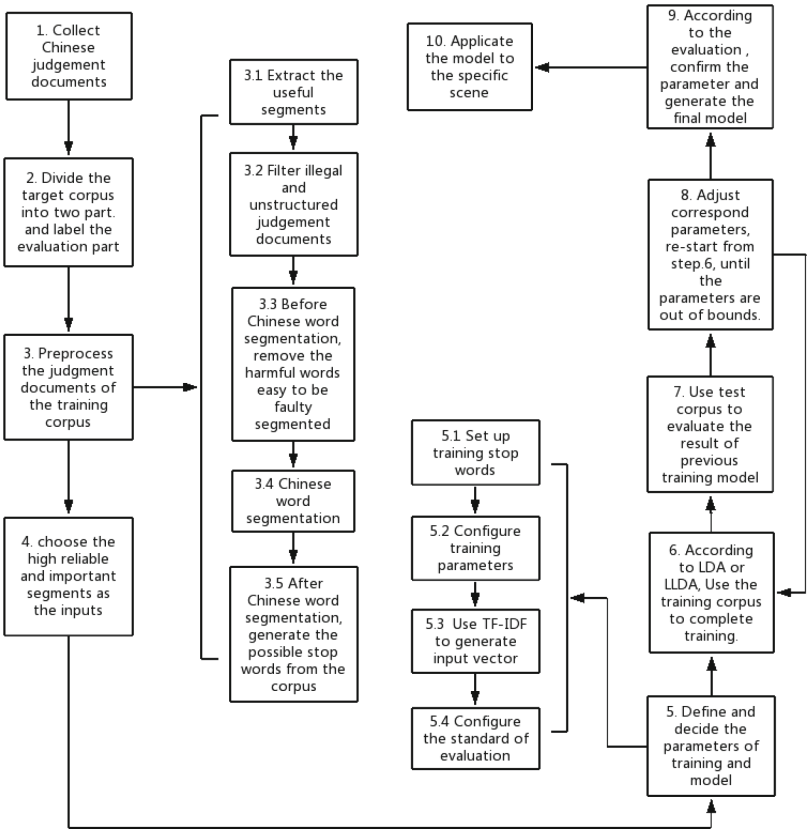

Fig. 1. Overview of the workflow for text similarity measure for Chinese judgment documents

8. Adjust correspond parameters, restart from step.6, until the parameters are out of bounds.
9. According to the evaluation of different parameters, confirm the parameter and generate the final model.
10. Applicate the model to the specific scene.

3.2 Preprocessing

The preprocessing step can be divided into four sub steps as follows: (1) Extract the useful paragraph; (2) Filter illegal and unstructured judgement documents; (3) Before Chinese word segmentation, remove the harmful words which are easy to be incorrectly segmented; (4) Chinese word segmentation; (5) After Chinese word segmentation, generate the possible stop words from the corpus. The purpose of this step is to further decompose the training corpus, minimize possible interference and prepare for training.

Chinese court has drew up a series of standards to define the structure of judgment document. It can help us to distinguish the corpus and identify low-quality judgement documents, as the content described in steps (1) and (2). With these steps, we can further filter the useless part of judgement document for semantic text similarity measure, just like the name of judge and the information of litigant participants, to focus on the case itself.

In Judgement document, a number of judicial specific words, just like prosecutor and defendant which occurs frequently, are not only meaningful, but also harmful for semantic similarity measure. Moreover, locale names and some ordinary names may also have influence on text similarity, especially the abbreviation of names for secrecy. Incorrect word segmentation of Chinese about the harmful words is another problem, which makes the target word split after word segmentation that can not be filtered by stop word list. The examples is shown in the following Table 1.

For these problems mentioned above, the steps (3) and (5) are necessary. In step (3), the main target is the name of litigant participants and some special judicial words. The formers can be extracted from judgement document by some rules and the specific words can be selected from the words list in step (5). In step (5), the most frequent terms are our candidates. Based on the segmented Chinese words, we can make statistics the frequency of terms and choose the stop words.

Table 1. Examples of incorrect word segmentation about the harmful words

Input	Correct Segmentation	Incorrect Segmentation	Main Affected Term
原告诉称	原告 | 诉称	原 | 告诉 | 称	告诉
李某某	李某某	李 | 某某	某某

3.3 Choice of Model

In step 6, we can choose LDA or LLDA to complete modeling. For Chinese judgement document, this choice is necessary. As a kind of text similarity measure, the approach generates different models for various similarity targets. The similarity targets based on statutes, which are explicit and finite tags of Chinese judgment documents, are an optional choice, which makes it possible to execute the process of evaluation automatically with little manual intervention. For this kind of similarity targets, such as statutes prediction, LLDA is more suitable than LDA, because of introduced supervision and fine grained topics.

With more manual intervention, such as the similarity targets based on manual classification or just the number of class, there will be different choices. In principle, if documents in corpus can be labeled explicitly, LLDA model will be recommended.

3.4 Topic Model Training

Each Chinese judgement document is associated with relatively fixed items, just like cause of action, category of case, codes and statutes. We can assume that Chinese judgement document is topic relevant, and based on this assumption, the topic model is an appropriate approach for semantic text similarity measure. We use LDA model to find the relation between number of topics and number of statutes. For verifying the conjecture and improving the accuracy of text similarity, we try to use supervised topic model named LLDA. This subsection including step 5 and 6 of our approach, which is hard to be separated described, is aimed to introduce the process of parameters choices and training.

There are four sub steps as follows in common: (1) Set up training stop words; (2) Configure training parameters, include the initial value and adjustment value, which is different for LDA and LLDA; (3) Use TF-IDF to generate input vector; (4) Configure the standard of evaluation. In the complete process, the training step and evaluation step should be execute repeatedly to confirm the parameters of model.

For LDA model, the most important parameter is topic number, which is also the difficulty for normal topic models. In this approach, we use a self-adaption method to choose the topic number. The brief steps is as follows: (1) Choose the initial topic number; (2) Start LDA training and evaluation the model; (3) Increase or decrease the topic number and return to step (2) until out of bound; (4) According the topic num and result of evaluation, choose the appropriate topic number. Besides the result about similarity, the perplexity [5] is also a important metrics.

For LLDA model, the additional work is to complete labelling. If the similarity target is based on statutes, the relevant statutes of Chinese judgement document will be the natural labels for supervised topic model. Because of the structuring of judgement document, the referenced statute can be extracted completely by fixed rules in most cases. Besides, the counted statutes should exclude

the low frequency statutes with a threshold, because the target statutes, as the performance of topic in the approach, must have a certain number of occurrence in the corpus, or they will have no statistically significant.

3.5 Choice of Inputs

In the process of text similarity measure, because the importance and reliability of different parts of Chinese judgement document is various, which is attributed to the structure of judgement document, we should choose the appropriate segments as input.

For a Chinese judgement document, the core content is consist of evidence, fact, statute and judgement [21]. Corroborate evidence with each other; Deduct evidence or facts to facts; Relate facts to statutes; Generate judgement from statutes and facts. The structure of Chinese judgement document is as showed in Fig. 2. The judgement is result, the statute is explicit but the evidence and fact is full of uncertainty. For example, judgement document records a list of evidence provided by plaintiff and defendant, including not only the accepted evidence, but rejected part; The fact in fact finding segment is more credible than the fact recorded in judgement document from plaintiff.

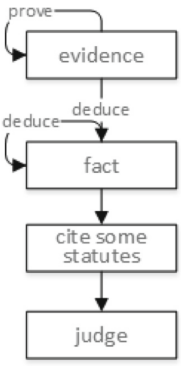

Fig. 2. Judgement document core structure

Ideally, the reliability of each evidence and fact can better reflect the content of Chinese judgement document. But considering the difficulty of this job and the redundancy of judgement document, we choose the segments as input which is confirm by judge, just like the fact finding segment and evidence segment of the base of case.

3.6 Result Evaluation

In step 2, as the concrete content presented in Fig. 3, some documents should be extracted from target corpus and labeled for evaluation. After the process of training, the model need been evaluated by designed evaluation method, which is described in detail in this subsection.

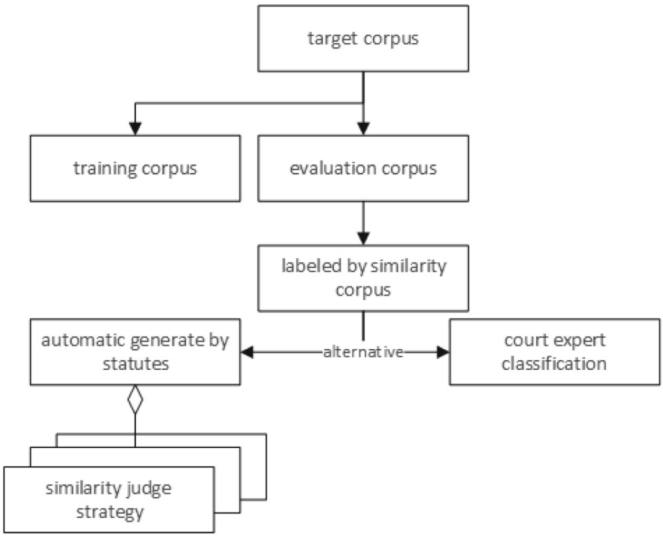

Fig. 3. Classification and labelling method for corpus

The process of similarity labelling is decided by two dimension. Labelling method as the first dimension, which means implementation methods of labelling, includes automatic labelling and artificial labelling. The former needs to formulate and implement correspond similarity strategy. The later demands the expert of court to finish the labelling. The other dimension includes digital labelling and non-digital labelling which is used to present if the label is represented by numbers called labelling granularity. For example, Comparing two judgement documents with the overlap of referenced statutes is an automatic digital labelling method, because the statutes can be extracted automatic and the result can be quantification; In another aspect, the classified judgement document handled by court expert is an artificial non-digital labelling method. The characteristics of the dimensions is showed in Table 2.

In this paper, we mainly focus on the labelling method with statutes, and the evaluation method is described below in general: The base of case in judgement document is the input of evaluation, and the result of evaluation depends on the accuracy of prediction on statutes. The reason for the choice is intuitive and universal for different model, includes TF-IDF, LDA and LLDA. The workflow of evaluation based on statute prediction is presented in Fig. 4.

Table 2. Advantage and disadvantage of labelling

	digital	non-digital	advantage	disadvantage
automatic	✓	✓	accurate, automatic	different standard for different target
artificial	✗	✓	currency	Lack of mathematical basis, Uncertain

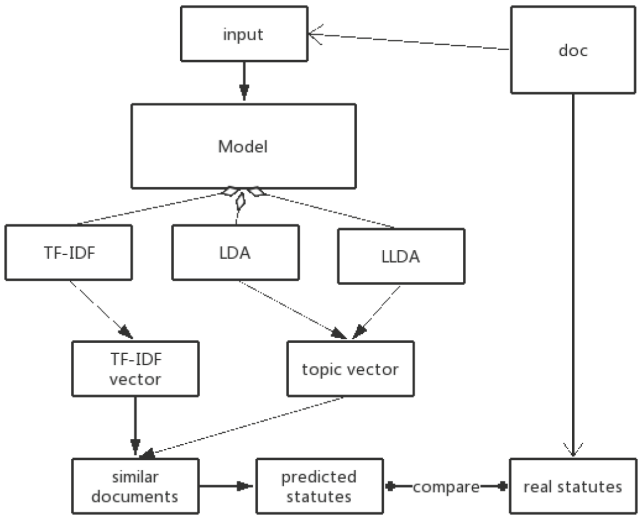

Fig. 4. Workflow of evaluation based on statute prediction

Based on various trained model, we can get the similar documents in training corpus from input of test document. For each test document, we sum the frequency of each unique statute by scanning all similar documents. And then sort the statutes in descending order. Without sufficient basis for truncating the sorted statutes, we choose top N to be the predicted statutes. The N is the number of real relevant statutes.

Perplexity, the exponentiation of the entropy, as a common metrics to evaluate the language model, is a reference metric for this approach. In Chinese judgment document, the statutes, which are associated with topic in our approach, can not be assumed independent. The correlation of statutes is common in judgment and the overlap is allowed. In this approach, we use perplexity to determine iteration number and to assist the evaluation of models.

4 Application and Evaluation

The purpose of this section is to implement topic model based text similarity measure for Chinese judgment documents, evaluate the result of experiments and provide support for our approach. The concerned points include: (1) the applicability of LLDA in this approach, (2) the performance of this approach in practical application, (3) The influence of specific preprocessing of corpus for Chinese judgment document on text similarity.

4.1 Preprocessing and Dataset

In lack of common corpus about Chinese judgement document, we collected documents in China Judgment Online System, the official website of Chinese

court. To reduce the complexity of problem, In this experiment, we chosen the same type of documents and totally extracted 53000 first-instance civil judgment documents. 50000 of all is used to training corpus and last is used to evaluation.

In the programming tool aspect, we use jieba module to segment Chinese word, gensim library to implement TF-IDF and LDA, Stanford Topic Modeling Toolbox to implement LLDA.

The evaluation method is same as the Subsect. 3.6 described. The evaluation metrics is F-measure, For each category, F-measure is calculated by $2 * \text{precision} * \text{recall}/(\text{precision} + \text{recall})$. F-measure represents the balance between precision and recall, the higher the F-measure of a category is, the better the performance of the classifier on this category is. In this situation, F-measure is same as precision and recall.

About the training parameters, $\alpha = \overrightarrow{l}/sl, \beta = 1/V$. The $\overrightarrow{l}$ is the vector of statutes frequency, the sl is the sum of statutes frequency and the V is number of words. The iteration number is depend on the perplexity of model.

4.2 Experiment and Result

For different threshold of statutes frequency, the number of counted statutes is showed in Fig. 5.

Based on the accounted statutes with different threshold of statutes frequency, we can generate corresponding LLDA models. In this scene, the topics of LLDA, as the output of model, are named after statutes in Chinese judgment documents. In another word, From LLDA, we obtain the probability distribution of statutes over the words in corpus, which can be used to predict statutes directly. Using the same idea as evaluation method, we can obtain the experiment results showed in Fig. 6.

The accuracy of direct statutes prediction based on LLDA is not ideal. For Chinese judgment document, one of the reasons, which is easily associated with,

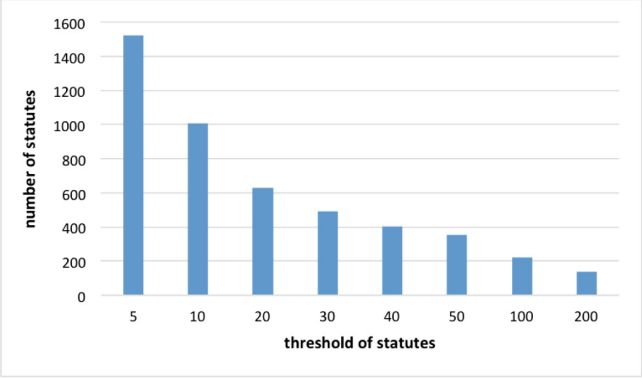

Fig. 5. Number of statutes with different threshold of statutes frequency

is the restriction of bag of words. Though being processed by TF-IDF, the model can not represent the logical relations between words, which is the source of statutes deduction. In another word, some kinds of statutes may not be predicted in current models, which is the emphasis of future work. As evidence, we manually culled some statutes which is hard to be predicted and obtained the better result.

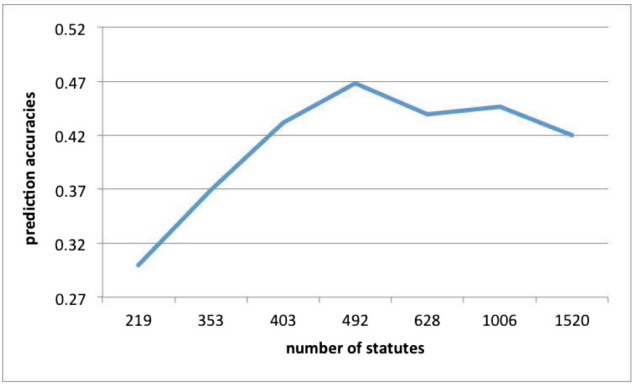

Fig. 6. Accuracies of direct statutes prediction based on LLDA

With the evaluation method mentioned in Subsect. 3.6, we experimented the statutes prediction on LLDA, LDA and independent TF-IDF. Figure 7 represents the comparison of LDA and LLDA. At least for statutes prediction in this scene, LLDA has better performance than LDA. Overall, the accuracy of statutes prediction based on LLDA has similar trend with the result of LDA model. Based on LLDA, the improvement of accuracies from direct statutes prediction to statutes prediction may shows that the statutes are not independent.

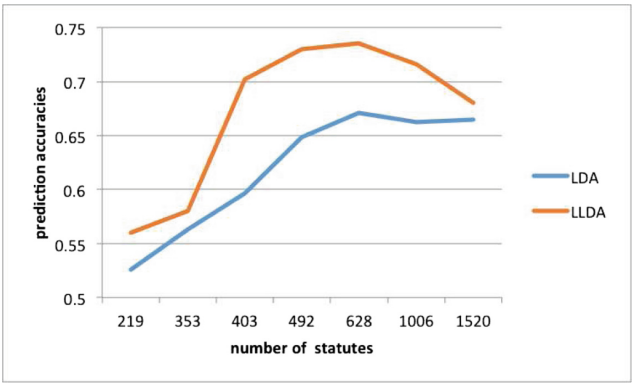

Fig. 7. Comparison of statutes prediction accuracy between LDA and LLDA

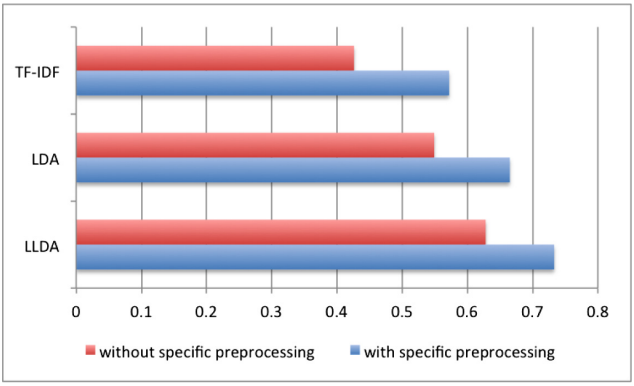

Fig. 8. Overall accuracies of statutes prediction with specific preprocessing and without specific preprocessing

As described in Sect. 3.2, besides the normal preprocessing, the approach asks more specific step for Chinese judgment document. In Fig. 8, it shows the overall accuracies of statutes prediction with specific preprocessing and without specific preprocessing. With specific preprocessing, the accuracies of statutes prediction improve obviously. Both LLDA and LDA has better performance than independent TF-IDF.

5 Conclusion and Future Work

In this paper, we propose a topic model based approach to measure the text similarity of Chinese judgement document, which is based on Latent Dirichlet Allocation (LDA) and Labeled Latent Dirichlet Allocation (LLDA), combining the characteristic of judgement document. In the experiments, we compared the result of statute prediction among TF-IDF based, LDA based and LLDA based approach. Both LDA and LLDA model have better performance than TF-IDF, and compared with LDA, LLDA improve a certain extent. The appearance also shows the word in Chinese judgement document has topic relevance on statutes.

However, the approach itself exits some defects as follows: (1) It is not a completely automatic approach. Manual intervention is required in the preprocess of corpus and the calculation of topic model parameters. (2) The whole model has some simplified assumption which need to be improve and perfected. (3) Some statutes generated from the logical relationship words, which can not be solved in word bag model or TF-IDF model, need the further research. These problems mentioned above left spaces for the future work.

Acknowledgement. This work was supported by the Key Program of Research and Development of China (2016YFC0800803).

References

1. Gomaa, W., Fahmy, A.: A survey of text similarity approaches. Int. J. Comput. Appl. **68**, 13–18 (2013)

2. Zhang, Z.: The construction of legal system in transitional China. China Legal Sci. **140**, 93 (2009)
3. Deerwester, S., Dumais, S., Furnas, G., et al.: Indexing by latent semantic analysis. J. Assoc. Inf. Sci. Technol. **41**, 391–407 (1990)
4. Hofmann, T.: Probabilistic latent semantic indexing. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
5. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Blei, D., Mcauliffe, J.: Supervised topic models. Adv. Neural Inf. Process. Syst. **3**, 327–332 (2010)
7. Ramage, D., Hall, D., Nallapati, R., et al.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Conference on Empirical Methods in Natural Language Processing, vol. 1, pp. 248–256. Association for Computational Linguistics (2009)
8. Perotte, A., Bartlett, N., Elhadad, N., et al.: Hierarchically supervised Latent Dirichlet Allocation. Adv. Neural Inf. Process. Syst. **24**, 2609–2617 (2011)
9. Li, F., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 524–531. IEEE Computer Society (2005)
10. Sivic, J., Russell, B., Efros, A., et al.: Discovering objects and their location in images. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 370–377. IEEE (2005)
11. Wang, C., Blei, D., Li, F.: Simultaneous image classification and annotation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1903–1910. IEEE (2009)
12. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: ACM Conference on Information and Knowledge Management, pp. 375–384. ACM (2009)
13. Lukins, S., Kraft, N., Etzkorn, H.: Bug localization using latent Dirichlet Allocation. Inf. Softw. Technol. **52**, 972–990 (2010)
14. Rasiwasia, N., Vasconcelos, N.: Latent Dirichlet Allocation models for image classification. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2665–2679 (2013)
15. Misra, H., Jose, J., Cappe, O.: Text segmentation via topic modeling: an analytical study. In: ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, pp. 1553–1556. DBLP, November 2009
16. Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N.: On finding the natural number of topics with Latent Dirichlet Allocation: some observations. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010. LNCS, vol. 6118, pp. 391–402. Springer, Heidelberg (2010). doi:10.1007/978-3-642-13657-3_43
17. Teh, Y., Jordan, M., Beal, M., et al.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**, 1566–1581 (2006)
18. Zirn, C., Stuckenschmidt, H.: Multidimensional topic analysis in political texts. Data Knowl. Eng. **90**, 38–53 (2014)
19. Li, W., Sun, L., Zhang, D.: Text classification based on labeled-LDA model. Chin. J. Comput. **31**, 620 (2008). Chinese Edition
20. Si, J., Mukherjee, A., Liu, B., et al.: Exploiting social relations and sentiment for stock prediction. EMNLP **14**, 1139–1145 (2014)
21. Zhou, F.: Reason, Jurisprudence, Sense and Writing. Shandong Justice (2007)