Chinese News Hot Subtopic Discovery and Recommendation Method Based on Key Phrase and the LDA Model

Bin GE[1],Chun-hui HE[1,*],Sheng-ze HU[1],Cheng GUO[1]

[1]Science and Technology on Information Systems Engineering Key Laboratory , National University of Defense Technology,Changsha, Hunan, P.R. China, 410073

Keywords: Hot subtopic discovery, Recommendation, Key phrase extract, LDA.

Abstract. The discovery and recommendation of Chinese news hot subtopic is a growing research area, and currently technology in this area is not yet mature. This research, inspired by traditional LDA model, uses the key phrase as feature to construct a "bag-of-phrase". On this basis, we propose the Chinese news hot subtopic discovery and recommendation method based on key phrase and the LDA model. Hot subtopic discovery and recommendation is the mainly innovation in this paper. During the cluster of hot subtopics, we select using the Longest Common Sequence (LCS) value as the similarity distance. To evaluate the proposed method, We used a mixed Chinese news dataset to experiment, and adopting: (1) time consumption of the training model, and (2) Perplexity value, and (3) quality of the discovery hot subtopic, three index to evaluate the performance of the method. The experimental results show that the proposed method can accurately discover the news hot subtopics, and also efficiently recommend relational hot subtopics in various fields.

1. Introduction

With the rapidly development of Internet and media technology, the influence of web news is increasing. Compared with traditional media news, web news score higher in terms of timeliness and convenience, and has gradually become the main medium for people to obtain useful information. In order to help users get valuable information, public opinion monitoring, information security, and other fields it has become a hot area of research.

In traditional text mining, the vector space model (VSM) [1] is often used for document representation. On this basis, the researchers put forward the Latent Semantic Indexing (LSI) [2] method based on the co-occurrence of lexical entry rules to find the semantic relations between words. LSI comes with an immediate benefit of dimension reduction, but the low-frequency part of useful lexical entry lead to impaired performance. Topic modeling has lots of applications in the field of text mining and information retrieval [3]. The Latent Dirichlet Allocation (LDA) [4] model is an unsupervised probability generation model. The probability distribution on multiple potential topics is used to represent the document features, and each topic is represented by the probability distribution of the features.

The method of parameter reasoning in LDA model usually uses Expectation propagation [5], VEM [6] and Gibbs Sampling among others. Radim [7] used LDA models to discover topics and implement engineering with python gensim tools package. Furthermore, a Locally Consistent Latent Dirichlet Allocation (LC-LDA) model was proposed to learn collective motion patterns through the use of low-level tracklet and bag-of-words features [8]. Yang et al. Proposed a novel approach based on a hierarchical latent topic model to learn and recognize scene and places [9]. Cha et al. proposed a method incorporating popularity into topic models for social network analysis [10]. Topic analysis can be further classified into two categories [11,12]. Hawes et al. [13] and Abbott et al. [14] have employed topic detection and tracking to extract signs of approval and disapproval from a corpus. Adams [15] reported a vector space model that produces documents and queries as eigenvectors and searches for topic corresponding to the content of conversation by computing similarity between vectors. An open-domain ICSI Meeting Corpus in 2008 emphasized the labeling and analysis of topics [16]. Lane et al. [17] exploited support vector machines (SVMs) to classify utterances into multiple topics and determine out-of-domain utterances. Time status can be combined with LDA for a dynamic model that is also called dynamic LDA (DLDA) [18]. Samuel and Noemie [19] considered

each utterance as a document through which they analyzed views and emotions about the content of online comments. Zhao et al. [20] exploited a Twitter-LDA model to analyze topics on Twitter, and they compared the results with the content of traditional news media. The Twitter-LDA model integrates documents on Twitter by extracting topical keywords or sentences [21]. Lu et al. adopted self-adaptive LDA modeling for topic extraction [22]. Liu et al. has proposed an attribute-restricted latent topic model that performs bests by imposing semantic restrictions onto the human-specific attributes [23].

Although the LDA topic model is one of the most efficient modeling methods for large-scale text corpus topic discovery, it also has some limitations, such as low accuracy and poor readability of topic discovering. Huang et al. [24] preliminary discussion the effect with word assignment for LDA model discovery topic. In order to solve and improve the above problems, Chinese news hot subtopic discovery and recommendation method based on key phrase and the LDA model is proposed. The proposed method is unique from others: (1) in the aspect of the selection of the LDA model, using key phrases instead of independently word as the document features. (2) considering the traditional LDA model discovery topic accuracy and readability is too low. Based on the key phrase LDA model to train corpus, and get topic-phrase distribution. Through hot subtopic discovery and recommendation method to discover and recommend high quality hot subtopics. (3) during clustering, value of the longest common sequence between two strings is regarded as the distance metric of similarity between topics.

The remainder of paper is organized as follows: Section 2 describes the TF-IDF weight calculate and key phrase extract principle. Section 3 describe based on key phrases LDA model. Section 4 describe hot subtopic discovery and recommendation method. Section 5, we analyze the experimental performance of our proposed method. Finally, Section 6 give the conclusions and future direction.

2. TF-IDF Weight Calculate and Key Phrase Extract

To analyze a document, At first, we through word segmentation preprocessing to get candidate key words. And then, using TF-IDF algorithm to calculate each candidate term weight. Next, we adopting word co-occurrence relationships to construct nodes and edges topological diagrams. The most important thing is using TextRank algorithm to calculate the score of the candidate word and give the descending results. Finally, according to candidate words descending result to select Top N words as key words, and the keywords that have a co-occurrence relationship in the original document are specified as key phrase. In this way, the extracted key phrase can be used to construct the vector matrix of the original document as the next input. Flow Chat of Extracting Key phrase from a document is shown Figure 1.

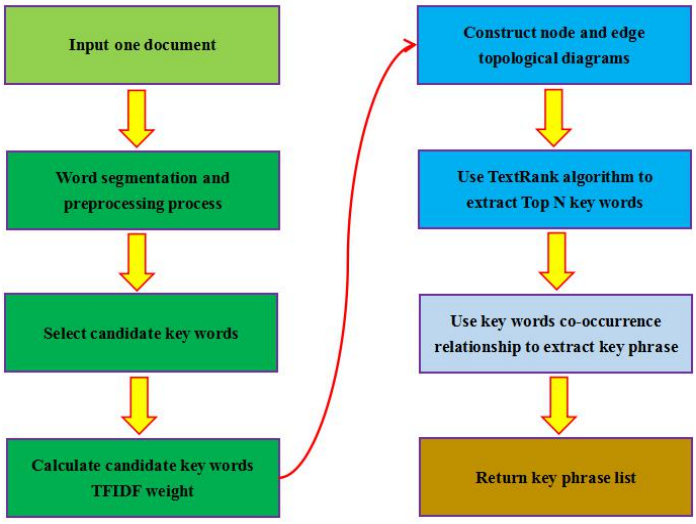

Figure 1. Flow Chat of Extracting Key Phrase from a Document

2.1 TF-IDF Weight Calculate

TF-IDF [25] is a statistical method used to evaluate the importance of a term to one of the files in a corpus. The importance of a term increases in proportion to the number of times it appears in the file. But at the same time, it falls inversely as the frequency of it appears in the corpus. Term frequency refers to the number of times a given word appears in the file. This number is usually normalized. For each word in a particular file, the TF weight is defined as below:

$$TF_{\mathrm{tj}} = \frac{n_{tj}}{\sum_k n_{kj}} \tag{1}$$

In formula (1), n_{tj} is the total number of times term t appears in document i, and the $\sum_{\mathrm{k}} n_{kj}$ is the total number of all terms k appeared in document i.

The main idea of inverse document frequency (IDF) is that the fewer the number of documents containing the t, the higher the value of IDF, which shows that the terms has a good classification ability. The IDF of a particular word can be calculate with the number of total files in the corpus divided to the total number of all the files containing the term t. The IDF is defined as below:

$$IDF_{\mathrm{t}} = \log(\frac{D}{\sum D_t + 1}) \tag{2}$$

In formula (2) D is total number of documents in the corpus, D_{t} is the number of documents in which t appeared.

High TF and low IDF value for term t can produce a high weight of TF-IDF. Therefore, it can effectively filter out common word and retain important word. So the TF-IDF is defined as below:

$$TF\text{ - }IDF = TF * IDF \tag{3}$$

2.2 Key Phrase Extract

TextRank is a well-known method to extract key phrase. The core formula of TextRank is defined as formula (4) [26]:

$$WS(V_{\mathrm{i}}) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{4}$$

Here $WS(V_{\mathrm{j}})$ is the score of node V_{i}. Constant d is damping factor, and the value is set to 0.85 in our method. w_{ji} is weight of the edge from the previous node V_{j} to the current node V_{i}. $In(V_i)$ is the set of nodes that point to it (predecessors). $Out(V_j)$ is the set of nodes that node V_{i} points to (successors). $\sum_{V_{\mathrm{k}} \in Out(V_j)} w_{jk}$ is the summation of all edge weights in the previous node V_{j}. w_{ji} is defined as the frequency that V_{j} and V_{i} appear in a window of maximum L words in the associated text, here L is 3 in our method. At the key phrase extraction stage, we use python tools package TextRank4ZH to implement it.

3. Based on Key Phrases LDA Model

LDA [27] is a three layer Bayesian model, and the three layers are the document layer, topic layer and word layer. The model is based on the follows assumptions:

(1)There are exist number of k independent topics in all documents.

(2)Each topic is a polynomial distribution on all phrases.

(3)Each document is a polynomial distribution on all topics.

(4)The priori distribution of each document is Dirichlet distribution.

(5)The prior probability distribution in each topic is Dirichlet distribution.

The process of generating one document is as follows:

(a) For the document set M, the phrase distribution parameter Ø is generated by sampling from the topic Dirichlet distribution parameter β.

(b) For one document m from M, the topic distribution parameter θ is generated by sampling from the document Dirichlet distribution parameter α.

(c) For the phrase N, Wmn in document m, we first extract a hidden topic Zm of document m according to θ distribution, and then sample phrase Wmn with topic Zm according to the Ø distribution.

Therefore, the joint probability distribution of the model is defined as below:

$$P(z,w,\theta,\phi|\alpha,\beta)=P(w|\phi,z)\cdot P(z|\theta)\cdot P(\theta)\cdot P(\phi) \tag{5}$$

After removing some of the hidden variables, the joint distribution integral is:

$$P(z,w|\alpha,\beta)=\iint P(z,w,\theta,\phi|\alpha,\beta)d_\theta d_\phi=\int P(w|\phi,z)\cdot P(\phi)d_\phi\int P(z|\theta)\cdot P(\theta)d_\theta$$
$$=\prod_{k=1}^{K}\frac{\Delta(n_k+\beta)}{\Delta(\beta)}\cdot\prod_{m=1}^{M}\frac{\Delta(n_m+\alpha)}{\Delta(\alpha)} \tag{6}$$

The transfer probability of indirect calculation can eliminate the intermediate parameter θ and Ø, in this way we can use the Gibbs sampling to each round of iteration, iterative process: first produced in a uniform random number, then according to the calculation on the transfer probability with each topic, in which new topic at random judgment by the cumulative probability, update parameter matrix, end iterate until convergence.

Based on key phrases LDA (KPLDA) model is different from the traditional LDA model. The features of traditional LDA model is usually composed by independently word. But the feature of KPLDA model is composed by extracted key phrases from each document. In KPLDA model, α=50/K, β=0.1 is used in the paper.

Finally, through KPLDA model we can train corpus to get the Document-Topic distribution and Topic-Phrase distribution.

The three layer structure diagram of the KPLDA model is shown in Figure 2.

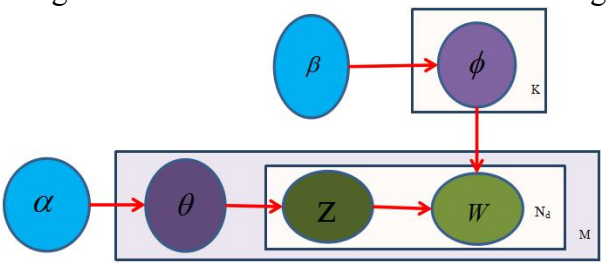

Figure 2. the Structure Diagram of the KPLDA Model

The definition of corresponding variables in Figure 2 are shown in Table 1.

Table 1. definition of variables in the LDA model

Variable	Definition of Variables
M	Number of documents in corpus
K	Number of topics in all documents
W	Phrases Vocabulary
N_d	Number of phrases in one document
Z	Topic for a phrase
α	Parameter of the Dirichlet prior on θ
β	Parameter of the Dirichlet prior on Ø
θ	document topic probability distribution
Ø	topic phrase probability distribution

4. Hot Subtopic Discovery and Recommendation Method

According to the method of section 3, using KPLDA model train corpus, we can get the topic-phrase distribution. The phrases have better reflects meaning for the topic than independently word [24]. Based on topic-phrase distribution, a subtopic discovery and recommendation method is proposed by us. This method is divided into the following steps:

Step1: According to the topic-phrase distribution, a rough topic set T_1 is made up by extracting the top K phrases with biggest weight under each topic.

Step2: For T_1, using the set union and difference to remove duplicate object get set T_2.

Step3: Calculating the Longest Common Sequence value between all elements in the set T_2 as a measure of the similarity distance D.

Step4: According to the similarity value to do Cluster, get the list of subtopic as T_3.

Step5: Word segmentation for each subtopic in T_3, and using word POS to filter subtopic that do not contain nouns or verbs get discovery hot subtopic list T_4.

Step6: Recommended Top M relational hot subtopics in T_4 for different users.

In step 3, the Longest Common Sequence (LCS) [28] recursion calculate formula for two strings $S=(s_1,s_2,...,s_i)$ and $T=(t_1,t_2,...,t_j)$ is defined as below:

$$LCS[\mathrm{i},j]=\begin{cases}0, if\,(i=0) or (j=0)\\ LCS[i-1,j-1]+1, if\,(i,j>0, \mathrm{s}_i=t_j)\\ \max\{LCS[i,j-1], LCS[i-1,j]\}, if\,(i,j>0, s_i\neq t_j)\end{cases} \tag{7}$$

LCS[i, j] is a nonnegative integer, it was expression the value of longest common sequence for $S=(s_1,s_2,...,s_i)$ and $T=(t_1,t_2,...,t_j)$. the range of similarity distance D is >1 in our method.

In step 6, value of M usually is from 10 to 30, it was set 20 in the method.

Process of hot subtopic discovery and recommendation method is shown in Figure 3.

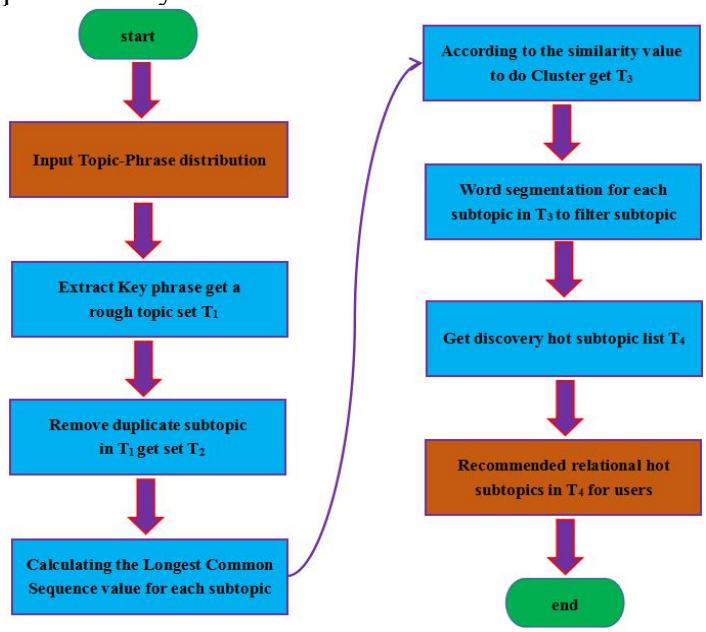

Figure 3. Flow Chat of hot subtopic discovery and recommendation method

5. Experimental Setup

5.1 Dataset Introduction

We used the KPLDA model to discover Chinese news hot subtopic. And select the text Dataset [29] with version of Mini published in the Sogou laboratory [30] to evaluate the performance of the model. The Dataset contains 1791 news documents, as well as 9 major topics, such as Internet, Finance, Health, Education, Military, Tourism, Sports, Culture and Recruitment. The number of documents contained in each topic is shown in Table 2.

Table 2. Topic distribution state with the Dataset

Topic name	Number	Topic name	Number	Topic name	Number
Internet	199	Education	199	Sports	199
Finance	199	Military	199	Culture	199
Health	199	Tourism	199	Recruitment	199

Table 2 reveals that the dataset was balanced among all of the topics.

5.2 Experimental Analysis

In order to objective verify performance of the KPLDA model, three core indicators were used to evaluate the model performance: (1) time consumption of the training model, and (2) Perplexity value, and (3) quality of the hot subtopic discovery.

By training different models and recording the corresponding time consumption, it can effectively reflect the advantages and disadvantages of the model through comparative analysis. The number of topics selected in all experiments is between 5 and 100. The relationship between different topic number of KPLDA and LDA models with the time consumption of training was shown in Figure 4.

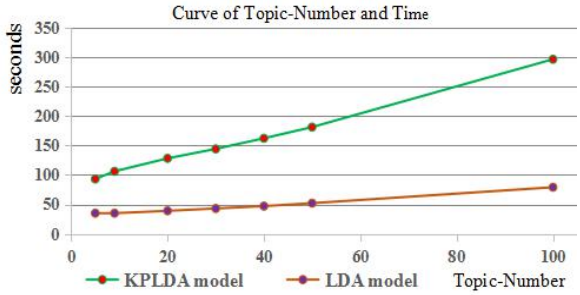

Figure 4. Curve of Topic-Number and Time consumption

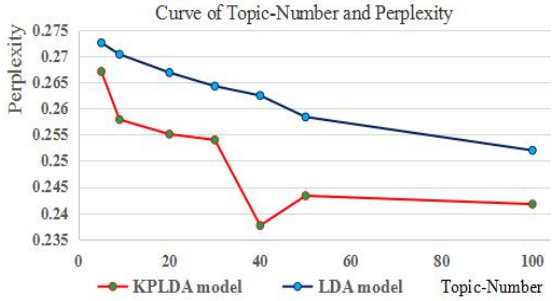

Figure 5. Curve of Topic-Number and Perplexity value

According to the results of Figure 4, LDA model training process has a great advantage relatively to KPLDA model. As the number of topics increase, the LDA model time consumption becomes less, ranging from 33 to 79 seconds, but the KPLDA topic model training time rapidly grows. When the number of topic reaches 100, time consumption is 296 seconds, about 3.7 times higher than LDA.

Perplexity value is a common indicator used to evaluate the performance of LDA model. In the evaluation of LDA model, it was initially introduced by Blei[4], and the smaller the Perplexity value the better is the performance of model. To evaluate gains of the KPLDA model in term of coherence in prediction of the unseen documents, we varied the number of topics from 5 to 100, and used the perplexity value as the evaluation criterion.The perplexity value was calculated as below:

$$Perplexity = \exp\left\{-\frac{\sum_{d=1}^{M}\ln(P(w_d))}{\sum_{d=1}^{M}N_d}\right\} \tag{8}$$

Where P(wd) is the generative probability of the phrase with respect to the feature, respectively in the document d. Figure 5 shows how the perplexity value changes for different number of topics with LDA and KPLDA model on the testing Dataset.

It shows that Perplexity value will decrease with increasing the number of topics, LDA model perplexity value reduced trend is relatively flat. When the number of topic reached 100, the Perplexity value reaches the minimum. It means that LDA model performance will get the best when the number of topic is 100; however, the changes of the KPLDA model is more obvious, when the number of topic reached 40, the Perplexity value will reach the minimum. Means that the best performance of KPLDA model is with 40 topics on the dataset. KPLDA model Perplexity value is less than the LDA model in all different number of topic, which fully illustrates that the KPLDA is better than LDA model on the task of Chinese News hot subtopic discovery.

In order to evaluate the quality of news hot subtopic discovering, different news topics we use separately with LDA and KPLDA model, combining the hierarchical clustering method to discover hot subtopic. Through experiments with the train Dataset, it will get some hot subtopics. LDA model

will get independent words under different topics, which have an incomplete meaning, as subtopic will missing a lot of information. Use LDA to discover Top 20 hot subtopics will shown in Table 3.

Table3. LDA model discovered Top 20 hot subtopics

server	drone	internet	radio station
system	day trip	investment rating	travel agencies
anti-dumping	Lakers	equity	ShenHua
avian influenza	intellectual property	AIDS	sudoku
surgery	interview	teacher	intern

According to the results of Table 3, the meaning of expression for one word under each topic is incompletely, so it is not suitable as a subtopic. However, the KPLDA model uses the key phrases to form the feature of the bag-of-phrase model. Therefore, the subtopic is more completely after training, and discovered Top 20 hot subtopics will shown in Table 4.

Table4. KPLDA model discovered Top 20 hot subtopics

storage engine	Indian navy	video surveillance system	gust fighter
Foreign exchange reserves	individual radio	Kelly index	elderly tourism
galaxy fund	countryside tour	community hospital	Belgium league
laparoscopic surgery	world heritage	breast hyperplasia	campus recruitment
international students	external recruitment	self-examination	world manager

According to table 4, it shows that the above of phrase has a perfectly meaning, and it is more suitable to be as the hot subtopic under the different news topics. Finally, we according to the result of hot subtopic discovery to recommend some relational hot subtopics for the user.

6. Conclusions and Future Work

The discovery and recommendation of Chinese News hot subtopics is a valuable research work. At present, the available technology is very poor. In order to improve the situation, through experiments and exploration, we recommend Chinese news hot subtopic discovery, based on key phrase and LDA model. Through experimental comparison, we make the following conclusions: for training of the model, LDA model time consumption is less than KPLDA model. But the quality of KPLDA model hot subtopics discovery is more better than LDA model. In addition, the relational hot subtopics can be accurately recommended to the user. There are some limitations to our proposed method, though: the experiment Dataset is small, and we just select TextRank to extract key phrases in this method. As an extension to this research, experimental validation on the large-scale Dataset is recommended as well as feature optimization. In addition,we can attempt using machine learning methods to extract key phrases.

7. Acknowledgement

This research was financially supported by the National Natural Science Foundation of China under Grant No. 71331008 and the Science Foundation of Hunan under Grant No. S100505.

References

[1] Salton, G.: A vector space model for automatic indexing. *Communications of the Acm*, 18(11), 613-620 (1975)
[2] Deerwester, S.: Indexing by latent semantic analysis. *Journal of the Association for Information Science & Technology*, 41(6), 391-407 (1990)
[3] Chen, Y., Yin, X., Li, Z., Hu, X., & Huang, J. X.: Promoting Ranking Diversity for Biomedical Information Retrieval Based on LDA. *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 456-461. IEEE (2012)
[4] Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. *J Machine Learning Research Archive*, 3, 993-1022 (2003)
[5] Minka, T.: Expectation-propagation for the generative aspect model. *Proc. Conference on Uncertainty in Artificial Intelligence*(UAI) (2002)

[6] Zhao, Z., Xu, W., & Chen, D.: EM-LDA model of user behavior detection for energy efficiency. *IEEE International Conference on System Science and Engineering* ,pp. 295-300. (2014)
[7] Radim., Rehurek., & Petr Sojka.: Software framework for topic modelling with large corpora,pp.45-50 (2010)
[8] Zou, J., Ye, Q., Cui, Y., Wan, F., Fu, K., & Jiao, J.: Collective motion pattern inference via locally consistent latent dirichlet allocation. *Neurocomputing*, 184, 221-231 (2016)
[9] Yang, J., Zhang, S., Wang, G., & Li, M.: Scene and place recognition using a hierarchical latent topic model. *Neurocomputing*, 148, 578-586 (2015)
[10] Cha, Y., Bi, B., Hsieh, C. C., & Cho, J.: Incorporating popularity in topic models for social network analysis. *International ACM SIGIR Conference on Research and Development in Information Retrieval*(pp.223-232). ACM (2013)
[11] Allan, J., Lavrenko, V., Frey, D., et al.: UMass at TDT 2000, in: *Proceedings of the Topic Detection and Tracking workshop*,pp.109-115 (2000)
[12] Liu, X., Tao, D., Song, M., Zhang, L., Bu, J., & Chen, C.: Learning to track multiple targets. *IEEE Transactions on Neural Networks & Learning Systems*, 26(5), 1060 (2014)
[13] Hawes, T., Lin, J., & Resnik, P.: Elements of a computational model for multi-party discourse: the turn-taking behavior of supreme court justices. *Journal of the American Society for Information Science & Technology*, 60(8), 1607–1615 (2009)
[14] Abbott, R., Walker, M., Anand, P., Tree, J. E. F., Bowmani, R., & King, J.: How can you say such things?: recognizing disagreement in informal political argument. *The Workshop on Languages in Social Media*,pp.2-11 (2012)
[15] Adams, P. H., & Martell, C. H.: Topic Detection and Extraction in Chat. *IEEE International Conference on Semantic Computing* ,pp.581-588. IEEE Computer Society (2008)
[16] Georgescul, M., Clark, A., & Armstrong, S.: A comparative study of mixture models for automatic topic segmentation of multiparty dialogues. Acl-08: Hlt (2013)
[17] Lane, I., Kawahara, T., Matsui, T., and Nakamura, S.: Out-of-domain utterance detection using classification confidences of multiple topics, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, No. 1, pp.150-161 (2007)
[18] Blei, D. M., Lafferty, J., D.: Dynamic topic models, in: *Proceedings of the 23rd International Conference on Machine Learning*, pp.113-120 (2006)
[19] Brody, S., & Elhadad, N.: An Unsupervised Aspect-Sentiment Model for Online Reviews. *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, June 2-4, 2010, Los Angeles, California, USA,pp.804-812. DBLP (2010)
[20] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., & Yan, H., et al.: Comparing twitter and traditional media using topic models. *European Conference on Advances in Information Retrieval*(Vol.6611/2011, pp.338-349). Springer-Verlag (2011)
[21] Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., & Lim, E. P., et al.: Topical keyphrase extraction from Twitter. *Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol.1, pp.379-388). Association for Computational Linguistics (2011)
[22] Lu, F., Shen, B., Lin, J., & Zhang, H.: A Method of SNS Topic Models Extraction Based on Self-Adaptively LDA Modeling. *Third International Conference on Intelligent System Design and Engineering Applications* (Vol.37, pp.112-115). IEEE Computer Society (2013)
[23] Liu, X., Song, M., Zhao, Q., Tao, D., Chen, C., & Bu, J.: Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, 45(12), 4204-4213 (2012)
[24] Huang, C. M., & Wu, C. Y.: Effects of Word Assignment in LDA for News Topic Discovery. *IEEE International Congress on Big Data*,pp.374-380. IEEE Computer Society (2015)
[25] Zhai, C., & Lafferty, J.: A study of smoothing methods for language models applied to Ad Hoc information retrieval. *International ACM SIGIR Conference on Research and Development in Information Retrieval*(Vol.22, pp.334-342).ACM. (2001)
[26] Wang , Z. , Feng, Y. , & Li, F.: The improvements of text rank for domain-specific key phrase extraction. *International Journal of Simulation Systems, Science & Technology*,17(20), 111–115 (2016)
[27] Yeh, J. F., Tan, Y. S., & Lee, C. H.: Topic detection and tracking for conversational content by using conceptual dynamic latent dirichlet allocation. *Neurocomputing*, 216, 310-318.(2016)
[28] Apostolico, A., & Guerra, C.: The longest common subsequence problem revisited. Algorithmica, 2(1-4), 315-336.(1987)
[29] Information on https://github.com/ustbprir1005gao/topic-model/ tree/master/main/resources/mini
[30] Information on http://www.sogou.com/labs/resource/list_pingce.php