A Method of Optimizing LDA Result Purity Based on Semantic Similarity

Zhu Jingrui[1], Wang Qinglin[1], Liu Yu[2] &Li Yuan[1]

1. School of Automation, Beijing Institute of Technology, Beijing 100081, China
E-mail: 535995225@qq.com; wangql@bit.edu.cn; liyuan@bit.edu.cn

2. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
E-mail: yu.liu@ia.ac.cn

Abstract: The result purity of traditional LDA (Latent Dirichlet Allocation) is uninterpretable because it is always difficult to summarize the meaning of each LDA result topic which contains multiple irrelevant words. To solve the problem, a method of optimizing LDA result purity based on semantic similarity in streaming news processing is proposed. In this method, the Category Cluster Density (CCD) of each topic is calculated first, and those topics with lower CCD value were dropped to optimize the overall LDA result purity. The news clustering experiment results show that the vague news can be removed effectively and the reserved topics are interpretable than traditional method, which can significant optimize the LDA result purity automatically.

Key Words: LDA, Purity, Semantic Similarity, Category Cluster Density

1 INTRODUCTION

LDA is the most used theme model, which is proposed by Blei et al. in 2003[1], has been widely used in text categorization, text clustering, abstract extraction, emotion analysis etc. [2-11]. The topics generated by LDA model are usually represented by a group of words. In many cases, the semantics of the words in one group are very different, and it is difficult to understand what the specific meaning of the group of words is.

For this problem, a LDA semantic model based on PMI-LDA (Point-wise Mutual Information) is proposed by Zhao B. [12], aiming at the advantages and disadvantages of the topic extracted from the topic model; Kaptein R. et al. improved the LDA semantic model [13]; Zhou D. et al. [14] proposed a new generation model related to social label, in order to combine the model of social tag with the language model based on information retrieval; Jonathan et al. [15] introduced the association between documents; Wang X. R. [16] took into account the order of words; Lu Y. et al. made improvements on the basis of the optimization formula, adding a priori information to distinguish between different topics[17]; Mei Q. Z. et al. [18] increased the regularization factor and introduced some correlation information and authentication information; Y. W. Teh et al. [19] proposed the HDP model which is deformation of Dirichlet Process, can automatically learn the number of topics. The topic extraction can be associated with the metadata of the document and other features (such as author, time, etc.), according to Andrew et al., by introducing a log-linear prior to the document topic distribution [20].

These methods mainly improved the process of building LDA model. However there are few methods are proposed on improving the results of LDA model directly. In this paper, a method based on semantic similarity is proposed aiming at improving the purity of LDA results. Preserving the higher purity topics that people can understand by discovered and deleted the bad topics. Experiment results show that this method can improve the readability and the interpretability of LDA Effectively.

2 LDA MODEL

2.1 Basic of LDA

LDA is a document-topic generation model proposed by Blei et al. in 2003, also known as the three-layer Bayesian probability model, which expresses the document into a three layer probability model consisting of text, topics and feature words. Each word of an article is first by a certain probability to choose a topic, and then from the topic of a certain probability to choose a word to get. The LDA model is shown in Fig. 1 [1]:

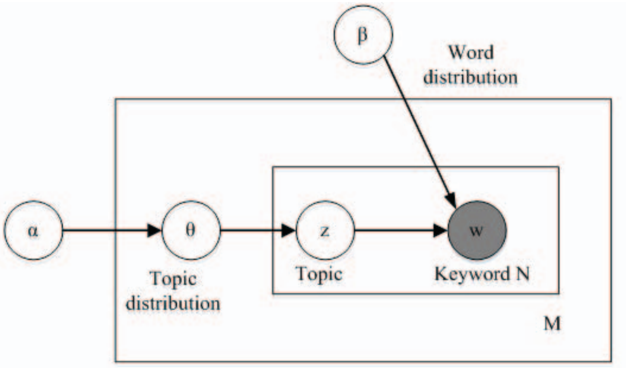


Fig 1. LDA model

This work was supported in part by the National Natural Science Foundation of China under Grants 61379099.

where sign θ and z represent the implicit variable, w is the unique characteristic word indicating the observable value. The sign α and β are the hyperparameter of the Dirichlet distribution. The sign α can be understood as the number of the topic is sampled before the document is obtained from the dataset, β can be understood as the number of the characteristic word frequency is sampled before finding the feature words in the dataset.

2.2 The Optimal Number of Topics

The number of topics T has a great influence on the performance of the LDA model. So we need to set a reasonable number of topics in advance before building a LDA model. It is a common method to set the number of topics based on perplexity [12]. Perplexity is often used in the language model to measure the language model's ability for the test corpus. The smaller the perplexity, the better model representation capability the topic model has. The perplexity formula as follows [12]:

$$perplexity = \exp\left\{ -\frac{\sum_{m=1}^{M} \log(\mathrm{P}(\mathrm{d}_m))}{\sum_{m=1}^{M} N_m} \right\} \qquad (1)$$

where M is the number of document in the dataset, N_m is the length of the m-th document, P(d$_m$) is the probability that the LDA model produces the m-th document, the formula as follows:

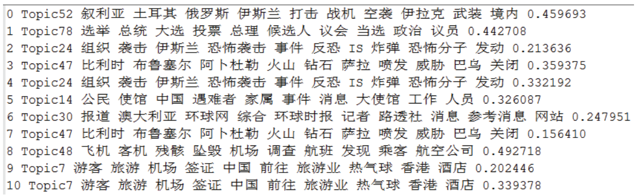
$$p(\mathrm{d}_m) = \prod_{i=1}^{n} \prod_{j=1}^{T} P(\mathrm{w}_i \mid \mathrm{z}_i = \mathrm{j}) P(\mathrm{z}_i = \mathrm{j} \mid \mathrm{d}_m) \qquad (2)$$

3 AN OPTIMIZATION METHOD BASED ON SEMATIC SIMILARITY

In this study, we propose a method of optimizing LDA result based on semantic similarity, and present the LDA result before and after optimization. Firstly, we cleaned the news and remove the stop words, prepositions, adverbs and other nonsense words by doing a series of preprocessing. Then, we deleted the news which poorly expresses the meaning of topic in each topic to optimize the single topic purity. Finally, the topics which are poor readability and poor interpretability are removed according the CCD to optimize the purity of LDA result.

3.1 The Input to Be Optimized

The corpus of this study is a news dataset which has be processed by many steps. Training the LDA model by using the news dataset, we can get two files, the theta file which is the topic probability distribution of each article and the phi file which is the probability distribution of words under each topic. And then each news can be assigned to the only topic through the two files, as shown in Fig. 2, which just the input to be optimized.



```
0 Topic52 叙利亚 土耳其 俄罗斯 伊斯兰 打击 战机 空袭 伊拉克 武装 境内 0.459693
1 Topic78 选举 总统 大选 投票 总理 候选人 议会 当选 政治 议员 0.442708
2 Topic24 组织 袭击 伊斯兰 恐怖袭击 事件 反恐 IS 炸弹 恐怖分子 发动 0.213636
3 Topic47 比利时 布鲁塞尔 阿卜杜勒 火山 钻石 萨拉 喷发 威胁 巴乌 关闭 0.359375
4 Topic24 组织 袭击 伊斯兰 恐怖袭击 事件 反恐 IS 炸弹 恐怖分子 发动 0.332192
5 Topic14 公民 使馆 中国 遇难者 家属 事件 消息 大使馆 工作 人员 0.326087
6 Topic30 报道 澳大利亚 环球网 综合 环球时报 记者 路透社 消息 参考消息 网站 0.247951
7 Topic47 比利时 布鲁塞尔 阿卜杜勒 火山 钻石 萨拉 喷发 威胁 巴乌 关闭 0.156410
8 Topic48 飞机 客机 残骸 坠毁 机场 调查 航班 发现 乘客 航空公司 0.492718
9 Topic7 游客 旅游 机场 签证 中国 前往 旅游业 热气球 香港 酒店 0.202446
10 Topic7 游客 旅游 机场 签证 中国 前往 旅游业 热气球 香港 酒店 0.339378
```

Fig 2. The input to be optimized

The first column is the news ID, the second column is the topic ID, the 10 words in the middle are the words that express the topic ID, the last column is the probability that this news belongs to the topic ID.

3.2 Remove Vague News

The vague news is the news which has no apparent relation with the labeled topic by LDA. Each news has a similarity to its topic. Then we set a threshold, the news whose similarity to its topic lower than the threshold is vague news. The threshold is determined by the average category similarity [14] and the loss degree.

The category similarity measures the clarity of the boundaries between categories, calculated by the class center vector method. Specifically, the lower the similarity between categories is, the higher clarity between the categories in the corpus is, so the better the classification performance has. The formula is as follows [21]:

$$\delta_H(\mathrm{z}_k,\mathrm{z}_m) = \frac{1}{2N}((\sum_{i=1}^{N} sim(\vec{\mathrm{d}}_i^{z_k}, \vec{\mathrm{d}}_o^{z_m})) + (\sum_{j=1}^{N} sim(\vec{\mathrm{d}}_j^{z_m}, \vec{\mathrm{d}}_o^{z_k})) \qquad (3)$$

$$V_{acs}(\theta) = \frac{1}{\mathrm{K}(\mathrm{K}-1)} \sum_{i=1,j=1}^{K} \sum_{j \neq i}^{K} \delta_H(\mathrm{z}_i, \mathrm{z}_j) \qquad (4)$$

where $\delta_H(z_k, z_m)$ represents the degree of similarity between category z_k and category z_m, N represents the number of documents, $\vec{\mathrm{d}}_i^{z_k}$ represents the *i*-th document vector in category z_k, $\vec{\mathrm{d}}_o^{z_m}$ represents the centroid vector in category z_m, $V_{acs}(\theta)$ represents the average category similarity when the threshold is θ, K represents the number of categories.

The loss degree is characterized by the ratio of the number of deleted news to the total number of cases, and the greater loss degree is, the more news will be deleted. The formula is as follows:

$$V_{ld}(\theta) = \frac{n_d}{N} \qquad (5)$$

where V_{ld} represents the loss degree when the threshold is θ, n_d represents the number of deleted news, N represents the total number of news.

The principle of setting the threshold is that the average category similarity will be better after the news which the similarity lower than the threshold deleted, and the loss degree of the news is not excessive. So we calculated the balanced score based on the average category similarity and the loss degree. The formula for calculating the balanced score is as follows:

$$BS(\theta)=\frac{V_{acs}(\theta)V_{ld}(\theta)}{V_{acs}(\theta)+V_{ld}(\theta)} \tag{6}$$

where the $BS(\theta)$ represents the balanced score, $V_{acs}(\theta)$ represents the average category similarity, V_{ld} represents the loss degree when the threshold is θ.

Then we calculated the BS for each threshold according to formula 6, and the threshold corresponding to the maximum BS value is the final set threshold. And then the news whose probabilities below the threshold are vague news which will be deleted to optimize the topics they belong to.

3.3 Optimizing LDA Result Purity

The news that not clear on the topics which they labeled will be deleted after the threshold is determined, and then calculated the CCD [16] of each topic. The CCD is used to measure the similarity of the document feature and the topic. If the average similarity of the corresponding set of documents is higher, the more the document set expresses the information of the category, the topic of the category is better. The CCD is expressed by the average similarity between each document in the category. The formula is as follows [21]:

$$\zeta_H(\mathrm{z})=\frac{1}{N(\mathrm{N}-1)}\sum_{i=1,j=1}^{N}\sum_{j\neq i}^{N} sim(\vec{\mathrm{d}}_i^{(\mathrm{z})},\vec{\mathrm{d}}_j^{(\mathrm{z})}) \tag{5}$$

where ζ_H(z) represents the CCD of category Z, N indicates the number of documents under the category, $\vec{\mathrm{d}}_i^{(\mathrm{z})}$ represents the *i*-th document vector in category z, $sim(\vec{\mathrm{d}}_i^{(\mathrm{z})},\vec{\mathrm{d}}_j^{(\mathrm{z})})$ represents the cosine similarity of the document vector $\vec{\mathrm{d}}_i^{(\mathrm{z})}$ and the document vector $\vec{\mathrm{d}}_j^{(\mathrm{z})}$.

Calculating the CCD of each topic, and then sorting the topics from large to small according to the value of CCD. The smaller the CCD is, the lower the average similarity of the newsgroups in the topic is, and the worse the expression of the topic is, which should be removed.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Set The Number of Topics

We conducted the following experiment to set a reasonable number of topics. First, we built the LDA model by setting the number of topics to 5-100 (interval 5). Then, we calculated the perplexity of different number of topics. The results are as follows:

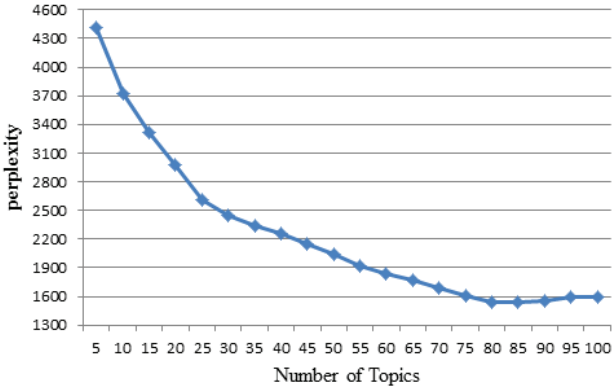
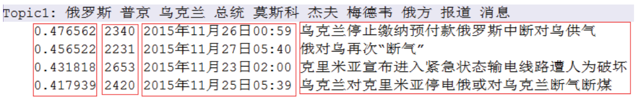


Fig 3. The relationship between the perplexity and the number of topics

As can be seen from Fig. 3, when the number of topics is 80 the perplexity minimum. Therefore, the number of topics in this experiment is set to 80.

4.2 Experiment Results and Analysis Without Optimization

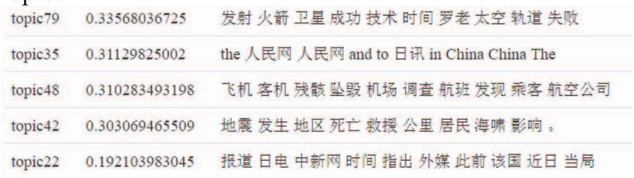
Following the steps described in section 3.1, the experimental corpus is shown as Fig. 4. The first column is the probability that the news belongs to the topic, and the second column is the news ID number, and the third column is the time of the news, and the fourth column is title of the news.



```
Topic1: 俄罗斯 普京 乌克兰 总统 莫斯科 杰夫 梅德韦 俄方 报道 消息
   0.476562 2340 2015年11月26日00:59 乌克兰停止缴纳预付款俄罗斯中断对乌供气
   0.456522 2231 2015年11月27日05:40 俄对乌再次“断气”
   0.431818 2653 2015年11月23日02:00 克里米亚宣布进入紧急状态输电线路遭人为破坏
   0.417939 2420 2015年11月25日05:39 乌克兰对克里米亚停电俄或对乌克兰断气断煤
```

Fig 4. The experimental corpus

The CCD of each topic can be calculated before optimizing after all news was classified to the most probable topic, as shown in Fig. 5. The middle column is the CCD of each topic.



topic79	0.33568036725	发射 火箭 卫星 成功 技术 时间 罗老 太空 轨道 失败
topic35	0.31129825002	the 人民网 人民网 and to 日讯 in China China The
topic48	0.310283493198	飞机 客机 残骸 坠毁 机场 调查 航班 发现 乘客 航空公司
topic42	0.303069465509	地震 发生 地区 死亡 救援 公里 居民 海啸 影响 。
topic22	0.192103983045	报道 日电 中新网 时间 指出 外媒 此前 该国 近日 当局

Fig 5. The example of LDA topic

As can be seen from Fig. 5, some topics are poor readability before optimizing, such as topic35. Very little information we can get from the topic, so it should be deleted. As can be seen from Table1, the average CCD and the average category similarity of the topics are low before optimization. However, the higher the two values are, the better the topics are.

Table1. The average CCD and the average category similarity of the topics before optimization

The average CCD	The average category similarity
0.412	0.37796

4.3 Remove Vague News and Result Analysis

As can be seen from section 3.2, it is necessary to set a threshold for optimizing the single topic to remove the news that is too similar to the topic. It will affect the average category similarity after deleting the news whose similarity of the topic below the threshold inevitably, meanwhile it will lose part of news. The magnitude of the threshold is determined from the effect of the threshold on the average category similarity and the loss degree. The effect of the threshold on the two indicators and BS is shown in Table 2.

Table2. The effect of the threshold on the V_{acs}, V_{ld} and BS

Threshold(θ)	V_{acs}	V_{ld}	BS
0.7	0.0927	0.985	0.0848
0.6	0.1266	0.9416	0.1117
0.5	0.1533	0.8557	0.13
0.4	0.2475	0.719	0.1697
0.3	0.2616	0.5041	0.1722
0.28	0.2927	0.4522	0.1707
0.26	0.3345	0.397	0.1706
0.24	0.2975	0.337	0.158
0.22	0.3329	0.2783	0.16
0.2	0.3513	0.2199	0.1353
0.18	0.3475	0.1633	0.1112
0.15	0.3568	0.0911	0.0725

We can see the influence of the threshold on the average category similarity and the loss degree more intuitively from Fig. 5 and Fig. 6.

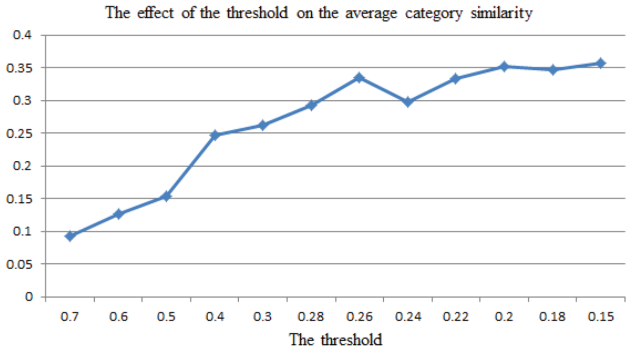


Fig 5. The effect of the threshold on the average category similarity

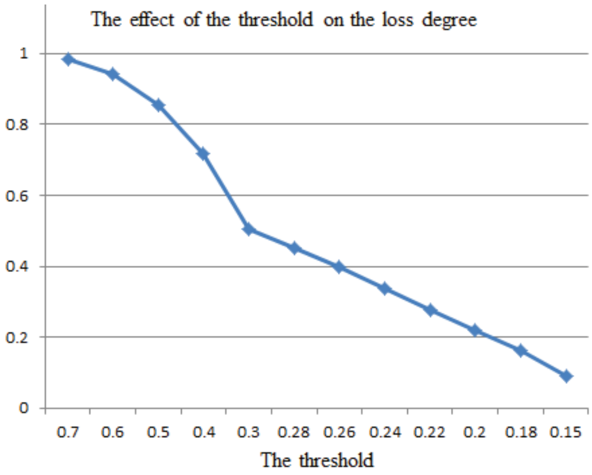


Fig 6. The effect of the threshold on the loss degree

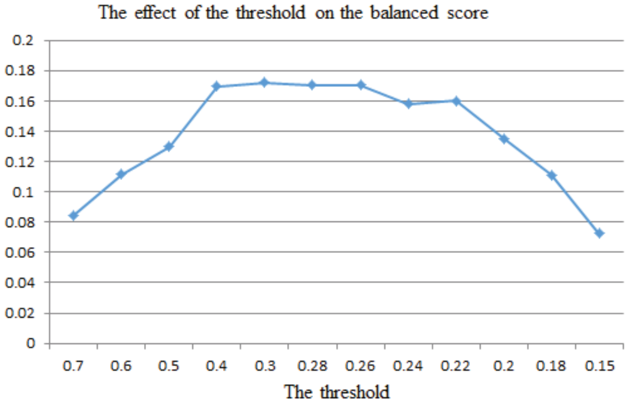
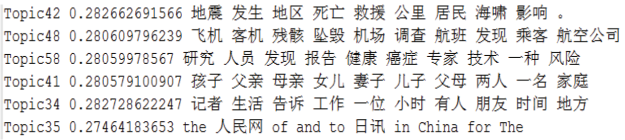


Fig 7. The effect of the threshold on the balanced score

It can be seen clearly from Fig. 5 that the average category similarity gradually increases with the decrease of the threshold when the threshold is higher than 0.3. The result shows that the threshold is higher, the more news deleted, and the lower the similarity between the topics is, and the higher category similarity is, which proves that the results are good. Further indicating that, it is necessary to remove the impurity news. However, when the threshold is less than 0.3, the average similarity of the category fluctuates with the decrease of the threshold, which may be related to the decrease of the number of news and the similarity between the news and the topic. As can be seen from Fig. 6, the number of deleted news is about half of the total when the threshold is 0.3. The loss tends to be flat with the decrease of the threshold, when the threshold is less than 0.3. Moreover, it can be seen from Fig. 7 that the balanced score is highest when the threshold is 0.3. In summary, the above experimental analysis shows that the threshold should be 0.3.

4.4 Remove Optimizing LDA Result Purity

The above results show that the average category similarity is better when the threshold is 0.3, and the similarity of news and topics is 0.3 has been regarded as a relatively small value. If the threshold is less than 0.3, the similarity of news and its topic is too low; in contrast, the loss degree will be too high. Sorting all the topics according to the CCD when the threshold is 0.3, the several topics of lowest score are shown in Fig. 8:



```
Topic42 0.282662691566 地震 发生 地区 死亡 救援 公里 居民 海啸 影响 。
Topic48 0.280609796239 飞机 客机 残骸 坠毁 机场 调查 航班 发现 乘客 航空公司
Topic58 0.28059978567 研究 人员 发现 报告 健康 癌症 专家 技术 一种 风险
Topic41 0.280579100907 孩子 父亲 母亲 女儿 妻子 儿子 父母 两人 一名 家庭
Topic34 0.282728622247 记者 生活 告诉 工作 一位 小时 有人 朋友 时间 地方
Topic35 0.27464183653 the 人民网 of and to 日讯 in China for The
```

Fig 8. The topics of lowest CCD when the threshold is 0.3

From Fig. 8 we can see that those topics are poor readability and poorly explanatory, they are difficult to express a specific topic, which should be deleted. And through Table 3 can be drawn, the average CCD is higher and the average category similarity reduced, which makes the topics more purity.

Table3. Comparison of the average CCD and the average category similarity between pre-optimization and optimized

The average category cluster density		The average category similarity	
pre-optimization	optimized	pre-optimization	optimized
0.412	0.455	0.37795	0.26169

In summary, the experimental analysis shows that the method of optimizing purity of LDA result based on semantic similarity proposed in this paper is effective and is of practical significance.

5 CONCLUSION

A method based on semantic similarity is presented in this paper to improve the purity of LDA results in streaming news processing, and a comparative analysis is carried out on the topics before and after optimization by using the average category cluster density and the average category similarity. The experimental results show that this method can identify and remove vague news that affect the purity of LDA topics, and effectively remove the topics with poor interpretability, which provides a way to increase LDA results purity for LDA topic analysis and LDA text clustering applications.

REFERENCES

[1] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent dirichlet allocation. Journal of Machine Learning Research, No.3, 993-1022, 2003.

[2] X. D. Li, Z. C. Ba and H. Li, Study on the influences of text categorization performance based on corpus information measurement, Journal of intelligence, No.9, 157-162, 2014.

[3] Z. Z. Wang, M. He and Y. P. Du, Text similarity computing based on topic model LDA, Computer science, 229-232, 2013.

[4] X. Zhao and X. M. Li, The Application of Theme Model in Text Mining, Report, Peking University, Beijing, 2011.6

[5] Y. Zhang, The research of text classification Technology Based on the part of speech and LDA topic model, Anhui University, Hefei, 2016.

[6] W. Li and Andrew McCallum, Pachinko allocation: Dag-structured mixture models of topic correlations, International Conference, 577-584, 2006.

[7] Z. M. Rosen, T. Griffiths and M. Steyvers, The author-topic model for authors and documents, Conference on Uncertainty in Artificial Intelligence, 487-494, 2004.

[8] A. Mccallum, X. Wang and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email, Journal of Artificial Intelligence Research, Vol30, No1, 249-272, 2007.

[9] Y. Xin, J. Yang and Z. Q. Xie, K-topic increment training algorithm based on LDA, Journal of Jilin University(Engineering and Technology Edition), Vol.45, No.4, 1242-1252, 2015.

[10] C. Zhang, L. Chen and Q. Li, Chinese text similarity algorithm based on PST_LDA, Application research of computers, Vol.33, No.2, 375-377, 2016.

[11] X. D. Li, F. Gao and C. Ding, Study on influences of different Chinese word segmentation methods to text automatic classification based on LDA model, Application research of computers, Vol.34, No.1, 62-66, 2017.

[12] B. Zhao, Topic optimization method based on pointwise mutual information, Harbin Institute of Technology, Harbin, 2012.

[13] R. Li, R. Kaptein and D. Hiemstra. Exploring Topic-based Language Models for Effective Web Information Retrieval, Proceedings of the 8th Dutch-Belgian Information Retrieval Workshop (DIR 2008), Vol.27, No.5, 65-71, 2008.

[14] D. Zhou, J. Bian and S. Zhengl, Exploring social annotations for information retrieval, International Conference on World Wide Web, WWW 2008, Beijing, China, 715-724, 2008.

[15] J. Chang and D. Blei, Relational topic models for document networks, Conference on Ai and Statistics, 2009.

[16] X. R. Wang, M. Andrew and X. Wei, Topical n-grams: Phrase and topic discovery, with an application to information retrieval, IEEE Computer Society. Seventh IEEE International Conference on Data Mining, 697-702, 2007.

[17] Y. Lu and C. X. Zhai, Opinion integration through semi-supervised topic modeling, International conference on World Wide Web, WWW 2008, New York, NY, USA, 121-130, 2008.

[18] Q. Z. Mei, D. Cai, D. Zhang and C. X. Zhai, Topic modeling with network regularization, International Conference on World Wide Web, WWW 2008, Beijing, China, 101-110, 2008.

[19] Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei. Hierarchical Dirichlet processes, Journal of the American Statistical Association, Vol.101, No.476, 566-1581, 2006.

[20] D. M. Blei and J. D. Lafferty, A correlated topic model of science, Annals of Applied Statistics, Vol.1, No.1, 17-35, 2007.

[21] X. Zu and F. Xie, A review of LDA topic Model, Journal of Hefei Normal University, Vol.33, No.6, 55-58, 2015.