A LDA Based Model for Topic Evolution: Evidence from Information Science Journals

Maoran Zhu, Xiaopeng Zhang* and Hongwei Wang
School of Economics and Management, Tongji University, Shanghai, China
*Corresponding author

***Abstract*—Mining evolution of topic from papers plays an important role in learning about the trend and monitoring the hot topic of research. The paper proposes a model based on Latent Dirichlet Allocation (LDA) for the purpose of mining evolution of topic. Firstly, we deal with the collection of all the papers using LDA to find out topics and their key words, and get probability distribution of document - topic on different time windows so that we can figure out the trend of topic intensity. Secondly, we apply LDA in papers on every single time window to get probability distribution of topic - word, through which we can compute similarity of topics from different time windows, and the words probability of similar topics can help us figure out trend of topic content.**

Keywords-evolution of topic; topic intensity; topic content; Latent Dirichlet Allocation

I. Introduction

Scientific journals are the vehicle for showing research achievement and spreading knowledge. Recently, scientific journals have been increasing exponentially so that it is hard for researchers to collect and deal with massive text from journals. We are supposed to dynamically track evolution of topic in certain fields to find out the appearance and development of new knowledge. Scientific journals consist of title, author, abstract, keyword, main body and reference, which contain a lot of high-value information, such as corporation between authors, Citation Analysis, and Co-occurrence Words Analysis. Especially, the occurrence of text mining and natural language processing provide technology support for research of topic mining and evolution based on massive scientific journals [1].

II. Related Work

We study topic evolution from two aspects: (1) evolution of topic intensity that reflects change of concern degree; (2) evolution of topic content that reflects the transference of concerns. Combining these two aspects, the stability, dynamics and development of topic can be revealed [15]. We are supposed to take time into consideration when studying evolution of topic, there are three ways to bring the element of time into topic model.

First one, we can take time as one parameter of topic model. Supposed that not only generation of words is affected by topic, but also affected by time. We regard time as continuous and observable variable to describe evolution of topic intensity overtime. Wang et al. [3] propose model named TOT (Topic Over Time), He et al. [4] propose model named dJST (dynamic joint sentiment- topic). These models don't require dividing timeline into time windows, ignore evolution of topic content and only focus on evolution of topic intensity. Besides, they deal with text offline which makes their extendibility not so strong.

Second one is prior discretization. We can mine topic on the collection of all documents using topic model, scatter documents to different time windows according to the generative time of document, and analyze topic intensity on different time windows. Griffiths et al. [5] mine topic on the collection of all documents using LDA, and compute distribution of document – topic on different time windows to analyze topic intensity. Hall et al. [6] study papers from ACL, COLING and so on, and compute probability of the topic on whole documents of certain time windows which is effective on describing research trends of certain fields. This way mines topic on the collection of all documents, so that there aren't problems such as topic alignment. However, it also deals with text offline and the extendibility is supposed to be strengthened.

Third one is post discretization. We firstly scatter documents to different time windows according to the generative time of document, and then mine topic on every time documents to study topic evolution. Blei et al. [7] propose Dynamic Topic Model which scatters documents to different time windows and supposes that number of topics generated from documents in every time windows is K. Wang et al. [8] bring timestamp information into evolution of parameter using Brownian Motion Model, building Continuous Time Dynamic Topic Model. Wei et al. [9] propose Dynamic Mixture Model, in which documents arrive at a time window according to time sequence and the model supposes that topics of two continuous documents have the evolution relation. Song et al. [10] propose a model named Incremental Latent Dirichlet Allocation, in which number of topics on every time windows is determined by unique Bayesian Model and distribution of words on topics describes evolution of topic content. L. AlSumait [11] propose a model named Online Latent Dirichlet Allocation which records existing topic, monitoring new topic, and describe evolution of topic content and topic intensity through evolution matrix. Besides, the model can update by itself with new documents arriving. This way scatters documents to different windows firstly, for which there is no need to train models again when new documents arrive at the collection of documents. However, it ignores the effect of documents' amounts to number of topic.

In conclusion, there are some shortcomings of researches nowadays: (1) for the way of prior discretization, topic alignment can't be guaranteed; (2) for the way of post discretization, the effect of documents' amounts to number of topic and distribution of topic - word on different time windows are ignored; (3) there are few researches about mining Chinese journals and classification of topics, and professional dictionary of certain fields is hard to be built.

This paper is to provide solutions of these two aspects: (1) describing topic intensity through proportion of documents containing certain topics in the collection of all documents on different time windows; (2) describing topic content from the dimensions of words and distribution of topic – word.

III. Evolution of Topic Using LDA Based on Scientific Journals

A. Research Frame

We can see the research frame in Figure I, the process is: (1) collecting abstract of papers of certain fields online; (2) pretreatment of documents collected; (3) figuring out the optimal number of topics when using LDA; (4) mining key words of topics from documents; (5) scattering documents to different time windows and getting distribution of topic – document on different time windows; (6) computing similarity of topics from different time windows and topic intensity on different time windows.

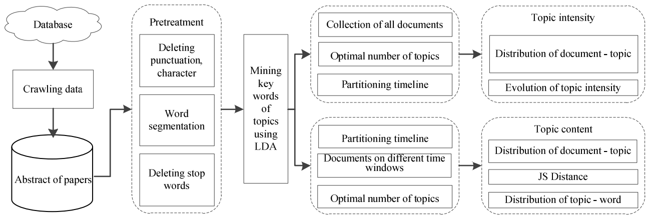
Database
Crawling data
Abstract of papers
Pretreatment
Deleting punctuation, character
Word segmentation
Deleting stop words
Mining key words of topics using LDA
Collection of all documents
Optimal number of topics
Partitioning timeline
Topic intensity
Distribution of document - topic
Evolution of topic intensity
Partitioning timeline
Documents on different time windows
Optimal number of topics
Topic content
Distribution of document - topic
JS Distance
Distribution of topic - word

FIGURE I. RESEARCH FRAME OF TOPIC EVOLUTION USING LDA BASED ON SCIENTIFIC JOURNALS

B. Latent Dirichlet Allocation

Before studying topic evolution, it's necessary to mine topics which can be regarded as dimension reduction of documents. Blei et al. [2] propose a model named Latent Dirichlet Allocation (LDA) in which K-dimension latent random variable obeying distribution of Dirichlet indicates topics probability distribution of documents to simulate generation of documents, as we can see in Figure II. LDA is three layer Bayesian model whose parameter is variable. The model supposes that a document can be indicated by polynomial distribution of some latent topics and a topic can be indicated by polynomial distribution of some words.

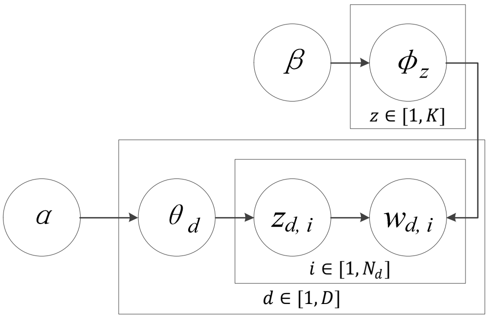

FIGURE II. THE GRAPH MODEL OF LDA

TABLE I. SYMBOL USED IN LDA

Symbol	Description of Symbol
D	Collection of Documents
K	Collection of Topics
N_d	Length of Document *d*
$w_{d,i}$	The *i* th Word of Document *d*
$z_{d,i}$	The *i* th Topic of Document *d*
α	Dirichlet Prior Distribution of Topics on Documents in LDA
β	Dirichlet Prior Distribution of Words on Topics in LDA
θ_d	Polynomial Distribution of Topics on Documents *d*
φ_z	Polynomial Distribution of Words on Topic *z*

Table I shows the symbol used in the LDA. θ_d and φ_z follow the Dirichlet Distribution, as (1). In the function, $0 \le \mu_k \le 1$, $\sum_k \mu_k = 1$, $\alpha_0 = \sum_{k=1}^{K} \alpha_k$, and Γ is gamma function. The generated process of documents using LDA is as Table II.

$$Dir(\mu \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \prod_{k=1}^{K} {\mu_k}^{\alpha_k - 1} \quad (1)$$

TABLE II. THE GENERATED PROCESS OF DOCUMENTS USING LDA

```
for all topics $z \in [1, K]$ do
    sample mixture components $\varphi_z \sim Dir(\beta)$
end for
for all documents $d \in [1, D]$ do
    sample mixture proportion $\theta_d \sim Dir(\alpha)$
    sample document length $N_d \sim Poiss(\xi)$
    for all words $i \in [1, N_d]$ do
        sample topic index $z_{d,i} \sim Mult(\theta_d)$
        sample term for word $w_{d,i} \sim Mult(\varphi_{z_{d,i}})$
    end for
end for
```

C. Evolution of Topic

Mining topics of documents and studying topic evolution involve two aspects: (1) evolution of topic intensity, which is measured by distribution of document – topic based on collection of documents on different time windows; (2) evolution of topic content, which is measured by distribution of topic – word among similar topics from different time windows.

1) Optimal number of topics. On the one hand, it's necessary to determine the optimal number of topics when using LDA and the more the documents are, the more the topics are in general. On the other hand, it's necessary for analyzing evolution of topic using LDA to evaluate generalization ability of the model, so that we can measure prediction ability of the model when dealing with data unobserved. This paper uses a widely recognized indicator that calls perplexity to measure generalization ability of the model and the littler the perplexity is, the stronger the generalization ability is. When the number of topics is different, the perplexity is also different, for which the optimal number of topics can be determined by computing perplexity on different number of topics. The calculation formula is as (2):

$$Perplexity(D) = exp\left\{-\frac{\sum_{d=1}^{D} logp(w_d)}{\sum_{d=1}^{D} N_d}\right\} \quad (2)$$

In the formula, $p(w_d)$ means probability generating document *d*, and the calculation formula is as (3):

$$p(w_d) = \prod_{i=1}^{N_d} \sum_z p(w_{d,i} \mid z) p(z \mid d) \quad (3)$$

2) Calculation of topic intensity. Topic intensity shows how much certain topics are paid attention on certain time windows. In other words, the more documents containing certain topics are, the stronger topic intensity of these topics is. We set θ_z^d as proportion of topic z in document d, set Dt as collection of documents on time window t, and set θ_z^t as topic intensity of topic z on time window t. The calculation formula is as (4):

$$\theta_z{}^t = \frac{\sum_{d=1}^{D_t} \theta_z{}^d}{D_t} \quad (4)$$

After getting θ_z^t, topic intensity of topic *z* on time window *t*, we draw a line chart about topic intensity of topic *z* changing by time to analyze evolution of topic intensity.

3) Calculation of topic similarity. In related work about evolution of topic, when mining topics of documents on different time windows separately, topic alignment can't be guaranteed. For this reason, we are supposed to compute similarity among topics from different time windows to satisfy topic alignment [13]. After computing similarity among topics from different time windows, we analyze evolution of topic content through distribution of topic – word among similar topics [14]. Widely used ways of calculation include Cosine Distance, Kullback-Leibler Difference Distance, and Jenson-Shannon Distance.

We set Z_m and Z_n as topics mined on collection of documents, which can be expressed as Z_m^{ti}, Z_m^{ti+1}, Z_n^{ti} and Z_n^{ti+1} on neighbor time windows t_i and t_{i+1}. Documents on time window *i* and time window *i+1* are combined into collection *V*. We set *p* as probability of Z_m^{ti} on *V*, and *q* as probability of Z_n^{ti+1} on *V*. The formula of KL Distance between Z_m^{ti} and Z_n^{ti+1} is as (5):

$$KL(Z_m{}^{t_i}, Z_n{}^{t_{i+1}}) = D(p \mid\mid q) = \sum_i^V p_i log \frac{p_i}{q_i} \quad (5)$$

It shows the difference between two topics on V, and the littler the difference is, the more similar the two topics are. The similarity is supposed to be symmetrical, but KL Distance is not symmetrical, for which Jenson-Shannon Distance is used to compute similarity between two topics, as formula (6):

$$\mathrm{Sim}(Z_m{}^{t_i}, Z_n{}^{t_{i+1}}) = JS(p,q) = \frac{1}{2}(D(p \mid\mid m) + D(q \mid\mid m)) \quad (6)$$

In the formula, $m = \frac{1}{2}(p+q)$. The interval of JS Distance is [0, 1], and the littler the JS Distance is, the more similar the two topics are.

IV. Design of Experiment

A. Data Source

This paper focuses on journals of Informatics in China which are Journal of The China Society for Scientific and Technical Information, Journal of Chinese Information Process, Journal of Information, Information and Documentation Services, Modern Information, and Information Studies: Theory & Application. We crawl abstract of papers on these journals from wanfangdata.com.cn (Wanfang Data, an affiliate of the Chinese Ministry of Science & Technology, provides access to a wide range of database resources) between 2000 and 2015, whose amount adds up to 29,552, and scatter all the documents into different time windows by year.

B. Pretreatment of Documents

Documents from those journals contain lots of numeric character, English character, and high-frequency meaningless words and it's necessary to filter noise in the documents. Through programing with python and professional dictionary of stop words, we remove these meaningless components and filter the documents into words using the package jieba. We divide timeline into continuous time windows by year for which papers from same year but different journals are put together as a collection of documents.

C. Evolution of Topic from Journals

1) Evolution of topic intensity. When analyzing evolution of topic intensity, we use the way of post discretization, for which we compute perplexity of LDA on different number of topics based on the collection of all documents. The experiment shows that the perplexity and number of topics could be balanced while the number of topics was 35, as we can see in Figure III. So we set number of topics as 35 and set α and β as default value. We mine topics based on the collection of documents using LDA and get distribution of document – topic. The documents are scattered into 16 time windows from 2000 to 2015 and topic intensity on every time windows is computed. We draw the hot map of topics as Figure IV and we choose some active topics to show key words as Table III.

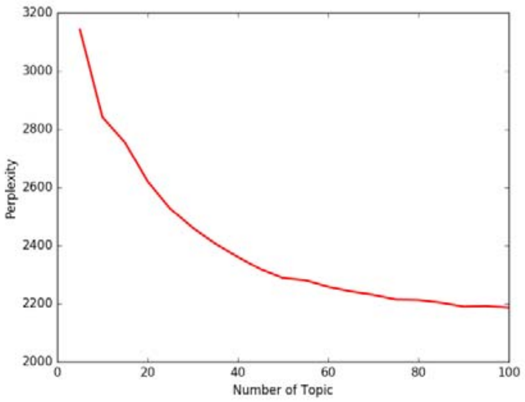

FIGURE III. PERPLEXITY WHEN USING DIFFERENT NUMBER OF TOPICS

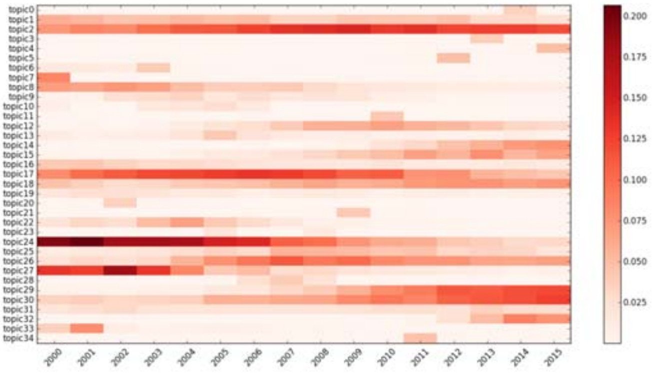

FIGURE IV. HEAT MAP OF TOPICS

TABLE III. TOPICS AND THEIR KEY WORDS

Topic	Key Words
Topic1	Intelligence, China, Society, Design, Science, Research, Website, Intellectual Property, Technology, Software
Topic2	Information, Knowledge, Service, User, Characteristic, Resource, Algorithm, Internet, Relationship, Structure
Topic8	Database, Content, Society, Principle, Conception, Market, Informatics, Internet, Strategy, Trend
Topic17	Management, Enterprise, Digitization, Knowledge, Model, Competition, Internet, Strategy, Data, Economy
Topic24	Information, Development, Resource, Literature, China, Technology, Retrieval, System, Platform, Digital
Topic26	Evaluation, Model, System, University, Library, Demand, Conference, Frame, Literature, Knowledge, Text
Topic27	Library, University, Internet, Service, Economy, Information, Innovation, Demand, Knowledge, Journals
Topic29	Internet, Literature, Development, Model, Means, Structure, Subject, Calculation, Hotspot, Recommendation
Topic30	Technology, Analysis, Fields, Information, Content, Semantics, Cooperation, Experiment, Effect, Industry

From Table III, Topic1 is about Intellectual Property Protection, Topic2 is about Internet Service and Resource, and Topic17 is about Enterprise Intelligence Management. We can get trend of topic intensity by computing probability of topic based on documents on different time windows, as Figure V.

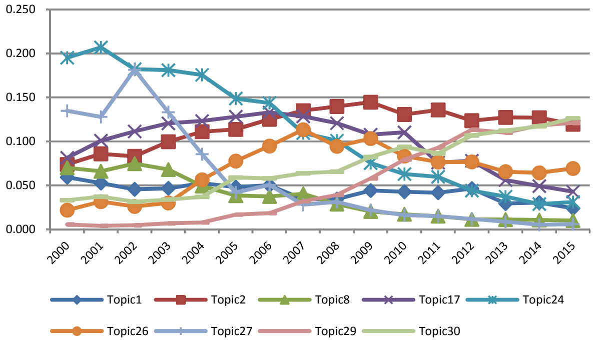

FIGURE V. EVOLUTION OF TOPIC INTENSITY

In Figure V, the horizontal axis represents time window, the vertical axis represents intensity, and points of line chart represent topic intensity of certain topics on certain time windows. The change of focus in Informatics fields can be seen in Figure V. Topic intensity of certain topics is on decline, such as Topic24 and Topic27. Topic24 is about Literature Retrieval Technology and Topic27 is about Network and Information Construction of University Library. At the beginning of 2000, the two topics were highly focused on, which can be explained by responding to national policies of information construction. With the improvement of related research, the degree of attention was decreasing year by year. Topic intensity of certain topics is stable, such as Topic1 which is about Intellectual Property Protection. Intellectual Property Protection is easily affected by policies in different times and background, for which it is to get continuous attention. However, the topic intensity of such topic isn't so high overall which means the topic doesn't get too much attention. Topic intensity of certain topics is on increase, such as Topic30 which is about Semantics Analysis. Nowadays, with explosive growth of User Generated Content (UGC), Semantics Analysis has great significance on Network Opinion Monitoring and Emotion Analysis, which is to be an important branch of Informatics and get more and more attention.

2) Evolution of topic content. We use way of prior discretization which deals with documents on different time windows separately using LDA and gets distribution of document – topic and topic – word. We get similar topics among topics from different time windows by means of JS Distance and Specialty Word Dictionary. The distribution of topic – word can help us analyze evolution of topic content.

This paper takes topic about Enterprise Intelligence Management as an example, and Table IV shows key words of similar topics on different time windows. Combining the distribution of topic – word, we draw hot map of key words, as Figure VI.

TABLE IV. TOPICS AND THEIR KEY WORDS ON DIFFERENT TIME WINDOWS

time/topic	Key Words
2000/Topic9	Enterprise, Competition, Intelligence, Information, Market, Patent, Environment, Talent, Innovation, Model, Circulation, Strategy
2001/Topic15	Enterprise, Competition, Digital, Intelligence, Management, Publication, Model, Circulation, Strategy, Operation, System, Advantage
2002/Topic10	Enterprise, Competition, Market, Intelligence, Production, E-business, Marketing, Value, Strategy, Commodity, Economy, Investment
2003/Topic0	Enterprise, Competition, Intelligence, Strategy, System, Technology, Competition, Development, Patent, Informatization, Management, Model
2004/Topic7	Intelligence, Competition, Enterprise, Development, Theory, Business, Practice, Risk, E-business, Advantage, Government
2005/Topic5	Enterprise, Competition, Intelligence, Strategy, Competition, Legend, Capital, Model, Value, Brand, Value Chain, Research
2006/Topic5	Knowledge, Enterprise, Management, Competition, Intelligence, Strategy, Innovation, Organization, Sharing, System, Value Chain, Information
2007/Topic15	Competition, Enterprise, Intelligence, Patent, Strategy, Knowledge, Technology, Value, Environment, Process, Innovation
2008/Topic13	Enterprise, Competition, Intelligence, Strategy, Technology, Market, System, Innovation, Service, Research, Management
2009/Topic3	Enterprise, Competition, Intelligence, Risk, Project, Strategy, Warning, Relationship, Information, Industry, Mechanism
2010/Topic6	Intelligence, Competition, Enterprise, Research, Industry, System, Analysis, Relationship, Service, Risk, Track
2011/Topic6	Enterprise, Competition, Innovation, Intelligence, Industry, Strategy, Core, legend, Technology, Management, Model, Advantage
2012/Topic2	Enterprise, Competition, Intelligence, Innovation, Cluster, Model, Industry, Decision, Legend, Cooperation, Strategy, Demand
2013/Topic4	Intelligence, Competition, Product, Analysis, Value, Industry, Service, Research, Enterprise, Content, Institution
2014/Topic19	Intelligence, Competition, Enterprise, Analysis, Strategy, Decision, Product, Research, Industry, Management, Demand
2015/Topic17	Intelligence, Competition, Enterprise, Analysis, Strategy, Research, Think-tank, Technology, Construction, System

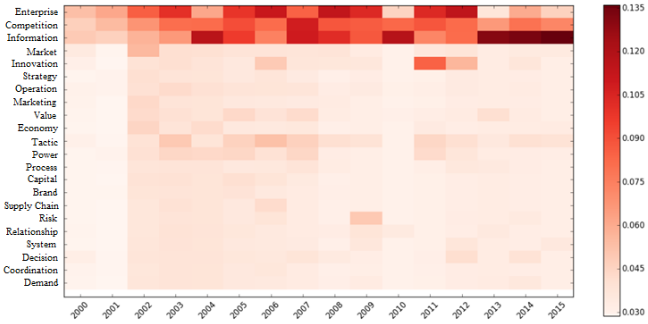

FIGURE VI. HOT MAP OF KEY WORDS

In Figure VI, the horizontal axis represents time window, the vertical axis represents key words contained by the topics, and the intensity of color represents probability of the word under the topic. On different time windows, content contained by certain topics will change. This paper measures change of content using change of words under certain topics. Figure VI is about evolution of topic content on Enterprise Intelligence Management. As we can see, the probability of enterprise and competition has always been high, which supposes that Enterprise Competitive Power or competition among enterprises has always been hotspot under the topic. Besides, the probability of Intelligence is also high, and becomes higher and higher, which means there is new breakthrough on aspect of Intelligence.

V. CONCLUSION

This paper focuses on evolution of topic using LDA based on Chines journals on Information Science. The combination of prior discretization and post discretization make it more comprehensive when analyzing evolution of topic based on journals. It has great significance for tracking trend and hotpot of certain academic fields. It can also be promoted to commercial application, such as generation, development and disappearance of certain topics on certain internet forum and Public Opinion Monitoring.

Based on better words segmentation, future work will combine evolution of topic based on journals with Citation Analysis to describe evolution path of topic. We will also refer to Author Relationship Network, which can help recognize the authors' position in the network, to improve analyzing trend of evolution.

ACKNOWLEDGEMENT

This work is partially supported by the Natural Science Foundation of China [71371144, 71601119, 71601082] and

Shanghai Philosophy and Social Science Planning Projects [2013BGL004].

REFERENCES

[1] Y. Chen, H. Amiri, Z. Li, and T. S. Chua, "Emerging topic detection for organizations from microblogs." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 43-52, July 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation." Journal of machine learning research. 3, pp. 993-1022, Jan 2003.

[3] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 424-433, 2006.

[4] Y. He, C. Lin, W. Gao, and K. F. Wong, "Dynamic joint sentiment-topic model." ACM Transactions on Intelligent Systems and Technology (TIST). 5(1), 6, 2013.

[5] T. L. Griffiths, A. N. Sanborn, K. R. Canini, and D. J. Navarro, "Categorization as nonparametric Bayesian density estimation." The probabilistic mind: Prospects for Bayesian cognitive science. pp. 303-328, 2008.

[6] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. pp. 363-371, 2008.

[7] D. M. Blei, "Probabilistic topic models." Communications of the ACM. 55(4), pp. 77-84, 2012.

[8] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models." arXiv preprint arXiv: 1206. 3298. 2012.

[9] X. Wei, J. Sun, and X. Wang, "Dynamic Mixture Models for Multiple Time-Series." Ijcai. Vol. 7, pp. 2909-2914, 2007.

[10] X. Song and B. L. Tseng, "Methods and systems for utilizing content, dynamic patterns, and/or relational information for data analysis." U.S. Patent. No. 7,853,485, 14, Dec 2010.

[11] L. Alsumait, D. Barbará, J. Gentle, and C. Domeniconi, "Topic significance ranking of LDA generative models." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, pp. 67-82, 2009.

[12] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, pp. 391-402, 2010.

[13] L. Ping, and H. Weidong, "Event Topic Evolution of Network Public Opinions: An Analysis Based on LDA Model." Journal of Intelligence. 12, 005, 2013.

[14] Y. Hu, L. Bai, and W. Zhang, "OLDA-based method for online topic evolution in network public opinion analysis." Journal of National University of Defense Technology. 34(1), pp. 150-154, 2012.

[15] B. Shan, and F. Li, "A Survey of Topic Evolution Based on LDA." Journal of Chinese Information Processing. 6, 008, 2010.