Topic Detection and Evolution Analysis on Microblog

Guoyong Cai, Libin Peng, Yong Wang

Topic Detection and Evolution Analysis on Microblog*

Guoyong CAI, Libin PENG, Yong WANG

(*Guilin University of Electronic Technology,Guangxi Key Lab of Trusted Software,541004 China*)

ccgycai@guet.edu.cn ccsupeng@163.com hellowy@126.com

Abstract: The study on topic evolution can help people to understand the ins and outs of topics. Traditional study on topic evolution is based on LDA model, but for microblog data, the effect of this model is not significant. An MLDA model is proposed in this paper, which takes microblog document relation, topic tag and authors relations into consideration. Then, the topic evolution in content and intensity is analyzed. The experiments on microblog data have shown the effectiveness and efficiency of the proposed approach to topic detection and evolution analysis on Microblog.

Keywords: Microblog, LDA model, Topic Evolution, Topic Detection

1.Introduction

Microblog is the shortened form of micro blog (MicroBlog). It is a type of platform which allows users to disseminate/obtain/share information based on relationship among users. Users can set up personal community through WEB, WAP and other client terminals, update the information with no more than 140 words of text, and achieve instantly information sharing. However, the content of microblog is diverse and changing rapidly. It is a challenging problem of how to discover automatically an effective topic and to analyze the evolution of the discovered topic.

Topic is defined as a number of related events caused by a set of seed in Topic Detection and Tracking[1]. The topic model represented by LDA (Latent Dirichlet Allocation) [2] is an important technology in the field of text mining in recent years. LDA model has a good ability of dimension reduction and scalability. It has achieved great success in mining traditional network news topic. Topic evolution is referred to the migration of topic content and intensity over times[3]. Researches on topic evolution in general are based on LDA model to extract topics, and then the topic evolution of content and intensity are analyzed. However, compared with the traditional network text, Microblog has distinct characteristics, such as a short text (usually less than 140 characters), sparse data, noise data, mixed and disorder content, etc. Besides, there also exist social relations, structural social network information. Therefore, it is not effective to use LDA model for microblog text. An MLDA model is proposed in this paper, which takes microblog document relation, topic tag and authors' relationships into consideration. The topic evolution in content and intensity are analyzed based on MLDA model.

2.Related Works

Due to its good scalability, LDA model, which proposed by Blei in 2003[2], is extended by many scholars. Researches on topic evolution based on the scalable LDA model are divided into the

* This work is supported by the NSFC(#61063039)，Guangxi Key Lab of Trusted Software(#kx201202)

following three categories:

(1) Continuous time models. The information of time is taken into topic model as a variable to study the topic evolution. For example, a TOT (Topic Over Time) model is proposed by Wang [4]. In TOT model, each document is a mixture of topics that is influenced jointly by its time stamp. The main disadvantage of TOT is that it uses the beta distribution to model the topic development trends. Therefore, for documents that don't release time, it will predict the release date of the document. For documents with given released date, it can predict the document distribution.

(2) Pre-discretization methods. In this type of method, documents are divided into some parts according to the time windows before modeling topics, and then documents are processed and the topic distribution is generated in each time window. Song and etc.[5] proposed a ILDA model (Incremental Latent Dirichlet Allocation) to study the content evolution of a topic and solve ILDA model with Gibbs sampling approach[6]. Online LDA model is proposed by AlSumait[7], history data are used as the prior distribution of the proposed model and LDA model are used to study topic evolution for the arrived data in each time interval. Hu Yanli and etc. [9] have also implemented the online topic evolution analysis based on online LDA model, but they consider the inter-connection of topic distribution in each time slice.

(3) Post-discretization methods. In this type of method, the affect of time is not considered first. After topics are extracted based on LDA in the document set, the topics are divided into time slice according to its time stamp. Grillffiths and Steyvers[8] have proposed a post-discretization method based on LDA model, they use the intensity of topics in each time slice to indicate the topic trends.

3.Topic Evolution Analysis on Microblog

3.1 LDA model

LDA model is a three-layer probability model. It consists of word, topic and document. The key idea behind the LDA model is to assume that the words in each document were generated by a mixture of topics, where a topic is represented as a multinomial probability distribution over words[9]. Each document has a specific mixture of topics, which generated from the Dirichlet distribution. The specific idea behind LDA is to assume that each document correspond to the multinomial distribution θ of T topics, each topic correspond with the multinomial distribution φ of N_d words, θ is the prior of Dirichlet distribution with parameters α , φ is the prior of Dirichlet distribution with parameters β . For each word in a document d, topic z is extracted from the distribution θ of the document, and the word w is extracted from the distribution φ of topic z .This process will be repeated N_d times, then the document d is generated. The generative process of LDA is shown in figure1.

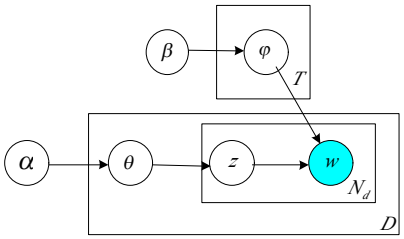

Figure 1 LDA model

The shaded circles in Figure 1 represent the observed variable and the unshaded circles represent the latent variables. The arrows between two variables represent the conditional-dependence and boxes represent repeated sampling, the number of repetitions is located at the lower right corner of the box. There are two parameters need to be inferred in this model, one is the distribution θ of document-topic and the other is the distribution φ of topic-word. Since it is difficult to obtain precise parameters, VEM[2], Gibbs Sampling[6] and Expectation propagation[10] are often applied to estimate the parameters. Gibbs Sampling has been widely used for its simple implementation.

3.2 Topic Discovery Model ----MLDA model

Compared with other texts, microblog texts have special symbols, such as "@"、"# " and "retweet". @ indicates the author's relationship of a microblog post. For example, a message "@Sandy Congratulations, you get a good job." and another message "@Sandy Can you teach me some IT knowledge used in a work". When considering the author's relationship, we can set a connection between these two seemingly unrelated microblogs and consider that “job” in the first message is related to "IT" in the second message. # indicates topic tag in a microblog. For example a message "Sunyang， come on! #Olympic Games#". If considering the topic tag, "Sunyang" and "Olympic Games" are related. "retweet" indicates microblog documents’ relation. For example, a message "Chinese Dream, retweet @Sandy the best popular new vocabulary …". It is difficult to analyze the specific meaning for "Chinese Dream", but compared with the original microblog, we know that "Chinese Dream" is a kind of "new vocabulary".

An MLDA model, extended from LDA, is proposed in this paper, which takes microblog document relation, topic tag and authors’ relationships into consideration. The parameters of MLDA model is shown in table 1. The Bayesian network of MLDA is showed in figure 2. c indicates the author's relationship, α_r indicates microblog document relation. t indicates topic tag in mircoblog. α_c is the parameter of distribution, θ_c associate with authors' relation. θ_c is computed by Dirichlet distribution with parameter α_c. α is the parameter of distribution θ_d of document-topic. α_t is the parameter of distribution θ_d associate with special topic. α_r is the parameter that decide whether it is a retweet message. The distribution θ of topics in microblog is shown as follow.

$$P(\theta \mid \alpha, \alpha_c, \alpha_t, \pi_c, t) = [P(\theta_c \mid \alpha_c)^{\pi_c} P(\theta_d \mid \alpha)^{1-\pi_c}]^{1-t} P(\theta_d \mid \alpha_t)^t \quad (1)$$

If there exists a symbol "@" in microblog, the value of π_c is 1, otherwise π_c is 0. If there exists a symbol "#" in microblog, boolean variable t is 1, otherwise t is 0. If there exists "retweet", the distribution of $z_{d,i}$ over topics depend on θ_{dRT} and θ_d, sampling from a Bernoulli distribution with parameter a_r to decide θ_{dRT} and θ_d, and based on them we extract the current topic $z_{d,i}$. Otherwise, we extract topics $z_{d,i}$ directly from Multinomial distribution with parameter θ_d.

For a microblog, the joint probability distribution over all words and their topics is computed as following equation (2).

$$P(w,z|\alpha_r,\theta,\beta)=P(\theta_{dRT}|\alpha_r)P(z|\theta)P(w|z,\beta)$$
$$=P(\theta_{dRT}|\alpha_r)P(z|\theta_d)^{1-\alpha_r}P(z|\theta_{dRT})^{\alpha_r}P(w|z,\beta) \quad (2)$$

Table 1：Parameters in MLDA

Parameter	Definition
α ， α_t α_c ，β	The prior parameters of $\theta_c,\theta_d,\varphi$
θ_c	Topic distribution associate with authors' relation
θ_d	Topic distribution over microblog d
φ	The distribution of topic-word
$z_{d,i}$	The ith topic in document d
$w_{d,i}$	The ith word in document d
D	The number of documents
T	The number of topics
N_d	The number of words
π_c	Bool parameter, decide whether there exists a @ message. If so, π_c is 1, else π_c is 0.
α_r	Decide whether it is a retweet message

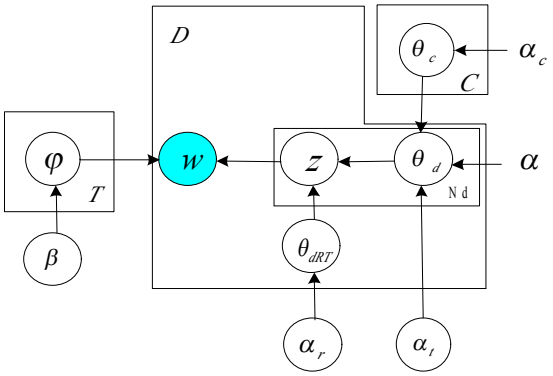

Figure 2 Bayesian network of MLDA

The description of generative process in MLDA is as follows:

(1) For each document $d \in D$, if there exists "#topic#", we compute $\theta_d=\theta_d \sim Dri(a_t)$. If there exists a symbol "@", we compute $\theta_d=\theta_c \sim Dir(\alpha_c)$. In other cases, we compute $\theta_d=\theta_d \sim Dir(\alpha)$.

(2) For each topic $z_{d,i}$, we judge document relation with "retweet". If there exists “retweet”, then $z_{d,j} \sim Multiomial(\theta_{dRT})$. Based on $z_{d,i}$, we generate $\varphi \sim Dir(\beta)$.

(3) For each word $w_{d,i}$ in a document, firstly, we choose a topic $z_{d,j} \sim Multiomial(\theta_d)$; and then we choose a word with $w_{d,j} \sim Multiomial(\varphi)$.

3.3 Algorithm Implementation

We apply Gibbs sampling to estimate the parameters θ, φ. Gibbs sampling is one simple realization method of MCMC (Markov chain Monte Carlo), which is a fast and effective algorithm for estimating the parameters[11]. Gibbs sampling is based on posterior distribution of words given a topic $p(z \mid w)$. Repeated iterations on the probability distribution, the parameters are derived. The process is computed with equation (3).

$$P(z_i = j \mid z_{-i}, w_{d,i}, \alpha, \beta) \propto \frac{n_{-i,j}^{(w)} + \beta}{n_{-i,j}^{(\cdot)} + N_d \beta} * \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + N_d \alpha} \tag{3}$$

Where, $n_{-i,j}^{(w)}$ is the word counts assigned to topic j and w. $n_{(-i,j)}^{(\cdot)}$ is the counts of words assigned to topic j. $n_{-i,j}^{(d)}$ is the counts of words assigned to topic j in document d. $n_{-i,\cdot}^{(d)}$ is the counts of words assigned to topic j. All counts do not include the current iteration. After iteration finished, we can estimate θ and φ from the value z using equation (4).

$$\theta = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + T\alpha} \qquad \varphi = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + N_d \beta} \tag{4}$$

Where, $n_j^{(d)}$ denotes the number of words that assigned to topic j in document d. $n_{\cdot}^{(d)}$ denotes the number of words appeared in document d. $n_j^{(w)}$ denotes the number that word w is assigned to topic j. $n_j^{(\cdot)}$ is the number of words that assigned to topic j.

The Gibbs sampling procedure of MLDA is implemented and shown in Fig.3. The Gibbs sampling algorithm consists of three parts: initialization, burn-in and sampling.

Algorithm: MLDA_Gibbs Sampling

Input: $V, T, \alpha, \alpha_c, \alpha_t, \lambda, \beta$

Output: associated topic Z, the distribution θ of document-topic, the distribution φ of topic-word.

//initialization

$n_j^{(d)} = 0, n_{\cdot}^{(d)} = 0, n_j^{(w)} = 0, n_j^{(\cdot)} = 0;$

for ： all document $d \in [1, D]$ do

for: all word $n \in [1, N_d]$ in document d do

$z_{d,j} \sim Multiomial(\theta_{dRT})$ //computed by Equation 1，Equation 2

$n_j^{(d)}$ +=1;

$n_{\cdot}^{(d)}$ +=1;

$n_j^{(w)}$ +=1;

$n_j^{(\cdot)}$ +=1;

//burn-in and sampling period

While (not reach maximum iteration) do

for ： all document $d \in [1, D]$ do

for: all word $n \in [1, N_d]$ in document d do

//decrement counts and sums:

$n_j^{(d)}$ -=1; $n_\cdot^{(d)}$ -=1; $n_j^{(w)}$ -=1; $n_j^{(\cdot)}$ -=1;

$Z \sim P(z_i = j \mid z_{-i}, w_{d,j})$ //multinomial sampling according to equation 3

//for the new assignment Z to the term t for word $w_{d,j}$

$n_j^{(d)}$ +=1; $n_\cdot^{(d)}$ +=1; $n_j^{(w)}$ +=1; $n_j^{(\cdot)}$ +=1;// increment counts and sums

If(finish sampling) then

Output θ, φ // according to equation 4.

Figure 3. Gibbs sampling algorithm

3.4 Topic Evolution

Topic evolution indicates that the same topic shows dynamism and difference with time going. The evolution of topic reflects in two aspects. First, the topic intensity changes over time. For example, the topic is the Olympic Games which take place every four years, which are active for Olympic years, other years is low. On the other hand, the topic content changes with the passage of time. For example, on the eve of the Olympic Games, when we pay more attention to the Olympic preparations, on the middle of Olympic, gold medal topics will be focused. In the end of Olympic Games, more attention should be paid to the summary and inventory of the Olympic Games. We use the distribution of topic j with time t to define the intensity of topic.

$$\delta_j^t = \frac{1}{D^t} \sum_{d \in D^t} \theta_{d,j} \qquad (5)$$

Where δ_j^t is the intensity of a topic, D^t is the number of document, $\theta_{d,j}$ is the distribution of topic j.

The content of topic evolution is characterized as the degree associated with topics. The differences of the distribution of two topics is described with KL(Kullback-Leibler) distance[12]. The smaller the difference is, the higher degree of the two topics will be associated. Assuming that the probability distribution of topic A is $A = (A_1, A_2 ..., A_n)$, and B is $B = (B_1, B_2 ..., B_n)$, the KL distance between topic A and B is computed as follow.

$$KL(A, B) = D(A \| B) = \sum_i^N A_i \log \frac{A_i}{B_i} \qquad (6)$$

4.Experiments

The experimental data are captured from a large microblog site named Sina microblog. Through API provided by Sina, about 8 million original microblog data from July 1th 2012 to August 29th 2012 are captured. We mark five topics, i.e., "London Olympics", "heavy rains in Beijing", "Price War", "Diaoyu Island Event", "The Voice of China". In order to save space or to speed up sampling, the punctuation and stop words in the original microblog dataset must be removed before experiments. We complete this preprocessing work by using a punctuation list and a stop words dictionary.

4.1 Topic Discovery

Effectiveness experiments are conducted on dataset mainly to examine the performance of MLDA. The parameters are set up with $\alpha = 0.2$, $\alpha_c = \alpha_t = \lambda = 1$, $\beta = 0.1$, $T = 5, N_d = 716$. The key words of each topic are extracted by MLDA models. The results are shown in table 2.

Table 2. The list of representative words for each topic.

topics	Most representative words
Topic1: London Olympic	Olympic Games, participation, London, champion, gold metal
Topic2: Heavy rains in Beijing	heavy rains, disaster, rescue, flood protection, emergency
Topic3: Price War	Sunning, jingdong, price war, E-commerce, fund
Topic4: Diaoyu Island Event	Diaoyu Island, Japan, China, Sovereignty, protection
Topic5: The Voice of China	Voice, China, music, pleasant music, jury

4.2 Intensity evolution of Topics

We analyse the topics from July to August in 2012. The different change of the intensity of topics will be calculated in different times, which is divided into eight time slice by each a week. The intensity of each topic is shown in figure 3.The ordinate represents the intensity of the topic at each time slice, which is calculated with equation 5.

Figure 3 shows us that the intensity of topic 1 is higher than other topics during each time slice. This demonstrates that the influence of topic 1 is higher and is a hot topic during the two months. Besides, for each topic, the intensity of topics will change with time. This shows that each topic will go through a evolutional process that is consistency with reality.

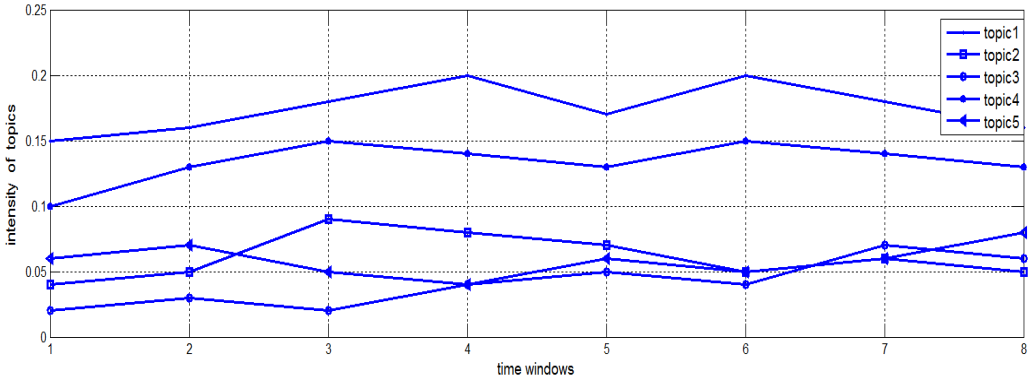

Figure 3 the intensity evolution of topics

4.3 Content evolution of Topics

For topic 1, the intensity is higher over time. The similarity distance of a topic during the adjacent time slice indicates the content evolution of the topic. KL distance is calculated according to the probability distribution of words in eight time intervals. Figure 4 presents the KL distance of topic 1 changes over the intervals, which indicates that the content of the topic changes over intervals. As can be seen from the figure 4 that the KL distance of topic 1 enlarge abruptly during the time slice 4 and 5, which indicates that the content of topic 1 changes quite a few. Figure 5 shows that the content of topic 1 have changed from "gold metal" into "open ceremory". During time slice 6 and7, the KL distance enlarge abruptly, which indicated that the content of topic1 have changed into "closing ceremony". The topic changing process is shown in figure 5.

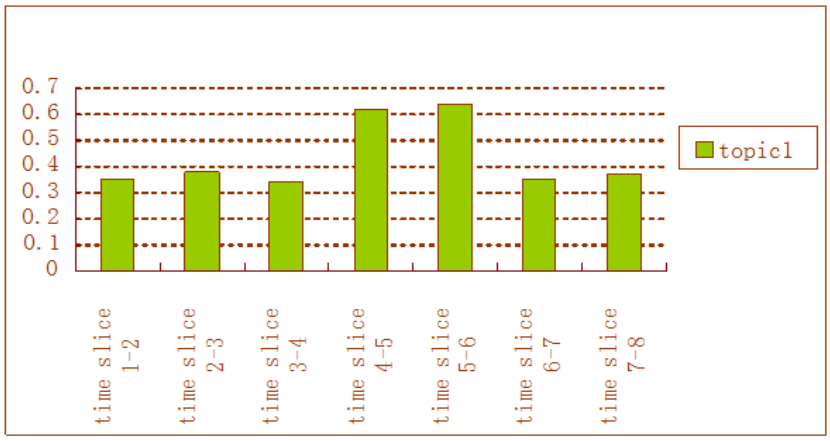

Figure 4 the KL distance of topic 1

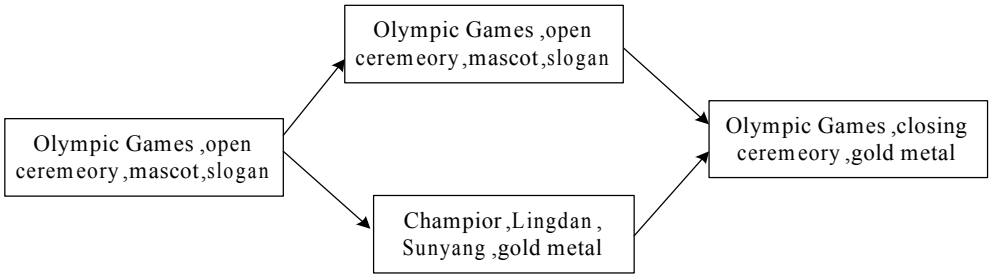

Figure 5 the chart of topic 1 evolution

In summary, the trend of topic evolution can be captured quite accurately with MLDA model. It

indicates not only the intensity change of the topic over time, but also the content change of the topic over time.

4.4 Comparisons of algorithm performance

We conduct a comparative experiment between MLDA and LDA, which is a baseline model in the field of topic modeling. The metric Perplexity [13] is a standard measure of performance for statistical models, which indicates the uncertainty in predicting a single word; the model with lower value is better in performance. Perplexity is defined as follows:

$$Perplexity(M_{test}) = \exp\{-\frac{\sum_{m} \ln p(w_m)}{\sum_{m} N_m}\} \quad (7)$$

where M_{test} is a test set with m documents, w_m and N_m indicate the observed words and number of words in the test document respectively.

Perplexity is used to measure the performance of LDA and MLDA under the same hyperparameters setting, and the result is shown in figure 6. Figure 6 show that the perplexity of MLDA is always lower than LDA, which show that the performance of MLDA is much better than LDA.

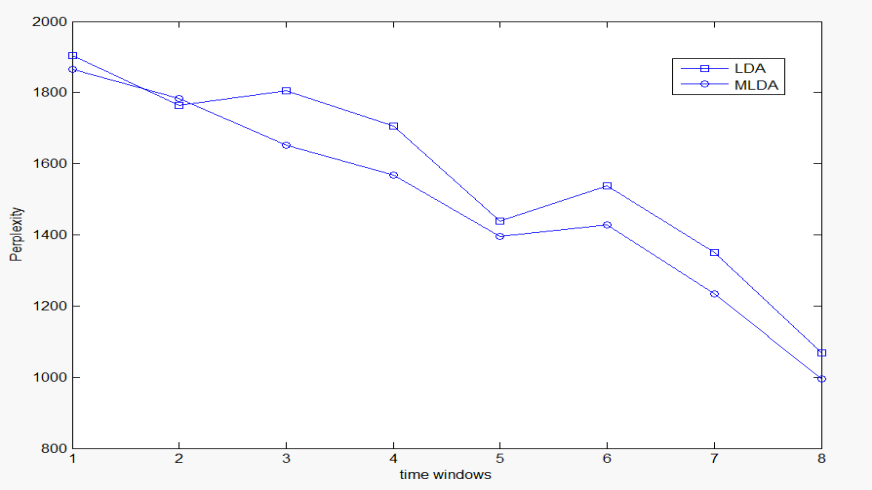

Figure 6 the performance of MLDA and LDA

5.Conclusion

An MLDA model is proposed in this paper, which takes microblog document relation, topic tag and authors' relation into consideration. Based on MLDA, the topic evolution in content and intensity is studied. The experiments on microblog data have shown the effectiveness and efficiency of the approach to depict topic evolution. In future work, we will improve efficiency of this algorithm. Besides, the scalability of the proposed model will be further investigated.

REFERENCES

[1] James Allan,RonPaPka,VietorLavrenko. Online new event detection and tracking [C]. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melboume:ACM Press.37-45(1998)

[2] Blei D,Ng A,Jordan M. Latent Dirichlet allocation[J].Journal of Machine Learning Research : 993-1022 (2003)

[3] Keming Chu ,Fang Li.Topic Evolution Based on LDA and Topic Association[J]. Journal of Shanghai Jiao tong University ,44(11): 1496-1500 (2010)

[4] Xuerui Wang， Andrew McCallum. Topic over time: A Non-Markov Continuous-Time Model of Topical Trends [C]. ACM SIGKDD :424 -433 (2006)

[5] Song X, Lin C Y, Tseng B L, et al. Modeling and predicting personal information dissemination behavior [C] Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.479-488 (2005)

[6] S.Mark,G.Tom.Probabilistic Topic Models[M]. Latent Semantic Analysis: A Road to Meaning. 2006.

[7] AlSumait L, Barbará D, Domeniconi C. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking[C]. Data Mining,ICDM'08.3-12(2008).

[8] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 5221-5228 (2004)

[9] Kai Cui, Bin Zhou,Yan jia, Zhen Liang. LDA-based Model for Online Topic Evolution Mining[J]. Computer Science, 37(11): 156-193 (2010)

[10] Thomas Minka，John Lafferty． Expectation-Propagation for the Generative Aspect Model [C] Uncertainty in Artificial Intelligence (UAI) (2002)

[11] Blei D., Lafferty J. Visualizing Topics With Multi-Word Expressions. Stat, 1050-1055 (2009)

[12] Yanli Hu, Liang Hu,Weiming Zhang.OLDA-based Method for Online Topic Evolution in Network Public Opinion Analysis[J].Journal of National University of Defense Technology,34(1);150-154 (2012)

[13] Blei D M， Lafferty J .Dynamic topic models [C].Proceeding of the 23rd In Conference on Machine Learning．New York：ACM.113-120 (2006)