Adaptive Topic Tracking Based on Dirichlet Process Mixture Model

Chan Wang, Xiaojie Wang, and Caixia Yuan

Beijing University of Posts and Telecommunications, Beijing, China
{wchan,xjwang,yuancx}@bupt.edu.cn

Abstract. This paper proposes a Dirichlet Process Mixture Model (DPMM) considering relevant topical information for adaptive topic tracking. The method has two characters: 1) It uses DPMM to implement topic tracking. Prior knowledge of known topics is combined in Gibbs sampling for model inference, and correlation between a story and each known topics can be estimated. 2) To alleviate topic excursion problem and topic deviation problem brought by existing adaptive tracking methods, the paper presents a new adaptive learning mechanism, the basic idea of which is to introduce tracking feedback with a reliability metric into the topic tracking procedure and make tracking feedback influence tracing computation under the condition of the reliability metric. The empirical results on TDT3 evaluation data show that the model, without a large scale of in-domain data, can solve topic excursion problem of topic tracking task and topic deviation problem brought by existing adaptive learning mechanisms significantly even with a few on-topic stories.

Keywords: adaptive topic tracking, ATT; traditional topic tracking, TTT, DPMM, Gibbs sampling, known topics.

1 Introduction

With the rapid development of Internet, the real-time, high-volume data stream resources increase rapidly, such as newswires, news broadcast, TV news, IM records, chat room messages, emails, twitter posts, etc. How to discover and track topics across such real-time streams is an urgent and practical problem.

The research of Topic detection and tracking aims to automatically organize and locate relevant stories from a continuous feed of news stories. There are several subtasks defined for the TDT evaluation. Among them, topic tracking task aims to associate incoming stories with topics that are known in advance. A topic is "known" by its association with stories that describe it. Thus each known topic is defined by one or more on-topic sample training stories (i.e., sample stories) [1].

The typical process of tracking system is: 1) building models of known topics and stories; 2) estimating correlation between them; 3) getting the tracking result of the story according to the correlation. In traditional topic tracking (TTT) methods, the known topic is represented using 1-4 sample stories given in advance [1] and keeps unchanged during tracking process. It is well known that the contents of topic will be

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 237–248, 2012.

enriched and the topic focuses transfer gradually with newly incoming data, which is called topic excursion. Thus, adaptive topic tracking (ATT) which has self-learning ability becomes a new research trend. ATT rich the topic model through considering the additional on-topic stories during tracking process, which can improve topic tracking performance.

This paper proposes an adaptive topic tracking method based on DPMM. The remainder of the paper is organized as follows: Section 2 discusses related work firstly. Section 3 proposes TTT and ATT based on DPMM. Section 4 presents experiments and result analysis, finally conclusions are given in Section 5.

2 Related Work

Researchers put forward many correlation estimation methods according to different representation of topics and stories. Most topic tracking methods based on vector space model use Hellinger distance [2], cosine similarity [3] to measure correlations. Topic tracking methods based on language model express correlations of story and topic as a probability model. Taking unigram model for example, the correlation of story S and topic Z_i can be calculated as:

$$P(Z_i \mid S) = \frac{P(S \mid Z_i)P(Z_i)}{P(S)} \approx \{\prod_{w_j \in S} \frac{P(w_j \mid Z_i)}{P(w_j)}\} P(Z_i) \qquad (1)$$

Where w_j is the *jth* word in S, $P(Z_i)$ and $P(w_j)$ is prior probability of Z_i and w_j, $P(w_j \mid Z_i)$ is probability of w_j under Z_i. Yamron [4], Lo [5], Spitters [6] used unigram model to conduct topic tracking and obtained good performance.

However, topic tracking task only have 1-4 sample stories to describe per known topic, so serious data sparse problem exists in tracking task. To alleviate this problem, representation method based on language model uses data smoothing technique to reestimate parameters, which needs a large quantity of background corpus as training data [5]. Moreover, existing topic tracking methods always need pre-set some parameters [7], such as correlation threshold, which also need training process. If the scale of training data is not large enough, or training data don't have the same words distributions as stories on question, parameters reestimation may have errors and lead to poor tracking performance.

Latent Dirichlet Allocation [8] (LDA) and DPMM are most commonly used topic models. In LDA, the number of topics must be preassigned, while the attractive advantage of DPMM is the number of mixture components is determined by the model and the data. They open new possibilities for parameter estimation in topic tracking task and solution of data sparseness problem. So this paper uses DPMM to estimate the correlation of stories with known topic.

Besides above problem, because of data sparseness and topic excursion, the topic model built by TTT should be poor and not accurate enough [9]. To solve above problems, ATT methods update the topic model based on tracking feedback during tracking process and following topic tracking process starts from the updated topic models, which make ATT have self-learning ability of topic tracking. The common updating methods are: add new correlated features to topic model, or continually adjust feature weights of topic model [4], or use two above methods simultaneously. This kind of ATT can alleviate imperfection of topic model in TTT caused by data sparseness. However, the tracking feedbacks add plenty of off-topic information into known topic model, causing that the updated topic will deviate far from the original topic. The problem becomes more and more serious in the tracking process. In general, this kind of ATT cannot improve the tracking performance to a great degree. This paper presents a new adaptive learning mechanism, the basic idea of which is to endow tracking feedback with a reliability metric. In tracking process, initial topic model keep unchanged. The correlation between story and known topic is estimated both under initial topic model and tracking feedback with the reliability metric. In our method, initial topic model don't contain off-topic information, and always influence tracing computation through a bigger influence metric. Thus, this method can reduce errors brought by off-topic stories and alleviate topic deviation problem.

3 ATT Based on DPMM

3.1 Task Description

To facilitate describing, we assume a collection of k known topics, $\{Z_1, Z_2, ..., Z_k\}$, and every topic is described by 1-4 sample stories which compose prior knowledge of the known topics. Topic tracking task aims to associate incoming stories with known topics one by one and detect all on-topic stories from following news stories.

With the assumption that prior probabilities of every word are equal, formula (1) can be simplified as:

$$P(Z_i \mid S) \approx \{\prod_{w_j \in S} P(w_j \mid Z_i)\} P(Z_i) \tag{2}$$

Formula (2) contains two parameters: $P(Z_i)$ and $P(w_j \mid Z_i)$. As mentioned above, parameter estimation of existing methods is easy to be affected by selection of data. To solve the problems, our paper takes DPMM to estimate formula parameters. The attractive advantage of DPMM [10] is topic information can be determined by the model and the data directly.

3.2 DPMM

The proposed method regards news texts as being generated by a sequence of underlying topics inferred using DPMM. The generation process of a text can be described as: for each word w in the text, firstly choose a component (topic) Z from a distribution θ. Topic Z is then associated with a distribution over words, φ. Finally, the word is chosen from φ. Notice we do not need pre-set the number of topics. Fig. 1 shows the graphical model depiction of DPMM, where *N* refers to the total number of words in text. Assume that θ and φ have Dirichlet prior with concentration parameter α and β respectively.

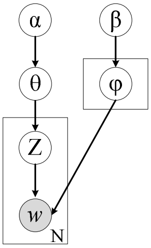

Fig. 1. Graphical model depiction of DPMM

In the paper, we use a Gibbs sampling procedure to infer parameters of the model [11]. Let w_j be the *jth* word of text. In Gibbs sampling, we need to sample Z_j, the topic of w_j. The relevant conditional distribution is:

$$P(Z_j \mid Z^-, W) \propto P(Z_j \mid Z^-) P(w_j \mid Z, W^-) \tag{3}$$

Where W^- denotes the words except w_j. The prior for assigning w_j to either an existing topic or to a new one conditioned on other topic assignments (Z^-) is:

$$P(Z_j = z \mid Z^-) \propto \{ \begin{matrix} \alpha, \ if \ \mathrm{z} = \mathrm{z}_{\mathrm{new}} \\ n_{-,z}, otherwise \end{matrix} \tag{4}$$

Where $n_{-,z}$ is the number of words assigned to topic z excluding w_j.

$$P(w_j = w \mid Z, W^-) \propto n_{w,z} + \beta \tag{5}$$

Where $n_{w,z}$ is the number of times we have seen *w* associated with topic index *z* in (Z, W^-).

3.3 Model Description of ATT Based on DPMM

As analyzed in section 2, in most existing ATT algorithms, integrating plenty of off-topic information into known topic model will likely lead to topic deviation. To solve the problem, we propose an ATT system based on DPMM with a "reliability" metric. Reliability is defined as the dependent degree of the tracking feedback. As TTT, our ATT system preserves the initial topic models unchanged. But the significant difference is that we update a tracking feedback with the reliability metric, denoted by M_reli, and simultaneously use initial topic model and tracking feedback to compute correlation between stories and known topics.

The graphical model depiction of ATT based on DPMM is shown as fig. 2. Assume that S_t is the story at time *t*. ATT model introduces a new parameter: guidance information. Use GI_t to denote guidance information of model at time *t*, and guidance information at time *0* means prior knowledge of known topics.

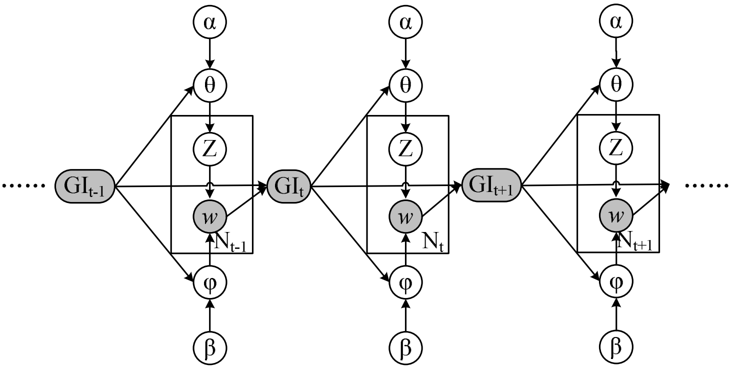

Fig. 2. Graphical model depiction of ATT

The generation process of S_t is determined by DPMM as described in section 3.2, but parameters $\theta$and φ are jointly affected by GI_t, GI_t and tracing result of S_t decide guidance information at time *t+1*. Thus, guidance information consists of two parts: prior knowledge composed of sample stories and tracking feedback. In ATT model, they guide topic tracking process by different ways respectively.

3.4 Algorithm Flow

We first create a corresponding topic Z_i^+ for every topic. $col\_Z_i^+$ refers to set of on-topic stories which belong to topic Z_i in tracking feedback. Obviously, at the start of ATT, $col\_Z_i^+$ is an empty set.

At time *t*, the implementation process can be described as:

1. Implement Gibbs sampling, and combine prior knowledge of known topics during sampling, which can be detailed as follows.
   (a) Random Initialization:

   Assign a known topic to every word of story S_t randomly.

   (b) Gibbs sampling combining with prior knowledge of known topics:

   Use Gibbs sampling on every word of S_t. Based on procedure described in section 3.2, we can use formula (3) to obtain parameters of the model. Improved Gibbs sampling procedure not only take account of current texts, but also consider the effect of prior knowledge of known topics on current word. Based on this thought, formula (4) can be rewritten as:

$$P(Z_j = z \mid Z^-) \propto \{ \begin{matrix} \alpha, & if\ \mathrm{z} = \mathrm{z}_{\mathrm{new}} \\ n_{-,z} + n_{col\_z}, & otherwise \end{matrix} \tag{6}$$

Where $n_{-,z}$ also refer to the number of words assigned to topic *z* excluding w_j in the word sets of S_t. col_z denotes word set of sample stories which belong to topic *z*. n_{col_z} is the number of words in col_z. If *z* do not belong to the known topic set, $col\_z = \Phi, n_{col\_z} = 0$. Likewise, after adding prior knowledge of known topics, formula (5) can be revised as:

$$P(w_j = w \mid Z, W^-) \propto n_{w,z} + n_{w,col\_z} + \beta \tag{7}$$

In the formula, $n_{w,z}$ is the number of times we have seen word *w* associated with topic index *z* in the words sets of S_t excluding current word *w*. n_{w,col_z} refers to the number of *w* in word set col_z.

(c) Reach a steady state, end sampling procedure.

The improved Gibbs sampling shows that every sampling step is affected by prior knowledge of known topics. This step obtains word-topic distribution of S_t and realizes guiding role of prior knowledge of known topics in topic tracking.

2. This step add corresponding topic Z_i^+ to known topics set. There are *2k* known topics, $\{Z_1, ..., Z_k, Z_1^+, ..., Z_k^+\}$. S_t-topic information can be obtained by:
   (a) Based on sampling results of step 1, estimate parameters in formula (2), get the correlation of S_t and every known topics. Using formula (7) for reference, estimation formula of $P(w_j \mid Z_i)$ can be rewritten as:

$$P(w_j \mid Z_i) \propto N_{w_j, Z_i} + n_{w_j, col\_Z_i} + \beta \tag{8}$$

Where, N_{w_j, z_i} denotes the number of w_j associated with topic index Z_i in the word sets of S_t after sampling procedure.

Likewise, estimation formula of $P(Z_i)$ is:

$$P(Z_i) \propto N_{Z_i} + n_{col\_Z_i} \tag{9}$$

Where, N_{Z_i} denotes the number of words with topic index Z_i in the word sets of S_t after sampling procedure.

(b) Combine formula (2), (8) and (9) to compute correlation between S_t and every known topic. Estimation formula of correlation between S_t and known topic Z_i , $P\_Adaptive(Z_i \mid S_t)$, can be rewritten as:

$$P\_Adaptive(Z_i \mid S_t) = (1 - M\_reli) * P(Z_i \mid S_t) + M\_reli * P(Z_i^+ \mid S_t) \quad (10)$$

Initial known topic models are built according to prior knowledge, but tracking feedback contains off-topic stories, thus M_reli is always less than 0.5. Based on formula (10), this step realizes guiding role of tracking feedback in topic tracking.

3. Assign the topic corresponding to the maximum correlation to S_t. Finally, add S_t to the corresponding stories set. Based on formula (6), sampling procedure allows the appearance of new topic and S_t may be assigned to a new topic. Under this situation, S_t is not associated with whichever known topic.

Repeat above steps for incoming news stories. Fig. 3 shows the flow chart.

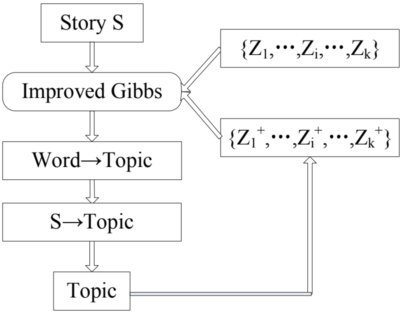

Fig. 3. Flow chart of DPMM based ATT

The characters of ATT based on DPMM:

1. In ATT model, guidance information consists of two parts: prior knowledge composed of sample stories and tracking feedback. In the implementation process, Gibbs sampling of step 1 realizes guiding role of prior knowledge, and formula (10) of step 2 realizes guiding role of tracking feedback.
2. From step 2, DPMM can compute relevant information of topics from the model and the data. Therefore, compared with existing topic tracking methods, ATT based on DPMM can directly estimate the correlations between story and every known topic, not requiring the correlation threshold comparison which needs to be trained via a large scale of in-domain data.
3. Because initial known topic models are reliable, our system ensures that initial known topic models always have bigger influence on correlation calculation than tracking feedback through setting M_reli . Thus, this method can reduce errors

brought by off-topic stories and alleviate topic deviation problem effectively brought by existing adaptive learning mechanisms.

3.5 TTT Based on DPMM

According to graphical model depiction of ATT described in section 3.3, this section cancels guiding role of tracking feedback in topic tracking and obtains graphical model depiction of TTT based on DPMM. In this model, guidance information only contains prior knowledge composed of sample stories, which keep invariant. TTT based on DPMM is similar to ATT, the difference is TTT based on DPMM don't need create a corresponding topic for every topic Z_i. The method computes correlation between story and every known topic via formula (2), (8) and (9) directly.

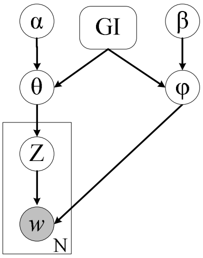

Fig. 4. Graphical model depiction of TTT

4 Experiments and Result Analysis

In TDT bakeoff [12], tracking performances are measured by error cost, C_{Det}, which is a weighted sum of the miss and false alarm probabilities. C_{Det} is usually transformed to the interval [0,1], $(C_{Det})_{Norm}$. The paper use $(C_{Det})_{Norm}$ to examine the tracking performances.

4.1 Results

We use TDT3 Chinese data as experiments test set. The premise of all experiments is that every known topic only has one sample story. The experiments examine the effectiveness of TTT and ATT based on DPMM separately.

4.1.1 Experiments of TTT

This part is a comparison between performances of TTT based on unigram model (B_TTT) and DPMM (D_TTT), which investigates influence of text feature selection simultaneously. B_TTT applies add-one smoothing to topic tracking task.

Our experiment designs four features to represent a known topic: feature set composed of content words, nouns and verbs, nouns, verbs are denoted by term_c, term_n+v, term_n, term_v respectively.

Firstly, we investigate the influence of model parameters of DPMM and different feature selection methods on topic tracking performances. In all experiments, parameter β is set at 0.01. Fig. 5 shows the relationship of tracking performance of D_TTT against influencing factors above.

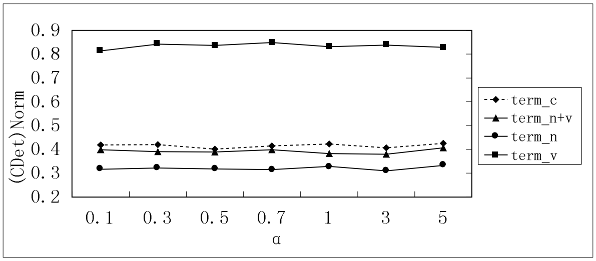

Fig. 5. Results of D_TTT

From fig. 5, we can find out:

1. When changing parameter α, $(C_{Det})_{Norm}$ values of term_v, term_c, term_n+v, term_n systems scatter in the range [0.81, 0.85], [0.40,0.43], [0.37,0.40], [0.30,0.34] respectively. The results verify that parameter α has little influence on tracking performance of D_TTT under a fixed feature sets.
2. Among different features, term_n contribute most to the performance, while term_v least. One likely reason is that verbs cannot represent the topic of stories. Among the results, when α is 3.0, and system choose nouns as feature, system obtain the best performance and smallest $(C_{Det})_{Norm}$, 0.3095.

Likewise, fig. 6 shows performance comparison between B_TTT and D_TTT.

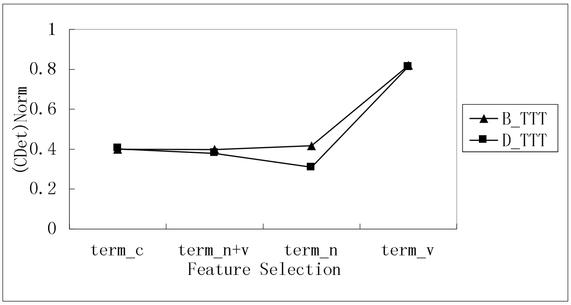

Fig. 6. Results of D_TTT and B_TTT

From fig. 6, we can find out:

1. Both B_TTT and D_TTT obtain poorest performance when choosing verbs as feature, which verify importance of feature selection in topic tracking task.
2. Under four feature selection conditions, all $(C_{Det})_{Norm}$s of D_TTT are smaller than that of B_TTT system. Compared with B_TTT, the smallest $(C_{Det})_{Norm}$ of D_TTT reduces to 0.3095 from 0.3989. Therefore, using DPMM to implement topic tracking can improve the performance of topic tracking.

4.1.2 Experiments of ATT

Firstly, we investigate results of ATT system based on DPMM (D_ATT) with different reliability metrics. Referring to results of TTT, the experiment chooses nouns as system feature. Results are shown as fig.7.

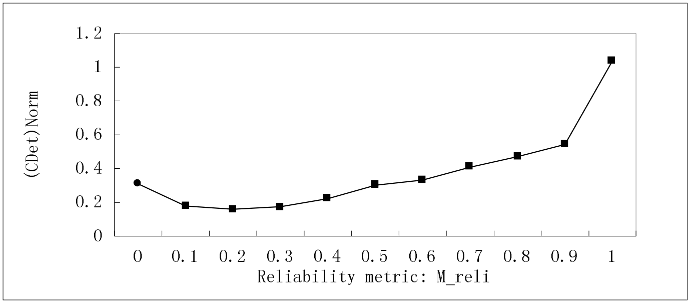

Fig. 7. Results of D_ATT

4.1.1 shows that best tracking performance of D_TTT is 0.3095, which is expressed by dot (the first point) in fig.7. Fig.7 shows that:

1. When $M\_reli < 0.5$, all $(C_{Det})_{Norm}$s of D_ATT are smaller than that of D_TTT system. When $M\_reli = 0.2$, system obtains the best tracking performance. Compared with D_TTT, $(C_{Det})_{Norm}$ of D_ATT reduces to 0.1599 from 0.3095. Therefore, our ATT method can solve topic excursion problem to some extent.
2. When $M\_reli > 0.5$, the $(C_{Det})_{Norm}$s of D_ATT increase obviously, and even are much bigger than those of D_TTT. Based on formula (10), initial known topic models and tracking feedback influence tracking results simultaneously, the influence degrees of them are $(1 - M\_reli)$ and M_reli respectively. The initial known topic models are built via prior knowledge. Inversely, tracking feedback may contain off-topic stories. Thus, when $M\_reli > 0.5$, tracking feedback have bigger influence on tracking computation than initial known topic models, which lead to bigger error of final tracking results.

Likewise, to assess the effectiveness of adaptive algorithm, this part uses a classical adaptive algorithm as a baseline: adding new correlated features to topic model, expressed by B_ATT. B_ATT system still uses DPMM for topic tracking.

Table 1. Best results of B_ATT and D_ATT

System type	B_ATT	D_ATT
$Min\{(C_{Det})_{Norm}\}$	0.2260	0.1599

Table 1 shows D_ATT has the much better performance than B_ATT. Compared with B_ATT, $(C_{Det})_{Norm}$ of D_ATT reduces to 0.1599 from 0.2260. The results verify effectiveness of adaptive algorithm proposed in our paper, and our adaptive algorithm can alleviate topic deviation problem effectively brought by existing adaptive learning mechanisms.

4.2 Result Analysis

Via experimental results, it can be concluded that:

1. DPMM is suitable for topic tracking task, and improves the tracking performance significantly compared with commonly used language model.
2. Results verify the importance of topic representation, and optimization of text feature selection algorithm can improve the tracking performance effectively.
3. Results show the influence of parameter of DPMM, α, on tracking computation can be neglectable. Based on this conclusion, topic tracking models (TTT and ATT) based on DPMM proposed in this paper don't contain any unknown system parameters, thus avoiding optimizing model parameters using additional data. The empirical results show that just with a few on-topic sample stories, TTT and ATT based on DPMM can achieve high topic tracking performance.
4. Results shown in section 4.1.2 verify the two characters of ATT based on DPMM: 1) D_ATT has the much better performance than D_TTT, which prove that our ATT method can solve topic excursion problem to a satisfactory extent. 2) D_ATT has much better performance than B_ATT, which verify our adaptive algorithm can alleviate topic deviation problem effectively brought by existing adaptive learning mechanisms.

5 Conclusion

Dirichlet Process Mixture Model (DPMM) considering relevant information of known topics is proposed for adaptive topic tracking. The method has two characters: 1) it uses DPMM to implement topic tracking and the basic idea is to implement Gibbs sampling to estimate correlation between a story and each known topic. Prior knowledge of known topics is exploited in Gibbs sampling procedure. Experimental results prove DPMM can improve tracking performance significantly. Results also verify importance of text feature selection in topic tracking task. Moreover, topic tracking methods based on DPMM can determine topic information via the model and the data directly, which can avoid parameter training process, reduce the errors and process times, and implement topic tracking task with a few on-topic sample stories effectively. 2) The paper presents a new adaptive learning mechanism, which can alleviate topic excursion and topic deviation problems simultaneously. The basic idea of our adaptive learning mechanism is to endow tracking feedback with a reliability metric. Our method makes initial topic model and tracking feedback influence computation of correlation between story and known topic under the condition of the reliability metric. Initial topic model which keep unchanged don't contain off-topic information, and always influence tracing computation through a bigger influence metric via influence metric setting. Thus, the method can reduce errors brought by off-topic stories and alleviate topic deviation problem. Results verify that our adaptive method can not only solve topic excursion problem to some extent, but also alleviate topic deviation problem effectively brought by existing adaptive learning mechanisms.

However, one major criticism of original DPMM is "Bag-of-Words" assumption by ignoring dependencies between words and neglecting word order, while in real data, each word is mutually related with other words and word order is also extremely important in text modeling applications. Thus, we will investigate how the dependencies between words and word order have impact on the model performance in the future work.

References

1. Allan, J., Carbonell, J., Doddington, G., et al.: Topic detection and tracking pilot study: final report. In: Proceedings of DARPA BNTU Workshop, pp. 194–218. DARPA, Lansdowne (1998)
2. Makkonen, J., Anonen-Myka, H., Salmenkivi, M.: Simple semantics in topic detection and tracking. Information Retrieval 7(3/4), 347–368 (2004)
3. Chen, F., Farahat, A., Brants, T.: Multiple similarity measures and source-pair information in story link detection. In: HLT-NAACL, Boston, pp. 313–320 (2004)
4. Yamron, J.P., Knecht, S., van Mulbregt, P.: Dragon's Tracking and Detection Systems for the TDT 2000 Evaluation. In: The Topic Detection and Tracking Workshop (2000)
5. Lo, Y., Gauvain, J.: The limsi topic tracking system for TDT 2001. In: The TDT Workshop. DARPA, Gaithersburg (2001)
6. Spitters, M., Kraaij, W.: Using language models for tracking events of interest over time. In: Proceedings of LMIR 2001, Pittsburgh, pp. 60–65 (2001)
7. Qiu, J., Liao, L.: Add Temporal Information to Dependency Structure Language Model for Topic Detection and Tracking. In: Proceedings of the International Conference on Machine Learning and Cybernetics, pp. 1575–1580. IEEE Press, Kunming (2008)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3(5), 993–1022 (2003)
9. Hong, Y., Zhang, Y., Liu, T., et al.: Topic detection and tracking review. Journal of Chinese Information Processing 21(6), 71–87 (2007)
10. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Annals of Statistics 1(2), 209–230 (1973)
11. Neal, R.M.: Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics 9(2), 249–265 (2000)
12. Luo, W., Liu, Q.: Development and Analysis of Technology of Topic Detection and Tracking. In: JSCL, Beijing, pp. 560–566 (2003)