# Distinguishing Specific and Daily Topics

Tao Ge, Wenzhe Pei, Baobao Chang, and Zhifang Sui

MOE Key Laboratory of Computational Linguistics,
School of EECS, Peking University, Beijing, 100871, China
Collaborative Innovation Center for Language Ability, Xuzhou 221009, China
{getao,peiwenzhe,chbb,szf}@pku.edu.cn

**Abstract.** The task of distinguishing specific and daily topics is useful in many applications such as event chronicle and timeline generation, and cross-document event coreference resolution. In this paper, we investigate several numeric features that describe useful statistical information for this task, and propose a novel Bayesian model for distinguishing specific and daily topics from a collection of documents based on documents' content. The proposed Bayesian model exploits mixture of Poisson distributions for modeling probability distributions of the numeric features. The experimental results show that our approach is promising to solve this problem.

**Keywords:** specific and daily topics, numeric features, Bayesian model, mixture of Poisson distribution

## 1 Introduction

Among techniques attempting to alleviate the impact of information overload, topic models can serve to organize information to users according to topics, helping them easily acquire knowledge. For people who want to learn about important events in the past, they can just glance over the topics discovered from a news archive without reading every document in the collection. However, it is common that some of the topics found are time-general; in other words, they concern daily and trivial topics which are considered having little retrospective value. A concrete example of daily topics is the topic document $D1$ talks about.

(*D1*) The Shenzhen composite index rose slightly 0.2 points to 615.39 on Friday.
(*D2*) An Alaska Airlines jet with at least 60 people on board crashed late Monday afternoon.

As we can see, $D1$ talks about a topic involving ups and downs of stock index, which is reported every day or every week regularly. From a retrospective viewpoint, the topic about temporary stock index has little value for general audience. In contrast, $D2$ concerns a specific topic about an important event – air crash. If one wants to learn about important events in some year (e.g. 2000), what he is interested in should be the specific topics like $D2$ discusses.

Distinguishing specific topics (time-specific) and daily topics (sometimes called time-general, routine or recursive topics) is a useful task for many applications in natural languages processing and data mining. One typical example is event chronicle generation [6] and timeline generation [8, 7]. As we know, chronicles and timelines have been adopted as a way to organize information by many websites such as Wikipedia and Facebook due to their readability and briefness. For generating a chronicle/timeline, it is necessary to distinguish specific topics from daily topics since only specific topics should be included.

Also, distinguishing specific and daily topics can help improve cross-document event coreference resolution systems which aim to find coreferential events in different documents. As Ge et al. [5] considered, the performance of cross-document event coreference resolution systems is likely to be affected by daily topics shown as follows:

(*D3*) The Germany's DAX index was down by 0.2 percent yesterday.
(*D4*) The Germany's DAX index was down by 0.2 percent on Wednesday.

Note that $D3$ and $D4$ were written in 2001 and 2004 respectively. If a cross-document event coreference resolution system can distinguish specific and daily topics, it will not consider these two events coreferential.

In this paper, we focus on distinguishing daily and specific topics without depending on documents' timestamps . The reasons why we do not use documents' timestamps are: First, the timestamp of documents, especially on the web, is unavailable (missing) or incredible due to arbitrary copy-paste behaviors such as retweet and reprint on the web, as [1, 3, 5] reported; Second, analyzing time distribution of topic is not feasible or reliable unless the documents are uniformly sampled from a text data stream; in other words, if the dataset used is not constructed by uniformly sampling from a text stream, the time distribution of a topic does not make sense. We investigate several numeric features for this task. For modeling the numeric features, we propose a novel Bayesian model with mixture of Poisson distributions. The experiments evaluate our model in terms of temporal distribution and semantic coherence and show that our model is more effective to discover and distinguish daily and specific topics from a collection of documents than conventional Bayesian topic models in an unsupervised manner.

The main contribution of this paper is: **(1)** we propose a Bayesian model for effectively discovering and distinguishing specific and daily topics; **(2)** we propose a general framework for incorporating numeric features into Bayesian models; **(3)** we give some measures for evaluating models for this task.

## 2 Methodology

### 2.1 Features for distinguishing specific and daily topics

As discussed in section 1, we attempt to distinguish between daily and specific topics using only textual features. For representing textual information, categorical features (also called nominal features) such as n-grams are usually adopted in NLP tasks. However, categorical features alone are not enough for describing statistical information such as the count of numerals in a document, which is important for this task. We introduce numeric features describing useful statistical information of documents for this task.

**#PERSON named entities:** Named entities are often discussed during specific time period. Intuitively, if a topic involves a number of PERSON names then this topic is less likely to be a daily topic. Hence, **the count of PERSON named entities** in a document is selected as a feature.

**#Numerals:** It is easy to understand that the documents concerning daily topics usually involve something that frequently changes such as temporary stock price, which should be updated in time. The most salient features for describing such variations and fluctuations are numerals. If a document contains too many numerals, it is very likely that the article may talk about a daily topic. Based on the intuition, **the count of numerals** in a document is selected as a feature.

**#IDF:** In addition to document-based features, we also investigate an important corpus-based feature – inverse document frequency (IDF) which is a measure of whether a term is common or rare across a collection of documents. Intuitively, if IDF of salient words of a topic tend to be high, this topic will be less likely to be frequently mentioned all the time. According to our analysis of corpus of Gigaword, we find that the terms whose idf is greater than 60% of MAXIDF is distinguished for determining a topic to some extent. Thus, we refer terms in set $H$ which is defined as follows to high-idf terms.

$$H = \{w | idf(w) > \text{MAXIDF} \times 0.6\}$$

Given that **the counts of high-idf terms** in a document can indicate whether the document concerns a daily topic or a specific topic to some extent, we select it as a feature for the task.

### 2.2 A Bayesian model with mixtures of Poisson distributions

For distinguishing specific and daily topics, we adopt a Bayesian model which is proved to be suitable for tasks concerning topics due to its ability of discovering topics hidden in a text collection and its flexibility that it can introduce various features with their dependencies.

In section 2.1, a variety of integer numeric features used for distinguishing specific and daily topics are discussed. Effective to reflect some statistical information as the features seem, it is somewhat difficult to directly incorporate these features such as into a Bayesian model. Although numeric features can be discretized into several categorical bins in advance which can be modeled by a multinomial distribution, that appears not very applicable. On one hand, if the numeric variables are discrectized into too many bins, data sparsity problem would arise, which may have an adverse effect on the performance. On the other hand, if the number of bins is too small, the discretization might be unreasonable.

To deal with this problem, we assume that the integer numeric features follow Poisson distributions which are proved to be very suitable for expressing the probability distribution of the integer numeric variables such as the count of a term in a document, and use the poisson distributions in our Bayesian model.

**Table 1.** Notations used in our model

| **Symbols** | **Descriptions** |
|---|---|
| $M, N, K$ | The number of documents in a corpus, tokens in a document and topics respectively |
| $F$ | The number of features used. If we use all the three features discussed in section 2.1, then $F = 3$. |
| $w$ | A word in a document |
| $z$ | The topic assigned to a document |
| $p$ | Topic type label indicating whether a topic is daily. $p \in \{0, 1\}$ |
| $f$ | Features discussed in section 2.1, which could be the count of PERSON named entity, numerals or high-idf tokens in a document |
| $\theta$ | The topic distribution |
| $\pi$ | The Bernoulli distribution for $p$ |
| $\Phi$ | The distribution for words given a topic |
| $\lambda$ | The parameter of poisson distributions. |
| $\alpha$,$\beta$ | The hyperparameters for Dirichlet prior of $\theta$ and $\Phi$ respectively |
| $\gamma$ | The hyperparameter for Beta prior of $\pi$ |

The plate diagram of our Bayesian model is given in figure 1 and notations of our model are summarized in table 1.

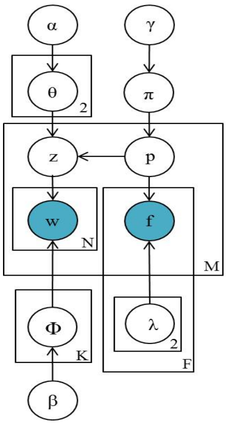

**Fig. 1.** Graphical illustration of our model.

The left part of the plate diagram of our model shown in figure 1 is similar to general latent dirichlet allocation (LDA) [2]. But unlike the general LDA model in which a document is a mixture of topics, our model assumes that a document corresponds to only one topic. Moreover, our model introduces another latent variable $p$ which indicates the type of a topic (daily or specific), and poisson distributions which aim to express the probability distribution of the count of PERSON named entities, numerals and high-idf terms in a document respectively. As shown in figure 1, the numeric features can be considered as being generated by a component of the mixture of Poisson distributions.

In addition, our model assumes that the topic distributions under different topic types are different, which is reasonable. While constructing topics, not only does our model consider the word distribution, but also takes into consideration the numeric features. In this way, we can even distinguish topics through the difference of the numeric features even if they have the similar word distribution. For example, a document about temporary stock price and another document talking about a specific stock market crash might be assigned to two different topics in our model due to differences of the numeric features while they might be in the same topic cluster in conventional Bayesian topic models because their word distributions are similar. As a result, specific topics found by our model can be more time-specific and daily topics found by our model could be more time-general than those found by conventional Bayesian topic models.

```
Draw $\theta$ ~Dirichlet($\alpha$) for each topic type $p$
Draw $\varPhi$ ~Dirichlet($\beta$) for each topic $k$
Draw $\pi$ ~Beta($\gamma$)
For each document $m$:
    Draw $p$ ~Bernoulli($\pi$)
    For each feature $f_i$:
        Draw $f_i$ ~Poisson($\lambda_i^p$)
    Draw $z$ ~Multi($\theta^p$)
    For each token $w$ in $m$:
        Draw $w$ ~Multi($\varPhi^z$)
```

**Fig. 2.** The generative story of our model

The generative story of our model is presented in figure 2. Note that $f_i$ in figure 2 is the $i$th feature of a document and $i \in \{1, 2, ..., F\}$, $\lambda_i^p$ denotes the parameter of the Poisson distribution for the $i$th feature of a document whose topic type is $p$. In this way, daily and specific topics can be distinguished while constructing topics.

Model inference and the method for estimation of parameters $\lambda$ of the Poisson distributions are to be discussed in detail in section 2.3.

### 2.3 Model Inference and Parameter Estimation

For model inference, we use Gibbs sampling approach to sample latent variables $p$ and $z$. Specifically, for a given document $m$, the conditional probabilities of its latent variable $p$ and $z$ are shown in (1) and (2) respectively:

$$\begin{aligned} & P(p_m|\boldsymbol{p}_{\neg m}, \boldsymbol{z}, \boldsymbol{f}(m); \gamma, \alpha) \\ & = \frac{c_p + \gamma}{\sum_p (c_p + \gamma)} \times \frac{c_{z,p} + \alpha}{\sum_z (c_{z,p} + \alpha)} \times \prod_{i=1}^{F} P(f_i(m)|\lambda_i^p) \end{aligned} \tag{1}$$

$$\begin{aligned} & P(z_m|\boldsymbol{z}_{\neg m}, \boldsymbol{p}, \boldsymbol{w}; \alpha, \beta) \\ & = \frac{c_{z,p} + \alpha}{\sum_z (c_{z,p} + \alpha)} \times \prod_{w \in W_m} \frac{c_{z,w} + \beta}{\sum_w (c_{z,w} + \beta)} \end{aligned} \tag{2}$$

where $W_m$ denotes tokens in document $m$, $c_{z,p}$ is the number of documents whose topic and topic type are $z$ and $p$ respectively, $c_{z,w}$ is the number of words $w$ in documents whose topic is $z$, $c_p$ is the number of documents whose topic type is $p$ and $\boldsymbol{f}(m) = [f_1(m), f_2(m), ..., f_F(m)]$ in which $f_i(m)$ is the $i$th feature of document $m$. $P(f_i(m)|\lambda_i^p)$ can be easily computed using the probability mass function of Poisson distributions.

Now, the problem for our model is how to estimate the parameters $\lambda$ of the Poisson distributions. As mentioned in section 2.2, the features can be regarded as being generated by their corresponding mixture of poisson distributions; thus, we use an EM-based method to estimate the parameters of the Poisson distributions. As is known, the only parameter $\lambda$ of a Poisson distribution is the expectation of the distribution. Therefore, whenever we finish sampling the latent variable $p$ for document $m$, we re-estimate $\lambda$ of the Poisson distribution $\lambda_i^p$ using maximum likelihood estimation (MLE) as shown in (3).

$$\lambda_i^p = \frac{\sum_{j:p_j=p} f_i(j)}{\sum_{j:p_j=p} 1} \tag{3}$$

where $f_i(j)$ denotes the $i$th feature of document $j$.

In our method of parameter estimation, (1) and (3) can be considered as E-step and M-step respectively. Different from the general EM algorithm, the latent variable $p$ sampled in the E-step is a hard estimation instead of a soft one for efficiency. In this way, the parameter $\lambda_i^p$ of the Poisson distributions can be estimated during sampling.

## 3 Experiments and Evaluations

### 3.1 Experimental Setting

Since there is no standard benchmark dataset for this task, we use all the news articles written in the year 2000 of the Xinhua News Agency in Gigaword English Corpus to evaluate our model. This dataset contains 99,538 news articles involving a variety of subjects.

As preprocessing, we use Stanford CoreNLP toolkit [10] to do lemmatization, named entity extraction and POS tagging. In this way, we can obtain PERSON named entities and numerals whose count is used as features in our Bayesian model as mentioned in section 2. Also, we compute *idf* of all terms in the vocabulary except numerals whose *idf* is set to 0. We normalized the count of PERSON named entities, numerals and high-idf terms in document $m$ as follows:

$$f_{norm}(m) = \lfloor \frac{f(m)}{length(m)} \times 100 \rfloor$$

where $f(m)$ denotes a feature of document $m$ which could be the count of PERSON named entities, numerals or high-idf tokens in document $m$ and $length(m)$ is document $m$'s length.

**Table 2.** Estimated parameters $\lambda$ of poisson distributions of different feature combinations. $\lambda_1$, $\lambda_2$ and $\lambda_3$ are parameters of Poisson distributions for the count of named entities, numerals and high-idf terms respectively

| Feature | p | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | topic type |
|---|---|---|---|---|---|
| per | 0 | 8.09 | - | - | specific |
| | 1 | 0.79 | - | - | daily |
| num | 0 | - | 3.59 | - | specific |
| | 1 | - | 25.07 | - | daily |
| idf | 0 | - | - | 20.34 | specific |
| | 1 | - | - | 7.25 | daily |
| per+num | 0 | 3.43 | 3.51 | - | speccfic |
| | 1 | 2.26 | 24.65 | - | daily |
| per+idf | 0 | 7.65 | - | 20.72 | specific |
| | 1 | 1.12 | - | 8.94 | daily |
| num+idf | 0 | - | 3.70 | 13.47 | specific |
| | 1 | - | 25.36 | 9.82 | daily |
| All | 0 | 1.12 | 25.48 | 7.67 | daily |
| | 1 | 3.72 | 4.14 | 14.09 | specific |

We empirically set hyperparameters $\alpha = 0.05$, $\beta = 0.01$, $\gamma = 0.5$. The number of topics $K$ is set to 100. We simply select the topic with the largest probability for a document as the document's topic. Formally, for document $m$,

$$topic(m) = argmax_z P_m(z)$$

where $P_m(z)$ denote the probability of a topic for document $m$ and it can be estimated as follows:

$$P_m(z=k) = \frac{\sum_{i=1}^{S} \delta(z^{(i)} = k)}{S}$$

where $z^{(i)}$ is the topic sampled for document $m$ at the $i$th iter after burn-in, $\delta(.)$ is an indicator function and $S$ is the number of iterations of after burn-in.

In order to verify the effectiveness of features we use, we try different combinations of features. Parameters $\lambda$ of poisson distributions after the model converges is shown in table 2. It should be noted that even though the topic type of a given topic can be indicated by the topic type label $p$ ($p \in \{0, 1\}$), the meaning of $p$'s value is unknown since our Bayesian model is an unsupervised approach. In other words, we do not know the topic type label $p$ of a daily topic should be 0 or 1. Therefore, we must make clear what the exact meaning of $p$'s values are. Based on the intuition discussed in section 2.1, we can use the estimated parameter $\lambda$ of the mixture of Poisson distributions of the integer numeric features under different topic type $p$ to help understand $p$'s value's meaning. Intuitively, as for the daily topic type, its average count of PERSON named entities and high-idf terms per document should be less while its average count of numerals per document should be much more than the counterpart of the specific topic type. After making clear the meaning of $p$'s values (as shown in table 2), we set a high confidence threshold (0.9) for determining the daily topics. Specifically, if the probability of a topic $z$ to be a daily topic is larger than 0.9 (i.e.

**Table 3.** Temporal perplexity of daily topics and specific topics under different feature combinations. Intuitively, for daily topics, the larger temporal perplexity, the better; for the specific topics, the smaller, the better. As for $\Delta_{avg}$, the larger, the better.

| Feature | topic type | #topics | *max* | *min* | *avg* | $\Delta_{avg}$ |
|---|---|---|---|---|---|---|
| *per* | daily | 36 | **362.04** | 5.78 | 192.67 | 65.55 |
| | specific | 64 | **308.69** | 5.66 | 127.12 | |
| *num* | daily | 15 | **362.04** | 51.63 | 183.55 | 32.38 |
| | specific | 85 | 324.03 | 6.06 | 151.17 | |
| *idf* | daily | 18 | **362.04** | 145.01 | 250.73 | 131.3 |
| | specific | 82 | 317.37 | 5.82 | **119.43** | |
| *per+num* | daily | 18 | 357.05 | 18.77 | 181.02 | 29.85 |
| | specific | 82 | 319.57 | 11.96 | 151.17 | |
| *per+idf* | daily | 27 | **362.04** | 37.79 | 218.27 | 93.77 |
| | specific | 73 | 319.57 | 8.11 | 124.5 | |
| *num+idf* | daily | 13 | **362.04** | **174.85** | **257.78** | **132.41** |
| | specific | 87 | 319.57 | **4.63** | 125.37 | |
| *All* | daily | 16 | **362.04** | 140.07 | 232.32 | 85.29 |
| | specific | 84 | 319.57 | 7.06 | 147.03 | |

$P(p = 0|z) > 0.9$, assuming that $p = 0$ indicates that the a topic is daily), the topic $z$ will be considered as a daily topic; otherwise, the topic is considered specific. The probability $P(p|z)$ can be estimated as follows:

$$P(p|z) = \frac{(c_{z,p} + \alpha)}{\sum_p (c_{z,p} + \alpha)}$$

### 3.2 Experimental Results

In this section, we evaluate and compare our method with other Bayesian models in terms of temporal perplexity and log-likelihood per document.

**Temporal Perplexity** Since there is no golden standard for evaluating whether a topic or a document concerns a daily topic or not, we alternatively use an indirect way to evaluate our model – using temporal distribution to measure a topic's distribution over time. If the temporal distribution of a topic is almost uniform, then the topic might be a daily topic. In contrast, if the temporal distribution of a topic fluctuates significantly over time or the number of articles involving the topic surge during a short period of time, the topic is more likely to be a specific topic. Therefore, it is possible to use temporal distribution to help evaluate our model. Inspired by the perplexity measure in information theory, we define *temporal perplexity* (*TP* for short) for measuring the temporal distribution of a topic, as shown in (4).

$$TP(z) = 2^{-\sum_t \frac{c_{z,t}}{c_z} \times \log_2 \frac{c_{z,t}}{c_z}} \tag{4}$$

where $t$ denotes a time epoch whose granularity can be either a day, a week or a month, $c_{z,t}$ denotes the number of documents involving topic $z$ at $t$ and $c_z$ is

**Table 4.** Comparison between NB, LDA and our model in terms of temporal perplexity

| Model | topic type | #topics | *max* | *min* | *avg* | $\Delta_{avg}$ |
|---|---|---|---|---|---|---|
| LDA | daily | 13 | 315.04 | 132.27 | 248.26 | 21.86 |
| | specific | 87 | 362.04 | 33.20 | 226.40 | |
| NB | daily | 13 | 303.40 | 30.41 | 198.91 | 70.75 |
| | specific | 87 | 362.04 | 9.81 | 128.16 | |
| ours | daily | 13 | **362.04** | **174.85** | **257.78** | **132.41** |
| | specific | 87 | **319.57** | **4.63** | **125.37** | |

the number of documents involving $z$ across the collection. According to (4), the more uniform the temporal distribution of a topic is, the larger the TP of the topic will be. Therefore, the temporal perplexity of daily topics should be larger than that of specific topics.

Table 3 reports the temporal perplexity of the identified daily topics and the specific topics under different combinations of features. Note that the temporal granularity is day. In table 3, *max*, *min* and *avg* for a topic type denote the maximal, minimal and average temporal perplexity of topics of the topic type (daily or specific). $\Delta_{avg}$ denotes the difference between the average temporal perplexity of the identified daily topics and that of the specific topics:

$$\Delta_{avg} = avg_{daily} - avg_{specific}$$

As shown in table 3, the features we used (i.e. count of PERSON named entities, numerals and high-idf terms in one document) are all capable of distinguishing daily and specific topics, which is reflected by a positive $\Delta_{avg}$ value. Among these features, the count of high-idf terms seems to be the most effective, which achieves the largest $\Delta_{avg}$ since IDF is a corpus-based feature, just as temporal perplexity which is also a measure based on a corpus. Hence, compared with the other document-based features, the count of high-idf terms appears to be more correlated with temporal perplexity.

As for the combination of features, it is not difficult to find that the *num+idf* combination performs best. This combination achieves the highest temporal perplexity for the identified daily topics in terms of *max*, *min* , *avg* as well as $\Delta_{avg}$. As discussed in section 2.1, the count of numerals in a document is an important feature for distinguishing between specific topics and daily topics. Although the feature alone may not result in a large $\Delta_{avg}$, it is very helpful in improving the performance when it is combined with *idf*. However, the combination of all the three features does not achieve a better result than *num+idf* and *per+idf* combinations. The possible reason is that the combination of *per* and *num* may affect the Bayesian model. As mentioned above, the task is to distinguish between the documents about daily topics which tend to contain few PERSON named entities and many numerals, and the documents about specific topics which tend to contain many PERSON named entities and few numerals. Nevertheless, it is not uncommon that a news article contains both few PERSON named entities and few numerals, or contains both many PERSON named entities and many numerals (e.g. list of top Premier League goal scorers). When *per* and *num* are simultaneously selected as features, the two features may interfere each other, which perhaps leads to a poor performance. In contrast, other feature combinations seem less likely to suffer from such a problem.

In addition, we compare our model with two typical Bayesian topic model – Naive Bayes (NB) and Latent Dirichlet Allocation (LDA). For NB and LDA, we first detect topics and then use idf+num feature to distinguish specific and daily topics. Formally, we define a score for topic $z$ as follows:

$$score_z = \#numeral_z - \#highidf_z$$

where features (e.g. $\#numeral_z$) are average of their counterparts of documents whose topic is $z$.

According to the intuition in section 2.1, the higher score, the more likely the topic is to be daily. Since our model identifies 13 daily topics with idf+num features, we identify the 13 topics with the highest score as daily topics for NB and LDA model. The performance of NB and LDA is given in table 4. It can be easily seen that our model performs much better than LDA and NB in terms of temporal perplexity. One main reason is that our model considers the numeric information while constructing topic clusters, as discussed in section 2.2. When a document contains many high-idf terms and few numerals, it would be more likely to be assigned to a specific topic cluster. In contrast, NB and LDA consider only word distribution when constructing topic clusters. Hence, in NB and LDA, a specific topic (e.g. stock market crash) and a daily topic (e.g. temporary stock price) might be in the same topic cluster owing to similar word distribution while they are less likely to be the same cluster in our model due to difference of numeric features. Therefore, specific topics found by our models tend to be more time-specific and daily topics found by our models tend to be more time-general. Figure 3 also verifies the claim. It is shown that our model can find more extremely daily(TE>300) and specific(TE<50) topics than those found by NB and LDA.

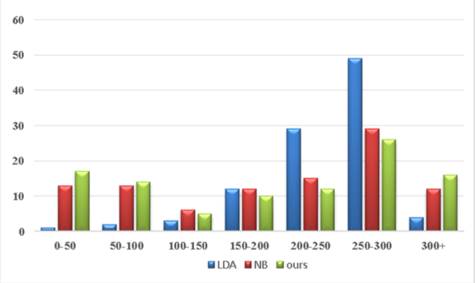

**Fig. 3.** The number of topics in intervals of temporal perplexity

**Max Log-likelihood per document** In addition to temporal perplexity which evaluates the temporal distribution of topics found by our model, we also use Max log-likelihood per document to compare our model with NB and LDA in terms of semantic coherence which is an important measure for evaluating topic models.

As is known, topic models assume a collection of documents are generated by mixture of language models. Thus, we can use the idea for evaluating language models for evaluations. Log-likelihood per document is one for such measures for topic model evaluation [4]. Since our work assumes that a document is associated with only one topic, therefore, we use max log-likelihood per document instead of log-likelihood for evaluations. Formally, max log-likelihood per document ($MLLPD$) for a test set $T$ is computed as (5):

$$MLLPD(T) = \frac{\sum_{d \in T} \sum_{w_i \in d} logP(w_i | z_d^*)}{|T|} \quad (5)$$

where $z_d^* = argmax_z \prod_{w_i \in d} P(w_i | z)$

**Table 5.** Max Log-likelihood per documents under different number of topics

| #topics | 20 | 50 | 100 |
|---|---|---|---|
| LDA | -2779.36 | -2870.44 | -2967.11 |
| NB | -2600.58 | -2584.08 | -2594.29 |
| ours | -2597.41 | -2587.22 | -2594.90 |

We use the news articles written during June 2000 by Associated Press Worldstream as test set which contains 4,392 news articles. Table 5 shows the comparison of $MLLPD$ between NB, LDA and our model on this test set. It can be seen that our model performs almost the same with NB and better than LDA, which verifies that incorporating the numeric features while constructing topic clusters do not affect the semantic coherence of topics.

At last, we list the top 5 daily and specific topics identified by our model (using *num+idf* feature combination) as well as their top words in table 6 for comparing the two topic types. As we can see, the most of daily topics concern economic and financial issues such as temporary stock prices and fluctuations of exchange rates, which are actually the most common daily topic in news genre. In addition, some daily topics are weather forecasts and air pollution reports. By contrast, the specific topics identified by our model seem to correspond to one or more specific events (e.g. Eritrea's border issue), which are usually time-specific.

## 4 Related Work

Distinguishing specific and daily topics based on text has not been well studied so far. As far as we know, Ge et al. [5] is the most related work to ours, which identifies daily events based on text for avoiding arbitrarily considering them coreferential. Unlike our unsupervised approach, they used an extra collection of documents with timestamps to generate a training set (but not golden standard) and trained a maximum entropy classifier with several categorical features like unigrams for identifying daily events. Another similar work is done by Li and Cardie [7], who used a hierarchical dirichlet process (HDP) model [9] to recognize time-specific and time-general topics for generating timelines for individuals from Twitter data.

**Table 6.** Examples of daily and specific topics identified by our approach (idf+num)

| Daily | | | | | Specific | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| dollar | index | degree | dollar | pollution | Iraq | Ethiopia | China | disease | drug |
| index | stock | breeze | exchange | air | oil | Eritrea | trade | EU | seize |
| rupee | fall | max | U.S | city | OPEC | UN | minister | ban | police |
| gold | rise | min | pound | level | sanction | peace | economic | U.S. | kilogram |
| turnover | NASDAQ | gentle | hongkong | report | minister | border | development | trade | heroin |
| silver | company | cloudy | British | pollution | Kuwait | Taliban | WTO | animal | myanmar |

Different from the previous work, our model only uses textual information and thus can be applied to any collection of documents regardless of the availability of their timestamps and the way of sampling from the text data streams.

## 5 Conclusion

In this paper, we investigate several numeric features and propose a novel Bayesian model with mixtures of Poisson distributions for distinguishing daily and specific topics. Our proposed model can be easily generalized to other tasks for incorporating numeric features in Bayesian models.

## Acknowledgements

We thank the anonymous reviewers for their valuable comments. This paper is supported by National Key Basic Research Program of China 2014CB340504 and NSFC project 61375074. The contact author of this paper is Zhifang Sui.

## References

1. Alonso, O., Gertz, M., Baeza-Yates, R.: Clustering and exploring search results using timeline constructions. In: CIKM (2009)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research (2003)
3. Chambers, N.: Labeling documents with timestamps: Learning from their time expressions. In: ACL (2012)
4. Doyle, G., Elkan, C.: Accounting for burstiness in topic models. In: ICML (2009)
5. Ge, T., Chang, B., Li, S., Sui, Z.: Event-based time label propagation for automatic dating of news articles. In: EMNLP (2013)
6. Ge, T., Pei, W., Ji, H., Li, S., Chang, B., Sui, Z.: Bring you to the past: Automatic generation of topically relevant event chronicles. In: ACL (2015)
7. Li, J., Cardie, C.: Timeline generation: Tracking individuals on twitter. In: WWW (2014)
8. Swan, R., Allan, J.: Automatic generation of overview timelines. In: SIGIR (2000)
9. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing clusters among related groups: Hierarchical dirichlet processes. In: NIPS (2004)
10. Toutanova, K., Klein, D., Manning, C., et al.: Stanford core nlp (2013)