https://doi.org/10.20965/jaciii.2016.p0467

Paper:

Topic Model Based New Event Detection Within Topics

Yaoyi Xi, Bicheng Li, and Yongwang Tang

Zhengzhou Information Science and Technology Institute
Zhengzhou 450002, China
E-mail: WIM_GY@163.com

[Received January 10, 2016; accepted March 22, 2016]

Traditional new event detection is first proposed by Topic Detection and Tracking and it is actually first event detection. However, one topic usually consists of many events. The automatic instant detection of each event in one topic, not only the first event but also the second, the third and so on, is very useful for users to correctly understand the main development trend of the topic. In this paper, we address the problem of new event detection in one single topic and propose a novel topic model to detect new events along with the topic evolution. Our topic model treats new event detection as novel semantic aspect identification in one topic, rather than measuring the analog degrees between content items by lexical congruence. Besides, it can automatically determine the appropriate number of aspects needed and can naturally adapt dynamic change in the vocabulary along with the topic evolution. We use a sequential Gibbs sampling algorithm for posterior inference, which well realizes the online new event detection. Experiments are presented to show the performance of our proposed technique. It is found that our proposed technique outperforms the comparable techniques in previous work.

Keywords: new event detection, topic evolution, hierarchical dirichlet process, sequential Gibbs sampling

1. Introduction

New Event Detection (NED) is first proposed by the Topic Detection and Tracking (TDT) program. It is one of the five tasks in TDT,[1] and is defined to detect stories about previously unseen topics in a stream of news stories [1]. New Event Detection in TDT (hereafter T-NED for short), also called First Story Detection, targets at a chronologically ordered stream of stories from multiple sources (and in multiple languages), involving multiple topics, and detects the first story that discusses a topic.

In TDT, a topic is defined to be a seminal event or activity, along with all directly related events and activities. Therefore, event in T-NED refers to the seminal event and it must appear in the first story about a topic.

1. The other four tasks are Story Segmentation, Topic Tracking, Topic Detection, and Link Detection respectively.

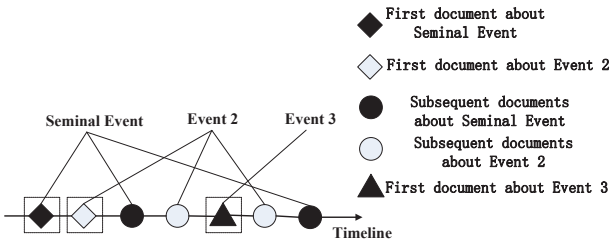

Fig. 1. Graphical representation of E-NED.

Compared with T-NED, this paper not only detects the seminal event, but also detects other topic-related events (hereafter E-NED for short). E-NED differs from T-NED in two ways, namely data objects and application requirements. From data objects, E-NED only involves one specific topic. From application requirements, E-NED aims at discover all new events that appear along with the evolution of the topic, not just seminal event. **Fig. 1** provides a graphical representation of E-NED. The topic in **Fig. 1** contains three events, namely seminal event, event 2 and event 3. E-NED aims to detect the first document for each of the three events along the timeline, namely the three documents within the dotted boxes, while T-NED only needs to detect the document in the leftmost dotted box, namely the first document about seminal event.

As far as we know, E-NED has never before been researched. This paper defines E-NED to detect the first document about an event in one specific topic. The formal definition of E-NED is as follows: Given a topic, suppose that there are D relevant documents arriving at time epoch T, which have been labeled as D_T and sorted in ascending order by their time of publication (As illustrated in **Fig. 1**, D_T contains seven documents). E-NED requires to sequentially detecting the first document d_{New} which reports a new event (As illustrated in **Fig. 1**, there are three new events and the first document about each of them is indicated by the box). All the documents detected comprise I_T (In **Fig. 1**, I_T contains three documents, namely the first document about seminal event, the first document about event 2 and the first document about event 3), which is a subset of D_T and satisfies $I_T \subseteq D_T$.

Although E-NED involves only one topic and defines a far smaller universe of documents than encountered in the T-NED task, it is still a difficult problem. In general, terms are different in different topics. So, there is always

a big difference between one seminal event and the other. However, documents which belong to the same topic tend to use similar terms. For instance, in the actual documents, especially news stories, the authors usually like to introduce other related information when they report a new event, such as background information. As a result, it is more difficult to distinguish between the relevant documents of the same topic. This is the key challenge E-NED faces.

The most prevailing approaches of T-NED are incremental clustering [2] and text classifying [3]. If the new story matches or can be classified into an existing cluster, it describes a known topic, otherwise it describes a new topic. The problem with the traditional approaches for T-NED is that they are mainly based on the computation of lexical similarities and fail to explore semantic information. In addition, they always retrospect all of the historical documents to improve the accuracy of new event detection [4]. As the number of relevant documents grows, this mechanism will suffer severe efficiency and accuracy declining. Topic model,[2] which can be less affected by synonymy and polysemy, is a hot research field in recent years. It has been applied successfully in TDT [5], multi-document summarization [6] and topic evolution analysis [7, 8]. In topic evolution analysis, researchers use topic model to model longitudinal data, and capture the trend of one topic by mining its aspects in different epochs. Although this approach can help to analyze topic evolution semantically, it only obtains the change of collections of keywords over time, but not novel relevant documents. In addition, the existing researches on topic evolution are mainly offline. Under the online environment, there are three problems must be considered in order to use topic model to analyze topic evolution: 1) the number of aspects will increase over time; 2) the vocabulary is dynamic, such as birth of new terms and disappearance of old terms; 3) the complexity of topic model's posterior inference should maintain a constant. Lau et al. [9] proposed a new online topic model for analyzing topic evolution, but the three problems are not properly settled. For the first problem, their model is based on LDA (Latent Dirichlet Allocation). The number of aspects not only needs to be pre-determined, but also does not change with the time. For the second problem, Lau et al. re-generated the vocabulary over time by removing terms that fall below a frequency threshold and adding new terms that satisfy it. This approach is too mechanical and is easy to lose terms. For the third problem, Lau et al. adopted a new posterior inference mechanism, which backtracked $(L-1)$ time slices as documents in a new time slice arrive and re-sampled the aspect assignments for all documents in L time slices. However, Lau et al. did not give a determination method of parameter L. It would add complexity if L is too high. Instead, it would easily misjudge redundant information as novel.

2. "Topic" in TDT refers to a collection of documents, whereas in "Topic Model," it means latent semantic in the document collection, such as topical aspects. In this paper, we use "aspect" for short and to differentiate it from "topic" in TDT.

To meet the requirement of online E-NED, this paper proposes an online-HDP model for E-NED. Compared with LDA, hierarchical Dirichlet process (HDP) model [10] does not need to predefine the number of aspects and can infer the number automatically. Online-HDP extends the standard HDP to the scenario of E-NED. The extensions include:

1) Extend the batch setting to the sequential one and propose a sequential Gibbs sampling algorithm for the online environment.

2) Translate E-NED into chronological novel aspect discovery and use the aspects mined by online-HDP in previous time epochs as the prior distribution for one specific time epoch. This mechanism not only avoids re-sampling old documents like [9], but also no longer requires retrospecting all of the historical documents.

The rest of the paper is organized as follows. In Section 2, we present related work. Section 3 gives a formal description of our E-NED approach, while Section 4 describes the experiments and reports the results. Finally, Section 5 summarizes our conclusions.

2. Related Work

In this section, we review the related work on topic evolution analysis, new event detection, novelty detection and update summarization.

2.1. Topic Evolution Analysis

Presently there are three ways to analyze topic evolution. The first class of methods is based on topic tracking [1, 11], the second one is based on storyline [12, 13] and the third one is based on topic model [1, 7, 8]. Methods based on topic tracking view a topic as a flat collection of stories, and are inefficient for a user to understand the topic quickly. Methods based on storyline build a storyline to help readers quickly grasp the general information of the topic, and their performances depend on the accuracy of event detection. Methods based on topic model capture the topic evolution trend by mining aspects in different epochs. Although topic model can analyze aspects in different epochs semantically, it only obtains the change of collections of keywords over time and cannot show the evolution of topic intuitively. In addition, methods based on topic model usually assume the full document collection as input and do not process documents dynamically as they arrive, so it is not the real sense of online topic evolution analysis [14, 15].

2.2. New Event Detection

In recent years, there has been a surge of interest in event detection, due to the ready accessibility of document streams from newswire sources and social media. Petrović et al. [16] used locality sensitive hashing (LSH)

to detect new event from a stream of Twitter posts, which can be scaled to massive volumes of data. Wang et al. [17] utilized topic model for new event detection in Twitter. They proposed a mixture Gaussian model for bursty word extraction and then employed a novel time-dependent HDP model to detect new event. Luo et al. [6] tried to develop a practical online new event system. They focused more on engineering implementation, such as user interface design and optimal allocation of computation resources, but less on improving algorithm performance.

2.3. Novelty Detection

Novelty detection is the task of identifying novel relevant information given a set of already accumulated background information. The TREC novelty track was conducted from 2002 to 2004. In the TREC novelty track, the task was to highlight sentences containing relevant and new information in a short, topical document stream [18]. The TREC novelty track is sentence-level novelty detection and is query-oriented. Zhang et al. [19] studied document-level novelty detection. Novelty detection is about user specified domain, and user information is available. It belongs to supervised learning. Compared with novelty detection, E-NED does not need to know the user information in advance and belongs to unsupervised learning.

2.4. Update Summarization

Update summarization aims to generate a short and concise summary about the novel information for the latest updating topic-related documents, under the assumption that the user has already read the earlier documents about the same topic [20]. The novel information refers to the information that could not be inferred by any previously documents. Summaries generated by such techniques consist of sentences extracted from the document collection. Li et al. [21] used h-uHDP model to explore the birth, splitting, merging and death of aspects for a given topic and proposed a new model for update summarization. Their method is state-of-the-art at present. Different from update summarization, E-NED only needs to detect document-level novel information.

3. Methodology

We first describe our proposed online-HDP model in Section 3.1. In Section 3.2, we show how we deal with the "rich get richer" phenomenon to improve the performance of the model. Next we explain how our topic model can be used to detect new event in Section 3.3.

3.1. Online-HDP Model

This section clarifies why and how we propose our online-HDP model for E-NED.

3.1.1. Model Description

Online-HDP model adopts three-layer structure, namely corpus layer, document layer and word layer. It mines the latent aspects in each document at each time epoch in chronological order. Online-HDP can solve the three problems mentioned in Section 1, and the reasons are the following.

Firstly, at one specific time epoch, online-HDP works like HDP. Therefore, it allows for unbounded number of aspects: aspects can be born at any epoch.

Secondly, most existing topic model based research about topic evolution requires to analyze the evolution of the aspects' term distribution and popularity. So they used state space model [22] or Markovian dynamics [15] and so on, to model the evolution process. One of the drawbacks of this way is that it assumes the vocabulary is static and unchanging across time. This is inappropriate. In contrast, E-NED mainly focuses on the identification of novel aspect and does not care how it evolves. Therefore, online-HDP avoids the above drawback by simply increasing the size of the vocabulary when there are new terms appearing and updating each aspect's term distribution by setting the newly added dimensions to be the prior weight. This way can naturally adapt dynamic change in the vocabulary without other complex processing. More formally, at the initial time epoch T_0, suppose each aspect's prior term distribution is $Dirichlet(V_{T_0}, \lambda)$. It is a symmetric Dirichlet distribution. V_{T_0} is the size of the vocabulary at T_0 and λ is every component's prior weight. At the next epoch T_1, the size of the vocabulary is V_{T_1}, then the k-th aspect's prior term distribution becomes an asymmetric Dirichlet distribution. The front V_{T_0} components' weights are the same as the ones of the k-th aspect's posterior term distribution in T_0, and the latter ones have the same prior weight λ. Note that for the novel aspects, online-HDP still adopts a symmetric Dirichlet distribution as their prior.

Thirdly, HDP belongs to nonparametric Bayesian model. The Bayesian paradigm provides a natural formalism for optimal learning from data in a sequential manner, with the posterior distribution at one time point becoming the prior distribution for the next. Online-HDP model makes use of this characteristic and retains the inference results in previous time epochs. In this way, every aspect's term distribution and popularity will be preserved and updated when it changes. This mechanism avoids resampling old documents like [9] and reduces the complexity of posterior inference.

As shown in **Fig. 2**, G_0^T denotes the global measure for the document set at time epoch T, G_j^T denotes the local measure for document j, $x_{j,i}$ is the i-th word in document j and $\theta_{j,i}$ is its aspect assignment. In the graphical model formalism, each node in the graph is associated with a random variable, where shading denotes an observed variable, such as word $x_{j,i}$ in **Fig. 2**. Rectangles denote replication of the model within the rectangle and the number of replicates is given in the bottom right corner of the rectangle, such as M^T and N_j^T in **Fig. 2**. M^T denotes the number

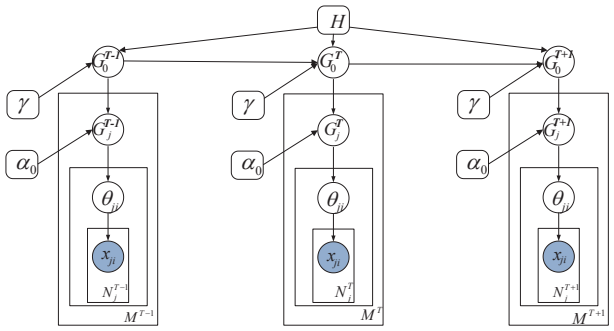

Fig. 2. Graphical model of online-HDP.

of documents at time epoch T and N_j^T denotes the number of words in the j-th document at time epoch T.

At time epoch T, the generation process of online-HDP model is as follows:

1. At the corpus layer, draw an overall base measure $G_0^T \sim DP(\gamma+\sum_k m_k^T, \sum_k \frac{m_k^T}{\sum_l m_l^T+\gamma}\delta(\phi_k)+\frac{\gamma}{\sum_l m_l^T+\gamma}H)$, which denotes the overall aspect distribution for the topic-related document set D_T at time epoch T, where $\{\phi_{1:k}\}$ are the aspects available in the previous epochs, m_k^T denotes the number of parameters associated with component k before epoch T, γ is the positive concentration parameter and H is the base probability measure.

2. At the document layer, for the document d_j in D_T, draw a local measure $G_j^T \sim DP(\alpha_0, G_0^T)$, where α_0 is the positive concentration parameter.

3. At the word layer, for the word $x_{j,i}$ in d_j, first draw the aspect assignment $\theta_{j,i}$ for $x_{j,i}$, then sample $x_{j,i}$ from the aspect corresponding to $\theta_{j,i}$.

3.1.2. Inference via Gibbs Sampling

In this section, we give a sequential Gibbs sampling algorithm for posterior inference in the online-HDP model. We use the Chinese Restaurant Franchise (CRF) to construct the model. Thus, we begin with an analog of the CRF process for online-HDP: a document d_j corresponds to a restaurant, the word $x_{j,i}$ corresponds to the i-th customer in restaurant j, and the latent aspect corresponds to the dish. CRF assumes that the number of tables in each restaurant can grow indefinitely and each table is served only one dish. The restaurant franchise has a shared menu across the restaurants. At each table of each restaurant one dish is ordered from the menu by the first customer who sits there, and it is shared among all customers who sit at that table. Multiple tables in multiple restaurants can serve the same dish. When a customer enters a restaurant, he can sit at an existing table with probability proportional to the number of customers already seated there, or sit at a new table with probability proportional to α_0. Due to limited space, we only give the calculation method for the aspect assignment of each document. More detailed information about Gibbs sampling can be found in [10]. The notations are summarized in **Table 1**.

Table 1. Table of symbols for online-HDP and Gibbs equations.

$n_{jt\cdot}$	the number of customers in restaurant j at table t
$m_{\cdot k}^T$	the number of tables serving dish k in all restaurants until epoch T
$m_{\cdot\cdot}^T$	the total number of tables in all restaurants until epoch T
$n_{\cdot\cdot k}^T$	the total number of customers eating dish k until epoch T
V_T	the size of the vocabulary at T

The conditional distribution of customer $x_{j,i}$ sitting at table t is

$$p(t_{ji}=t|\mathbf{t}^{-ji},\mathbf{k}) \propto \begin{cases} n_{jt\cdot}^{-ji} f_{k_{jt}}^{-x_{ji}}(x_{j,i}) & \text{if } t \text{ is previously used} \\ \alpha_0 p(x_{j,i}|\mathbf{t}^{-ji}, t_{ji}=t^{new}, \mathbf{k}), & t=t^{new} \end{cases} \quad . \quad (1)$$

When customer $x_{j,i}$ choose a new table to sit,

$$p(x_{j,i}|\mathbf{t}^{-ji}, t_{ji}=t^{new}, \mathbf{k}) = \sum_{k=1}^{K} \frac{m_{\cdot k}^T}{\gamma+m_{\cdot\cdot}^T} f_k^{-x_{ji}}(x_{j,i}) + \frac{\gamma}{\gamma+m_{\cdot\cdot}^T} f_{k^{new}}^{-x_{ji}}(x_{j,i}) \quad . \quad (2)$$

The conditional probability of generating word $x_{j,i}$ given a specified aspect k is

$$f_k^{-x_{ji}}(x_{j,i}=v) = \begin{cases} \frac{n_{\cdot\cdot k}^{T,-x_{j,i},v}+\lambda}{n_{\cdot\cdot k}^{T,-x_{j,i}}+V^T\lambda}, & \text{if } k \text{ is previously served} \\ \frac{1}{V^T}, & k=k^{new} \end{cases} \quad . \quad (3)$$

Once the customer chooses table t to sit, he then eats the dish served at this table. The conditional probability of serving dish k at table t is

$$p(k_{jt}=k|\mathbf{t},\mathbf{k}^{-jt}) \propto \begin{cases} m_{\cdot k}^{T,-jt} f_k^{-x_{jt}}(\mathbf{x}_{jt}), & \text{if } k \text{ is previously served} \\ \gamma f_{k^{new}}^{-x_{jt}}(\mathbf{x}_{jt}), & k=k^{new} \end{cases} \quad (4)$$

When a superscript is attached to a set of variables or a count, this means that the variable corresponding to the superscripted index is removed from the set or from the calculation of the count.

After Gibbs sampling, we can get the dish served for each customer, namely the latent aspects in the document.

3.2. Deal with "Rich Get Richer" Phenomenon

Dirichlet exhibits a "rich get richer" or clustering behavior, causing that few aspects quickly dominate over time. That is, most aspects are not updated with new observations and new observations are more easily assigned

to few existing dominated aspects. This is similar to the notorious particle degeneracy issue, which causes performance to degrade quickly over time.

When a new epoch's documents come, their aspect assignments are mainly affected by the dominating terms in the previous epochs. With the constantly development of the topic, the number of dominating terms should also be growing. These terms are that which play dominant roles in causing the "rich get richer" phenomenon. To alleviate this problem, this paper proposes an approach to identify these dominating terms and then filter out them.

Dominating terms generally appear frequently and account for a large proportion in each aspect's term distribution. To continually identify the dominating terms, online-HDP updates the dominating terms set at each epoch. The identification of dominating terms in one epoch is as follows:

1) For term t in time epoch T, which has not appeared in dominating terms set, its appearing frequency $AF_T(t)$ is the total number of its appearance in the epoch's documents. That is, the sum of its term frequency in each document. More formally,

$$AF_T(t) = \sum_{d_i \in D_T} TF(t, d_i)$$

where, D_T represents the number of documents in epoch T and $TF(t, d_i)$ denotes the term frequency of t in document d_i.

2) Sort the terms according to their appearing frequencies and choose the top N frequently appeared terms in the epoch to update the dominating terms set.

After the identification of dominating terms, online-HDP filters out them when the next epoch's new documents come. That is, online-HDP only keeps the terms which are not in the dominating terms set as the new epoch's model input.

By chronologically identifying and filtering out dominating terms, the influence of dominating aspects from any previous epochs will be reduced, which enhancing the model's sensitivity to new appearing terms.

3.3. New Event Detection Based on Online-HDP

When modeling topic-related document collections D_T by online-HDP at time epoch T, we can get the aspects and the aspect distribution of each document. Suppose there are K_T aspects in D_T, according to the publication time and aspect distribution of each document, we can sort the K_T aspects in chronological order and denote the sorted result as $\phi_1, \ldots, \phi_{K_T}$. Suppose that $\phi_{J+1}, \ldots, \phi_{K_T}$ only appear in the new incoming documents, the paper thinks that each of the $(K_T - J)$ aspects contains novel topic-related information. Denote the $(K_T - J)$ aspects as Nov_T, the documents containing these aspects as ND_T. The process of new event detection based on online-HDP is conducted as follows:

Table 2. Online-HDP based new event detection within topics.

Input:	Document collections coming in time order $D_{T_0}, D_{T_1}, \ldots$
Output:	New event document collection $I = \{I_{T_0}, I_{T_1}, \ldots\}$
Step 1:	For the initial document collection D_{T_0} coming at time epoch T_0 1.1 Get the aspect distribution of each document through modeling D_{T_0} by online-HDP; 1.2 Sort all the documents in D_{T_0} in chronological order and generate the timestamp of each aspect according to the publication time of the first document containing it; 1.3 Detect new event document according to the latent aspect it contains; 1.4 Identify and update the dominating terms set.
Step 2:	For the document collection D_T coming at time epoch T 2.1 Generate the input document collection at T by filtering out the terms which are in the dominating terms set; 2.2 Repeat Step 1 to complete new event detection.
Step 3:	Repeat Step 2 until the end of topic.

Firstly, sort the new incoming documents at time epoch T by their publication time;

Secondly, sequentially judge whether a document reports a new event or not. The judgment method is as follows: If document d contains a subset of Nov_T, namely $sNov_T$, it is considered as a new event document and is added to new event collection I_T. Remove $sNov_T$ from Nov_T, i.e., $Nov_T = Nov_T \setminus sNov_T$. If $Nov_T = \emptyset$, stop processing the documents at time epoch T and wait for new ones at next time epoch to arrive.

The process of detecting new event for topic evolution based on online-HDP is shown in **Table 2**.

4. Experiments

4.1. Datasets

There is no existing standard test set for E-NED methods. We randomly choose 4 bursty news topics from 4 selected Chinese news websites. Details statistics are listed in **Table 3**. We choose these sites because all of them provide special topic of news edited by professional editors.

After crawling all linked news stories for each topic, we hired two human annotators to label the new event documents for each topic independently. The annotators were asked to read the news stories of each topic several times to form a general picture on its development. In the next step, each annotator was asked to identify the events for each topic, and annotate new event documents independently. The two annotators then met together, reviewed the new event document collection constructed individually for each topic, and revised them to arrive at a "consensus" new event document collection for the topic.

For the online-HDP model, we set the topic Dirichlet

Table 3. Detailed information of 4 topics.

Source	Topics	Explosions in the Russian city of Volgograd, December, 2013	Crash of Asiana Airlines Flight 214 in San Francisco, July, 2013	Kunming terror attack, March, 2014	Boston Marathon bombing, April, 2013
Sina News		31	120	0	229
Tencent News		32	108	270	209
Phoenix News		74	322	0	224
Netease News		0	78	0	65
Sum		137	628	270	727

parameter $\lambda = 0.01$, and we sample the concentration parameters γ and α_0 from a gamma prior Gamma(10.0, 1.0). The sampling method is the same as that in [10].

For the purpose of modeling, we divide each of the topics into a set of epochs according to the article publication date and time. Each epoch contains 10 documents, which is the same as the setting in TDT2004 [23].

It should be mentioned that all the experiments were conducted on a PC with a Intel Pentium 4 CPU of 2.2 GHz and 4 GB memory running Microsoft Windows 7 operating system. We remove common stop-words and tokens which are neither verbs, nor nouns, nor adjectives from the news articles with the help of NLPIR[3]. Although the experiments are conducted for Chinese data sets, the method proposed by the paper also can be applied in English data sets. The only difference lies at that the preprocessing stage. For English, the English words have the spaces between the words, so it is easy to separate and the verbs, nouns, and adjectives can be easily identified with the help of Part-of-Speech tagging tools, such as NLTK[4], coreNLP[5] and so on.

TDT uses a cost function C_{Det} that combines the probability of missing a new story P_{Miss}, the probability of seeing a new story in the data P_{target}, the cost of missing a new story C_{Miss}, the probability of a false alarm P_{FA}, the probability of seeing an old story $P_{non\cdot target}$, and the cost of a false alarm C_{FA} in the following way:

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non\cdot target}. \quad (5)$$

Because the cost function values vary with the application, it is usually normalized by the minimum expected performance of a system that either answers YES or NO to all decisions. This normalization is defined as:

$$(C_{Det})_{Norm} = \frac{C_{Det}}{\min(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non\cdot target})}. \quad (6)$$

Different NED systems are compared based on their minimal cost $C_{\min}$, which is the minimal value of $(C_{Det})_{Norm}$ over all threshold values:

$$C_{\min} = \min_{\theta}(C_{Det})_{norm}(\theta). \quad (7)$$

In NED evaluation, the lower the $C_{\min}$ score reported by a system on test data, the better the system.

Because NED evaluation uses many topics to evaluate the performance, we have to somehow report average performance of the system. This is done by averaging $C_{\min}$ across all the topics, and it is called the topic weighted cost. The topic weighted cost is similar to the macro averaging technique that is standard in IR evaluation.

Besides the cost function, another measure of a NED system's performance is the F1-Measure [4] which is defined based on precision and recall. F1-Measure differs from C_{Det} in that a higher value indicates a better system.

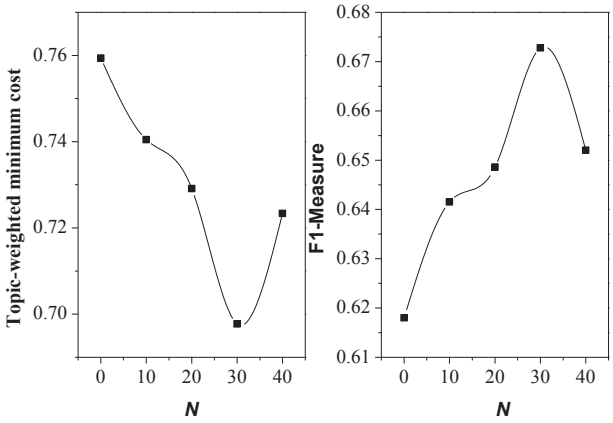

Fig. 3. Tuning parameter N.

4.2. Parameter Tuning

To complete E-NED using online-HDP, we still need to determine one parameter: dominating term frequency threshold N. Considering that the dominating terms set will be updated continually and the dominating terms are few in number at the beginning of a new topic, we gradually change N from 0 to 40 at the step of 10 to examine the effect in **Fig. 3**.

From **Fig. 3**, we can see that when $N = 0$, the performance is worst on the both evaluation metrics. $N = 0$ means that no dominating terms are filtered out, namely that the model does not dealing with the "rich get richer" phenomenon. When $N > 0$, the performance increases. We find that the minimal cost reaches minimum when $N = 30$. Likewise, the F1-Measure scores reach their peak at around 30 and drop afterwards. The experimental results conform to our expectation and verify that the "rich get richer" phenomenon can be alleviated by filtering out the dominating terms. We set $N = 30$ in our experiments.

3. http://ictclas.nlpir.org/
4. http://www.nltk.org/
5. http://nlp.stanford.edu/software/corenlp.shtml

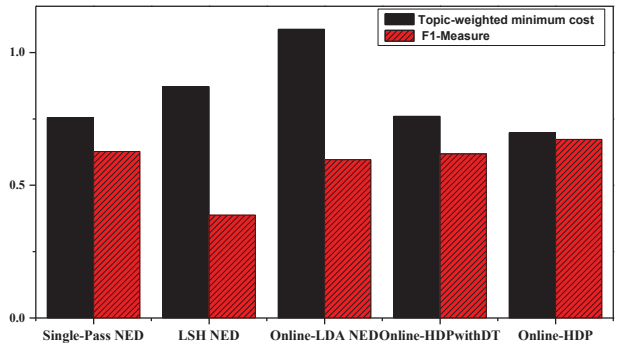

Fig. 4. Overall performance on 4 datasets.

4.3. Comparison of E-NED Results

To verify the E-NED approach, we compare it to a state-of-the-art NED system in the TDT5 competitions and other latest researches concerned. The UMass NED system [24] took use of an incremental clustering technique that is similar to Single-Pass. Single-Pass, which is also adopted by other research [1] declares a story to be new when the story's similarity to the nearest neighbors exceeds a threshold. Petrović et al. [16] adapted locality-sensitive hashing to NED to deal with high volume of data. Lau et al. [9] proposed an online-LDA model to capture the trending topics. Their model could also detect new event in one topic.

We implement the above three methods and denote them as Single-Pass NED, LSH NED and online-LDA NED, respectively. At the same time, we implement one online-HDP model without dealing with the "rich get richer" phenomenon and denote it as online-HDPwithDT. Because of the non-deterministic nature of Gibbs sampling, the online-HDP and online-LDA results reported here are the average of five runs.

Figure 4 shows topic-weighted minimum normalized costs and F1-Measure for our systems and other NED systems. From **Fig. 4**, we can see that our approach obviously outperformed the other systems.

LSH based approach is worse than Single-Pass, this has been verified in [16]. The advantage of this method is that it improves the speed of large quantity of data processing.

Online-LDA model's detection cost is highest. This is mainly because it restricts the number of aspects to be static, which cannot keep up with the actual. Under the online environment, the number of aspects should increase over time. In addition, different topics generally have different number of aspects.

Single-Pass is a classical method in T-NED, it also works fine in E-NED. For Single-Pass, it needs to predetermine the similarity threshold. Previous T-NED experiments indicated that the optimal threshold was around 0.2 [7]. However, the optimal similarity found in this paper is around 0.3. This is in accordance with the actual situation. Because E-NED focuses on the new event detection in the same topic, while T-NED only detects the seminal event of different topics. Intuitively, documents within a topic are more similar to each other than they are to a document belonging to a different topic.

As shown in **Fig. 4**, online-HDP model is better than online-HDPwithDT, which verifies that the identification and filtering out dominating terms can promote the performance of E-NED. Even without dealing with the "rich get richer" phenomenon, we can see the online-HDPwithDT approach is comparable to the other methods, indicating the effectiveness of applying topic model to E-NED. However, it also has its drawbacks. Topic model uses word co-occurrence to mine latent aspects. Latent aspects often result from higher-order co-occurrence, i.e., t1 co-occurring with t2 that co-occurs with t3 represents a second-order co-occurrence between t1 and t3, and so on. For some dominating terms, many other terms co-occur with them. Therefore, new observations are easily to be assigned to the dominated aspects which dominating terms belong to. By continually removing dominating terms, online-HDP alleviates this problem. Take the following two documents as an example. They are all about the topic "Explosions in the Russian city of Volgograd, December, 2013." The first document d_1 appeared on December 29 and the second document d_2 appeared on December 30. Both of them report a new event. The first document reports that the first explosion in the train station of Volgograd, while the second document reports that the second explosion on the bus in Volgograd. However, because the terms used in d_2 are similar with d_1, it is easy to identify d_2 as a old event. In our experiments, we find that all of the four comparison methods make the same mistake, but online-HDP can easily identify that d_2 reports a new event. The main reason is that online-HDP not only can mine the latent semantic information, but also it can remove the dominating terms, such as kill and *Volgograd and so on,* which highlights the contributions of new terms and reduces the probability of missing new event.

d_1 : *At least 14 people were killed and dozens wounded on Sunday when a female suicide bomber blew herself up in a train station in the southern Russian city of Volgograd.*

d_2 : *A day after a suicide bomber attacked a train station in the southern Russian metropolis of Volgograd, killing 17 and wounding scores more, a second bombblasted through one of the city's trolley buses during the Monday morning rush hour. At least 14 people were killed and more than two dozen wounded.*

The result details on each topic are listed in **Tables 4**–**7**.

For topics 2–4, online-HDPwithDT performs worse than Single-Pass NED, but for topic 1, its performance is far better than Single-Pass NED. The main reason is that topic 1 only lasts for a short time and online-HDPwithDT is less affected by the "rich get richer" phenomenon. On the whole, however, the performance of online-HDPwithDT is comparable to that of Single-Pass NED, indicating the effectiveness of our proposed methods.

Table 4. Performance comparison on Topic 1.

1. Explosions in the Russian city of Volgograd, December, 2013		
Approach	Topic-weighted Minimum Cost	F1-Measure
Single-Pass NED	0.7950	0.5682
LSH NED	0.8327	0.4795
Online-LDA NED	1.0828	0.5263
Online-HDPwithDT	0.5486	0.7216
Online-HDP	**0.5269**	**0.7523**

Table 5. Performance comparison on Topic 2.

2. Crash of Asiana Airlines Flight 214 in San Francisco, July, 2013		
Approach	Topic-weighted Minimum Cost	F1-Measure
Single-Pass NED	0.6817	0.6653
LSH NED	0.9139	0.2970
Online-LDA NED	1.0151	0.6595
Online-HDPwithDT	0.7005	0.6125
Online-HDP	**0.5867**	**0.6932**

Table 6. Performance comparison on Topic 3.

3. Kunming terror attack, March, 2014		
Approach	Topic-weighted Minimum Cost	F1-Measure
Single-Pass NED	0.7737	0.6725
LSH NED	0.7890	0.5178
Online-LDA NED	1.2237	0.6000
Online-HDPwithDT	0.9155	0.6462
Online-HDP	**0.7131**	**0.6867**

Table 7. Performance comparison on Topic 4.

4. Boston Marathon bombing, April, 2013		
Approach	Topic-weighted Minimum Cost	F1-Measure
Single-Pass NED	0.7654	0.5789
LSH NED	0.9484	0.2558
Online-LDA NED	1.0277	0.5983
Online-HDPwithDT	0.8729	0.4918
Online-HDP	**0.7643**	**0.6012**

From **Tables 4–7**, we can see that the assessment results are sometimes inconsistent by using the different evaluation methods. For instance, online-LDA NED performs worst according to topic-weighted minimum cost, but its performance is not bad at all according to F1-Measure. The main reason for this is that cost function not only considers the prior probability of target and non-target document, but also takes the cost of miss probability and false alarm probability into account. In general, the prior probability of non-target document is much higher than the target one. Therefore, if a NED system reports too many non-target documents, its minimum cost will be high. Different from cost function, F1-Measure only considers the probability of precision and recall. So even if there are many non-target documents, its F1-Measure will be not much lower as long as enough target documents are detected. Compared with F1-Measure, cost function is more widely used in research on NED.

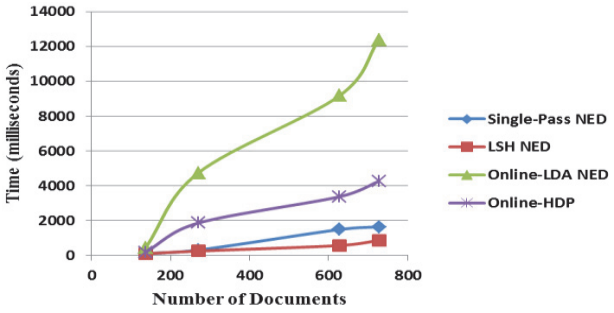

Fig. 5. New event detection time of different approaches.

Figure 5 shows the running time of the various new event detection approaches. LSH NED and Single-Pass NED are the fastest methods among all approaches because no iterations are involved. On the contrary, the running time of topic models grows significantly as the number of documents increases. More specifically, the running time of Online-LDA increases the most dramatically. Compared with Online-LDA, the running time of Online-HDP grows, but not obvious. Considering that Online-HDP performs best and the time it takes can be expected to remain within acceptance criteria, we conclude that our proposed Online-HDP method is efficient for new event detection within topics.

5. Conclusion

In this paper, research on E-NED is formally proposed for the first time and the necessity, emphasis and difficulty for the research are discussed. We treat E-NED as novel aspect identification and uses topic model to capture the semantical changes of a topic, which avoids the drawback of traditional approaches. Besides, aiming at the three problems of topic model-based online E-NED, we propose a new online-HDP topic model. This model comes up with solving measures from two angles: allows for unbounded number of aspects and retains the inference results of previous time epochs as the prior knowledge for next one. Besides, online-HDP alleviates the "rich get richer" phenomenon by continually identifying and filtering out the dominating terms.

Although online-HDP has achieved better performance than other approaches in T-NED, there are issues to resolve. A major issue is that the detection performance decreases over time. The main reason for it is the influence of the "rich get richer" phenomenon and the error made by approximate posterior inference. This is the point of our further research.

Acknowledgements

We thank Dr. Jey Han Lau for sharing the online-LDA code. This research is supported by the National Social Science Foundation of China No.14BXW028.

References:

[1] J. Allan, Topic detection and tracking: event-based information organization, Springer, 2002.
[2] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," Proc. of the 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, ACM, pp. 330-337, 2003.
[3] G. Kumaran and J. Allan, "Using names and topics for new event detection," Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 121-128, 2005.
[4] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," Proc. of the 21st Annual Int. ACM SIGIR Conf. on Research and development in information retrieval, ACM, pp. 28-36, 1998.
[5] X. Guo, Y. Xiang, Q. Chen, et al., "LDA-based online topic detection using tensor factorization," J. of Information Science, Vol.39, No.4, pp. 459-469, 2013.
[6] G. Luo, C. Tang, and P. S. Yu, "Resource-adaptive real-time new event detection," Proc. of the 2007 ACM SIGMOD Int. Conf. on Management of data, ACM, pp. 497-508, 2007.
[7] Y. Hu, L. Bai, and W. Zhang, "Modeling and Analyzing Topic Evolution," ACTA AUTOMATICA SINICA, Vol.38, No.10, pp. 1690-1697, 2012.
[8] Y. Hu, L. Bai, and W. Zhang, "OLDA-based method for online topic evolution in network public opinion analysis," J. of National University of Defense Technology, Vol.34, No.1, pp. 150-154, 2012.
[9] J. H. Lau, N. Collier, and T. Baldwin, "On-line Trend Analysis with Topic Models: #Twitter Trends Detection Topic Model Online," Proc. of the 24th Int. Conf. on Computational Linguistics, pp. 1519-1534, 2012.
[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Beli, "Hierarchical Dirichlet processes," J. of the American Statistical Association, Vol.101, No.476, pp. 1566-1581, 2006.
[11] M. Serizawa and I. Kobayashi, "Topic Tracking Based on Identifying Proper No.of the Latent Topics in Documents," J. of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.16, No.5, pp. 611-618, 2012.
[12] L. Huang and L. Huang, "Optimized Event Storyline Generation based on Mixture-Event-Aspect Model," Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing, pp. 726-735, 2013.
[13] S. Xu, S. Wang, and Y. Zhang, "Summarizing Complex Events: a Cross-modal Solution of Storylines Extraction and Reconstruction," Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing, pp. 1281-1291, 2013.
[14] J. Li and S. Li, "Evolutionary Hierarchical Dirichlet Process for Timeline Summarization," Proc. of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 556-560, 2013.
[15] A. Ahmed and E. P. Xing, "Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream," Proc. of the 26th Int. Conf. on Uncertainty in Artificial Intelligence, 2010.
[16] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 181-189, 2010.
[17] X. Wang, F. Zhu, J. Jiang, and S. Li, "Real time event detection in twitter, Web-Age Information Management," Springer Berlin Heidelberg, pp. 502-513, 2013.
[18] Soboroff and D. Harman, "Novelty detection: the trec experience," Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 105-112, 2005.
[19] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," Proc. of the 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, ACM, pp. 81-88, 2002.
[20] X. Li, L. Du, and Y. Shen, "Update Summarization via Graph-Based Sentence Ranking," IEEE Trans. on Knowledge and Data Engineering, Vol.25, No.5, pp. 1162-1174, 2013.
[21] J. Li, S. Li, X. Wang, Y. Tian, and B. Chang, "Update Summarization Using a Multi-level Hierarchical Dirichlet Process Model," Proc. of the 24th Int. Conf. on Computational Linguistics, pp. 1603-1618, 2012.
[22] D. M. Blei and J. D. Laerty, "Dynamic topic models," ICML, pp. 113-120, 2006.
[23] The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan [H], version 1.2, http://www. nist. gov.
[24] J. Allan, V. Lavrenko, D. Malin, and R. Swan, "Detections, bounds, and timelines: Umass and tdt-3," Proc. of Topic Detection and Tracking Workshop, pp. 167-174, 2000.

Name:
Yaoyi Xi

Affiliation:
Department of Data Processing Engineering, Zhengzhou Information Science and Technology Institute

Address:
Box 112, No.62, Science Avenue, High-Tech Zone, Zhengzhou City, Henan Province, China

Brief Biographical History:
2008- Joined Zhengzhou Information Science and Technology Institute

Main Works:
• "Temporal summarization based on biterm Dirichlet process," Acta Automatica Sinica, Vol.41, No.8, pp. 1452-1460, 2015.
• "ZZISTI at TREC2013 Temporal Summarization Track," TREC2013.
• "A Semantic Aspect-Based Vector Space Model to Identify the Event Evolution Relationship within Topics," J. of Computing Science & Engineering, Vol.9, No.2, pp. 73-82, 2015.

Name:
Bicheng Li

Affiliation:
Department of Data Processing Engineering, Zhengzhou Information Science and Technology Institute

Address:
Box 109, No.62, Science Avenue, High-Tech Zone, Zhengzhou City, Henan Province, China

Brief Biographical History:
1998- Joined Zhengzhou Information Science and Technology Institute

Main Works:
• "Intelligent Image Processing," Publishing House of Electronics Industry, 2005.
• "Efficient combination rule of evidence theory," Multispectral Image Processing and Pattern Recognition, Int. Society for Optics and Photonics, pp. 237-240, 2001.
• "A *k*-nearest neighbor text classification algorithm based on fuzzy integral," ICNC, Vol.5, pp. 2228-2231, 2010.

Name:
Yongwang Tang

Affiliation:
Department of Data Processing Engineering, Zhengzhou Information Science and Technology Institute

Address:
Box 109, No.62, Science Avenue, High-Tech Zone, Zhengzhou City, Henan Province, China

Brief Biographical History:
2008- Joined Zhengzhou Information Science and Technology Institute

Main Works:
• "Classifying Articles in Chinese Wikipedia with Fine-Grained Named Entity Type," J. of Computing Science & Engineering, Vol.8, No.3, pp. 137-148, 2014.