Computing and Informatics, Vol. 36, 2017, 1283–1311, doi: 10.4149/cai_2017_6_1283

LDA-BASED TOPIC STRENGTH ANALYSIS

Jiamiao Wang, Lei Li

School of Computer Science and Information Engineering
Hefei University of Technology
Hefei 230009, P.R. China
e-mail: wjmzjx@163.com, lilei@hfut.edu.cn

Xindong Wu

School of Computing and Informatics
University of Louisiana at Lafayette, USA
e-mail: xwu@louisiana.edu

Abstract. Topic strength is an important hotspot in topic research. The evolution of topic strength not only indicates emerging new topics, but also helps us to determine whether a topic will produce some fluctuation of topic strength over time. Thus, topic strength analysis can provide significant findings in public opinion monitoring and user personalization. In this paper, we present an LDA-based topic strength analysis approach. We take topic quality into our topic strength consideration by combining local LDA and global LDA. For empirical studies, we use three data sets in real applications: film critic data of "A Chinese Odyssey" in Douban Movies, corruption news data in Sina News, and public paper data. Compared to existing approaches, experimental results show that our proposed approach can obtain better results of topic strength analysis in detecting the time of event topic occurrences and distinguishing different types of topics, and it can be used to monitor the occurrences of public opinions and the changes of public concerns.

Keywords: LDA (latent Dirichlet allocation), topic strength

1 INTRODUCTION

With the rapid development of the Internet, information on various topics has an explosive growth. Hence, unsupervised learning based topic models which require no human interaction are becoming increasingly important. Most traditional data mining algorithms adopt supervised analysis, which requires human involved manual processing and analysis. However, when dealing with huge amounts of data on the Internet, relying on manual processing of web data is not realistic. In contrast, unsupervised topic models allow computers to deal directly with unlabeled data sets, thus ensuring the possibility of dealing with a larger data set. Moreover, a topic model itself has a good scalability, such as LDA (Latent Dirichlet Allocation). For example, in Citation-LDA [25], words can be replaced with the citation information. This is not only more suitable for academic research areas, but also reduces the computational complexity. In addition, correlated topic models (CTM) [3] can be obtained by taking the correlation between topics into consideration, and the model fits the real situation that there are certain links between topics in real life. Furthermore, coupled LDA [30] takes time and user's interest as iteration parameters in LDA to capture the change of a user's inherent interest over time directly. Therefore, because of their own merits and the growing mass of information, topic models have been widely used in social networks [5, 9, 15, 19], public opinion monitoring [8], user personalization analysis [24, 30], news reports [6, 12] and other fields.

In traditional topic detection and tracking (TDT), more attention is paid to the evolution of topic content, i.e. a change of focus in topic content over time. Different from the emphasis of TDT, this paper concentrates on the change of topic strength trends, with the following considerations. First, the concern about a topic in a certain period can be illustrated by the changes of topic strength very well. For example, compared to the results of topic content evolution, it is better to use topic strength in public opinion monitoring for a certain topic evolution. Second, the trends of topic strength can be used to distinguish different types of topics. During strength analysis, we can locate an event topic to analyze quickly and accurately, which can be applied to a field that needs a real-time analysis, such as online computing and timely personalized recommendations [14, 32]. Therefore, when facing with the vast amounts of data on the Internet, analysis techniques of topic strength can provide a very broad positive effect on research and can deliver very broad positive impact on a variety of applications. Existing approaches of topic strength analysis can detect the occurrences of events to a certain degree, but for practical applications, there are still many open problems.

1. The current definition of topic strength [18, 25, 24] is too simple what depends too much on the number of documents covering a topic. When the distribution of topic related data in the time slot is uneven, this definition is prone to adverse effects on results. Therefore, in this paper we redefine the topic strength, and add more factors when measuring topic strength in order to reduce the impact caused by the uneven distribution of topic related data.

2. Existing works make no difference between different types of topics in topic strength analysis. Hu et al. [12] pointed out that we should use a different time granularity when we analyze different types of topics. However, this approach not only wastes time and energy especially under a large number of topics, but also leads to errors when handling the different types in the same ways.

With the above observations, this paper has the following contributions:

1. Considering the uneven distributions of data sets, we introduce a definition named *topic recency*. In addition, local LDA is used to alleviate the problems caused by over-reliance on the number of documents.
2. We distinguish topics into event topics and routine topics. An event topic refers to a current event; in contrast, a routine topic is like a background topic, which is determined by the features of data sets, and it has nothing to do with the occurrences of events. Hence, in event analysis, we can ignore routine topics [10]. It will be easy to filter out event topics according to the topic strength analysis with *topic distinctiveness* introduced in this paper, so we can focus more on the analysis of event topics.
3. *Topic recency* and *topic distinctiveness* mentioned above are collectively referred to as topic quality, and in this paper topic quality is taken into the consideration of topic strength by combining post-discretization and pre-discretization; considering that the topic quality can acquire a good human-interpretability [13]. According to [29], LDA with a post-discrete process (also named as global LDA) is more accurate, because more data are used. In contrast, LDA with a pre-discrete process (also named as local LDA) can better discover an occurence of a new topic and a demise of an old topic. Our proposed approach combines the global LDA and the local LDA as follows: when calculating topic strength of a time slice, we analyze the relationship between a global LDA topic and a local LDA topic at the time slice. If the global LDA topic is closely related to a local LDA topic of the time slice, then the global LDA topic strength can be believed to be high in this time slice.

The paper is organized as follows. In Section 2 we review topic detection and tracking, topic strength and the LDA model. Section 3 introduces the definition of topic strength and its corresponding calculation. In Section 4, we choose three data sets in real applications to demonstrate the effectiveness of our approach, analyze some typical topics in those data sets, summarize the experimental results and show the effectiveness of our topic strength. Finally, the conclusion of our work is presented in Section 5.

2 RELATED WORK

In this section, three types of related work are reviewed, including topic detection and tracking (TDT) and topic strength.

2.1 Topic Detection and Tracking (TDT)

The Topic Detection and Tracking project was put forward in 1996 by the Defense Advanced Research Projects Agency. This project aimed to automatically identify new topics and keep track of known topics from flows of news media information, in order to deal with information flood on the Internet. This technology has gradually become a hotspot in information processing. The main target of TDT, since the very beginning, is to fulfill a task that analyzes news data without excessive manual intervention. In addition, TDT technology has been applied to personalized recommendations, and other practical applications such as commercial market research and public opinion monitoring, over those years.

Topic detection and tracking has two main sub-problems: topic detection and topic evolution.

1. Topic detection includes topic modeling, traditional vector space modeling and term co-occurrence approaches. Among these approaches, vector space modeling was proposed by Salton [20] in the 1970s. With it, a popular open source project named SMART (System for the Mechanical Analysis and Retrieval of Text) text retrieval system was established. There are several extended models of the original vector space model: such as multi-level vector space modeling with an optimized hyperlink selection [27] and single pass incremental clustering for online event detection [11]. The main idea of term co-occurrence is to boil down words with a high frequency to one topic, and this approach can be used for summary generation and keyword extraction. For instance, Madani et al. [16] used the term co-occurrence diagram for a more fine-grained topic. Toda et al. [24] constructed a graphical structure by combining term co-occurrences and document similarity, and generated topics by the graphical structure.
2. Compared to topic detection, studies on the topic evolution are relatively few, and they can be divided into the naive topic evolution model and the topic evolution model with different ways of time slice division. Tang et al. [21] proposed the naive topic evolution model, in which topic detection algorithms were used in every time slice, and then topics of adjacent time slices were compared to get keyword similarity so as to analyze the evolution of topics. The HDP model [1] proposed later is a little different, and this model selected overlapping time slices to analyze the topic evolution.
3. LDA is the most widely used topic model, and according to the characteristics of different areas, there are many different types of topic models. For example, Yin et al. [28] proposed latent periodic topic analysis (LPTA) for dealing with a problem that some topics appear periodically, and determining whether there was a cycle in the potential topic space. Tang et al. [23] took user points of interest into consideration and proposed a topic-user-trend model (TUT), in which they believed the topic model combining user interests allowed the model to extend to invisible data better. The eTOT model and the eDTM model took emoticons as major data sets for LDA, and they could analyze the situation of

topic evolution better according to the public attitude to an event. Dubey et al. proposed the npTOT model [9] that took timestamp as an iteration variable of LDA during topic evolution analysis, and this model could determine the number of topics adaptively and obtain an excellent result of topic evolution analysis.

Overall, topic detection and tracking is a highly comprehensive technology, and it focuses on both topic modeling and traditional statistical analysis techniques. Moreover, the technology has a wide application prospect in many fields, such as business decisions, public opinion monitoring and information retrieval, so there are a lot of approaches from different fields.

2.2 Topic Strength

Compared to topic evolution that focuses on topic contents, research efforts on topic strength are mainly about analysis approaches, not the definition of topic strength. As for analysis approaches, there are three main approaches: term frequency, term co-occurrence, and the LDA model [2] and its extended models.

1. As for the first technique, term frequency can be divided into word level frequency and sentence level frequency [7]. Cataldi et al. [4] proposed a model to find the life cycle of a topic according to a theory of aging. In essence, this technique relies on term frequency in computing topic strength. What draws our attention is that Cataldi et al. [4] considered topic quality when analyzing the historical information. So it is necessary to put topic quality into our definition of topic strength. Cataldi took topic quality into account, but their main idea was to calculate topic strength by term frequency. It is not bad to analyze traditional documents by term frequency, but when facing with information from social networks, there are some problems caused by text features, such as short length, noise and so on.
2. In the second technique, topic strength is defined by co-occurrence of terms and it assumes term co-occurrence is proportional to topic strength. Deng et al. [8] chose the sentence level to calculate the cosine of angle between public comments and a documents vector, and determined whether they appear together by thresholds. But there is an issue that we have to select the size of the window if we use the second technique. And using different sizes of windows to calculate term co-occurrence, at either the sentence level and paragraph level, will bring different results. In addition, different data sets also lead to different results. Therefore, the term co-occurrence approach has a poor versatility.
3. As for the third technique, LDA-based technology is the focus of the proposed research approach in our paper. As an LDA topic model changes the original word dimension into the topic dimension (specific details are described in Section 2.3), LDA-based technology can avoid the problem that the first technique

and the second technique have to face with. Liu et al. [18] proposed a simple definition about the LDA-based topic strength, in which only the document length and the topic-document distributions were used. TIARA [24] also used the LDA approach, but their definition was more complicated, and the topic strength was computed by the variance of a topic and the coverage of topic contents. Additionally, TIARA [24] proposed an adaptive algorithm for obtaining a time slice instead of fixed time intervals to divide data sets, and the time slice acquisition approach made a difference. As the application area of LDA modeling is very broad, there are many LDA extended models. For example, the correlation based ranking topic model [26] is based on correlated topic models (CTM) [3]. As it can be seen from the title, the correlation between the topics has been taken into consideration, and topic strength is calculated by the weighted averages of topic quality and topic correlation. Citation-LDA [25] is an extended LDA model, and the paper contents are replaced with citation information, thus, the computational complexity is greatly reduced. In Citation-LDA model, two parameters are taken into the topic strength definition: topic-document distributions and topic importance which indicates the probability of a topic being selected. In summary, LDA can avoid the issue that the term frequency approach and the term co-occurrence approach have to face with. Therefore, LDA topic modeling is a common technology in the research field of topic strength.

The utilization of topic strength is very broad. For instance, Citation-LDA [25] can reveal important milestone research papers and find the main research direction, when this model is applied in public scientific literature data sets. Topic strength analysis can be used in social networks, too. For example, the approach proposed by Ankan Saha et al. [19] had been applied in Twitter. An approach proposed by Deng et al. [8] was also applied in social media, and it could analyze user personalization and the public opinions in Weibo, such as a transition of public concerns over the epidemic outbreak, and the poll situation about the presidential election on Twitter, etc. TIARA [24] applies topic strength analysis into a personal E-mail information analysis system and a case analysis system, and it achieves good results. Coupled LDA modeling [30] is an application of Internet Protocol television (IPTV) for distinguishing different types of families according to habits that people watch TV in different time periods.

Topic strength is a subclass of traditional topic detection and tracking. But for now, most papers which study topic strength have adopted a simple topic strength definition, which can only handle a general situation. In Liu et al. [18], topic strength is defined as the quantity of documents covering this topic at particular time slices. The definition believes that a topic has a high strength if there are many documents covering this topic at the time slice. As for Citation-LDA [25], the emphasis of topic strength is the topic importance. For example, in public scientific literature data sets, topic strength is high when a milestone paper appears. Evolutionary Pattern Mining (EPM) considers the influence of documents and topic quality on topic strength separately, and both topic impact and topic attention are

taken into consideration when calculating topic strength. In addition, there are different emphases on definitions in public opinion monitoring. Deng et al. [8] took public attention as topic strength, public forwarding times and contents of news in Weibo as measures, and it was divided into five levels description: Warming Concern, Sustained Concern, Saturated Concern, Expected Concern and Unheeded Concern.

In addition, with the above approaches of topic strength, it is very difficult to distinguish different topics into routine topics or event topics. And it is hard to detect the time of occurrence, too. To solve these problems, we propose a new approach of topic strength analysis.

3 LDA-BASED TOPIC STRENGTH ANALYSIS

Our purpose in topic strength analysis is to learn how much attention a topic gets and to distinguish different types of topics. Technically, our main approach is to take topic quality into the consideration of topic strength, and topic quality here has two factors, *topic recency* and *topic distinctiveness*, which are calculated by combining post-discretization and pre-discretization. In particular, this approach calculates the relationship between topics obtained by global LDA and topics obtained by local LDA.

3.1 Local LDA and Global LDA

The difference between local LDA and global LDA is to select a pre-discrete process or select a post-discrete process to model the LDA topic. According to the definitions in Zhang et al. [29], local LDA refers to a pre-discrete process that divides documents according to time at first, and then models these documents by LDA. In contrast, global LDA refers to a post-discrete process that models documents by LDA at first, and then divides these documents according to time. There are some different characteristics between the two approaches. According to conclusions in Zhang et al. [29], pre-discrete topics by local LDA can find the occurrences of new events easily, and can also get more fine-grained topics, because local LDA focuses on local information. Meanwhile, post-discrete topics by global LDA can show the changes of trends better in the entire data sets.

We do not calculate the topic strength of the pre-discrete topics obtained from local LDA, so there is no need to record topic-document distributions. All we need is word-topic distributions, for calculating the similarity with post-discrete topics obtained from global LDA, and we use the similarity to evaluate the topic strengths of post-discrete topics in different time slices. In this paper, a pre-discrete topic is expressed by a triad $\{t, ts, \varphi\}$, where t represents the time when a topic appears, ts indicates topic strength of a topic at time t, and φ indicates word-topic distributions. The approach of local LDA is to divide the documents according to the time and to model each part of documents by LDA. For all documents D,

according to a certain time slice, it can be divided into $D = \{ld_1, ld_2, \ldots, ld_n\}$. Pre-discrete topics $lk_t = \{lk_{t1}, lk_{t2}, \ldots, lk_{tn}\}$ that belong to the time slice of ld_t are gained by modeling each ld_t, where $t = \text{time}(ld_t)$. The specific generation process of local LDA is as follows:

1. All documents D are divided into $D = \{ld_1, ld_2, \ldots, ld_n\}$ according to their time slice.
2. Determine the parameters α and β of the Dirichlet distribution.
3. For each topic lk_{ti} in pre-discrete topics $lk_t = \{lk_{t1}, lk_{t2}, \ldots, lk_{tn}\}$, choose a word-topic distribution $\varphi_{ti}^{lk} \sim Dir(\beta)$.
4. For each document ld_{tj} in documents $ld_t = \{ld_{t1}, ld_{t2}, \ldots, ld_{tm}\}$, choose topic distributions $\varphi_{ti}^{lk} \sim Dir(\beta)$, and for each word in a topic:
   (a) Choose a topic $z \sim Multi(\theta_{tj}^{ld})$.
   (b) Choose a word $w \sim Multi(\varphi_{ti}^{lk})$.

The topics analyzed by our approach are the post-discrete topics obtained from global LDA. Compared to local LDA, global LDA uses more data, so post-discrete topics are more accurate and representative, and it is better to describe developments of events that are concerned by people via post-discrete topics when calculating topic strength. In this paper, a post-discrete topic is expressed by a quad t, ts, φ, θ, where t represents the time that a topic appears, ts indicates the topic strength of a topic at time t, φ indicates word-topic distributions, and θ indicates topic-document distributions. The entire data sets are divided into training sets and test sets $D = \{D_{test}, D_{train}\}$, and we use the model inferred from the training sets to infer the topics from test sets. Global LDA is modeling on test sets, and then calculating the topic strength of different time slices according to the time of each document. After inferring the test sets $D_{test} = \{dtest_1, dtest_2, \ldots, dtest_{dn}\}$ by LDA, we obtain post-discrete topics $k = \{k_1, k_2, \ldots, k_m\}$ and their topic-document distribution θ. The specific generation process of global LDA is as follows:

1. The entire data sets are divided into training sets and test sets $D = \{D_{test}, D_{train}\}$.
2. Determine the parameters α and β of the Dirichlet distribution.
3. For each topic in post-discrete topics $k = \{k_1, k_2, \ldots, k_m\}$, choose a word-topic distribution $\varphi_{ki} \sim Dir(\beta)$.
4. For each document $dtrain_i$ in training sets $D_{train} = \{dtrain_1, dtrain_2, \ldots, dtrain_{dm}\}$, choose topic distributions $\theta_i^{dtrain} \sim Dir(\alpha)$, and for each word in the topic:
   (a) Choose a topic $z \sim Multi(\theta_i^{dtrain})$.
   (b) Choose a word $w \sim Multi(\varphi_{ki})$.
5. For each document $dtest_i$ in test sets $D_{test} = \{dtest_1, dtest_2, \ldots, dtest_{dn}\}$, choose a topic distribution $\theta_j^{dtest} \sim Dir(\alpha)$, and for each word in the topic:

(a) Choose a topic $z \sim Multi(\theta_j^{dtest})$.
(b) Choose a word $w \sim Multi(\varphi_{ki})$.

The difference between local LDA and global LDA is in using LDA at different time stages, and for this reason, there are different characteristics between pre-discrete topics and post-discrete topics. We combine pre-discrete topics and post-discrete topics, in order to obtain better analytical results of topic strength.

3.2 Event Topics and Routine Topics

Topics are divided into event topics and routine topics in this paper. A routine topic refers to a topic that emerges by the characteristics of data sets, and it can be understood as a background topic. A routine topic has nothing to do with events, i.e., whether events happen or not, the routine topic has always been there. However, an event topic refers to a topic where its strength changes dramatically if an event occurs, and the strength of the event topic is high in the period of this event. The difference between an event topic and a routine topic can be distinguished by topic strength.

Take film critics as an example, which is displayed in Table 1. We define a topic as a routine topic if its content is about the story or reviews. No matter whether the movie re-release events happen or not, a routine topic exists for a long time. We define a topic as an event topic if the keywords of the topic include "re-release" or "cinemas", and the topic strength of this topic is high when the event occurs.

Routine Topic	love, understand, miss, do not know how to cherish, live, well, give up, lose, life, had, with, no matter encountered, regret, when I, in a word, tell me
Event Topic	**A Chinese Odyssey**, **cinemas**, two, Pandora's Box, Wukong, then, the era, sequel, Journey to the West, Xian, shoot, understanding, shoe, **theater**, Romance, **re-release**, classic lines, the director, released

Table 1. An example of different types of topics

3.3 LDA-Based Topic Strength Analysis

In this section, we introduce LDA-based topic strength analysis in detail. First, two new concepts, *topic recency* and *topic distinctiveness*, are presented in Section 3.3.1 and Section 3.3.2, respectively. In Section 3.3.3, we introduce the definition of topic strength. The algorithm of topic strength calculation is shown in Section 3.3.4.

3.3.1 Topic Recency

In order to solve the situation that the distribution of some data sets is uneven, we introduce a new concept named *topic recency*. When an event occurs, data sets

are prone to uneveness due to a sudden rise of attention. In the movie critic data, for example, if a re-release event occurs, there will be a sharp rise of comments at the time of occurrence. However, the increasing comments are not all about a re-release, as there might be a large part of a discussion about the movie story. Without considering the important factor named *topic recency*, and just taking the number of documents covering the topic into consideration, the strength of a topic which describes the discussion about the movie story has a peak at the time of occurrence, too. As stories are mentioned during public comments, the strength of the topic which describes the discussion about the movie story is even higher than the re-release topic at the time of occurrence. We believe that the topic which describes the discussion about the movie story belongs to routine topics, so it is improper that the strength of this topic produces an obvious fluctuation with the development of events, and it is even more improper that this topic has an even higher topic strength than the re-release event topic at the time of occurrence. Taking *topic recency* into consideration can avoid the issue that topic strength relies too much on the number of documents when the data sets are unevenly distributed.

We define *topic recency* to make sure that an event topic has the higher topic strength at the time of occurrence. We tested several different similarity measures, and the results showed that different measures have almost the same effect on *topic recency*. *Topic recency* is different from topic similarity, and indicates the similarity between the global topic and the local topic. These two types of topics are obtained from different vocabularies. Hence, we choose cosine similarity as it is conventional and popular. When calculating the *topic recency* of topic k_i at time t, we first calculate cumulative cosine similarity. Cumulative similarity is shown in Definition 1, and the *topic recency* of topic k_i at time t is shown in Definition 2.

Definition 1. For a post-discrete topic k_i, it calculates cosine similarity $simi(k_i, lk_{tj})$ with each topic lk_{tj} in pre-discrete topics $lk_t = \{lk_{t1}, lk_{t2}, \ldots, lk_{tn}\}$ at first, and then accumulates the similarity according to Equation (1). The obtained results are cumulative similarity $Csimi(k_i, lk_t)$ of topic k_i at time t.

$$Csimi(k_i, lk_t) = \sum_{j=0}^{n} simi(k_i, lk_{tj}). \tag{1}$$

Definition 2. *Topic recency* $tr(k_i \mid t)$ is used to judge whether the post-discrete topic k_i happens at time t. Particularly, a specific approach of calculation is shown as Equation (2).

$$tr(k_i \mid t) = pow(p, Csimi(k_t, lk_t)). \tag{2}$$

In Definition 2, we add a control parameter p to control the impact of cumulative similarity on *topic recency*, and we set $p = 50.0$ here. The threshold of control parameter p depends on time slice granularity and data sets. As shown in Figure 1, when the control parameter p selects a larger value, such as 50.0, *topic recency* has a greater value. Thus, the value of control parameter p is inversely proportional to

the uniformity of data set distribution. Namely p is small if the distribution of data sets is even, and p is large if the distribution of data sets is uneven.

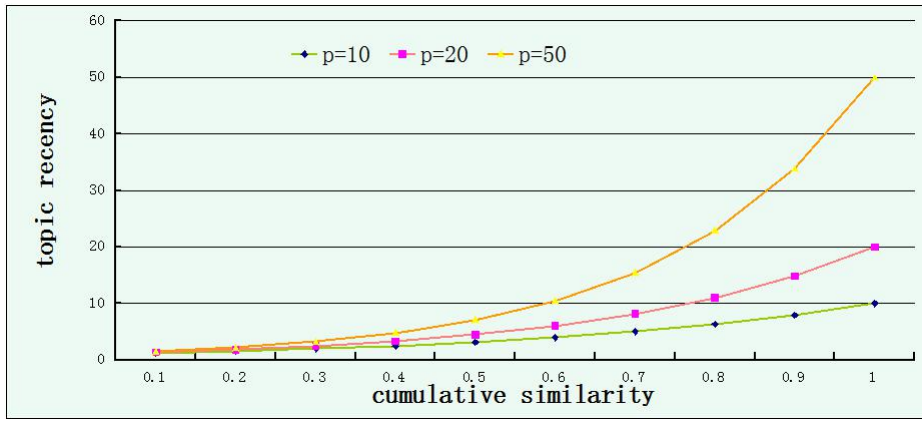


Figure 1. The influence of cumulative similarity on *topic recency*

The physical meaning of *topic recency* is easy to understand, as described in Section 3.1, i.e., a pre-discrete topic can detect the emergency of a new topic and disappearance of an old topic better. Therefore, a pre-discrete topic can detect topics which relate to current events accurately, and it has a finer granularity and a better timeliness. When a post-discrete topic k_i has a high similarity with a pre-discrete topic lk_t at a time slice, a post-discrete topic k_i has a high *topic recency* at the time slice. Thus, the topic strength at this time slice is high. Moreover, there is a control parameter p in the definition of *topic recency*, so that we can set a different value of p to control the influence between cumulative similarity and *topic recency* according to the data sets. Thereby, the control parameter p makes the definition of topic strength applicable to different fields.

3.3.2 Topic Distinctiveness

We introduce a new concept named *topic distinctiveness* for distinguishing event topics and routine topics. Literally, *topic distinctiveness* indicates how special a topic is, and topics can be distinguished by *topic distinctiveness* (a detailed description is given in Definition 3). *Topic distinctiveness* has a similar idea with TF-IDF which is used to describe category discrimination. For difference, *topic distinctiveness* is used to distinguish topics not documents. TF-IDF treats words equally. The key words of topics have weights and the key words with small weights are not important for a topic. Considering this situation, *topic distinctiveness* only takes top 20 key words into consideration. We use word frequency to measure *topic distinctiveness* which reduces the computing. Hence, we define *topic distinctiveness* as the number of words which have low frequencies in the topic, and the *topic distinctiveness* of

a topic is high when there are more low frequency words. Most routine topics are similar, such as routine topics of movie critics we mentioned in Section 3.2, and the routine topics of movie critics always have the same words owing to the discussions of the movie story, such as classic lines of the movie, actor names, etc. For this reason, the number of low frequency words of a routine topic is less. As for an event topic, the words related to the event appear rarely, so the number of low frequency words of a routine topic is higher. For example, the word "re-release" only appears in the event topic. So an event topic has a larger *topic distinctiveness*, and in contrast, routine topic has a smaller *topic distinctiveness*.

Definition 3. *Topic distinctiveness* $td(k_i \mid t)$ is used to measure the uniqueness of a post-discrete topic k_i, and the calculation is based on word frequency. The word frequency here refers to the collections which contain post-discrete topic k_i and pre-discrete topic $lk_t = \{lk_{t1}, lk_{t2}, \ldots, lk_{tn}\}$ at the current time slice. The specific calculation approach is shown in Equation (3).

$$td(k_i \mid t) = \#\langle TF_{k_i,lk_t} \rangle. \tag{3}$$

Here, TF_{k_i,lk_t} indicates the frequency of a word in the collections which contain post-discrete topic k_i and pre-discrete topic lk_t, and $\#\langle\cdot\rangle$ indicates the number of words satisfying $\cdot$ in $\langle\ \rangle$.

We take the post-discrete topic and the pre-discrete topic into consideration when we calculate *topic distinctiveness*. Thus, since a word in a routine topic is universal, the frequency of most words in the collections is high, and the *topic distinctiveness* is low. However, words in an event topic always relate to the event. Not like the situation that most routine topics are similar, an event topic has some unique words related to the event topic, so *topic distinctiveness* of an event topic is high.

3.3.3 Calculation of Topic Strength

Three kinds of parameters, cumulative topic-document distribution, *topic recency* and *topic distinctiveness* are used in the calculation of topic strength, which means that the number of documents and topic quality are taken into consideration for topic strength. We give the definitions of cumulative topic-document distribution and topic strength as follows.

Definition 4. Cumulative topic-document distribution shows the influence of document frequency on topic strength. Formally, for a post-discrete topic k_i, we accumulate the topic-document distribution of topic k_i at time t according to Equation (4).

$$Pr(t \mid k_i) = \sum_{time(d_j)=t} P(d_j \mid k_i). \tag{4}$$

Here, d_j indicates the j^{th} document in documents D, the time slice is denoted by t, and $P(d_j \mid k_i)$ indicates the topic-document distribution of topic k_i in document d_j.

Definition 5. Topic strength is affected by a combination of the number of documents which cover this topic and the quality of the topic. As for a post-discrete topic k_i, its topic strength $ts(k_i, t)$ at time t can be calculated by Equation (5).

$$ts(k_i, t) = Pr(t \mid k_i) * tr(k_i \mid t) + td(k_i \mid t). \tag{5}$$

Here, the topic-document distribution $Pr(t \mid k_i)$ is the cumulative result according to the time slice, and it explains the influences caused by the number of documents which cover this topic. Meanwhile, the influences caused by topic quality can be explained satisfactorily after the addition of *topic recency* and *topic distinctiveness*. The definition we proposed not only alleviates the problem caused by an uneven distribution of data sets, but also distinguishes different types of topics.

3.3.4 Algorithm for Topic Strength Calculation

The steps of the topic strength calculation are given in this section. As mentioned before, we take topic quality which is composed by *topic recency* and *topic distinctiveness* into the consideration of topic strength by combining post-discretization and pre-discretization.

Steps of the topic strength algorithm are as follows:

Step 1: Acquire pre-discrete topics $lk_t = \{lk_{t1}, lk_{t2}, \ldots, lk_{tn}\}$ by local LDA where $t = time(ld_i)$ (lines 1–6 in Algorithm 1).

Step 2: Acquire post-discrete topics $k = \{k_1, k_2, \ldots, k_m\}$ by global LDA. Considering the topics calculated by global LDA can cover the whole data sets, we calculate the topic strength on post-discrete topic (line 7).

Step 3: Calculate *topic recency* $tr(k_i \mid t)$ of topic k_i to alleviate the problem caused by the uneven distribution of data sets (lines 10–11).

Step 4: Calculate *topic distinctiveness* $td(k_i \mid t)$ of topic k_i to distinguish the different types of topics according to the features of words in the topic (line 12).

Step 5: Calculate topic strength $ts(k_i, t)$ of topic k_i (lines 13–14).

4 EXPERIMENTS

We present experimental results in this section, and analyze both the effectiveness and the time complexity. In the experiments below, we use the topic model in machine learning toolkit MALLET [17] to set up a LDA module.

Algorithm 1 Topic strength calculation

Input:
time t, data sets $D = \{ld_1, ld_2, \ldots, ld_n\}$ after divided according to the granularity of time slices.

Output:
topic strength $ts(k_i, t)$ of topic k_i at time t.

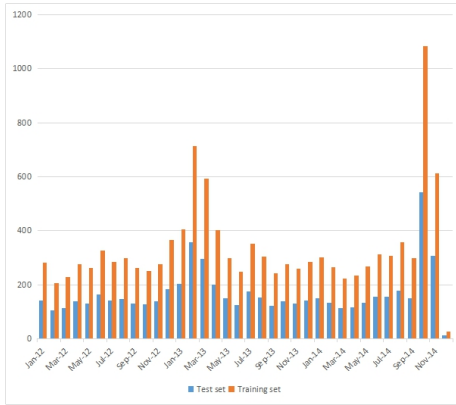
Divide data sets $D = \{ld_1, ld_2, \ldots, ld_n\}$ according to time slices.
for $i = 1$ to n **do**
  if $t = time(ld_i)$ **then**
    $lk_t = LDA(ld_i)$
  end if
end for
Acquire post-discrete topics $k = \{k_1, k_2, \ldots, k_m\}$ by global LDA.
for $i = 1$ to m **do**
  for $t = starttime$ to $endtime$ **do**
    $Csimi(k_i, lk_t) = \sum_{j=0}^{n} simi(k_i, lk_{tj})$
    $tr(k_i \mid t) = pow(p, Csimi(k_t, lk_t))$
    $td(k_i \mid t) = \#\langle TF_{k_i, lk_t} \rangle$
    $Pr(t \mid k_i) = \sum_{time(d_j)=t} P(d_j \mid k_i)$
    $ts(k_i, t) = Pr(t \mid k_i) * tr(k_i \mid t) + td(k_i \mid t)$
  end for
end for
return $ts(k_i, t)$

4.1 Data Sets

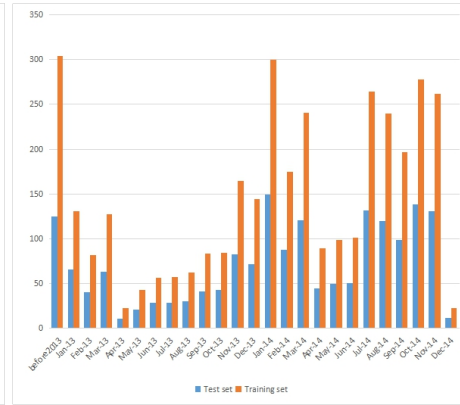
We select three data sets in real applications: film critic data of "A Chinese Odyssey" in Douban Movies[1], corruption news data in Sina News[2] and a public data set about papers used in [22]. These three data sets satisfy our requirements: all have a good time continuity; their size is moderate; there are some events happening with time elapsing, which satisfies the conditions for routine topics and event topics; and also, there is evidence to support true time occurrences. It is convenient for us to analyze these data sets.

The details of these data sets are shown in Figure 2. For every data set, we take 2/3 as the training set, and the remaining 1/3 as a test set. A point of attention is that the data distribution in each data set is uneven. For instance, in Figure 2 a) the uneven distribution of critic data in month granularity is caused by re-release events in October 2014 and sequel release events in February 2013. Using the approach proposed in [18] has a defect that it depends too much on the number of documents,

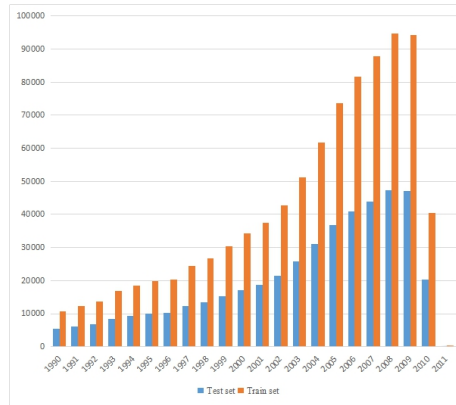
[1] http://movie.douban.com/
[2] http://news.sina.com.cn/



a) Film critic data in Douban Movies



b) Corruption news data in Sina News



c) Papers data

Figure 2. Three data sets

which makes every topic peaking in October 2014. It is difficult to distinguish the topics.

4.2 Baselines

In this section, the comparative methods are introduced and the time complexity is analyzed. Additionally, according to the characteristic of our method, we provide an idea on using parallel processing to optimize the time complexity.

The definitions of topic strength in [18] and [25] are chosen for comparative experiments. The definition of topic strength in [18] named "strength of the topic" (hereinafter referred to briefly as ST) emphasizes the number of documents which cover this topic, and it is a common topic strength approach. The strength of the

topic can be computed by Equation (6), where, $L(d_j)$ means the length of d_j, and $P(d_j \mid k_i)$ indicates the topic-document distribution of topic k_i in document d_j.

$$ST(t \mid k_i) = \sum_{time(d_j)=t} L(d_j) * P(k_i \mid d_j). \tag{6}$$

The definitions of topic strength in [25] named "topic temporal strength" (hereinafter referred to briefly as TTS) emphasizes on the influence of topic importance on topic strength, and it can be computed by Equation (7). Although TTS takes topic quality into the consideration of topic strength like we do, it only considers the average posterior distribution over topics, which is named as $\widehat{Pr}$, and its calculation is shown in Equation (8), where $\widehat{Pr}$ indicates the probability that k_i is chosen during the iteration.

$$TTS(t \mid k_i) = \widehat{Pr}(k_i) * P(k_i \mid d_j), \tag{7}$$

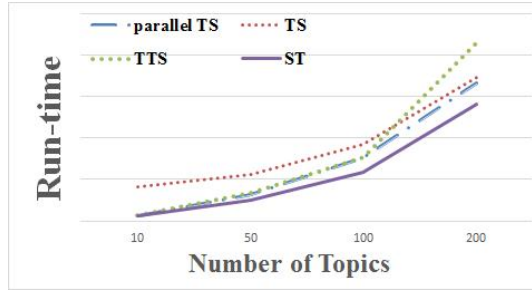
$$\widehat{Pr}(k_i \mid W) = Average\left(\frac{\#\langle z = k_i\rangle}{\sum_k \#\langle z = k_i\rangle}\right). \tag{8}$$

We choose these two definitions of topic strength for comparative experiments, because the definition of topic strength we have proposed considers both topic quality and the number of documents.

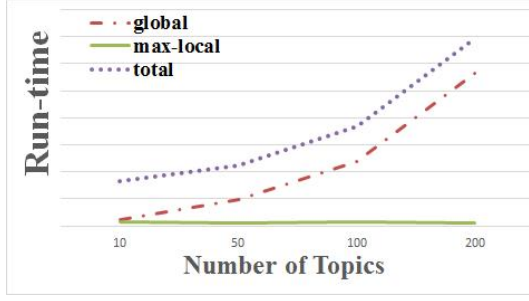
The performance of the proposed approach in terms of run-time is shown in Figure 3. As we can see in Figure 3 a), TS is the most time-consuming method as the number of topics is less than 100, because our approach needs to model the LDA on both the entire data set and every time slice. However, as the number of topics increases, TTS becomes more time-consuming, because TTS has to calculate $\widehat{Pr}$ in every LDA iteration, where $\widehat{Pr}$ indicates the probability that k_i is chosen during the iteration. So when the number of topics is large, TTS takes more time to get topic strength. In addition, the global LDA and the local LDA are processes independent to each other, thus this independence enables them to implement simultaneously. And local LDA focuses on a time slice dataset, so the run-time of the local LDA is much less than the global LDA, what is shown in Figure 3 b). Based on this observation, using parallel computing can optimize the time complexity of our approach.

4.3 Experiment with Film Critic Data in Douban Movies

We checked in a total of 27 812 critic reviews from 2006 to 2014, and we set the number of topics of our interest as 30. Overall, the results are good in detection of event topic occurrences and distinction of topics. We selected some typical topics to prove the effectiveness of our approach.



a) The time complexity



b) The run-time of different steps

Figure 3. The analysis of time complexity

4.3.1 Detection of Event Topic Occurrences

We selected an event topic to compare the detection of event topic occurrences, and verified whether the results are consistent with the actual situation. Topic 14 is an event topic, its contents are shown in Table 2 and its strength is shown in Figure 4. The sequel of film "A Chinese Odyssey" named "Journey to the West: Conquering the Demons" was released in February 2013, and the sequel shared the same theme song named "The Love in My Whole Life". So according to the actual situation, the strength of Topic 14 should have a peak in February 2013. As we can see in Figure 4, our results match the actual situation that its strength is 3 times higher than others. However, TTS [25] and ST [18] do not exhibit the obvious peak.

The contents of Topic 14	finally, the love in my whole life, hearts, A Chinese Odyssey, leaving, love, conquering the demons, music, hear sounds, sadly, scene, OST, really, beside him, to stay in, nice, there is not, mind

Table 2. The contents of Topic 14

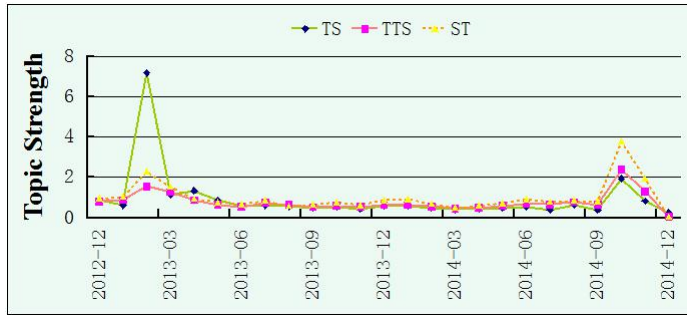


Figure 4. Comparative results of Topic 14 strength

At detection of event topic occurrences, our approach is better than TTS [25] and ST [18]. TTS and ST are affected by the uneven distribution of data sets, because they depend too much on the number of documents which cover this topic. The approach we put forward uses pre-discrete topics to alleviate the problem by considering *topic recency* for a better result.

4.3.2 Topic Strength of Routine Topics

In addition, topic strength of a routine topic also fits the actual situation. Topic 6 is a typical routine topic. More specifically, according to the topic contents given in Table 3, we know Topic 6 is about reviews of "A Chinese Odyssey". The topic strength of a routine topic has no relation with events, so its topic strength should not have an obvious fluctuation at the time of occurrence. The results of TTS and ST have an error peak in October 2014 because of the number of documents. Compared to the results of TTS and ST, our results are more realistic.

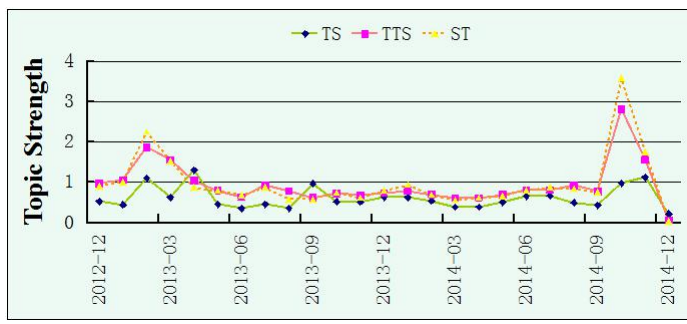


Figure 5. Comparative results of Topic 6 strength

In general, the results obtained by our approach can show the characteristics of routine topics and have a better performance in exhibiting the strength of routine

The contents of Topic 6	love, understand, miss, do not know how to cherish, live, love, well, give up, lose, life, had, with, no matter encountered, regret, when I, a sentence, tell me

Table 3. The contents of Topic 6

topics. As clearly shown in Figure 5, our strength of routine topics is smoother than the results of comparative approaches, and it does not have an obvious fluctuation over time. Visually, our results meet the real situation more properly.

4.3.3 Distinction of Topics

Our approach not only distinguishes event topics and routine topics, but also distinguishes different types of routine topics. For example, Topic 1, Topic 6 and Topic 20 belong to routine topics, and Topic 29 is an event topic, and their contents are shown in Table 4. Because of the re-release event in October 2014, some words like "re-release", "cinema" appear in the Internet users reviews. Topic 29 is an event topic, thus, its strength has a different fluctuation with three routine topics (shown in Figure 6). As for three routine topics, Topic 1 and Topic 6 have an approximate fluctuation (shown in Figure 6), but their strengths are different from strength of Topic 20. The reason is that although the three topics are all routine topics, they belong to different sub-classes. To be precise, Topic 1 and Topic 6 belong to the subclass which is about reviews of movies. However, according to the contents of Topic 20, it belongs to the subclass which is about lines of movies. The four topic strengths are shown in Figure 6, and what we can see from the figure is that the topic strength which is about classical lines of movies is high. The factor leading to this situation is that most comments tend to cite the film's classical lines.

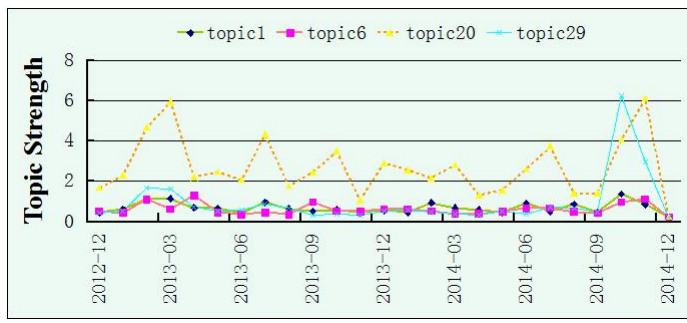


Figure 6. Different types of topics in TS

Judged on the topic strength which is clearly shown in Figure 6, some routine topics like Topic 1 and Topic 6, their topic strength is smooth, the variance of

strength is 0.07 and the average of strength is less than 1. This kind of routine topic does not have a fluctuation over time. However, the other kind of routine topic like Topic 20 has a high topic strength, i.e., the average of strength is 2.7, and Topic 20 has an obvious fluctuation. As for Topic 29 which is about the re-release of a movie, it has a peak in October 2014. In conclusion, besides the detection of event topic occurrences, results obtained by our approach can distinguish event topics and routine topics, and can also distinguish different types of routine topics.

Topic 1	see, movies, many times, read, sad, a few times, a lot, everyone, to see, to look, several, many times, too good, too much, Hong Kong films, somehow, good films, and now, the students, section
Topic 6	love, understand, miss, do not know how to cherish, live, well, give up, lose, life, had, with, no matter encountered, regret, when I, in a word, tell me
Topic 20	heroes, one day, married, loved, colorful, rainbow, clouds, stepping on the outcome, guessed, clouds, front, con not guess, clouds, foot, wearing, shining armor cloth, appear, driving, marching
Topic 29	A Chinese Odyssey, cinemas, two, Pandora's Box, Wukong, then, the era, sequel, Journey to the West, Xian, shoot, understanding, shoe, theater, Romance, re-release, classic lines, the director, released

Table 4. The contents of topics

According to the results of film critic data in Douban Movies, the TS algorithm we proposed have a good performance in detection of event topic occurrences, distinction of topics and calculation of routine topic strength. TS can get the accurate topic strength in the granularity that the distribution of data sets is uneven thanks to the consideration of topic quality.

4.4 Experiment with Corruption News Data in Sina News

News data sets are obtained from Sina News and we select corruption news as the experimental data sets. We crawled a total of 5386 news pages from 2005 to 2014.

We select some typical topics to verify the effectiveness of our approach. The event topic about the CPC fourth plenary session and the event topic about a special central patrol group are chosen to show the detection of event topic occurrences. Meanwhile the event topic about Zhou Yongkang and a routine topic are chosen to show the distinction of topics.

4.4.1 Detection of Event Topic Occurrences

We analyze the results according to the real time of event occurrences. The contents of Topic 2 which is about the CPC fourth plenary session are shown in Table 5, and its topic strength is shown in Figure 7.

According to the actual condition, we know that the fourth plenary session of the 18^{th} Communist Party of China Central Committee (CPC) was convened on

October 20, 2014. ST and TTS cannot detect the right time of the event occurrence, but our proposed TS can establish a peak at the real time of the event occurrence. Another point worth noting are the two lines of topic strength which substantially coincide, and the reason is that Topic 2 is about the CPC fourth plenary session. However, the lifecycle of the topic is not long enough, and the time of occurrence is rearward. Unfortunately, the definition of topic strength in TTS emphasizes the importance of the topic, an event topic which occurs rearward is affecting the whole data sets slightly, so the importance of the topic is low, and it leads to the situation that the topic strength of TTS and ST is close.

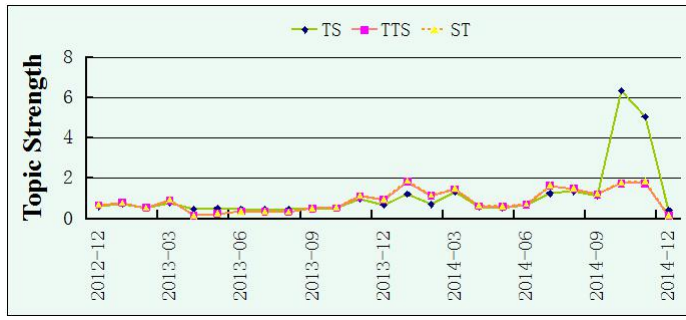


Figure 7. Topic 2 strength

The contents of Topic 2	law, Xi Jinping, military, anti-corruption, national, Fourth Plenary Session, corruption, the rule of law, the way, the original, strictly, according to the law, the country, the Central Military Commission chairman, heading, against, the President, how, use, legal

Table 5. The contents of Topic 2

As clearly shown in Figure 7, our proposed TS can detect the time of the occurrence more accurately. Meanwhile, there is a defect of TTS we can identify from the strength of Topic 2. TTS takes the number of documents and topic importance into consideration, but topic importance is affected a lot by the time of occurrence. Hence, it is difficult to acquire the accurate time by topic importance if the event takes place at a later time. Our approach takes pre-discrete topics into consideration, so that we can acquire the right strength of the event topic which takes place at a later time.

4.4.2 Detection of Multi-Time Event Topic Occurrences

A multi-time event topic is a subclass of event topic, and is very common in real life. We choose an event Topic 20, and its contents are shown in Table 6. What is different from Topic 2 is that Topic 20 is a multi-time event topic. Thus, Topic 20 has multiple peaks with the development of events, and its topic strength is shown in Figure 8.

As we can observe in Table 6, Topic 20 is about the central patrol group. In the light of real situation, we know that the patrol group has 3 patrols in March, July and October 2014, respectively. The approach we proposed can come to the conclusion at the real situation. However, a few patrols in 2013 are not founded, and we speculate that one reason is that the news of patrol group is small because the public and the media do not think the patrol group was significant and useful in 2013, and another reason is that the distribution of data is uneven. Although TS we put forward can alleviate the problem caused by uneven data, TS cannot alleviate the problem caused by missing data. It is impossible to avoid errors generated by missing data.

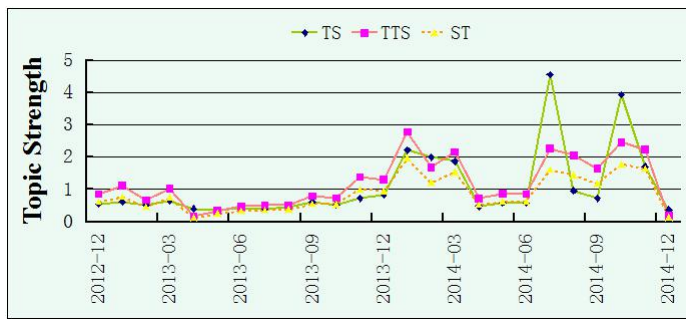


Figure 8. Topic 20 strength

The contents of Topic 20	Work, patrol, center, patrol group, implementing, clean government, anti-corruption, important, internships, general secretary, as the form, focus, deter, center-based, rectification, identify problems, one, around

Table 6. The contents of Topic 20

In face of a multi-time event topic, TS can acquire a better result of topic strength. Although TTS and ST have peaks at the time of occurrence, the peaks are not obvious as the peaks TS obtained. TTS and ST only acquire a time range

of occurrence instead of the accurate time, because the interval of the occurrence time is very short.

4.5 Experiment with Papers Data

Papers data is different with news and comments, so the definition of events is different. In this paper, the development of research papers and the appearances of milestone papers are treated as events. We selected some typical topics to interpret, including a topic about an open source tool ICA & BSS and a topic about the World Telecommunications Standardization Assembly (WTSA).

4.5.1 Development of an Event Topic about Research

Topic 27 is selected to reveal the development of open source tools, and the contents of Topic 27 is shown in Table 7. According to the key words, we can identify that this topic describes an open source tool about blind source separation, and a widely used tool of BSS (Blind Source Separation) is named ICA & BSS. The topic strength is shown in Figure 9.

As clearly shown in Figure 9, the topic strength gained by our approach increases since 2000 and has a peak in 2005. How can we verify that it is correct? Fortunately, the development of ICA & BSS can be found at its official website[3]. According to the information at this official website, the research started at 2000, and in the first time they only studied on ICA (Independent Component Analysis). So the topic strength showed increase since 2000, but not a sharp rise. And in 2005, they started to focus on BBS and released the source of ICA & BSS, and the topic strength obtained by our approach can pick up the important time points in the development of this research.

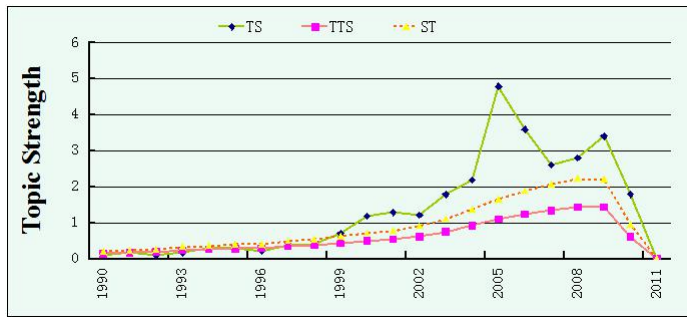


Figure 9. Topic 27 strength

[3] `http://research.ics.aalto.fi/ica/research.shtml`

The contents of Topic 27	open source processing signal based analysis blind separation music multiple single signals audio speech sources sound identification acoustic noisy code

Table 7. The contents of Topic 27

4.5.2 Detection of Event Topic Occurrences

Topic 33 is an event topic, but it is different from Topic 27. According to the contents of Topic 33 shown in Table 8, Topic 33 is not about the development of research and is related to the World Telecommunications Standardization Assembly (WTSA). This assembly involves internet services, QoS (Quality of Service), Multimedia management and so on.

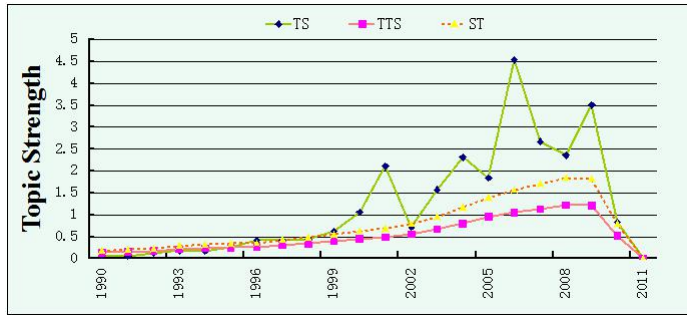


Figure 10. Topic 33 strength

The contents of Topic 33	service services architecture internet QoS mobile quality composition oriented multimedia mobility framework platform providing middleware supporting discovery ipv delivery management

Table 8. The contents of Topic 33

WTSA is held every four years, as we can see from Figure 10, the topic strength coincides the actual situation. The strength got by our approach has a peak near the year when the assembly was held and this result is better than with the other two approaches. What we can infer from this result is that much research literature related to telecommunication is proposed after WTSA was held. However, the time points are not very accurate. And the reason is that the paper data is different from the other data sets, and paper data set is not so real-time as the social network data

sets. So it is difficult to acquire the accurate time points of event occurrences from the paper data and it is good enough to obtain a periodicity.

5 CONCLUSIONS

Topic strength analysis is a significant problem in topic detection and tracking, and it has an important practical value in public opinion monitoring, information forecasting and user personalization, etc. However, the current definition of topic strength is so simple that there is an obvious shortcoming in the calculation of topic strength. Our main contribution in this paper is to improve the definition of topic strength. More specifically, the approach we proposed combines global LDA topics and local LDA topics, and not only takes the number of documents covering the topic into consideration, but also takes the topic quality into account. We used three data sets in real applications: film critic data of "A Chinese Odyssey" in Douban Movies, corruption news data in Sina News and public paper data. The experimental results showed that our approach is effective for detecting the time of event topic occurrences and distinguishing the different types of topics, and it can be used in many fields, such as public opinion monitoring and user personalization.

Acknowledgments

This research has been supported by the National Key Research and Development Program of China under grant No. 2016YFB1000901, the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education, China, under Grant No. IRT17R32, the National Natural Science Foundation of China (NSFC) under Grants Nos. 61503114 and 91746209, and the US National Science Foundation (NSF) under grant IIS-1613950.

REFERENCES

[1] Beykikhoshk, A.—Arandjelović, O.—Venkatesh, S.—Phung, D.: Hierarchical Dirichlet Process for Tracking Complex Topical Structure Evolution and Its Application to Autism Research Literature. Advances in Knowledge Discovery and Data Mining (PAKDD 2015). Lecture Notes in Computer Science, Vol. 9077, 2015, pp. 550–562, doi: 10.1007/978-3-319-18038-0_43.

[2] Blei, D. M.—Ng, A. Y.—Jordan, M. I.—Venkatesh, S.—Phung, D.: Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, 2003, pp. 993–1022.

[3] Blei, D. M.—Lafferty, J. D.: Correlated Topic Models. Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS '05), 2005, pp. 147–154.

[4] Cataldi, M.—Di Caro, L.—Schifanella, C.: Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. Proceedings of the Tenth International Workshop on Multimedia Data Mining (MDMKDD '10), 2010, Art. No. 4.

[5] Cataldi, M.—Aufaure, M.-A.: The 10 Million Follower Fallacy: Audience Size Does Not Prove Domain-Influence on Twitter. Knowledge and Information Systems, Vol. 44, 2015, No. 3, pp. 559–580, doi: 10.1007/s10115-014-0773-8.

[6] Chen, Q.—Gui, Z.: Research on Online Topic Evolutionary Pattern Mining in Text Streams. Journal of Multimedia, Vol. 9, 2014, No. 6, pp. 789–795, doi: 10.4304/jmm.9.6.789-795.

[7] Chenlo, J. M.—Losada, D. E.: Combining Document and Sentence Scores for Blog Topic Retrieval. Proceedings of the Spanish Conference on Information Retrieval, 2010, pp. 29–40.

[8] Deng, L.—Xu, B.—Zhang, L.—Han, I.—Zhou, B.—Zou, P.: Tracking the Evolution of Public Concerns in Social Media. Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service (ICIMCS '13), 2013, pp. 353–357, doi: 10.1145/2499788.2499826.

[9] Dubey, A.—Hefny, A.—Williamson, S.—Xing, E. P.: A Nonparametric Mixture Model for Topic Modeling over Time. Proceedings of the 13th SIAM International Conference on Data Mining, 2013, pp. 530–538, doi: 10.1137/1.9781611972832.59.

[10] Fujino, I.: Refining LDA Results and Ranking Topics in Order of Quantity and Quality with an Application to Twitter Streaming Data. 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014, pp. 209–216, doi: 10.1109/CyberC.2014.45.

[11] Gupta, Ch.—Grossman, R. L.: GenIc: A Single-Pass Generalized Incremental Algorithm for Clustering. Proceedings of the Fourth SIAM International Conference on Data Mining (SDM '04), 2004, pp. 147–153, doi: 10.1137/1.9781611972740.14.

[12] Hu, Ch.—Hu, Y.: Understanding Popularity Evolution Patterns of Hot Topics Based on Time Series Features. Web Technologies and Applications (APWeb 2014). Lecture Notes in Computer Science, Vol. 8710, 2014, pp. 58–68, doi: 10.1007/978-3-319-11119-3_6.

[13] Lau, J. H.—Newman, D.—Baldwin, T.: Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 530–539, doi: 10.3115/v1/E14-1056.

[14] Li, B.–Li, R.-H.—King, I.—Lyu, M. R.—Yu, J. X.: A Topic-Biased User Reputation Model in Rating Systems. Knowledge and Information Systems, Vol. 44, 2015, No. 3, pp. 581–607.

[15] Li, C.—Cheung, W. K.—Ye, Y.—Zhang, X.—Chu, D.—Li, X.: The Author-Topic-Community Model for Author Interest Profiling and Community Discovery. Knowledge and Information Systems, Vol. 44, 2015, No. 2, pp. 359–383.

[16] Madani, O.—Yu, J.: Discovery of Numerous Specific Topics Via Term Co-Occurrence Analysis. Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM'10), 2010, pp. 1841–1844, doi: 10.1145/1871437.1871743.

[17] McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`, 2002.

[18] Liu, S.—Zhou, M. X.—Pan, S.—Qian, W.—Cai, W.—Lian, X.: Interactive, Topic-Based Visual Text Summarization and Analysis. Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09), 2009, pp. 543–552, doi: 10.1145/1645953.1646023.

[19] Saha, A.—Sindhwani, V.: Learning Evolving and Emerging Topics in Social Media: A Dynamic NMF Approach with Temporal Regularization. Proceedings of the Fifth International Conference on Web Search and Data Mining (WSDM'12), 2012, pp. 693–702, doi: 10.1145/2124295.2124376.

[20] Salton, G.—Wong, A.—Yang, C. S.: A Vector Space Model for Automatic Indexing. Communications of the ACM, Vol. 18, 1975, No. 11, pp. 613–620, doi: 10.1145/361219.361220.

[21] Tang, J.—Wang, T.—Lu, Q.—Wang, J.—Li, W.: A Wikipedia Based Semantic Graph Model for Topic Tracking in Blogosphere. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11), Volume Three, 2011, pp. 2337–2342.

[22] Tang, W.—Zhuang, H.—Tang, J.: Learning to Infer Social Ties in Large Networks. Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011). Lecture Notes in Computer Science, Vol. 6913, 2011, pp. 381–397, doi: 10.1007/978-3-642-23808-6_25.

[23] Tang, X.—Yang, C. C.: TUT: A Statistical Model for Detecting Trends, Topics and User Interests in Social Media. Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12), 2012, pp. 972–981, doi: 10.1145/2396761.2396884.

[24] Toda, H.—Kitagawa, H.—Fujimura, K.—Kataoka, R.: Topic Structure Mining Using Temporal Co-Occurrence. Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication (ICUIMC), 2008, pp. 236–241, doi: 10.1145/1352793.1352843.

[25] Wang, X.—Zhai, C.—Roth, D.: Understanding Evolution of Research Themes: A Probabilistic Generative Model for Citations. The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13), 2013, pp. 1115–1123.

[26] Xiao, Z.—Che, F.—Miao, E.—Lu, M.: CorrRank: Correlation Based Ranking Topic Model. Journal of Computational Information Systems, Vol. 10, 2014, No. 7, pp. 3081–3088.

[27] Yang, X.—Sui, A.-N.—Tang, Z.-K.: Topical Crawler Based on Multi-Level Vector Space Model and Optimized Hyperlink Chosen Strategy. Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI), 2010, pp. 430–435.

[28] Yin, Z.—Cao, L.—Han, J.—Zhai, C.—Huang, T.: LPTA: A Probabilistic Model for Latent Periodic Topic Analysis. IEEE 11th International Conference on Data Mining (ICDM), 2011, pp. 904–913.

[29] Zhang, J.—Li, F.: LDA Topic Evolution Based on Global and Local Modeling. Journal of Shanghai Jiaotong University (Science), 2012, p. 1–11.

[30] Zhang, Y.—Chen, W.—Zha, H.—Gu, X.: A Time-Topic Coupled LDA Model for IPTV User Behaviors. IEEE Transactions on Broadcasting, Vol. 61, 2015, No. 1, pp. 56–65.

[31] Zhu, C.—Zhu, H.—Ge, Y.—Chen, E.—Liu, Q.: Tracking the Evolution of Social Emotions: A Time-Aware Topic Modeling Perspective. IEEE International Conference on Data Mining (ICDM), 2014, pp. 697–706, doi: 10.1109/ICDM.2014.121.

[32] Zou, H.—Gong, Z.—Zhang, N.—Li, Q.—Rao, Y.: Adaptive Ensemble with Trust Networks and Collaborative Recommendations. Knowledge and Information Systems, Vol. 44, 2015, No. 3, pp. 663–688.

Jiamiao Wang is a Ph.D. student in the School of Computer Science and Information Engineering at the Hefei University of Technology, China. Her research interests include data mining and social computing. She received her Bachelor of Engineering degree from the School of Computer Science and Information Engineering, Hefei University of Technology.

Lei Li is Associate Professor in the School of Computer Science and Information Engineering at the Hefei University of Technology, China. His research interests include social computing, data mining, and trust computing. He received his Ph.D. in computing from Macquarie University, Australia. He is a member of IEEE.

Xindong Wu is an Alfred and Helen Lamson Endowed Professor in Computer Science in the School of Computing and Informatics at the University of Louisiana at Lafayette, USA, and a Chang Jiang Scholar in the School of Computer Science and Information Engineering at the Hefei University of Technology, China. His research interests include data mining, knowledge-based systems, and web information exploration. He received his Ph.D. in artificial intelligence from the University of Edinburgh. He is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the Editor-in-Chief of Knowledge and Information Systems, and Editor-in-Chief of the Springer book series, Advanced Information and Knowledge Processing (AIKP). He is Fellow of IEEE and the AAAS.