Cogn Comput (2016) 8:577–586
DOI 10.1007/s12559-015-9380-6

Detection and Extraction of Hot Topics on Chinese Microblogs

Liang Yang[1] · Hongfei Lin[1] · Yuan Lin[2] · Shengbo Liu[2]

Received: 12 May 2015 / Accepted: 29 December 2015 / Published online: 17 February 2016

Abstract Peoples' perceptions of reality are conditioned on how others see the world. Unfortunately, with the vast amount of information made available through online media, such as microblog sites, it is impossible for people to absorb all information in a timely manner. Therefore, the detection of hot topics on a microblog platform is becoming increasingly important. The present paper proposes a new hot-topic detection and extraction approach based on language and topic models, which analyzes the differences in emotion distribution language models between adjacent time intervals to detect hot topics. According to the contents and repost degree of microblogs, we estimate the importance of each microblog and generate topic models. Experiments conducted on the Sina Microblog show that the proposed approach can detect and extract hot topics effectively and can thus assist the Sina Microblog platform in managing and monitoring hot topics.

Keywords Topic detection · Emotion distribution · Chinese microblog

✉ Liang Yang
yangliang@mail.dlut.edu.cn

Hongfei Lin
hflin@dlut.edu.cn

Yuan Lin
zhlin@dlut.edu.cn

Shengbo Liu
liushengbo1121@dlut.edu.cn

[1] School of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2] WISE Lab, Dalian University of Technology, Dalian, China

Introduction

With the rapid development of WEB 2.0, more and more people are acquiring knowledge from the Internet and expressing their personal opinions or attitudes about people and events actively. Microblogs update message in 140 words or less and share information using multi-publishing tools in a convenient way, and they have become an important social medium on the Internet and provide a suitable place for the generation and discussion of hot topics. Each microblog user can be an event reporter, using mobile devices to release news. Information on the microblog platform is displayed in a scattered and fragmented way; however, once some microblogs focus on a particular subject, the information flow forms a hot topic. Most microblogs contain elements of personal emotion, and we can thus use them as explicit indicators in identifying hot topics.

Recently, research on microblogs has become a popular domain in the fields of natural language processing and text mining, where the detection of hot topics in microblogs is a major area of interest and problems associated with topic detection thus need to be solved urgently. In this paper, a new approach of detecting hot topics is proposed employing an emotion distribution model and topic model. This is different from previous work on topic detection, which has commonly used seed topic words to guide the clustering process and may suffer from the overfitting problem. According to social cognitive theories, current word-use behaviors relate to past and future behaviors of users [1]. By analyzing the differences in the distributions of emotion words in adjacent time intervals, we can detect hot events in an unsupervised way and avoid the overfitting problem at the same time. Additionally, people tend to repost a microblog that contains a detailed description about a hot topic. In our work, the repost degree as well as the content of the microblog is taken into account in estimating the importance of each microblog and generating topic models.

In summary, our paper makes the following contributions. First, we propose the emotion distribution language model (ELM) for the simulation of the emotion distribution in microblogs, and use it to detect the potential time intervals for hot-topic detection. Second, we analyze the topics in detected time intervals, and estimate the importance of each microblog according to its content and repost degree to extract a hot topic by applying the topic model. Third, we apply our two-step method to Sina Microblog; experiment results show that the approach is effective.

This paper is organized as follows. First section introduces the background to and importance of our research. Second section briefly introduces related work. Third section presents our method of detecting and extracting hot topics. Fourth section describes the experimental process and analyzes experimental results. Final section concludes the paper and proposes further research.

Related Work

Research and Trends in the Microblog Domain

Microblogs have recently received attention from researchers. More and more researchers are investigating public opinion about various topics or news. Kwak et al. [2] analyzed Twitter as a social network and new form of media. Weng et al. [3] proposed the method TwitterRank to find a sensitive topic among influential Twitter users. Marchetti-Bowick et al. [4] used microblogs to forecast political polls by distance supervision. Bollen et al. [5] attempted to predict the stock market by analyzing daily microblogs. Researchers have also investigated the trustworthiness of microblog content [6]. In addition to these studies, topic detection, tracking, and extraction problems in the microblog research domain remain an attractive area of research.

Topic Detection and Tracking

Topic detection and tracking technology [7] is widely applied to topic detection. Information retrieval, text mining, and information extraction focus on a particular topic rather than a wide range of information, while topic detection and tracking emphasize the discovery of new information [8]. Topic detection usually applies clustering algorithms to different events. However, most of the content of microblogs is about the feelings of users, and the proportion of emotion words in microblogs is higher than that in traditional text messages. Additionally, fragmentation, timeliness, and mobility are explicit features of the microblog and increase the difficulty of topic clustering. For example, some topic-related words may not occur in microblogs, and this dramatically affects the cluster results to a large degree. Therefore, traditional topic tracking and detection technology is not suitable for hot-topic detection on the microblog platform. Emotion elements, which are important indicators of hot topics, should be taken into account in topic detection and tracking.

Considering the temporal changes in public opinion, Ku and Liang [9] used language characteristics to capture opinions and presented the changes in the overall sentiment about presidential candidates in an election. Akcora et al. [10] proposed a method of finding public opinion on Twitter by analyzing the emotion centroid (EC) and set space model (SSM). The above works addressed the detection of hot topics and public opinion in different domains and provide a background for the present study. On the basis of these works, we propose an ELM and apply it to the detection of hot topics on the Sina Microblog platform.

Opinion Mining

Sentiment analysis or opinion mining, from coarse to fine grained (e.g., from the document level to concept level), has attracted much attention from researchers [11, 12]. Online opinions on products and services not only affect choices made by consumers but also help merchants improve their services. Currently, text opinion analysis is based on an annotated corpus, using machine learning algorithms to analyze words, sentences, and chapter polarity. Owing to the shortness of a microblog, each microblog is similar to a sentence in an article; therefore, the technology of sentence-level opinion mining provides great support for this paper. Pang and Lee [13, 14] adopted Bayesian and maximum entropy methods to analyze the tendency of movie reviews with manual tagging training data. Pandarachalil et al. [15] used an unsupervised approach to analyze sentiment of Twitter. Additionally, researchers have focused on human emotion [16] and Chinese text characteristics [17, 18], such as the hourglass of emotions and Chinese recognition. All of these studies serve as an important basis for analyzing emotion fluctuation on Chinese Microblogs.

Methodology

The daily collection of microblogs can be seen as a document d, and the whole corpus is the document set D; therefore, $D = \{d_1, d_2,\ldots,d_n\}$. Each microblog is a sentence s of document d, and $d = \{s_1, s_2,\ldots,s_n\}$. Therefore, each term in the microblog can be regarded as a word *w* in the language model. There are three major steps in our paper.

The first step is the construction and recognition of emotion ontology. The recognition of emotion ontology is the basis of an ELM. In this paper, we use DUTIR Emotion Ontology [19] to recognize emotion words in microblogs. The process involves DUTIR Emotion Ontology and words that are specifically used as emotion words on the microblog platform; e.g., "Ink," which is not an emotion word when used generally but sounds like the word "humorous" in Chinese, which has an emotion factor.

The second step is the detection of hot topics using an ELM. When an event occurs, there is an expectation that the emotions of microblog users will fluctuate, and the distributions of emotion words will thus change. By constructing an ELM for each time interval, we compare emotion distributions between adjacent time intervals to detect hot topics.

The third step is the extraction of hot topics from the microblog platform. After detecting the potential hot-topic time interval, we consider the content and repost degree of each microblog and then generate topic keywords using the topic model, and use them as indicators in extracting the hot events.

Construction and Recognition of Emotion Ontology

The paper uses DUTIR Emotion Ontology [19] as the emotion lexicon resource. In emotion classification, there is still no standard for how many emotion classes there are. Presently, emotions can be divided into four, six, eight, ten or twenty categories. In DUTIR Emotions Ontology, emotions are classified into seven categories and 20 subcategories, which can be used in coarse or fine emotion computing.

To recognize an emotion word, mutual information is calculated between the word and ontology in DUTIR Emotion Ontology as

$$MI(w, S_{ui}) = \log \frac{P(w, S_{ui})}{P(w)P(S_{ui})} \tag{1}$$

where S_{ui} is the *i*th ontology in emotion *u*, $p(w)$ is the probability of occurrence of word *w*, and $p(S_{ui})$ is the probability of occurrence of the *i*th ontology in emotion *u*.

We also consider other rules, such as the co-occurrence rule, the part-of-speech rule, and the context rule, to expand the emotion ontology. A machine learning method is also adopted for automatic expansion of emotion ontology. In this paper, conditional random fields [20] are implemented for automatic collection according to

$$P_\theta(y|x) = \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} u_k g_k(v, y|_v, x) \right) \tag{2}$$

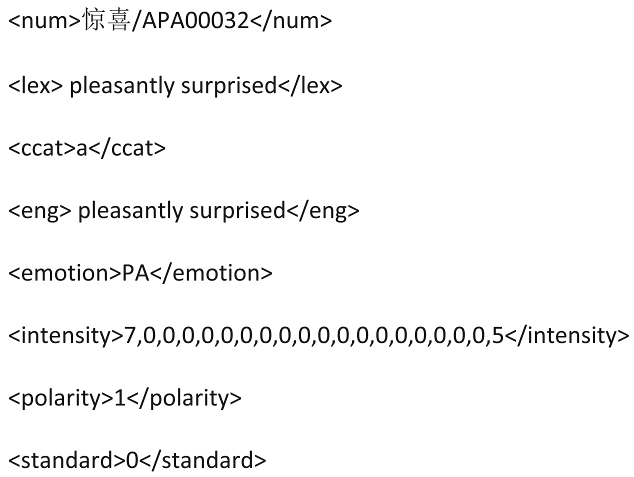
where graph $G = (V, E)$, with *V* being a vertex and *E* an edge, *x* is a data sequence, *y* is a label sequence, and $y|_s$ is the set of components of *y* associated with the vertices in sub-graph *S*. The features f_k and g_k are given and fixed, and λ_k is a weight parameter.

DUTIR Emotion Ontology is defined by 3-tuples as

$$\text{Lexicon} = (B, R, E) \tag{3}$$

where *B* represents basic lexical information including the serial number, entry, English meaning, part of speech, editor, and version; *R* represents synonymous relations; and *E* represents emotion information, which is the most important part of 3-tuples.

For example, "pleasantly surprised" is describes as follows.



```
<num>惊喜/APA00032</num>

<lex> pleasantly surprised</lex>

<ccat>a</ccat>

<eng> pleasantly surprised</eng>

<emotion>PA</emotion>

<intensity>7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5</intensity>

<polarity>1</polarity>

<standard>0</standard>
```

The value "PA" in the <emotion> tag is the symbol of the "happy" emotion, and the <intensity> tag includes the emotion and intensity of the word. The value in the "intensity" field is a vector, and the vector component ranges from 0 to 9 (where zero indicates no such feelings), which delegates a specific emotion. As an example, the words "pleasantly surprised" contain both "happiness" and "surprise" feelings. The intensity of emotions is given on a five-point scale having values of 1, 3, 5, 7 and 9. The "happiness" intensity is 7 and the "surprise" intensity is 5. The numbers 0, 1, 2 and 3 in the "polarity" field represent neutral, positive, negative and both positive and negative attributes, respectively. The framework includes a description of both the static and dynamic attributes, and emotion information is shown in quantitative and qualitative terms, which provides useful information for emotion analysis (Table 1).

Table 1 DUTIR emotion ontology

No.	Emotion	No.	Emotion
1	Happy	11	Yearning
2	Reassuring	12	Flustered
3	Honorific	13	Terrified
4	Praiseful	14	Shy
5	Trustful	15	Agonizing
6	Fond	16	Hateful
7	Angry	17	Disagreeable
8	Sorrowful	18	Jealous
9	Disappointed	19	Dubitable
10	Compunctious	20	Surprised

DUTIR Emotion Ontology is based on existing dictionaries such as *Dictionary of Chinese Praise and Blame Words* [21], *Dictionary of Chinese Adjective* [22], *Dictionary of Chinese Idiomatic Phrases* [23], *Dictionary of Chinese Idiom* [24], *New Century Dictionary of Chinese New Words* [25] and *Chinese Classified Dictionary* [26]. In addition, semantic network resources, including HowNet [27] and WordNet [28], are used. Network emotion words are also contained to improve precision. Therefore, DUTIR Emotion Ontology has a wide range of application in the emotion analysis of microblogs and similar network platforms, such as blogs and bulletin board systems. There are presently 27,243 entries in DUTIR Emotion Ontology, which is still being updated. We plan to introduce more semantic resources to enrich the ontology.

Hot-Topic Detection Based on an EML

A statistical language model [29], based on statistical methods for natural language processing, can be estimated using a multinomial distribution. The model provides a statistical way of scoring and ranking documents. The use of a statistical language model is a two-stage method; a language model is generated for each document in the first stage, and documents are ranked according to the query relevance score in the second stage. The relevance score is calculated as

$$P(Q|D) = \prod_{w \in v} P(w|D)^{q_w} \tag{4}$$

where Q is the query, D is the document, V is the word set, and q_w is the number of instances of the word w.

We then used relative entropy (Kullback–Leibler divergence) to measure the similarities between the two models. Their distance is a reflection of the difference between the learning model and real model. If the two models are the same, then the relative entropy is zero. Higher relative entropy corresponds to a greater difference between two models. The relative entropy calculation formula is defined as

$$D_{KL}((P(w|Q)),(P(w|C))) = \sum_{w \in V} P(w|Q) \log \frac{P(w|Q)}{P(w|C)} \tag{5}$$

where $p(w|Q)$ is the probability that the word w occurs in Query Q and $p(w|C)$ is the probability that the word w occurs in the experiment dataset C.

In the domain of information retrieval, according to "bag of words" theory, each word is independent and the distribution of words may be estimated from distributions. Emotion words on a microblog platform should also belong to a distribution. According to this distribution and considering the fragmented nature of microblogs, the present paper proposes an ELM approach. This approach analyzes the differences in ELMs between adjacent time intervals to detect a hot topic. We define the ELM of time period T_n as

$$P(t|DT_n) = \prod_{t \in E} P(t|C)^{q_t} \tag{6}$$

where E is the set of emotion ontology, DT_n denotes the microblogs in time period T_n, $p(t|C)$ is the probability that emotion ontology t occurs in experiment dataset C, and q_t is the number of occurrences of emotion ontology t in time period T_n.

Owing to the short and brief nature of microblogs, emotion words are sparse lexicon to some extent. We therefore apply Dirichlet smoothing to the experiment dataset [29]. The Dirichlet smoothing formula is defined as

$$P_\mu(w|d) = \frac{c(w;d) + \mu p(w|C)}{\sum_w c(w;d) + \mu} \tag{7}$$

where $P_\mu(w|d)$ is the probability of occurrence of word w after smoothing, $c(w;d)$ is the number of occurrences of word w in document d, μ is a smoothing parameter, and $p(w|C)$ is the probability of occurrence of word w in the experiment corpus C.

Relative entropy is an important evaluation metric of a statistical language model. By calculating ELMs for adjacent time intervals T_n and T_{n-1}, relative entropy can be used to measure the differences, where higher relative entropy corresponds to greater differences between adjacent time intervals, which provides a basis for detecting a potential hot topic. The equation for relative entropy is

$$D_{KL}(p(t|DT_{n-1}), p(t|DT_n)) = \sum_{t \in E} p(t|DT_{n-1}) \log \frac{p(t|DT_{n-1})}{p(t|DT_n)} \tag{8}$$

where E is the microblog Emotion Ontology, w is one emotion ontology in E, DT_n denotes the microblog dataset in time period T_n, and $P(t|DT_n)$ is the probability that emotion ontology w occurs in the microblog dataset in time period T_n.

According to observations in [10], when a hot event occurs, individuals tend to post more microblogs, and the microblogs include more emotion words. Thus, emotion expression patterns of microblogs are different from the emotion expression patterns that appeared in the previous period, but have a higher similarity with the emotion expression patterns of the following period. According to the above analysis, if the criteria are satisfied, then a potential hot event occurs in time period T_n.

$$D_{KL}(p(t|DT_{n-1}), p(t|DT_n)) > D_{KL}(p(t|DT_n), p(t|DT_{n+1})) \tag{9}$$

$$D_{KL}(p(t|DT_{n-1}), p(t|DT_n)) > D_{KL}(p(t|DT_{n-1}), p(t|DT_{n-2})) \tag{10}$$

After defining the equation, we calculate the relative entropy between adjacent time intervals to detect a potential hot topic.

Extracting Hot Topics from Microblogs

It has been observed that topic clustering can be helpful in the quick retrieval of desired information. Traditional text mining techniques have no special considerations for the short and sparse characteristics of microblog data, and it is therefore impossible to directly apply the traditional topic clustering technique to microblog posts. The topic detection process commonly uses the vector space model. However, for a short and sparse microblog, the vector space model (using words or terms as characters) cannot be used for the accurate calculation of the similarities between texts. To reduce data scarcity, we apply the latent Dirichlet allocation (LDA) model [30] to the data modeling and extract the hidden microblog topics. The high-dimensional sparse text vector is then reduced to a low-dimensional hidden-topic space.

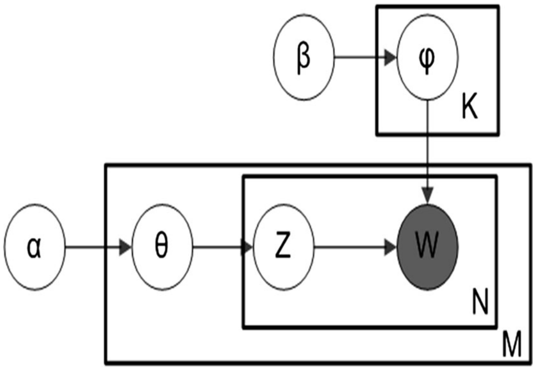


Fig. 1 Representation of latent Dirichlet allocation

The LDA model is a probabilistic graphical model that has three levels as shown in Fig. 1. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

LDA assumes the following generative process for each document w in a corpus D.

1. Choose $\theta_i \sim Dir(\alpha)$, where $Dir(\alpha)$ is the Dirichlet distribution for parameter α and $i \in \{1,\ldots, M\}$.
2. Choose $\varphi_k \sim Dir(\beta)$, where $k \in \{1,\ldots, K\}$.
3. For each of the words w_{ij}, where k $\in \{1,\ldots, N_i\}$,

   (a) choose a topic $z_{ij} \sim$ Multinomial (θ_i),
   (b) choose a word $w_{ij} \sim$ Multinomial $(\varphi_{z_{ij}})$.

The Markov chain Monte Carlo method [31] is a general method of obtaining samples from a complex distribution, with Gibbs sampling being a special case. After deduction, the final Gibbs sampling equation is

$$p(z_i = k|z_{\neg i}, w) \infty \frac{n^{(i)}_{k,\neg i} + \beta_i}{\left[\sum_{v=1}^{V} n_k^{(v)} + \beta_v\right] - 1} \cdot \frac{n^{(i)}_{m,\neg i} + \alpha_k}{\left[\sum_{z=1}^{K} n_m^{(z)} + \alpha_z\right] - 1} \tag{11}$$

where $n_k^{(v)}$ denotes the number of occurrences of word w_v in topic k and $n_m^{(z)}$ denotes the number of occurrences of topic z in document m.

Once we have the label of topic z for each word, we can estimate the values of the other latent variables according to

$$\varphi_{k,i} = \frac{n_k^{(i)} + \beta_i}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v}, \quad \theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{z=1}^{K} n_m^{(z)} + \alpha_z} \tag{12}$$

where $\varphi_{k,i}$ is the probability of word w_i being in topic k, and $\theta_{m,k}$ is the probability of topic z_k being in document m.

Based on the above deduction, we can produce the following probability and its weight score s for the word w in microblog m. In our experiment, the repost degree R is an important factor used to estimate the importance of the word w:

$$p(w_j|m_i) = p(w_j|\theta_{m_i}, \beta) = \sum_{k=1}^{K} R_{m_i} p(w_j|z_n, \beta) p(z_n|\theta_{m_i}) \tag{13}$$

$$S(w_j|d_{T_n}) = \sum_{n=1}^{N} s(w_j|m_i) = \sum_{n=1}^{N} R_{m_i} p(w_j|m_i) \tag{14}$$

In this part, we treat the microblogs posted in 1 h as a "document" d, and each microblog in that hour is a "sentence" in the "document" d. N is the number of microblogs in time period T_n. Because we have detected the potential hot topic in some time periods using the ELM, parameters α and β are related to the dataset, and they can be sampled once in the process of generating a dataset. The variables θ_{mi} are document-related variables, sampled once per "document" d, and the variables z_n and w_j are word-related variables that are sampled once for each word in each "document," which consists of microblogs published in a specific hour. After taking into account the repost degree, we can re-weigh the score of each word in microblogs, and the results of the LDA model guide us in extracting the topic.

Experiments and Analysis

Datasets and Work Flow

We create a dataset by collecting microblogs from Sina Microblog during the period from June 7, 2010 to June 13, 2010. A total of 52,500 microblogs are collected and stored in a uniform format. A microblog is defined as below.

```
<blog>

<name>逆风蝴蝶/Unwind Buffterfly</name>

<text>新年快乐！/Happy New Year! </text>

<rt>@大钟：元旦快乐！/@Big Clock: Happy New Year's Day! </rt>

<time>2010-1-1 00:00</time>

</blog>
```

Here *<name>* is the user name, *<text>* denotes the content that the user posts, *<rt>* is the content that the user reposts, and *<time>* is the timestamp of a microblog. As users sometimes post original messages without repost, the *<rt>* segment can be null.

To extract the ELM, we analyze the emotion words in a day's corpus to detect a hot topic on the microblog platform. A flowchart of our work is shown in Fig. 2.

Results of Hot-Topic Detection

The length of the time intervals is another important factor in our work. If the time intervals are shorter than 1 h, too few meaningful microblogs are included, leading to biased results. In contrast, more than one topic may occur in a time interval, which does not suit the present problem domain. Therefore, we set the time interval as 1 h. This is the most acceptable time interval for providing meaningful data, and furthermore, it allows us to detect hot topics with fine granularity.

After we define the time interval, we use the Sina Microblog data crawled through from June 7, 2010 to June 13, 2010 as the experiment dataset. We choose this time period because there were a number of hot topics in that week. Information relating to the dataset is presented in Table 2.

According to the hot topics provided by the microblog platform, we compare our method with that used in Ref. [10]. Results are presented in Fig. 3, and a detailed description of the topic detection results is given in Table 3. In the table, a number outside parentheses is the number of time periods for which detection using the method was correct, while a number inside parentheses is the number of time periods for which detection using the method was incorrect.

Figure 3 shows that all four methods can be used to find hot topics in our experiment dataset effectively and that the ELM has the highest precision and F1 score, demonstrating the effectiveness of our proposed approach.

The EC method has the highest recall rate but the lowest accuracy. Because the EC method only focuses on emotion words, if fewer emotion words are contained in the adjacent time intervals for different topics, then the EC is misleading in result detection. For the SSM, emotion words are only a part of the whole word set, and the Jaccard similarity may not depend on emotion words. The recall rate therefore drops obviously, but the precision is higher than that of the EC method for all kinds of words.

The EC&SSM method considers the advantages and disadvantages of the emotion centroid and set space model. It uses the Jaccard similarity to validate the hot topic detected by cosine similarity, which guarantees the basic recall rates and increases the accuracy. Figure 3 and Table 3 show that EC&SSM has higher precision and a higher F1 score compared with the EC and SSM.

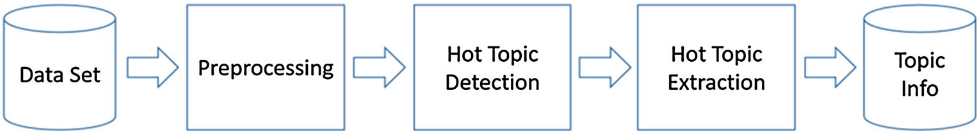


Fig. 2 Flowchart of our work

Table 2 Experimental dataset

Date	Occurrence time period of hot topics
June 7, 2010	9, 10, 15, 18
June 8, 2010	10, 15, 18
June 9, 2010	9, 10, 12
June 10, 2010	10, 11
June 11, 2010	9, 13, 14, 20
June 12, 2010	9, 20, 21
June 13, 2010	9, 13, 20, 21

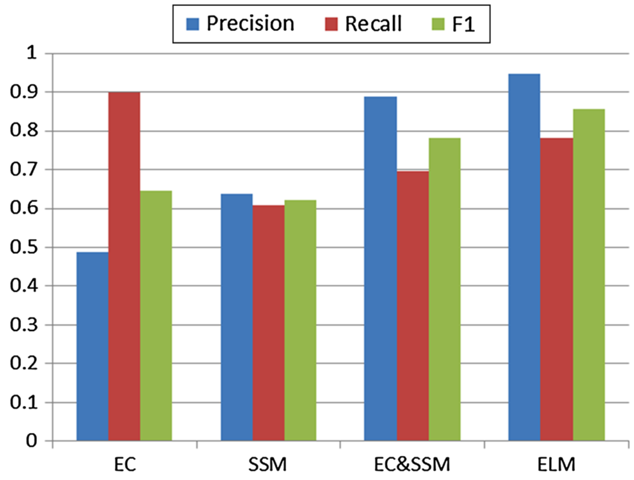


Fig. 3 Experiment results of different methods

The results of the ELM show that the ELM has not only higher precision but also higher recall and a higher F1 score when compared with EC&SSM.

The results reveal the importance of the relationship between language expression and the language model. Language expression, which is described by the language model, is more suitable for natural language processing, and the ELM therefore approximately reflects the facts of the experiment dataset and it has the highest precision and F1 score. The results show that the method that we proposed can be used to detect hot topics throughout the microblog dataset.

According to conditions (9) and (10), we find the hot-topic time intervals 9, 13, and 20 provided by the platform were detected by the ELM in Fig. 4. The ELM also has disadvantages, e.g., it does not take the relationship of emotion words into account, and when there are fewer emotion words occurrences, such as narrative microblogs, this could affect the performance to some degree.

To further validate our results, we perform statistical analysis. We use the D-values of the emotion word ratio for adjacent time intervals to confirm our conclusion; i.e., emotion words play an important role in topic detection and more microblogs contain emotion words when a hot event occurs. In Fig. 5, for example, we find that there is an emotion burst between time intervals 9 and 10, and this burst is also detected using our proposed method.

Results of Hot-Topic Extraction

After detecting the time periods using the ELM, the next step is to extract the hot topic using the topic model. The LDA model generates the probabilities of topics in the "document," which consists of microblogs published in 1 h, and provides a topic space for "documents." The results show that the method provides good results.

The topic number is an important parameter of the LDA model. If we set the topic number small, the model will not fully generate the "document," whereas if it is too large, the hidden topic will be separated, affecting the result. According to our analysis, we set the topic number parameter as three and obtain a better experimental result. The results of topic extraction are given in Table 4.

In this part of the experiment, we tend to detect and extract "new" topics that have not yet been detected. If the topic has been detected previously, we ignore the topic and select the next topic according to the probability ranking result. For example, one topic, which relates with the South Africa World Cup 2010, may continue all day, but if the

Table 3 Topic detection results

Date	Event nums	EC	SSM	EC&SSM	ELM
June 7, 2010	4	3 (3)	2 (1)	3	3
June 8, 2010	3	2 (4)	2 (1)	2	2
June 9, 2010	3	3 (2)	2 (1)	2 (1)	3 (1)
June 10, 2010	2	2 (3)	1 (2)	1 (1)	1
June 11, 2010	4	3 (3)	1 (1)	3	3
June 12, 2010	3	3 (3)	3 (1)	3	3
June 13, 2010	4	4 (3)	3 (1)	2	3
Total event number	23	20 (21)	14 (8)	16 (2)	18 (1)

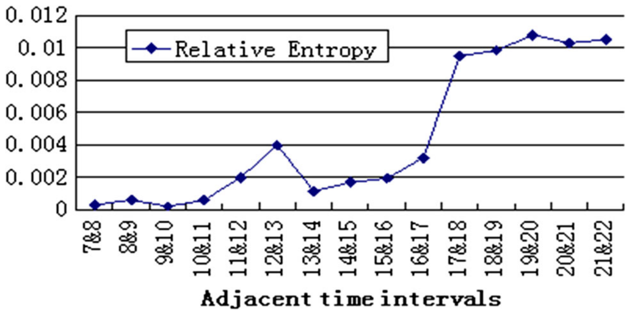


Fig. 4 Relative entropy of the ELM for adjacent time intervals on June 13, 2010

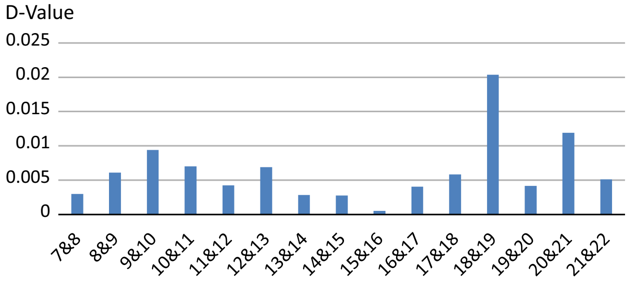


Fig. 5 Emotion word ratio *D*-values for adjacent time intervals on June 13, 2010

match has been detected, we focus on matches other than the detected one.

Two factors are taken into account. One is the content of the microblog, and the other is the repost degree. The former is the basis of detecting the topics, and the latter is used to re-weight the words related to the topics. People tend to repost the microblog with a detailed description about the hot topic, and therefore, the repost degree is used to estimate the importance of each microblog and its content. The hot topics in the extracted results for June 13, 2010 are listed in Table 5.

From the results in Table 5, we conclude that the ELM can detect a previously existing topic during the time period. It is clear that three topics are explicitly extracted, namely "the mistake made by England Goalkeeper Green in the match between England and the United States," "Duan Wu Festival," and "Fake Caocao Tomb" based on the topic words and their probability values.

The results are reasonable and encouraging. The model benefits from the potential advantages of LDA [32, 33] as a generative model for documents. We also considered the repost degree, which is a significant factor in the microblog research domain and indicates the microblog users' behaviors in emphasizing specific topics. Using the repost

Table 4 Topic extraction results

Topic	Time interval	Description words
NBA final game	9 A.M., June 7	凯尔特人/ Celtic, 决赛/Final
Entrance exam—Chinese	10 A.M., June 7	考试/Exam, 语文/Chinese
Composition topic	3 P.M., June 7	作文/ Composition, 题目/topic
–	6 P.M., June 7	–
Iphone 4 release	10 A.M., June 8	Iphone, 发布/Release
–	3 P.M., June 8	–
Ending of entrance examination	6 P.M., June 8	高考/Entrance Exam, 结束/End
South Africa World Cup	9 A.M., June 9	南非/South Africa, 世界杯/World Cup
Mourinho in Real Madrid	10 A.M., June 9	皇马/ Real Madrid, 穆里尼奥/
intention for university	12 A.M., June 9	志愿/ Intention, 填报/Fill
stock market rebounce	10 A.M., June 10	股市/Stock, 上涨/Rise
–	11 A.M., June 10	–
The opening of world cup tonight	9 A.M., June 11	开幕/Opening, 今晚/Tonight
Crash of Mandra's granddaughter	1 P.M., June 11	车祸/Crash, 悲剧/Tragedy
–	2 P.M., June 11	–
South Africa versus Mexico	8 P.M., June 11	南非/S. Africa, 墨西哥/Mex
Draw of last night matches	9 A.M., June 12	平局/Draw, 昨晚/Last Night
Korea versus Greece	8 P.M., June 12	韩国/Korea, 希腊/Greece
Argentina versus Nigeria	9 P.M., June 12	阿根廷/Arg, 尼日利亚/Nig
Mistake made by Green	9 A.M., June 13	黄油手/Butter Hand, 格林/Green
–	1 P.M., June 13	–
DuanWu festival	8 P.M., June 13	端午/DuanWu, 放假/Holiday
Fake Caocao Tomb	9 P.M., June 13	曹操/Caocao, 遗体/Body

Table 5 Topic words and their probabilities for June 13

9 AM June 13th	Probability	8 PM June 13th	Probability	9 AM June 13th	Probability
英格兰/England	0.1617	端午/DuanWu	0.0295	曹操/Caocao	0.0210
美国/ the United States	0.0459	端午节/DuanWu Festival	0.0277	脸部/Face	0.0105
格林/Green	0.0396	粽子/Zongzi	0.0206	遗体/Remains	0.0105
门将/Keeper	0.03038	假期/Holiday	0.0152	出现/Appear	0.0088
守门员/Goalkeeper	0.0092	快乐/Happiness	0.0125	现象/Phenomenon	0.0079
黄油/Butter	0.0092	今晚/Tonight	0.0116	非正常/Abnormal	0.0070
失误/Mistake	0.0092	放假/Recess	0.0098	文化/Culture	0.0061
脱手/Slip out of hand	0.0064	开心/Glad	0.0089	盗墓/Grave Robbery	0.0061
平局/Draw	0.0064	回家/Back home	0.0072	国家/Country	0.0061
低级/Low level	0.0057	节日/Festival	0.0063	出土/Unearth	0.0035

degree as a factor of the word probability score, we can re-weigh the score of each word in each microblog and use the re-weighted score to extract the topic precisely.

Conclusions and Future Work

In this paper, we proposed an approach of detecting and extracting hot topics using an ELM and topic model. The approach was used to analyze the differences in ELMs between adjacent time intervals in detecting hot topics.

According to a microblog content and its repost degree, we estimated the importance of each microblog and generated topic models and then used the topic keywords provided by the topic model to extract the hot topics. Experimental results show that the approach effectively detects and extracts hot topics on a microblog platform. The findings can be used to help Sina Microblog manage and monitor hot topics daily.

In our future work, we will continue to enrich DUTIR Emotion Ontology so that it can be widely used in Chinese text emotion analysis and to integrate social cognitive theory and thus quantify the emotion intensity. Meanwhile, we will attempt to improve the performance of the ELM and hot topic extraction with more semantic and statistic features on the microblog platform.

Funding This work is partially supported by grant from the Natural Science Foundation of China (Nos. 61402075, 61572102, 61277370), Natural Science Foundation of Liaoning Province, China (Nos. 201202031, 2014020003), and State Education Ministry and The Research Fund for the Doctoral Program of Higher Education (No. 20090041110002).

Compliance with Ethical Standards

Conflict of Interest Liang Yang, Hongfei Lin, Yuan Lin, and Shengbo Liu declare that they have no conflict of interest.

Informed Consent All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Additional informed consent was obtained from all patients for which identifying information is included in this article.

Human and Animal Rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Rochat P. Early social cognition: understanding others in the first months of life. London: Psychology Press; 2014. p. 2014.
2. Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media? In: 27th World Wide Web. In: Proceedings of the 19th international conference on World Wide Web. 2010. p. 591–600.
3. Weng J, Lim E, Jiang J, He Q. TwitterRank: finding topic sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. 2010. p. 261–70.
4. Marchetti-Bowick M, Chambers N. Learning for microblogs with distant supervision: political forecasting with Twitter. In: Proceedings of the 13th conference of the European Chapter of the Association for Computational Linguistics. 2012. p. 603–12.
5. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci. 2011;2(1):1–8.
6. Hsu C, Liu C, Lee Y. Effect of commitment and trust towards micro-blogs on consumer behavioral intention: a relationship marketing perspective. Int J Electron Bus Manag. 2010;8(4): 292–303.
7. Yu H, Zhang Y, Liu T, Li S. Topic detection and tracking review. J Chin Inf Process. 2007;21(6):71–87.
8. Li B, Yu S. Research on topic detection and tracking. Comput Eng Appl. 2003;17(1):133–6.
9. Ku L, Liang Y. Opinion extraction, summarization and tracking in news and blog corpora. In: AAAI spring symposium: computational approaches to analyzing weblogs. 2006. p. 100–7.
10. Akcora C, Bayir M, Demirbas M, Ferhaosmanoglu H. Identifying breakpoints in public opinion. In: Proceedings of the first workshop on social media analytics. 2010. p. 62–6.
11. Cambria E, Hussain A. Sentic computing: a common-sense-based framework for concept-level sentiment analysis. Cham: Springer; 2015.
12. Cambria E, Schuller B, Xia Y, Havasi C. Knowledge-based approaches to concept-level sentiment analysis. IEEE Intell Syst. 2013;28(2):12–4.

13. Pang B, Lee L. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing. 2002. p. 79–86.
14. Pang B, Lee L. A sentimental education: sentiment analysis using subjective summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on association for computational linguistics. 2004. p. 271–8.
15. Pandarachalil R, Sendhilkumar S, Mahalakshmi GS. Twitter sentiment analysis for large-scale data: an unsupervised approach. Cognit Comput. 2015;7:254–62.
16. Cambria E, Livingstone A, Hussain A. The hourglass of emotions. Cognitive Behavioral Systems (LNCS 7403). 2012, p. 144–57.
17. Wang QF, Cambria E, Liu CL, Hussain A. Common sense knowledge for handwritten Chinese recognition. Cognit Comput. 2013;5(2):234–42.
18. Chen Y, Zhou Q, Luo W, Du J. Classification of Chinese texts based on recognition of semantic topics. Cognit Comput. 2013;1–11.
19. Xu L, Lin H, Pan Y, Ren H, Chen J. Constructing the affective lexicon ontology. J China Soc Sci Tech Inf. 2008;27(2):180–5.
20. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: 18th International conference on machine learning. 2001. p. 282–9.
21. Wang G. A dictionary of Chinese praise and blame words. Beijing: Encyclopedia of China Publishing House; 2001.
22. Zheng H, Meng Q. A dictionary of Chinese adjective. Beijing: The Commercial Press; 2004.
23. Cheng Z. A dictionary of Chinese idiomatic phrases. Beijing: Encyclopedia of China Publishing House; 2003.
24. Yang X. A dictionary of Chinese idiom. Chengdu: SiChuan Lexicographical Publishing House; 2005.
25. Wang J. New century dictionary of Chinese new words. Shanghai: Great Chinese dictionary Press; 2006.
26. Dong D. A Chinese classified dictionary. Shanghai: Great Chinese dictionary Press; 1998.
27. HowNet. http://www.keenage.com/.
28. WordNet. http://wordnet.princeton.edu/.
29. Zhai C, Lafferty J. A study of smoothing methods for language models applied to information retrieval. Trans Inf Syst. 2004;22(2):180–216.
30. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. J Mach Learn Res. 2003;2003(3):993–1022.
31. Stuart G, Donald G. Stochastic relaxation Gibbs distributions and the Bayesian restoration of images. Pattern Anal Mach Intell IEEE Trans. 1984;6:721–41.
32. Lavrenko V, Croft W. Relevance-based language model. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. 2001. p. 120–7.
33. Liu X, Croft W. Cluster-based retrieval using language models. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. 2004. p. 186–93.