ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20190413112011
Original scientific paper

Hot Topic Discovery in Online Community using Topic Labels and Hot Features

Minjuan ZHONG

Abstract: With huge volumes of information on Internet, how to extract user-concerned hot topics quickly and effectively has become a fundamental task for information processing on Internet. Generally, hot topic detection includes two tasks, the first one is topic discovery and the other is its hotness evaluation. In this paper, we propose a hot topic detection method. For topic discovery, topics are identified by clustering based on extracted topic labels. For hotness evaluation, the proposed model has fully considered the internal and external dual features and combined them together. The experimental results over TianYa BBS demonstrate the efficiency of the proposed method: compared with topic discovery based on latent semantic indexing, the improved vector space model based on topic labels gets better results and the identified topics are more accurate. Moreover, the proposed hotness features could reflect the popularity of a topic, and hence have obtained better hot topic results finally.

Keywords: external features; improved vector space model; internal features; topic hotness; topic label

1 INTRODUCTION

As an important platform for message sharing and view dissemination, there are unprecedentedly large and complex data on the online community. Compared with traditional media, the community data has characteristics of mess, disorder and dispersion. Therefore, how to organize and extract meaningful public opinion from the massive community data, without being submerged in the knowledge ocean, is an open challenge.

To address the challenge, many public opinion analysis models have been proposed to extract important opinions. For example, the TDT (Topic Detection and Tracking) method which was originally presented by the US Defense Advanced Research Project (DARPA), combined both Natural Language Processing (NLP) and Data Mining (DM) to automatically recognize and continuously track important opinions from text streams. Under the initiative and support of DARPA, many research institutions and famous universities are actively involved in TDT evaluation every year, such as Carnegie Mellon University, IBM Watson Research Center, Massachusetts Institute of Technology. With the rapid development of social network in recent years, researchers get to focus on topic detection towards social medias, such as BBS, Blogs and Twitter.

General, a hot topic is defined as a popular or frequent topic in a given time interval. The key technology of it consists of two major tasks, one is topic detection, and the other is hotness evaluation. Topic detection is essentially equivalent to unsupervised clustering due to lack of a priori knowledge of topics. Therefore, how to express the text, i.e. the model expression, is a fundamental and critical problem in topic detection. The models are usually Vector Space Model (VSM) and Probability Model (PM). The Probability Model includes Language Model (LM) and Probabilistic Latent Semantic Analysis (PLSA). James Allan and Schiiltz [1] applied VSM to describe feature space of the news report, in which TF-IDF word frequency distribution is used for weight estimation, and cosine formula for similarity calculation. Blei [2] presented Latent Dirichlet Allocation (LDA) model to analyse the probability of occurrence of topics in the text. Y. Sun [3] combined term mutual information with probabilistic topic model for hot topic detection. Wang [4] proposed k-means clustering for large scale multidimensional data. Ye [5] proposed a topic discovery algorithm for pictures of microblog, video and other multimedia comment information. Yu [6] proposed a novel two-times single-pass clustering algorithm for "burst" hot topic detection, in which life circle model of extracted topics with aging theory is built. In Guojing's work [7], frequent pattern stream mining model was applied to detect hot topics from twitter streams. Numerous studies have shown that traditional VSM-based topic detection methods are widely used and simple to implement, but there are shortcomings, such as sparse data, semantic information loss. PLSA, LDA and other PM require a large number of corpuses, and the implementation method is more complicated.

For the hotness evaluation of topic, many scholars have presented many heat-related topic characteristics from different aspects. Feng [8] believed that topic popularity of micro blog evaluated four factors, the number of micro blog, number of forwarding, number of comments, and number of praise. Liu [9] pointed out that the heat of micro-blog texts and topic keywords got mutual reinforcement. They introduced the concept of topic keywords combination support confidence to represent hot topic. In literature [10], hot topics were determined by measuring the users participation, opinion sharing and user forgetfulness. Wangxiao Dong [11] used the value of keywords weight to evaluate the topic hotness. Chen Kuan-yu et al. [12] introduced two critical properties of a hot term, "pervasiveness" and "topicality". Pervasiveness refers to the frequency with which a term appears in a set of documents, while topicality is the variation in the frequency of usage of a term over time. Lan You [13] did a similar research to find hot topics on BBS. In his method, a BPNN (Back-Propagation Neural Network) based classification algorithm was used to judge the topic hotness according to its popularity, its quality as well as its message distribution over time. Ma Hui-fang [14] proposed a dynamic hot topic extraction model for news report data. The model introduced timing window based on the extensive and burst characteristics of the hot topic, and extracted hot term combining the weight calculation of TF-IDF from them. This method could find hot topic for a period of time, but has the limitation of tracking the topic hotness in real time. Zhengdonghui [15] applied aging theory to model the hotness of topics and considered a new

topic as a life form with stages of birth, growth, decay and death. Wang Can-hui et al. [16] ranked the topics in terms of both media focus and user attention. They used the concept of energy function to calculate and update the energy of topics in every time slot.

In this paper, we are interested in extracting hot topics from online community documents, and propose a hot topic detection model. The basic approach is to identify the topic based on their content, and select the top topics, ranked according to the internal and external dual features, as hot topics. The contributions of this paper are described as follows:

(1) For topic discovery, the documents must be exactly associated with the corresponding topic. However, due to the semantic fluctuations, this identification does not work particularly well. Here, we firstly focus on topic labels extraction. Different from the existing method, the proposed model fully considers the contextual characteristics of the term and combines it with their weight. The experimental results show that the obtained labels can express the topic of the document well. Secondly, k-medoid clustering algorithm is used to discover topics. Therefore, it is important how to measure the similarity between two documents. An improved VSM on basis of extracted topic labels is presented. The experiment data from TianYa community show that the proposed model can identify topics effectively, and hence get better topic results.

(2) For hotness evaluation, unlike other hotness evaluation methods, the proposed model has fully considered the internal and external dual features which can reflect their popularity. The inner features include number of clicks, reply, participating users and topic post. The external features refer to the duration time of topic, post source, number of released post and topic quality. We have conducted experiments and obtained satisfactory results.

As concerns the remainder of this paper, in Section 2 we describe the topic identification based on label extraction. Section 3 presents document model, in which the similarity measurement is pointed out. Section 4 proposes the hotness evaluation model, in which influence factors of topic hotness are analysed, followed by the evaluation model with multiple features. In Section 5, we present and discuss experimental results on TianYa community test collections. Finally, conclusions are provided in the last section.

2 TOPIC DISCOVERY

Topic detection is essentially equivalent to unsupervised clustering, in which each topic is treated as one cluster. Therefore, it is important how to measure the similarity between two documents. In traditional Vector Space Model, classic cosine formula is used to compute the similarity. However, the model has the following drawbacks: firstly, the number of terms in the document space is very large, sometimes reaching super dimension. On the other hand, it is weak ability to express the semantic of the document for single term. So, in this paper, we propose the improved Vector Space Model based on topic label extraction, in which labels are introduced as the unit of document expression instead of the single term.

Generally, a document contains at least one topic, and the author will discuss, describe, or report around them. Therefore, some vocabulary related to it will appear in the document. These words are called topic label, which is the most representative information in the document. Topic label can be expressed as a term or phrase. A term is usually a more important word in a document, while a phrase is a combination of several terms. Comparing the characteristics of the two, from the semantic point of view, although the phrase can express the main idea of the document more clearly than the individual term, the premise must be the correct form. That is to say, if the obtained phrase is not a correct phrase, the noise it contains is even larger, especially in the Chinese context; the lower accuracy rate will cause the subsequent processing performance to decrease. Therefore, we intend to use the term to express the topic label and select central terms from each post document.

How to determine whether a term is the central word of a document, we believe that it can be considered from the following two perspectives:

(1) Term Weight: In the traditional vector space model, term weight represents the contribution degree of the term to the document. The greater value of the term weight, the more important it is for the document.

(2) The Context of the Term: In document, merely selecting a term with high weight value to represent the topic is incomplete, and in fact its context also contains important information. For example, in a document about network virus events, terms such as "virus" and "trojan" often appear near the word "network". So, the term "trojan" should also be selected as the keywords of the document.

2.1 High Weight Term Selection

This paper first selects high weight terms as topic label. We extract the most important terms from each document based on Pivoted Normalization weight. The weight formula is expressed as follows:

$$w_{ik} = \frac{\ln(tf_{ik} + 1)}{1 - s + s\dfrac{pl}{avpl}} \tag{1}$$

where tf_{ik} is the frequency of tem t_k in the document d_i.pl refers to the length of the document and $avpl$ denotes the average length of all documents.

As we all known, the forum page is divided into an index page and a content page. The content page is also further divided into source post and reply post. Therefore, the term appearing in a different location should have different influences. In fact, users see the index page firstly, which displays the source post in the form of a title list. If the user is interested in a certain source post in the index page, he clicks the source links to open the document. So, in order to attract the attention of the users, the presentation of title is often very important, and the author often states the topic clearly in a very simple way. Words appearing in the title are more generalized than those in the source text, and should have greater weight.

On the other hand, the status of the source posts and replies should also be differentiated. The source posts

represent the initiation of a topic. Compared to the replies, the content description has the characteristic of more detailed, standard and lower noise. In most cases, the replies express their attitude towards the source post, or express its own opinion. Some of the replies even contain only one or two simple texts. Obviously, the terms appearing in the source post should be more important than the ones appearing in the reply. Based on the above consideration, term frequency in Eq. (1) not only represents the number of occurrences of the terms, but also the influence of location information. The specific equation is described as follows:

$$f_{ik} = w_{\text{place}} \cdot \sum_{j=1}^{4} w_p(t_k) \cdot tf(t_k, d_{ij}) \tag{2}$$

where w_{place} refers to the position weight (whether it is in the source or reply) of the term t_k, w_p also denotes the position weight of the term appearing in the post (head, first paragraph body, tail body and other parts of the body). Secondly, we ranked them in descending order and selected the top N terms as the candidate keywords for one post document.

2.2 Contextual Characteristic of the Term

Additionally, we investigate that the context of those high weight terms also included important terms and can express the topic of the document well. We believe that those terms with frequent co-occurrence in the same document have statistical correlation. Therefore, the context feature is introduced into the label extraction. Supposing that the collection is expressed as D, l refers to the slide window and the length is L (that is, L is the number of terms in the window). We are of the opinion that the following three factors should be considered in the term context relation: average frequency of co-occurrence to the high weight term under certain slide window, inverse co-occurrence frequency and co-occurrence distance.

(1) Average frequency of co-occurrence to high weight term ($\overline{C}(t_i, t_j)$). Generally, in large-scale corpus, the more frequently occurring word pairs (t_i, t_j) co-occur, the stronger is semantic relationship between them. The equation is as follows:

$$\overline{C}(t_i, t_j) = \frac{\sum_{d \in D} \sum_{l \in d} C_l \langle t_i, t_j \rangle}{\sum_{d \in D} W_{\text{d}}} \tag{3}$$

where $\overline{C}(t_i, t_j)$ is the number of co-occurrences between t_i and t_j in l. W_d is the number of moving windows of slide window l based on steps from left to right in posts document d.

(2) Inverse co-occurrence frequency (*ICof)*. Similar to inverse document frequency, it reflects the number of co-occurrence terms as a contribution to the context relation. The greater the number of different co-occurrence terms to t_i, the smaller the contribution to the context relation. The computation is shown as Eq. (4).

$$ICof(t_i) = \log \frac{|D|}{\left| C_l \langle t_i, t_j \rangle \right|} \tag{4}$$

where $|D|$ is the total number of different terms in the collection and $|C_l < (t_i, t_j) >|$ refers to the number of terms that co-occur with t_i under the window unit l.

(3) Co-occurrence distance (*Co_Distance)*. Generally we think the relevance between co-occurring term pairs results in exponential decay as the term distance decreases. The formula is shown as follows:

$$Co\_Dis\tan ce = e^{-\lambda s(t_i, t_j)} \tag{5}$$

where λ is the impact factor and $s(t_i, t_j)$ stands for the distance between the co-occurring term pairs t_i and t_j , calculated by the number of terms between them.

Combining the above three factors, we defined context weight as the following formula:

$$Context\_Value(t_i, t_j)) = \overline{C}(t_i, t_j) \cdot ICof \cdot \text{Co}\_dis\tan ce \tag{6}$$

We find all terms that have a contextual relationship to t_i with high weight value and compute their context weight based on Eq. (6), and then rank them in descending order. We select the top M terms, and then combine the above top N terms with high weights to constitute topic labels for one posts document.

3 DOCUMENT MODELING BASED ON TOPIC LABEL

Traditional document usually adopts vector space model, in which all the different terms are regarded as a dimension in the document space. This model has the following disadvantages: (1) the number of vocabulary in the document space is very large, and the lexical-based document space has super-high dimensions; (2) the individual term has weaker ability to express the main topic of the document. Therefore, this paper proposes topic label to replace the general term to construct document model. The reason is as follows: firstly, the number of topic label is much lower than terms, leading to the large reduction in dimension, which is more beneficial to follow-up processing; secondly, labels are more expressive to the topic document than the single term, and could eliminate the influence of polysemy.

Supposing document space $D = \{d_1, d_2, \ldots, d_n\}$, d_i is the i^{th} document, label collection $C = \{c_1, c_2, \ldots, c_m\}$, c_i refers to the i^{th} extracted label. The entire document D can be represented as $n \times m$ matrix, in which row vector $d_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\}$ denotes a document, and column vector $c_j = \{c_{1j}, c_{2j}, \ldots, c_{mj}\}$ indicates the distribution of a label in each document. If the label appears in the document, the weight is calculated from above Eq. (1) to (2), otherwise, the weight value is set 0, as shown in the following formula:

$$\begin{aligned} D &= (d_1, d_2, ..., d_n)^{\text{T}} \\ d_i &= (c_{i1}, c_{i2}, ..., c_{im}) \\ c_{ij} &= \begin{cases} w_{i,j} \\ 0 \end{cases} \end{aligned} \tag{7}$$

So, the similarity between post documents is transformed into the semantic similarity between topic labels, and the classical cosine formula in the vector space model is used to calculate the similarity value.

$$sim(d_i, d_j) = \cos\theta = \frac{\sum_{k=1}^{m} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2} \cdot \sqrt{\sum_{k=1}^{m} w_{jk}^2}} \quad (8)$$

4 HOTNESS EVALUATION OF THE TOPIC

The user's interest largely determines the popularity of the topic. Based on the above topic discovery, we get a list of topics that are presented in clusters. Due to the vast amount of information on the Internet and complexity of them, there are a large number of topics. If we give all these topics to the users, they need to spend a lot of time to read them one by one to find their interest. Therefore, it is far from enough to return to the user's topic, and further analysis of these topics should be conducted. Topic hotness evaluation is performed.

To measure the topic hotness, we must first understand what a hot topic is and what the features of it are. Researchers put forward their respective opinions on the characteristics of hot topic. Bun K. [17] believed that hot topics were those which frequently appear over time. Chen Kuan [12] thought that a hot topic should have the following four characteristics: (1) the news media channel has a large number of reports on it; (2) the topic appears in multiple news media channel; (3) stronger coherence, that is, many different events related to it have also been reported; (4) the popularity of the topic continues to change over time. Donghui Zheng [15] considered that the hot topic is always accompanied by the abundance of users' participation, and defined four distinct characteristics: (1) massive posts; (2) high quality posts; (3) high cohesion; (4) bursting.

In the following work, we will analyse the influence factors of the topic hotness, and based on it, the model for calculating the popularity of the topics is proposed.

4.1 High Influence Factors Analysis of the Topic Hotness

Compared with ordinary topic, hot topic has the characteristics of a large number of online information, longer reporting time, spreading wider range, stronger response by public and greater impact to society. Therefore, we believe the influence factors of the topic hotness have mainly two aspects: internal and external features. Internal features refer to the user's activity, while external features are the topic attention. Generally, user's activity is represented by the number of clicks, replies, participation and released post. Topic attention is mainly showed in the last time, post source and quality of the topic. Supposing $Topic\_Set = \{topic_1, topic_2, \ldots, topic_m\}$, in which each $topic_k$ represents one cluster based on k-medoid clustering algorithm. The analysis of both aspects is as follows

A) Internal Features

(1) Number of clicks. In the BBS forum, topics are usually showed in the list post form. If the user is interested in the title of the post for one topic, he will naturally click this post to browse. So, the number of clicks could directly reflect the extent to which the post attracts the netizen. Generally, a topic to become hot spot must be more concerned by public. Therefore, the amount of visiting to the post is also very frequent and the corresponding hits number will be large.

(2) Number of replies. Although the public often browse the forum post, they will not reply to them. In fact, a user has a behaviour of replying only when the postcontent caused his enough attention; for example, he agreed with or against the expressed opinion. So, the number of replies can really reflect participation degree and the topic hotness. In other words, the topic with large number of clicks is not necessarily a hot topic.

(3) Number of user participation. The user's feedback information is thought to be helpful for hot topic detection. Consequently, counting the number of the participation users in a unit time may be used to measure the hotness indirectly. We think the more participating users, the larger the group that is concerned about the topic, and the greater the influence and wider the range of the topic.

(4) Number of post for the topic. In general, the more number of post, the more likely the topic becomes a hot topic. Many network hot topics have a common feature, that is, during the event, the posts to the topic are overwhelming. If we could count the number of the released post for a topic over a period of time, we could see the popularity and activity of the topic.

B) External Features

(1) Duration of topic. We believe the longer the duration of the topic, the greater the influence, the deeper impression on people, and the more likely the hot topic is. Therefore, the difference between the earliest publishing time and the last time for $topic_i$ is computed and the formula is as follows.

$$Last\_Time_{topic_i} = \frac{t_{\text{begin}} - t_{\text{end}}}{T} \quad (9)$$

where t_{begin} is the earliest publishing time for $topic_i$, t_{end} is the last releasing time for $topic_i$, T refers to the time interval.

(3) post source. One hot topic usually spreads widely and appears in different media. On the other hand, online community with greater influence will gather more users to express various opinions, and is thus more likely to erupted hot topic. So, tracing the source of the post could investigate the influence of the forum for the topic. We rank the various forums according to their influence, and select the top 10 forums which are assigned weight artificially, ranging from 1 to 10.

(4) topic quality. The higher the quality of a topic, the greater the user's clicks and the greater the likelihood of becoming a hot topic. This quality is showed in the following two aspects:

a) Post text length. Usually, when you express opinions on a topic, you need a certain length of text to express clearly. Therefore, the longer the average length of post text is, the higher the quality. Based on this, the average length within one topic cluster is applied to measure the topic quality.

$$Ave\_Length(topic_i) = \frac{\sum_{d_i \in topic_i,\, i=1}^{|topic_i|} Length(d_i)}{|topic_i|} \quad (10)$$

where $Length(d_i)$ is the length of post text, $|topic_i|$ stands for the total number of posts in the $topic_i$.

b) Post content concentration. If the amount of posts is all talking about one topic, the more concentrated the content is, the higher quality of the post is. So, the text similarity between two posts is also used to measure the concentration. The formula is defined as follows:

$$Focus\_Degree(topic_i) = \sum_{j=1}^{|topic_i|} \frac{\sum_{j \neq k,\, k=1}^{|topic_i|} sim(post_j, post_k)}{|topic_i| - 1} \quad (11)$$

where $sim(post_j, post_k)$ refers to the similarity between the j^{th} post and the kth post in the $topic_i$.

4.2 The Model of Topic Hotness Evaluation

Through above factor analysis, we have presented a quantitative calculation model for the topic hotness. Each topic has its own life cycle. So, the hotness value is limited to the range of their life cycle and is determined by both internal and external features. The internal features for $topic_i$ are assigned as the following equation:

$$\begin{aligned} &Inter\_Feature(topic_i) = w_1 * Click\_Num + \\ &+w_2 * Re\,ply\_Num + w_3 * UserNum + \\ &+w_4 * Post\_Num \end{aligned} \quad (12)$$

Similarly, the external features are computed as follows:

$$\begin{aligned} &Exter\_Feature(topic_i) = w_5 * Duration\_Time + \\ &+w_6 * Post\_Source + w_7 * Avg\_Length + \\ &+w_8 * Focus\_Degree \end{aligned} \quad (13)$$

The topic hotness is defined as the following formula:

$$\begin{aligned} &HotValue(topic_i) = \lambda_1 * Inter\_Feature(topic_i) + \\ &+\lambda_2 * Exter\_Feature(topic_i) \\ &\lambda_1 + \lambda_2 = 1 \end{aligned} \quad (14)$$

Rank the topics according to their hotness value and select the top N as final hot topic.

5 THE EXPERIMENT RESULTS AND Y ANALYSIS

The purpose of this experiment is to verify the effectiveness of proposed hot topic detection model. Consequently, topic discovery based on improved vector space model is verified firstly, and subsequently, based on them, we test whether the proposed hotness features can effectively reflect the popularity of topic, and hence finally get better results.

5.1 Experimental Preparation

We first construct an experimental corpus. Specific steps are as follows:

Step1: collect and select data. We use train collector to crawl several influential forum data, such as "TianYa talk", "Voice of the People", "Kaidi Community". Due to the complexity and a large amount of data, this paper only collected them in December 2013 for TianYa talk, and from October to December 2013 for Voice of the People and Kaidi Community. Meanwhile, this paper limits the data size to reduce the time complexity of clustering algorithm:

(1) Posts with less than 30 clicks were filtered.
(2) Source posts with less than 50 words are filtered.
(3) Replies with less than 70 words are also filtered.

Step2: delete duplicate or low information posts. In the forum, some users may submit duplicate or low information posts for some purpose. For such posts, it should be filtered.

Step3: data preprocessing. The online community data is full of noise and needs to be preprocessed for the experimental datasets:

(1) Word segmentation. Directly use the Chinese morphological analysis system ICTCLAS for Chinese word segmentation and named entity recognition. In addition, considering the network specific words and fixed collocations, some unidentified words and collocations, such as "wechat", "weibo", are imported into the users' dictionary and perform the second word segmentation.

(2) Filter stop words. Construct a stop list manually to filter common meaningless terms.

(3) Filter emotion symbol tags. Users often publish their opinion by emotion tags, such as "! ! !", "^_", "re". These marks have no effect on the topic discovery and hence should be deleted.

(4) Term selection. The massiveness of forum information makes the collection to be huge, and if the collection is directly used to the model, the time and space complexity will be greatly increased. Therefore, it is necessary to further filter and select the important term to improve the accuracy of subsequent models, reduce the dimension and running time. The entropy of a term is the importance indicator of a term, which represents the distribution of terms in a document.

$$Entropy(t, d_i) = -\sum_{k=1}^{m} (p(t, s_k) * \log_2 p(t, s_k)) \quad (15)$$

where s_k is the k^{th} paragraph in the document d_i, $p(t, s_k)$ is the occurrence probability of term t in s_k

$$p(t, s_k) = \frac{tf(t, s_k) + 1}{tf(t, d_i) + m} \quad (16)$$

Set a threshold D_{min}. If the entropy value of the term is smaller than D_{min}, the term is regarded as low number of the occurrence and hence too small contribution to the document to be deleted. All remaining terms were sorted

in descent and constituted the document term vectors, as shown in the following form

$$d_i = \langle < term_1 >, < term_2 >, ..., < term_n > \rangle .$$

After the above four pre-processing, we obtain the experimental corpus, and some of its statistical characteristics are shown in Tab. 1.

Step 4: manual annotation. In order to measure the subsequently experimental results, we take a lot of time to manually label the topics of all the collection posts. Tab. 2 is the partial annotation topic, which are just hot events voted from October to December 2013 in TianYa forum.

Table 1 Partial Statistical Characteristics of Experimental Corpus

	TianYa Talk	Voice of the People	KaidiCommunity
The number of source posts	15128	5102	1081
The number of replies	664730	155440	9541
The number of users' id	109087	14481	5109
Click numbers	48214148	10748958	5360851
Average words of source post	1083	1148	1382

Table 2 Online Selected Hot Topics from October to December in 2013

Time	Event	Time	Event
Oct.	Tourism chaos of Golden week	Nov.	The third plenary session of the 18th CPC Central Committee
Oct.	1.9 billion people commemorating the birthday of Mao Zedong	Nov.	The bombing case near the Shanxi provincial party committee
Oct.	Flood into the Yuyao city	Nov.	Beijing violent terrorist attack
Oct.	Free medical in Russian	Nov.	Prime Minister's open class became popular
Oct.	New express reporter was arrested across provinces	Nov.	The rape case of Li Tianyi
Oct.	CCTV accuses Starbucks of profiteering	Nov.	Locke announced his resignation
Oct.	Pensions for delayed payment	Nov.	New two-child policy
Oct.	Guangxi drunken policeman shot the owner	Nov.	Evergrande AFC(Asia Football Confederation)won the championship
Oct.	Office building which spent hundreds of millions in Pei country, JiangSu province	Nov.	Zhanjiang official get a room with their female subordinates
Oct.	Demolition officers of Nanning fired at villager	Nov.	Assistance 100,000 to the Philippiness, but they said it was too little
Dec.	Zhangyimou having extra kids	Dec.	North Korea's Zhang Zecheng fell a cropper
Dec.	Official take the lead in delaying retirement	Dec.	The god reply of Guangxi's traffic police
Dec.	Beijing foreigner knocked down aunt	Dec.	Problematic vaccine-induced infant death
Dec.	New fire recruits of Inner Mongolia were beaten	Dec.	The lake of supervision in the express delivery industry caused by the death of express delivery
Dec.	Large-scale smog across the country	Dec.	An old man of Henan home-made cannon for combating forced house demolition

Table 3 Topics Labels Results

No.	Topic label
142	Corruption, discipline inspection commission, supervision, style, issue, discipline inspection, the masses, responsibility, clean government, party style, monitor
1247	ask for salary, migrant workers, government, hard-earned money, municipal committee, municipal government, petition, official, difficult
2202	Leukemia, hospital, patient, anticancer, doctor, treatment, help, fee, transplantation, gain
2273	Reform, development, china, socialism, economics, state council, system, construction, market economics, the central committee of the communist party of china, party
2718	Fang Kaipeng, in advance, retirement, stipulate, Fujian Province ,policy, party, constitution, volunteer, state
3753	Teacher, school, education, student, backbone, salary, vote, treatment, professional ranks and titles
6544	Transgenic, china ,food ,safety, grain, agriculture, science and technology, technology, gene, study
7537	Discipline inspection commission ,court , prison, procuratorate, legality , tip, justice, letter of accusation, democracy, china
8055	Li xx, Mengge, rape, young person, legality, sex, deal, bar, evidence, lawyer, organization
16021	Girl, baby, throw, elevator, small, media, news, government, forgive, Chongqing
16063	Old man, children, happiness, provide for aged, government, duty, society, respect the old, service, aging
16980	Smog, weather, pollute, Beijing, shanghai, air, city, serous, Yangzi River, delta, curb
17391	Demolition , expropriation, government , land , peasant , compensation , construction ,countryside , farmland , land
19711	Family planning, population, Zhang Yimou, law, china, society, family planning committee, citizen, right, nation

5.2 Experiment Results and Analysis on Topic Discovery

In our topic discovery model, the extracted topic label is very important to the performance. So, the labels are extracted firstly. Tab. 3 is the part of the results.

From the results, it is shown that the extracted topic label model gets better performance. The obtained label not only has completed meaning, but strong semantic relationship to each other, which reflects the main idea of document.

On the basis of above labels, the topic discovery is performed by k-medoid clustering algorithm, in which improved vector space model based on label extraction is used to measure similarity (TD_IVSLE). Therefore, in order to test the validity of the similarity measurement, a contrast experiment is carried out between the proposed model (TD_IVSLE) and topic detection based on latent semantic indexing model (TD_LSI). The experiment results are shown in Tab. 4.

Table 4 Overall Performance Comparison Results

Topic number	TD_LSI			TD_IVSLE		
	Recall	Precision	*F* Value	Recall	Precision	*F* Value
60	0.28	0.36	0.26	0.37	0.45	0.33
61	0.29	0.38	0.27	0.38	0.47	0.34
62	0.32	0.40	0.30	0.40	0.49	0.35
63	0.35	0.43	0.32	0.43	0.53	0.37
64	0.33	0.42	0.29	**0.45**	**0.55**	**0.40**
65	0.30	0.37	0.26	0.42	0.48	0.36

From the results, it is obvious that the best performance is obtained in 64 clusters. Compared with the TD_LSI model, the average recall of the proposed model is 0.45, the average precision is 0.55, and F value is also 0.40. The performance increased by 36%, 31% and 38% respectively, indicating that topics are identified well. The reason is mainly the adoption of the topic label extraction. From the data of Tab. 2, the extracted labels can basically reflect the main content of the document. So, it is more easy to cluster those related topic posts together based on these labels, and hence get better clustering results.

5.3 Experiment Results and Analysis on the Hotness Evaluation

Based on above extracted topics, hot evaluation model is employed and the top 20 hot topics are selected. In our experiment, we do not compare the proposed model with previous work since the existing methods for hot topic detection are aimed at different research purpose. We believe if the evaluated hot topic is the same as the online annual hot event, the evaluation model is proved to be reasonable and effective. Therefore, we compare our hotness ranking results with the online hot topic list. If the two sides overlap more, they indicate that the evaluation model has better performance. The data in table 1 is a hot event selected by TianYa forum from October to December 2013. The results are shown in Tab. 5.

From the experimental data, it can be seen that our evaluation model can identify some hot topics. Compared with the data of table1, there are nine hot events that coincide with the online selected hot topics, and the accuracy value is 0.45, indicating that our proposed internal and external features can better reflect the topic hotness. At the same time, we also observe that some hot topics have not been selected, and the reasons are as follows:

Firstly, some topics are not detected in the beginning while performing topic clustering task, for example, free medical in Russian, Evergrande AFC won the championship, Locke announced his resignation and CCTV accuses Starbucks of profiteering and other events. This is because the collected data is just only from two sections of TianYa forum, which is "TianYa Talk" and "Voice of the People". Furthermore, the coverage of download data from the section of "TianYa Talk" is only from December. Therefore, it causes a lot of original topic information not to be collected in the dataset.

Secondly, for some hot topics in December, for example, Beijing foreigner knocked down aunt, which happened in December, net users really discussed it in the next month. So, there are only small amount of data in the dataset. However, our hotness is measured by the features of released posts, reply number, which lead these hot events not to be ranked in the top.

Finally, topic detection based on clustering has influence on the subsequent ranking task. Based on the extracted topic, we can see that some of hot topics are not clustered together, but scattered in different cluster, which also affects the hotness ranking.

Table 5 Hot Topic Ranking List

Num	Event	Hot Value
1	Zhangyimou having extra kids	0.983420
2	Throw the baby down to the ground in Chongqing	0.966926
3	Chain marketing	0.730712
4	The rape case of Li Tianyi	0.696769
5	commemorating the birthday of Mao Zedong	0.437405
6	Transgenic battle between cui and zhou	0.409406
7	Henan red cannon anti strike	0.393693
8	Vaccine-induced infant death	0.388864
9	tourism chaos of Golden week	0.388302
10	New two-child policy	0.372789
11	Corruption in pingxiang, jiangxi province	0.371827
12	delaying retirement policy	0.370211
13	The third plenary session of the 18th CPC Central Committee	0.367566
14	Unjust case of Zhou Yanbing in Weihai	0.363061
15	Medical accident	0.359553
16	Illegal land occupation	0.358721
17	No holiday policy in New Year′s Eve	0.356371
18	Bank security issues	0.354022
19	Fujian surveying mapping institute persecute the party member named fang pengkai	0.353229
20	Migrant worker's Salary	0.352678

6 CONCLUSION

Hot topic detection from BBS is a challenge research problem. In this paper, we propose a model for hot topic detection. Firstly, topic discovery based on improved vector space model is presented, and subsequently, the factors that affect the hotness are analysed from the internal and external features, and based on them, an effective topic hotness evaluation model is proposed. Experiments have demonstrated its effectiveness of detecting hot topic in mining real-world BBS data.

Acknowledgements

This material is based upon work supported by the National Science Foundation of China under Grant Numbers 71861014, 71762017, 61662027, 61762042 and 61562032 and the postdoctoral fund of China (2017M612602).

7 REFERENCES

[1] Allan, J., Lavrenko, V., & Swan, R. (2002). *Explorations within Topic Tracking and Detection: Topic Detection and Tracking*, Kluwer Academic Publishers Norwell, MA, USA. https://doi.org/10.1007/978-1-4615-0933-2_10

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(1), 993-1022. Retrieved from http://papers.nips.cc/paper/2070-latent-dirichlet-allocation.pdf

[3] Sun, Y., Huifang, M. A., Yao, W., et al. (2016). Microblog hot topic detection based on positive point mutual information and probabilistic topic model. *Computer Engineering & Application, 52*(6), 61-63.

[4] Wang Heyong & Lan Jinjiong. (2015). Text Topic Mining Oriented toward Massive High-dimensional Data. *Journal of Intelligence*, 11, 162-167.

[5] Ye Chuan & Ma Jing. (2015). Research on Topic Discovery Algorithm of Multimedia Microblog Comments Information. *New Technology of Library and Information Service*, 11, 51-59.

[6] Ruiguo Yu, Xiaodong Xie, Yongming Li, et al. (2006).

Online Hot Topic Detection based on Segmented Timeline and Aging Theory. *International Journal of Hybrid Information Technology, 9*(2), 247-258. https://doi.org/10.14257/ijhit.2016.9.2.22

[7] Guo, J., Zhang, P., & Guo, L. (2012).Mining hot topics from twitter streams. *Procedia Computer Science*, 9, 2008-2011. https://doi.org/10.1016/j.procs.2012.04.224

[8] Feng Xu, Jue Liu, Ying He, et al. (2017). Hot Topic Trend Prediction of Topic based on Markov Chain and Dynamic Backtracking. *Advances in Multimedia Information Processing - PCM, Lecture Notes in Computer Science*, vol. 10736, Springer. https://link.springer.com/chapter/10.1007/978-3-319-77383-4_51

[9] Peiyu Liu, Xiuyan Hou, Zhenfang Zhu, et al. (2016). Micor-Blog Hot Topic Detection Based on Heat Co-Ranking. *Journals of Frontiers of Computer Science and Technology, 10*(4), 573-581. https://www.docin.com/p-1716523796.html

[10] Wang, X. Y. (2014). Hot Topic Detection in News Blog. *Applied Mechanics & Materials*, 513-517, 1114-1118. https://doi.org/10.4028/www.scientific.net/AMM.513-517.1114

[11] Xiaodong Wang. (2013). A Method of Hot Topic Detection in Blogs using N-gram Mode. *Journal of Software, 8*(1), 184-191. https://doi.org/10.4304/jsw.8.1.184-191

[12] Chen, K. Y., Luesukprasert, L., & Chou, S. (2007). Hot Topic Extraction based on Timeline Analysis and Multidimensional Sentence Modeling. *IEEE Transaction on Knowledge and Data Engineering, 19*(8), 1016-1025. https://doi.org/10.1109/TKDE.2007.1040

[13] Lan, Y., Yongping, D., Jiayin, G., Xuanjing, H., & Lide, W. (2004): BBS based Hot Topic Retrieval using Back Propagation Neural Network. *Proceedings of First International Joint Conference on Natural Language Processing (IJCNLP)*. Hainan Island, China. March 22-24, 139-148.

[14] Ma, H. F. (2011). Hot Topic Extraction Using Time Window. *Proceedings of International Conference on Machine Learning and Cybernetics (ICMLC)*. Guilin, China, July 10-13, 56-60. https://doi.org/10.1109/ICMLC.2011.6016664

[15] Donghui Zheng & Fang Li. (2009). Hot Topic Detection on BBS using Aging Theory. *Proceedings of the International Conference on Web Information Systems and Mining (WISM)*. Shanghai, China. November 7-8, 129-138. https://doi.org/10.1007/978-3-642-05250-7_14

[16] Wang, C., Zhang, M., Ru, L. et al. (2008). Automatic Online News Topic Ranking using Media Focus and User Attention based on Aging Theory. *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*. California, USA. October 26-30, 1033-1042. https://doi.org/10.1145/1458082.1458219

[17] Bun, K. K. & Ishizuka, M. (2002). Topic Extraction from News Archive using TF*PDF Algorithm. *Proceedings of third International Conference on Web Information System Engineering (WISE)*, Singapore, 73-82. Retrieved from http://www.doc88.com/p-1498501959523.html

Contact information:

Minjuan ZHONG
School of Information Technology and Management,
Hunan University of Finance and Economics,
Changsha, China
E-mail: lucyzmj@sina.com