Multimed Tools Appl (2016) 75:8843–8860
DOI 10.1007/s11042-014-2184-y

Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog

Yanyan Zhao · Bing Qin · Ting Liu · Duyu Tang

Received: 31 March 2014 / Revised: 20 June 2014 / Accepted: 7 July 2014 /
Published online:19 August 2014

Abstract As a new form of social media, microblogging provides platform sharing, wherein users can share their feelings and ideas on certain topics. Bursty topics from microblogs are the results of the emerging issues that instantly attract more followers and more attention online, which provide a unique opportunity to gauge the relation between expressed public sentiment and hot topics. This paper presents a Social Sentiment Sensor (SSS) system on Sina Weibo to detect daily hot topics and analyze the sentiment distributions toward these topics. SSS includes two main techniques, namely, hot topic detection and topic-oriented sentiment analysis. Hot topic detection aims to detect the most popular topics online based on the following steps, topic detection, topic clustering, and topic popularity ranking. We extracted topics from the hashtags using a hashtag filtering model because they can cover almost all the topics. Then, we cluster the topics that describe the same issue, and rank the topic clusters via their popularity to exploit the final hot topics. Topic-oriented sentiment analysis aims to analyze public opinions toward the hot topics. After retrieving the topic-related messages, we recognize sentiment for each message using a state-of-the-art SVM (Support Vector Machine) sentiment classifier. Then, we summarize the sentiments for the hot topic to achieve topic sentiment distribution. Based on the above framework and algorithms, SSS produces a real-time visualization system to monitor social sentiments, which is offering the public a new and timely perspective on the dynamics of the social topics.

Y. Zhao (✉)
Department of Media Technology and Art, Harbin Institute of Technology, Harbin, China
e-mail: zyyster@gmail.com

B. Qin · T. Liu · D. Tang
Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

B. Qin
e-mail: bqin@ir.hit.edu.cn

T. Liu
e-mail: tliu@ir.hit.edu.cn

D. Tang
e-mail: dytang@ir.hit.edu.cn

Keywords Sentiment analysis · Social media · Topic detection · Microblogging · Opinion mining

1 Introduction

Microblogging has become a very popular communication platform of social media among Internet users. The number of active Internet users using microblogging websites, such as Twitter, is rapidly increasing worldwide. Users share either information or opinions about personalities, politicians, products, companies, events, and so on, through these platforms. Sina Weibo has been the most popular Chinese microblogging website, which is widely used by over 30% of the Internet users and has a market penetration similar to what Twitter has established in the USA.[1] By providing a vast amount of user-generated content every day, Sina Weibo can be efficiently useful for Chinese marketing or social studies.

The big data from microblogging are attracting the attention of different communities interested in analyzing its content. For example, some works studied how social media content can be used to predict real-world outcomes, such as forecasting box-office revenues for movies [4] or exploiting topic-based twitter sentiment for stock prediction [21]. Other works that focus on finding bursty topics proposed a unified probabilistic latent variable model [8] or a topic model [9] to identify the topics or the events on Twitter. Other research discussed gender inference problem of Twitter users [7], or extracted and modeled durations for habits and events from Twitter [25] and so on. Among the applications, topic detection [6, 19], topic prediction [4], and sentiment analysis [1, 16], are the typical techniques.

As previously discussed, microblog is considered as a reflection of the reaction of the general public to social topics. Therefore, it can be treated as a social sensor. This paper constructs a system called **Social Sentiment Sensor (SSS)** to detect real-time sentiment tendency toward daily hot topics. We demonstrat the system using a map interface of China (Fig. 1). This system presents the hot topics and their corresponding sentiment distributions for entire China, especially for each province. In particular, SSS contains two main techniques such as follows:

- **Hot topic detection**, which aims to detect the popular topics that are being followed by a large number of users, for a better understanding of what is being discussed in microblogs. For example, recently, "马航失联" (Malaysia Airlines plane missing) and "两会" (two sessions), are the two hot topics in China in March 2014.
- **Topic-oriented sentiment analysis** is a very meaningful task, which aims to exploit the overall or general sentiment tendency toward the hot topics by analyzing the topic-related messages. Sentiment is the attitude, opinion, or feeling toward a person, an organization, a product or a location [17]. For example, the topic about "Malaysia airlines plane missing," stimulates the need for the government to perceive the emotions of the Internet users to channel public sentiments properly. The topic "two sessions," drives people's curiosity on how others feel about the proposals, and expects to get an overview about the public opinions.

Considerable studies have been previously conducted on topic detection, but they always worked on news data [14, 26]. Unlike news data, hashtag is one typical characteristic of

[1] http://en.wikipedia.org/wiki/Sina_Weibo

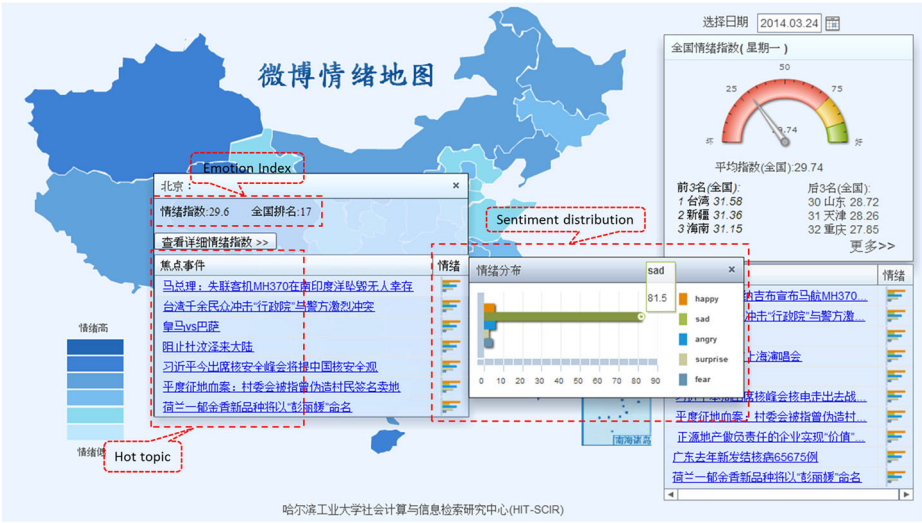


Fig. 1 Demo interface for social sentiment sensor

microblogging. In Sina Weibo, hashtag is represented as a word or an unspaced phrase, which is prefixed and suffixed with the hash symbol ("#"). Hashtag is used to group similar messages, such as "#马航失联#" (#Malaysia Airlines plane missing#), or "#两会#" (#two sessions#). In most cases, hashtags themselves can be treated as topics, because they can cover almost all the topics according to the expreiments. Thus, we can detect hot topics from hashtags. In this paper, the hot topic detection includes three steps: topic detection, topic clustering and topic popularity ranking.

To exploit user sentiment distribution toward each hot topic, we initially retrieved all the topic-related messages based on the topic word representation. We then build a state-of-the-art SVM classifier to recognize the sentiment of each message. Finally, we summarize the sentiments of all related messages and infer the final topic sentiment distribution.

Through hot topic detection and topic-oriented sentiment analysis on Sina Weibo, we can exploit the hot topics and their corresponding sentiment distributions in the entire China, especially in its provinces. Accordingly, exploring interesting social and economic values becomes possible.

Sentiment analysis tools on Twitter, such as Tweet Sentiment,[2] Social Mention,[3] and C-Feel-It [13], exist in the previous works. These tools focus on analyzing and visualizing the sentiment of tweets posted on Twitter, by typing a keyword (topic) into the input field. Unlike these systems, SSS can automatically detect and express hot topics. EMM (Europe Media Monitor) can monitor news from different sites and subsequently apply sentiment analysis to classify them into positive, negative, or neutral. Similarly, this system cannot detect bursty events or topics. Sina Weibo also has a sentiment analysis system called Moodlens [27],[4] which uses sentiment analysis techniques for real-time monitoring of the

[2] http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

[3] http://www.socialmention.com/

[4] http://gana.nlsde.buaa.edu.cn/hourly_happy/moodlens.html

messages. Unlike our SSS system, Moodlens uses keywords to represent abnormal events, which are more ambiguous than the hashtags used in SSS.

The main contributions of this paper are as follows:

- We build a visualization system called SSS to detect the hot topics and analyze the sentiment distributions toward these topics. In particular, China's map shows the functions of SSS, such as detecting the hot topics and sentiments for entire China and its provinces.
- We propose a new hot topic detection framework, which includes the following steps: topic detection, topic clustering, and topic popularity ranking.
- We use state-of-the-art SVM classifier to detect the sentiments of the messages and combine the topic-related message sentiments to acquire the sentiment distribution for a hot topic.
- SSS is a real time topic-sentiment monitoring system. The demo of SSS is now available at http://qx.8wss.com/.

The remainder of this paper is organized as follows. Section 2 introduces the overview of SSS. Section 3 shows the framework for hot topic detection. Section 4 shows the framework for topic-oriented sentiment analysis. Section 5 presents the display and visualization of SSS. Finally we conclude this paper in Section 6.

2 Overview of the framework

The main goal of SSS is to help users detect the hot topics, and the opinions posted on those topics for a period of time. Generally, SSS can be divided into two tasks, namely, hot topic detection and topic-oriented sentiment analysis. Figure 2 shows the framework in detail. The final results reveal that hot topics and the corresponding opinions can be represented vividly on China's map. Next we provide the specific definition for each step in each task.

Hot topic detection aims to find the most popular topics discussed by the Internet users. Because of the characteristics of microblogging, hot topic detection is different from the common topic detection, which has been studied in the previous work [2, 3]. Messages that report such topics are usually teemed with meaningless "babbles". Moreover, topic detection algorithm should be scalable given the sheer amount of messages [23]. There are quite a few studies in this direction in the recent years [6, 8, 19, 23]. Considering the works of these researchers, the task of hot topic detection in this study contains the following steps:

- **Topic detection**: Considering that hashtags can cover almost all topics, hashtags that appear in the messages can be extracted after a filtering model as the topic set T. Formally, $T = \{t_1, t_2, ..., t_n\}$, each topic t_i in T refers to a hashtag.

  The biggest advantage of using hashtags as topics is that hashtags themselves are perfectly organized, topic-related phrases, which are short, simple, and easy-to-understand. They are much more easy and accurate than the phrases extracted from the main body of the messages.
- **Topic clustering**: The topics (hashtags) are sponsored by different users. Thus, some topics describe the same issues. Topic clustering aims to solve this problem by clustering the topics into different clusters. And each cluster is factually a meaningful topic, such as the clusters in Fig. 2. This task can be defined as clustering T into $TC = \{tc_1\{t_1, ..., t_i\}, tc_2\{t_1, ..., t_j\}, ..., tc_k\{t_1, ..., t_p\}\}$.

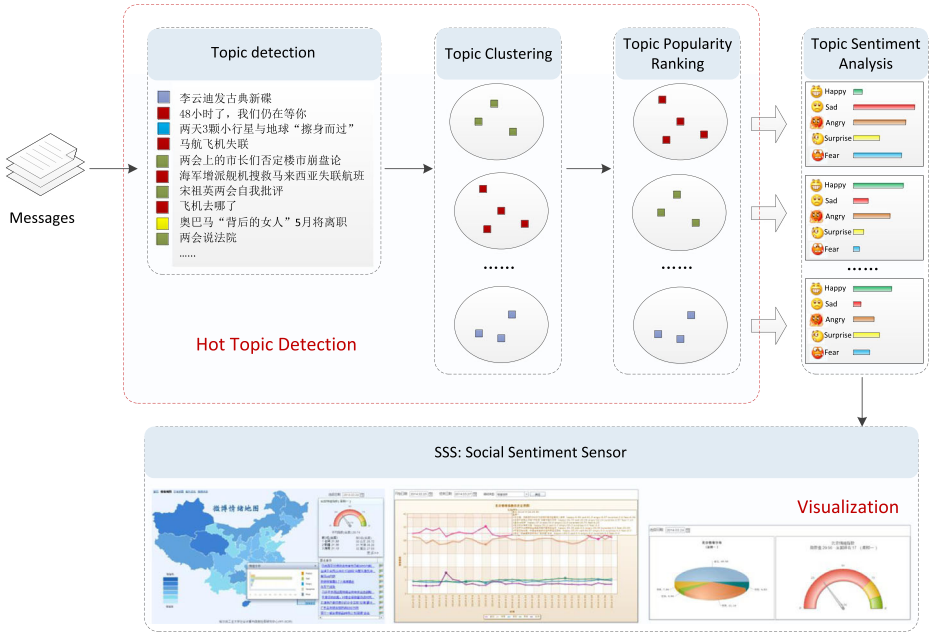


Fig. 2 Framework of social sentiment sensor

– **Topic popularity ranking**: Each topic cluster has different popularity. To exploit the hot topics, the topic clusters should be ranked according to their popularity. The hot topics can be extracted from the topic clusters TC and expressed as $HT = \{tc_1, tc_2, ..., tc_{k'}\}$, where each cluster tc_i in HT refers to a hot topic, and k' is set based on the threshold of popularity.

After discovering the hot topics, users may want to determine public opinions toward them. Topic-oriented sentiment analysis is conducted to address this problem. Given a hot topic, this task aims to detect sentiment distribution, which shows the ratios of five kinds of emotions (happy, sad, surprise, angry, and fear) for each hot topic. Sentiment analysis has long been a research topic [15, 17]. However, topic-oriented sentiment analysis is somewhat different because it requires the collection of topic-related messages before conducting sentiment analysis. Accordingly, the task of topic-oriented sentiment analysis in this study contains three steps:

– **Collecting topic-related messages**: For each detected hot topic ht_i in HT, we need to collect all the topic-related messages within a specified time. The message set for ht_i can be defined as $MS_i = \{m_1, m_2, ..., m_n\}$.
– **Detecting message sentiment**: Five kinds of emotions are used to tag the sentiments of the messages, namely, “happy,” “sad,” “surprise,” “angry,” and “fear”. This task aims to classify each message m_i into one of the five emotions. Many researchers focused on this task [16, 18], which classifies it as a typical task in the sentiment analysis.
– **Summarizing topic sentiment distribution**: For all messages in MS_i about the hot topic ht_i, this task aims to summarize the sentiments of all the messages and find the ratios of the five kinds of emotions. Figure 2 uses different colored bars that represent the sentiment distribution for each topic.

In addition to our study on the algorithms for the above two tasks, we construct a visualization SSS demo to show the topics and the sentiments. SSS can be viewed as a social sensor, where the users can find current topics, and people's emotions and reactions on a specific topic. SSS is also a good platform to polish our algorithms.

In the next two sections, we will introduce the detailed algorithms proposed for hot topic detection and topic-oriented sentiment analysis.

3 Hot topic detection

We propose a framework by using the following steps to detect the hot topics: topic detection, topic clustering and topic popularity ranking. We will introduce them one by one.

3.1 Topic detection

Existing topic detection algorithms can be broadly classified into two categories, namely, document-pivot and feature-pivot methods. The former detects events by clustering documents based on the semantic distance between documents, and the latter studies the distributions of words and discovers events by grouping words together [23]. However, we discovered two disadvantages of the two kinds of methods. First, the performances of these algorithms are not very good; and second, these algorithms always use top words to represent the topic, which are unclear for the users.

Hashtag is one of the characteristics of Microbloggings. Considering that hashtags can describe the topic in a short and easy-to-understand form, we treat all the hashtags extracted from the messages as topics, such as the topics in the "Topic Detection" step in Fig. 2. We hypothesize that the hot topics can be extracted from these topics. In other words, these topics can cover almost all the hot topics. Therefore, using hashtags is more intuitive than the use of top words to describe the topic.

To verify the hypothesis, we collect the hashtags daily, from March 16 to 25, 2014. We also gather real hot topics from Top Baidu,[5] which can be considered as the standard daily hot topics. An average of more than 90% standard hot topics can be detected in the hashtags in 10 days, which illustrates that all the hashtags can cover most of the hot topics. Therefore, we can detect the hot topics from hashtags, which are also suitable for describing topics.

A large amount of messages are being posted everyday, which yield to a considerable number of hashtags. For example, the hashtags in March 16 are more than 167,000. Only a few among the hashtags are meaningful topics. A large part of them are ads or the topics that are not about specific events, such as "Why reading?". Therefore, we need a topic filtering step before topic clustering. To tackle this problem, we utilize three kinds of filtering rules such as follows.

– Literal filtering: If the hashtag can satisfy one of the following conditions, we filter it out. (1) The length of the hashtag is less than 4 characters or more than 30 characters. For example, the topic "毕业季" which is too short should be filtered; (2) the hashtag contains the words in the filtering dictionary;[6](3) the hashtag contains some kinds of

[5]Top Baidu can get the hot topics everyday from the queries. http://top.baidu.com/?vit=1.

[6]We manually construct a dictionary that records the words that are irrelevant to the meaningful topics. This dictionary contains 381 words.

Table 1 Performances of hashtag filtering rules on 10 days' messages

Filtering method	# of hashtags	Hot topic recall (Top 5)	Hot topic recall (Top 10)
Before filtering	167,309	96%	97%
Literal filtering	87,338	94%	95%
Content filtering	102,144	94%	94%
Historical data filtering	107,256	96%	97%
All	25,966	92%	92%

punctuations, such as "—", "-", " " or "_"; (4) the hashtag contains two special forms of time information, such as "3月1日" or "3.01" (March 1st).

- Content filtering: If the messages that contain the hashtag can satisfy one of the following conditions, we assume that the hashtag is not meaningful and should be filtered out. (1) The messages contain the information of "@user"; (2) the messages contain the history time, such as "公元25年"; (3) the messages contain less than 4 characters, such as one message "强！"; (4) the messages contain the words in the ads dictionary,[7] such as "大促" (sales).
- Historical data filtering: If the hashtag has appeared in the historical data of 30 days ago, we filter it out. For example, the topic "周一心情记" (Mood in Monday) started from November 2013 and continued until today, therefore it can be considered as historical topic. We need to filter these topics since they are not the new topics even if hot.

Table 1 shows the performances of the three kinds of filtering rules, which are obtained from the messages observed in 10 days. After filtering, the amount of topics has been greatly reduced, from 167,309 to 25,966, which is useful for topic clustering. The statistics also show that the three kinds of filtering rules are complementary. In addition, in order to estimate the topic coverage ability of the filtered hashtags, we use "hot topic recall" that is the recall value of the real hot topics from Top Baidu, to evaluate the three filtering rules. Table 1 shows that the recall values are both 92% when evaluating with top 5 hot topics and top 10 hot topics. And the upper bound value is the result of the hashtags before filtering, which is 96% and 97% when using the two evaluation measures. All the above can illustrate that the three kinds of rules are effective and just lose very few real hot topics. This can prove that the hashtags after filtering can cover almost all the hot topics.

3.2 Topic clustering

Topic clustering is a typical clustering problem. We can use various kinds of clustering algorithms. This paper employs hierarchical clustering algorithm as a case of study. Similarity computation between two topics is the key technique during the clustering process. We can initially segment the topic and use a word vector to represent each topic. For example, we can use the words "马航" (Malaysia Airlines) and "失联" (missing) to represent the topic "马航失联" (Malaysia Airlines plane missing). Second, we use the cosine similarity

[7] We manually construct a dictionary that records the words that always appearing in the ads. The dictionary contains 412 words.

computation method to compute the similarity between two topics. Finally, we conduct hierarchical clustering algorithm based on the similarity matrix.

However, the mere use of words that appear on the topics to compute similarities is insufficient because the topic contains only a few words, which are so sparse for clustering. For example, the word vector of the topic "马航失联" (Malaysia Airlines plane missing) is totally different from the word vector of the topic "飞机去哪儿" (where is the airplane). They cannot be clustered into one group. Nevertheless, they are factually referring to the same topic.

We introduce the background knowledge for each topic to alleviate these problems. This means that we can use more knowledge to compute the similarity between two topics besides their literal similarity. This idea is based on the hypothesis that the background knowledge of two similar topics is similar. This paper expands the background knowledge for each topic by importing all the messages that contain the topic. In detail, we segment the topic-related messages for each topic into a word vector, where each dimension is computed by the word's TF*IDF. Therefore, more words are used to represent the meaning of the topic, which is helpful for topic clustering.

Formally, the topic background for a topic t_i can be expressed as a word vector $tb_i = \{w_1, w_2, ..., w_n\}$, where n is the word dimension computed from all the messages containing t_i. Here, we delete the dimensions of the words that appear less than two times. We use TF*IDF to represent each word. Thus, the topic word vector can be expressed as $tb_i = \{tfidf_1, tfidf_2, ..., tfidf_n\}$, where TF*IDF for each word is computed as follows.

$$TF(w,t) = 0.5 + \frac{0.5 \times f(w,t)}{max\{f(w',t) : w' \in t\}} \tag{1}$$

$$IDF(w,T) = log\frac{N}{|\{t \in T : w \in t\}|}, \tag{2}$$

where T is the topic set and N is the size of T and $f(w,t)$ refers to the occurrence number of word w in all the words of topic t.

The similarity between two topics, t_i and t_j, is converted to compute the similarity between two topic word vectors tb_i and tb_j. A Cosine similarity measurement is used to compute the similarity as follows.

$$Sim(t_i, t_j) = Sim(tb_i, tb_j) = \frac{tb_i \cdot tb_j}{\| tb_i \| \| tb_j \|} \tag{3}$$

The hierarchical clustering algorithm is conducted based on this similarity. First, we suppose each topic as a cluster, noted as $tc_1, ..., tc_i, ..., tc_n$. Next, we compute the similarity between each pair of clusters. If the similarity between tc_i and tc_j is the maximum and greater than a threshold θ, tc_i and tc_j are merged into a new cluster. The process is repeated until the amount of the clusters remains constant. Note that we use the longest hashtag in the cluster tc_i to represent tc_i.

We simply rank the topic clusters according to the number of the topic-related messages. We choose the top 5 and top 10 topic clusters for each day during March 16 to 25, 2014, and manually evaluate the accuracies of them using two measurements of $P@5$ and $P@10$. Here, $P@5$ and $P@10$ means the accuracy of the top 5 or top 10 clusters. Table 2 shows the performances of the topic clustering methods. When we choose top 5 clusters to evaluate, the accuracy is 94%, and when we choose top 10 clusters to evaluate, the accuracy is 86%. The practical performances can demonstrate the effectiveness of the clustering method.

Table 2 Performances of topic clustering method on 10 days' messages

Clustering method	P@5	P@10
Hierarchical	94%	86%

3.3 Topic popularity ranking

To select the hot topics from the topic clusters, we need to rank these topic clusters according to their popularity, such as the "Topic Popularity Ranking" step in Fig. 2.

There are many ranking methods on microblog. For example, some researchers proposed a multi-faceted brand tracking method to solve a ranking problem of data gathering and content analysis [10]. And some proposed a visual-textual joint relevance learning method to solve the ranking problem in social image search [11].

Different from image processing in social media, the intuitive factor that influences the topic popularity in this paper is the frequency of the topic cluster. This means that if a topic cluster appears more frequently, the cluster is more popular. This factor is called "Topic Frequency." However, although a specific topic cluster is frequent among all the topic clusters, it should slip in the rankings because it is not talking about the current topic. Therefore, comparing with the historical messages to determine whether the topic cluster is special, is important for topic popularity ranking. This factor is called "Topic Specificity." For example, the topic cluster "Malaysia Airlines plane missing" has lasted for more than 10 days. And this topic cluster continues to appear frequently in its 10th day. However, we use the "Topic Specificity" factor to pull down this topic cluster in the ranking because it is not new for the public.

Formally, for one topic cluster tc, we use the formula $Popu(tc)$ to compute its popularity, which is defined as follows.

$$Popu(tc) = Freq(tc) \times Spec(tc), \tag{4}$$

where $Freq(tc)$ and $Spec(tc)$ refer to the two popularity factors, and tc contains many similar topics (hashtags). Specifically, $Freq(tc)$ refers to the times of each topic t in tc appearing among all the messages. To judge the topic speciality, we compare tc with the hot topics (clusters) 10 days ago, which is defined as follows.

$$Spec(tc) = 1 - Sim(tc, tc'), \tag{5}$$

where the topic tc' is one of the top topic clusters 10 days ago, which is most similar to topic cluster tc. During the similarity procedure, each topic cluster is represented by the word vector introduced in Section 3.2.

In summary, if the topic cluster is more frequent and more special, it is more likely to be popular and to be a hot topic. We select the top hot topics by setting a threshold η for $Popu(tc)$. Generally, it is hard to examine the ranking results. In this study, we manually examine 10 days' ranking results by comparing the daily news in Sina,[8] which shows the results are reasonable.

[8] http://news.sina.com.cn/

4 Topic-oriented sentiment analysis

Generally, sentiment analysis aims to classify a document into positive and negative. Different from it, the topic-oriented sentiment analysis has two main characteristics. First, topic-related messages are classified into five emotions, such as happy, sad, angry, surprise and fear, instead of the commonly used positive and negative. Because the users want to know the public emotions toward the topics. Second, the topic-oriented sentiment analysis should collect all the topic-related messages as the first step.

Therefore, the topic-oriented sentiment analysis in this paper includes the following steps: collecting topic-related messages, detecting message sentiment and summarizing topic sentiment distribution. We will introduce them one by one.

4.1 Collecting topic-related messages

Section 3 indicates that each hot topic ht_i, which is also a topic cluster, contains several topic-related hashtags. Therefore, the messages containing these hashtags can be retrieved as one part of the hot topic-related messages. We call it "Hashtag based" method. However, this is insufficient because, in many cases, most of the topic-related messages do not use hashtags. For example, although the message "从马航事件发生到现在可以完全看透一些国家的脸孔和真正性格。" does not contain any hashtags, it still should be retrieved because it discusses about the topic "马航失联" (Malaysia Airlines plane missing).

To solve this problem, we reuse the word vector $\{tfidf_1, tfidf_2, ..., tfidf_n\}$ for each hot topic ht_i to retrieve the related messages. Specifically, we choose the top five words to represent each hot topic according to the TD*IDF values. If a message contains 2 of the top words, we consider the message as topic-related. Accordingly, we call it "Word vector based" method.

In summary, we collect the topic-related messages based on the two heuristic methods. Table 3 shows their performances on the top 10 topics for 10 days, from March 16 to 25, 2014. Compared to Hashtag based method, the Word vector-based method can retrieve more topic-related messages, which are 1,413 messages. When combining the two methods, we can obtain much more messages, which can make the following sentiment analysis results more reasonable.

In addition, to evaluate whether the messages obtained from word vector based method are topic-related, we use the"Accuracy" evaluation measurement, which means the ratio of topic-related messages in all messages. Experiments on 200 hot topics from 10 days show the Word vector based method is effective, achieving an accuracy of 84.5%.

Table 3 Performances of two heuristic methods for topic-related messages collection

Method	Average # of messages for each hot topic
Hashtag based	1,024
Word vector based	1,413
All	2,077

Table 4 Emotion lexicon

Lexicon	Emotion lexicon				
	Happy	Sad	Angry	Surprise	Fear
Size	2,394	3,121	1,929	288	1,113

4.2 Detecting message sentiment

There are many useful features for sentiment analysis in previous studies [5, 12, 18, 24], such as word unigram, POS tags, polarity word lexicon and so on. Inspired by previous studies, we combine the features from Tang et al.'s [22] and Mohammad et. al's work [16].[9] The basic features used in this paper are described as follows:

- Word unigram feature: Unigram features have been proven useful for sentiment classification [18]. To address the sparsity problem of the word vector, we use the 2000 most frequent words in the training data, similar to [20].
- Punctuation features: Some punctuation sequences reflect emotion, such as "!!!", "..." and "???". We manually collect those punctuation sequences and utilize them as binary features according to whether a predefined punctuation occurs in a message.
- Emotion lexicon feature: An emotion lexicon is introduced in Table 4 to map the emotional words in a message into predefined emotion category. Given a message, the lexicon identifies whether each emotion word exists in the message. Similarly, the emotion lexicon feature is used as a binary feature. For example, in the message "今天我真高兴啊" (I am very glad today), the word "glad" occurs in the emotion lexicon and its corresponding category is "happy." Thus, the emotion lexicon feature can be expressed as: Y N N N N, the values of which indicate whether the word appears in the five kinds of emotion categories.

A manually annotated corpus of 3,357 messages is collected for emotion classification, including 548 happy, 837 sad, 905 angry, 567 surprise and 500 non-sentiment messages. The accuracy of cross validation on the gold dataset is used as evaluation metric. We utilize LibLinear[10] to train models for emotion classification. Experimental results show that the performance can achieve an accuracy of 70.80%.

4.3 Summarizing topic sentiment distribution

For a given hot topic, we obtain the sentiment for each topic-related message using a state-of-the-art emotion classifier. We then summarize and show the topic sentiment distribution by computing the ratios of five kinds of emotions. The ratio is computed as follows:

$$Ratio(e) = \frac{N_e}{N}, \tag{6}$$

where N_e is the number of the messages showing the emotion e ($e \in \{happy, sad, angry, surprise, fear\}$), and N is the number of the topic-related messages.

[9] Because the proposed features of [16] are used for English tweets, some of them are not suitable for Chinese microblogs.

[10] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Figure 3 shows the sentiment distribution of a given hot topic. We use different colors to represent different emotions.

5 Display and visualization

SSS provides a good sketch to understand topic-sentiment behavior of the human society depending upon the algorithms of hot topic detection and topic-oriented sentiment analysis. We crawl and store about 120 messages each second because the use of the Sina Weibo API allows only the retrieval of a subset of messages. Finally, we store about 10,000,000 messages for each day. We process the data offline and refresh the display of SSS every 2 hours.

5.1 Topic-sentiment display and analysis

Figure 4 shows the two dashboards for topic-sentiment detection. For example, Fig. 4a displays the hot topics and the corresponding sentiment distributions of Beijing City on March 20, 2014. The hot topics are listed on the left side. Once we click on the hot topics, we can browse all the topic-related messages through the search engine of Sina Weibo. The sentiment distribution for each topic is shown by colored bar chart when clicking on the right side picture of each topic. Similarly, we can observe all the topic-sentiment results for every province in China. In addition, we can produce China's daily top 10 topics and corresponding sentiment distributions, as shown in Fig. 4b.

Besides the topics in Fig. 4a and b, SSS can also work out other good topics and their sentiment distributions. For example, for the topic "招远命案" (Homicide Case in Zhaoyuan), its sentiment distribution is shown in the left part of Fig. 4c, in which "angry" and "fear" are the main emotions. For the recent topic "世界杯德国对阵葡萄牙" (World Cup GERMANY VS PORTUGAL), its sentiment distribution is shown in the right part of Fig. 4c, in which "Happy" is the main emotion. And we can also observe that a few messages show the "sad" emotion, because these messages are posted by the Portugal fans who are unsatisfied with the match.

The topic-sentiment results are presented in another kind of dashboard, such as in Fig. 4d. This chart displays the emotion change as time goes on with a time window of recent 30 days. When clicking on the anchor for each day, we can observe the hot topics and their sentiment distributions. The chart in Fig. 4d presents the topic-sentiment results from February

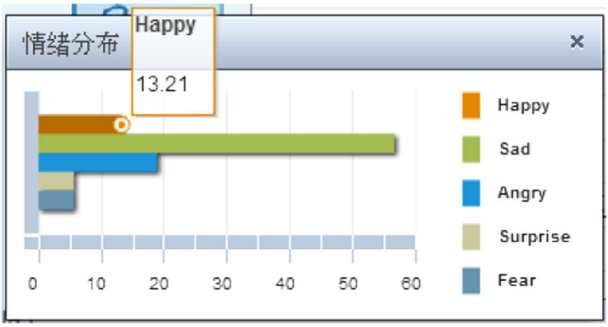
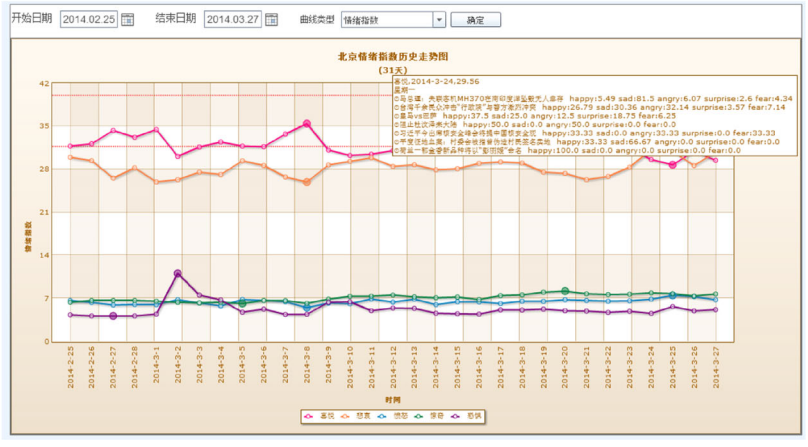


Fig. 3 Representation of sentiment distribution for a given hot topic

(a) dashboard1 for topic-sentiment detection of each province

(b) dashboard2 for topic-sentiment detection of entire China

(c) dashboard3 for sentiment distributions of some typical topics



(d) dashboard4 for topic-sentiment detection of a period of time

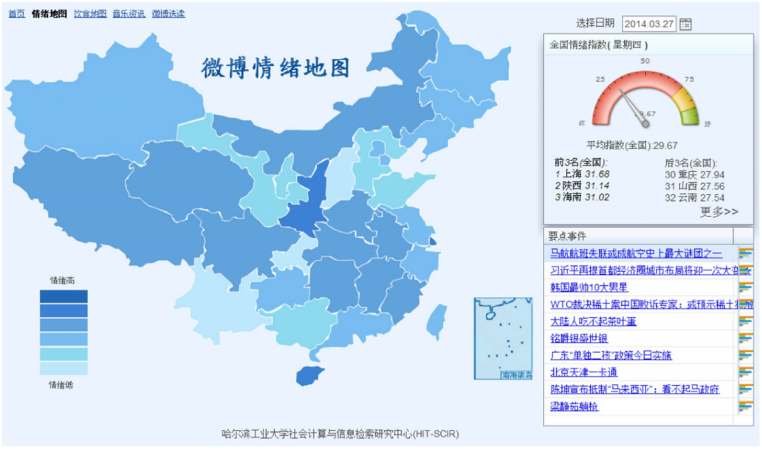
Fig. 4 Dashboards for topic-sentiment detection

25, 2014 to March 27, 2014. We also discover an interesting phenomenon. For example, "March 8 is the happiest day among the 30 days" because that day is Women's Day. "March 25th is the saddest day" because China was informed that non of the passengers of crew aboard MH370 (Malaysia Airlines plane) survived at about 10 p.m. in March 24 and the next day almost all of the people knew this event and were very sad. "March 2nd is the penultimate saddest day" because there was a terrorist attack in Kunming at about 9 p.m. in March 1st and lots of people talked about it on Sina Weibo the next day.

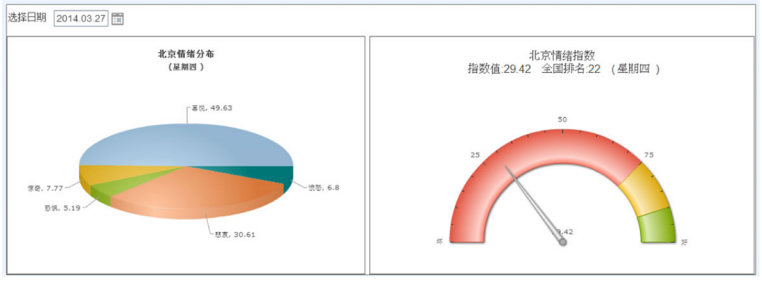
5.2 Nationwide emotion index display and analysis

SSS shows some interesting emotion index. Figure 5 displays two dashboards that represent nationwide emotion index. For example, Fig. 5a displays the nationwide emotion distribution with a range of different blues. Dark blue represents "more happy" and light blue represents "less happy." Thus, we can rank the happiness for all the provinces in China, which are listed at the top-right.

The emotion index for each province can be presented in different forms. The left chart of Fig. 5b shows the emotion distribution for Beijing City in March 24, and the right side shows the happiness index of Beijing City by the ratio of happy messages in all emotional messages.



(a) dashboard1 for nationwide emotion index



(b) dashboard2 for nationwide emotion index

Fig. 5 Dashboards for nationwide emotion index

6 Conclusion and future work

Information gathering and analysis over the Internet have become so important in providing more efficient and effective responses to social topics. Thus, this paper utilized SSS for real-time hot topic detection and sentiment analysis on Sina Weibo to obtain a full and accurate picture of online social landscape.

Our real-time data processing infrastructure includes two parts, namely, hot topic detection and topic-oriented sentiment analysis. SSS detects the most popular topics by the following steps: topic detection, topic clustering, and topic popularity ranking. SSS retrieves topic-related messages and subsequently performs sentiment analysis toward those topics using a state-of-the-art SVM classifier. Depending on the algorithms of hot topic detection and topic-oriented sentiment analysis, SSS presents graphically rich displays. SSS not only displays daily hot topics and their sentiment distributions, but also presents the emotion index for entire China and its provinces. Experiments for each step of the infrastructure show that the system works at a good level, giving a relatively accurate picture of the social media reactions to the hot topics.

We aim to extend the system in two aspects for future research. One aspect is by adding a new function to track the hot topic and analyze its public sentiment changes. Some hot topics, especially political events, may last for several days or more. Accordingly, sentiments for the topics may change every day. Topic tracking and sentiment analysis are important for public event monitoring. The other aspect is polishing the algorithms for hot topic detection and sentiment analysis. Some of the algorithms in this paper are heuristic, such as the algorithms for topic ranking, and should be improved further.

Acknowledgments We thank the anonymous reviewers for their helpful comments. This work was supported by National Natural Science Foundation of China (NSFC) via grant 2014CB340506, 61300113 and 61273321, and the Ministry of Education Research of Social Sciences Youth funded projects via grant 12YJCZH304.

References

1. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of twitter data. In: Proceedings of the workshop on languages in social media, LSM '11. Association for Computational Linguistics, Stroudsburg, pp 30–38. http://dl.acm.org/citation.cfm?id=2021109.2021114
2. Allan J (ed) (2002) Topic detection and tracking: event-based information organization. Kluwer Academic Publishers, Norwell
3. Allan J, Carbonell J, Doddington G, Yamron J, Yang Y (1998) Topic detection and tracking pilot study: final report. In: Proceedings of the DARPA broadcast news transcription and understanding workshop. Lansdowne, pp 194–218
4. Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology - Volume 01, WI-IAT '10. IEEE Computer Society, Washington, pp 492–499. doi:10.1109/WI-IAT.2010.63
5. Barbosa L, Feng J (2010) Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd international conference on computational linguistics: Posters, COLING'10. Association for Computational Linguistics, Stroudsburg, pp 36–44. http://dl.acm.org/citation.cfm?id=1944566.1944571
6. Cataldi M, Di Caro L, Schifanella C (2010) Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the 10th international workshop on multimedia data mining, MDMKDD '10. ACM, New York, pp 4:1–4:10. doi:10.1145/1814245.1814249
7. Ciot M, Sonderegger M, Ruths D (2013) Gender inference of twitter users in non-English contexts. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, pp 1136–1145. http://www.aclweb.org/anthology/D13-1114

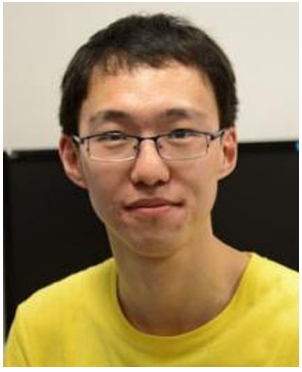
8. Diao Q, Jiang J (2013) A unified model for topics, events and users on Twitter. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Seattle, Washington, pp 1869–1879. http://www.aclweb.org/anthology/D13-1192
9. Diao Q, Jiang J, Zhu F, Lim EP (2012) Finding bursty topics from microblogs. In: Proceedings of the 50th annual meeting of the association for computational linguistics (Vol 1: Long Papers). Association for Computational Linguistics, Jeju Island, Korea, pp 536–544. http://www.aclweb.org/anthology/P12-1056
10. Gao Y, Wang F, Luan H, Chua TS (2014) Brand data gathering from live social media streams. In: Proceedings of International Conference on Multimedia Retrieval, ICMR '14. ACM, New York, pp 169–176. doi:10.1145/2578726.2578748
11. Gao Y, Wang M, Zha Zj, Jialie S, Li X, Wu X (2013) Learning for tag-based social image search. IEEE Trans Image Process 22(1):363–376
12. Johansson R, Moschitti A (2011) Extracting opinion expressions and their polarities - exploration of pipelines and joint models. In: ACL (Short Papers). The Association for Computer Linguistics, pp 101–106
13. Joshi A, R, B A, Bhattacharyya P, Mohanty RK (2011) C-feel-it: a sentiment analyzer for micro-blogs. In: ACL (system demonstrations). The Association for Computer Linguistics, pp 127–132. http://dblp.uni-trier.de/db/conf/acl/acl2011d.html#JoshiABM11
14. Kleinberg J (2002) Bursty and hierarchical structure in streams. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '02. ACM, New York, pp 91–101. doi:10.1145/775047.775061
15. Liu B (2012) Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. Morgan & Claypool
16. Mohammad S, Kiritchenko S, Zhu X (2013) Nrc-canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the 7th international workshop on semantic evaluation exercises (SemEval-2013). Atlanta, Georgia
17. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1-2):1–135. doi:10.1561/1500000011
18. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02, vol 10. Association for Computational Linguistics, Stroudsburg, pp 79–86. doi:10.3115/1118693.1118704
19. Petrović S, Osborne M, Lavrenko V (2010) Streaming first story detection with application to twitter. In: Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics, HLT '10. Association for Computational Linguistics, Stroudsburg, pp 181–189. http://dl.acm.org/citation.cfm?id=1857999.1858020
20. Salakhutdinov R, Hinton G (2009) Semantic hashing. Int J Approx Reasoning 50(7):969–978. doi:10.1016/j.ijar.2008.11.006
21. Si J, Mukherjee A, Liu B, Li Q, Li H, Deng X (2013) Exploiting topic based twitter sentiment for stock prediction. In: Proceedings of the 51st annual meeting of the association for computational linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Sofia, Bulgaria, pp 24–29. http://www.aclweb.org/anthology/P13-2005
22. Tang D, Qin B, Liu T, Li Z (2013) Learning sentence representation for emotion classification on microblogs. In: Proceedings of natural language processing and chinese computing, pp 212–223
23. Weng J, Lee BS (2011) Event detection in twitter. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2767
24. Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Lang Resour Eval 1(2):0. http://www.cs.pitt.edu/~wiebe/pubs/papers/lre05withappendix.pdf
25. Williams J, Katz G (2012) Extracting and modeling durations for habits and events from twitter. In: Proceedings of the 50th annual meeting of the association for computational linguistics: short papers - ACL '12, vol 2. Association for Computational Linguistics, Stroudsburg, pp 223–227. http://dl.acm.org/citation.cfm?id=2390665.2390720

26. Yang Y, Pierce T, Carbonell J (1998) A study of retrospective and on-line event detection. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '98. ACM, New York, pp 28–36. doi:10.1145/290941.290953
27. Zhao J, Dong L, Wu J, Xu K (2012) Moodlens: An emoticon-based sentiment analysis system for chinese tweets. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '12. ACM, New York, pp 1528–1531. doi:10.1145/2339530.2339772

Yanyan Zhao PhD in Computer Science, is an assistant professor at the Department of Media Technology and Art of Harbin Institute of Technology (China). Her interests include: sentiment analysis and text mining. She is the author of more than 10 research papers on sentiment analysis, including the NAACL, EMNLP papers and some international journals. She is member of ACL and CCF.

Bing Qin PhD in Computer Science, is a professor at the Department of Computer Science and Technology of Harbin Institute of Technology (China). Her interests include: text mining and natural language processing. She has published more than 50 papers in distinguished journals and conferences. She is member of ACL and CCF.

Ting Liu PhD in Computer Science, is a professor at the Department of Computer Science and Technology of Harbin Institute of Technology (China). His interests include: information retrieval, natural language processing and social computing. He has published more than 50 papers in distinguished journals and conferences. He is member of ACL and CCF.



Duyu Tang is a PhD candidate at the Department of Computer Science and Technology of Harbin Institute of Technology (China). His interests include: sentiment analysis and natural language processing. He has published many conference papers, such as ACL and Coling. He is member of ACL.