Neurocomputing 254 (2017) 53–58

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An algorithm for event detection based on social media data

Wenjuan Cui[a], Pengfei Wang[a,b], Yi Du[a], Xin Chen[a], Danhuai Guo[a], Jianhui Li[a], Yuanchun Zhou[a,*]

[a] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 1 February 2016
Revised 17 July 2016
Accepted 9 September 2016
Available online 6 March 2017

Keywords:
Event detection
Social media
Foodborne disease
Recommender system

ABSTRACT

Online social network applications such as Twitter, Weibo, have played an important role in people's life. There exists tremendous information in the tweets. However, how to mine the tweets and get valuable information is a difficult problem. In this paper, we design the whole process for extracting data from Weibo and develop an algorithm for the foodborne disease event detection. The detected foodborne disease information are then utilized to assist the restaurant recommendation. The experiments results show the effectiveness and efficiency of our method.

1. Introduction

With the development of information technology, people spend more time surfing on the internet. Lots of daily activities such as shopping, news reading, social network communication, information searching could be made online. While people's lives get convenient from the information services, the continuously generated huge volume of data make it difficult to easily get the useful information fulfilling people's requirements. Recommender system was designed aiming to overcome the information overload problem [1]. Except for the traditional recommender systems, the personalized recommendation has been developed and applied in various fields to meet the users' interest [2]. Lots of approaches have been proposed to the recommendation research, in which content based approach, collaborative filtering and hybrid models are most used methods [3–5]. Various recommender systems have been deployed in tremendous applications. Constructing user profiles from folksonomy systems is also useful for many applications such as personalized search and recommender systems [6]. However, most of the systems only consider one particular data source. For example, a music website will recommender music to users according to the behaviors of the users on the website, such as which songs they share, which singers they pay attention to. They seldom care about the information from other data sources like the social network.

Twitter, Facebook and other social network applications have been frequently used in recent years. People tend to express their opinions, feeling or just tell friends what they are doing on the social network. Twitter is a popular micro-blogging service which attracts much attention. People may post tweets at any place and during any time. In general, the length of one tweet has a limit of 140 characters. There may be only little information in a particular tweet, but the accumulated content can generate a vast amount of information and include important knowledge. Twitter has helped to provide valuable message for various applications. Latent user community could be discovered in social media such as twitter [7]. Tumasjan et al. tracked the public political opinions on Twitter and predict the election result [8]. Sakaki et al. investigated the real-time interaction of earthquakes in Twitter and proposed a method to monitor the tweets and to detect the earthquakes. They can detect the earthquakes with high probability and faster than the government [9].

Tweets can also be used in the area of public health. Users often post the messages like "I got a flu, getting a running nose" or "got a stomachache after eating pizza". Millions of such messages may give a direction for the influenza tracking or other public health problems. Aramaki et al. propose a machine learning method to detect the influenza epidemics from twitter data [10]. Tweets can be used to track the influenza and forecast future influenza rates with high accuracy [11,12]. Twitter data have been used on surveillance of other public health related problems like Dengue and foodborne disease [13].

In this paper, we will consider the relevance of foodborne disease and the restaurants with contaminated food. First, we will

* Corresponding author.
E-mail address: zyc@cnic.cn (Y. Zhou).

http://dx.doi.org/10.1016/j.neucom.2016.09.127
0925-2312/

crawl the foodborne disease related tweets from Weibo, which is a Chinese social network application similar to Twitter. Then the tweets are classified to filter the part which are the actual foodborne disease related ones. The key-phrases are extracted and a SVM classifier is designed to detect the foodborne disease events. Meanwhile, the locations are also determined and the restaurants with contaminated food are identified. Finally the foodborne disease factor is used in the restaurant recommendation.

2. Related work

Foodborne disease (or foodborne illness) refers to any disease resulting from the consumption of contaminated food. Foodborne Disease Outbreak (FBDO) is defined as the two or more cases of a similar illness resulting from the ingestion of a common food. According to CDC 2011 Estimates, each year roughly 1 in 6 Americans (or 48 million people) gets sick, 128,000 are hospitalized, and 3000 die of foodborne diseases [14]. In the past years, the tracking and detection of foodborne disease were mainly carried on by surveillance systems. But the traditional surveillance systems have the limit of time lag in the detection of FBDO. Recently, the social media data have been introduced to the surveillance of foodborne disease [15,16].

With the users for Twitter or Weibo getting more and more, the functions of these social network applications get more complex. People tend to express their feelings and describe what they are doing on the social network applications. And there is a common characteristic for these social network services which is the real-time nature. Users often post several messages on Twitter or Weibo in a single day. Each message describes what the users are doing or thinking and more specifically, each message contains the time and the location. The time is just when the message is updated, and the location may be included in the message when the users update with mobile devices or just be the registered location in the user's profile. Due to the real-time nature of Twitter or Weibo, the large number of messages result in numerous events.

There are a lot of public events hidden in the messages of Twitter or Weibo. The event detection from the short text data such as tweets has been of significant value. Lots of information could be used in the short text to infer the event occurrence. Topic-based profiles could be obtained from social media [17]. Verbal context could be utilized and the verbal context graph could model contents and interrelationships of verbal context in folksonomy and thus capture the user intention in the short text [18]. Many event detection algorithms based on short texts have been proposed. The detected events range from national competitions in sports to earthquake forecasting. In recent years, multimedia event detection has attracted extensive research attention because of the exponential increase in volume of video data on the web. Yu et al. propose a robust spatial-temporal deep model for multimedia event detection [19]. The most event-relevant videos could be detected from large datasets. Bin et al. develop an event detection system based on combination of multi-view representations and co-training algorithm [20]. Zhang et al. propose a new approach to detect burst novel events and predict their future popularity simultaneously. They utilize multiple types of information such as term frequency and user's social relation to detect events from online microblogging stream. Meanwhile, A diffusion model is used to predict the popularity of detected event which takes both the content and user information of the event into account [21]. The influenza and foodborne disease could also be detected by the Twitter or Weibo data. The response time is proved to be faster than the traditional surveillance systems.

As the foodborne diseases are mostly caused by contaminate food, the twitter data can also help with the identification of restaurants related with foodborne diseases. Sadilek et al. combine the Twitter data with foodborne disease and restaurants [22]. They collect the tweets posted in mobile devices and get the geolocation of the tweets, and then find the tweets which are near to some restaurants. After that, they design a human guided machine learning method to classify the foodborne disease events. Finally, the detected tweets are associated with the restaurants. Their application is deployed in Las Vegas [23]. Their method performs well comparing with the health department. However, they only collect the mobile data with specified geo-locations. For the data which do not contain specific geo-locations, their method does not work. In this paper, we will design an algorithm to detect the foodborne disease event from Weibo data. The algorithm is novel and efficient. We use not only the mobile data, but also all the Weibo data. We design a method to identify the location for the foodborne disease event. The method is novel. This is a new try to determine the location of the event for non-mobile users.

The paper is organized as follows. Section 1 introduces the background and Section 2 shows some related works. In Section 3, the process for the foodborne disease event detection will be illustrated step by step. Section 4 will show the experiment results and we will conclude in Section 5.

Table 1
The access frequency for the users of social network applications in China.

Frequency	Percentage(%)
Several times per day	38.2
One time per day	20.3
3–6 times per week	15.4
1–2 times per week	13.1
One time for several weeks	7.0
Less than one time per month	6.1

3. The method

3.1. Data preprocessing

Weibo, which is a Chinese social network application similar to Twitter, has a large number of users in China. Until December 2013, there are 45.8% Chinese people, that is, 618 million of them surf on the Internet. Weibo is popular in the Chinese people. All the contents of Weibo are user generated and the users are very active. Table 1 shows the access frequency for the users of social network applications in China. We can see that more than 50% users use the social network at least once in a single day. This makes sure that the real time data from social media is large and it is suitable for our event detection research.

To get the relevance of foodborne diseases and the restaurants with contaminated food, we should first get the tweets with the information of foodborne diseases. There is a government surveillance system for the foodborne diseases which collects foodborne disease cases in 13 provinces in China from 2013. Each case is reported by the doctors from the sentinel hospital. The reported cases contain the location, time, symptoms and the treatment for the diseases. We analyze the symptoms of these cases and get a list of keywords for them. To use the most representative symptoms to describe the diseases, the frequency of appearance for each keyword is calculated and the keywords with high frequency are selected. The keywords in the surveillance system are reported by the doctors and usually in medical terms. However, when people post messages in Weibo, they will usually use oral expressions. To overcome this problem and improve the accuracy of keyword identification, we transfer the keywords in medical terms into more oral words and ask some medical experts to do a double-check. These keywords for the symptoms of foodborne disease give the

hint for the tweet selection. In the following part of this paper, the keywords for the symptoms of foodborne disease mean the selected words which have been transferred into oral format.

With the help of Weibo API, we then crawl the tweets in Weibo which contain the keywords of the symptoms for the foodborne diseases. While the tweets contain huge amount of information, there are also lots of noisy data. The extracted tweets may be from the public accounts which offer health advices. For example, "diarrhea" is a keyword for foodborne disease symptoms. Users might make tweets such as "How to avoid diarrhea in summer". In this case, the tweet is noisy for the event detection. Moreover, although some tweets contain the keywords of the symptoms for the foodborne disease, they may not be real time. For example, tweets like "I got a diarrhea the day before yesterday" contains the keywords, but the real-time nature is not fulfilled. This case is also considered as noise in the foodborne disease event detection. Therefore, it is necessary to distinguish the real foodborne disease cases which is denoted as a positive class. To reduce the effect of the useless tweets, we build a support vector machine (SVM) classifier to filter the tweets, which is based on some properties of the tweets and the users. We first manually select examples of positive class and negative class as the training set. Then we use the ten features for each tweet to train the SVM classifier, which are the number of followings of the user, the number of followers of the user, the length of personal description, the number of all tweets of the user, the number of average retweeting,the number of average recommendation, the number of average comments, the average length of the tweets, the time of posting the tweet and the number of average links in tweets. When the SVM classifier is bulit, we use it to classify each tweet we crawl from Weibo.

After the Weibo data are filtered, we should identify the foodborne disease event.

3.2. Foodborne disease event detection

As the contents of tweets are texts, we should first convert the natural language into mathematical format in order to process the tweets using machine learning methods. Constructing the vector representation for the words is a good way to capture the syntactic and semantic word relationships. The toolkit word2vec is an open source software developed by Google, which is used to convert the words into vectors. It is efficient while billions of words could be trained in a day, which makes it possible for us to train our data from Weibo. We construct the vector representations for the words in tweets using word2vec. Then the sematic similarities between the tweets could be calculated by the similarities in the vector space [24,25].

There are huge amount of data continuously generated in Weibo and the topics change fast. As the content of one tweet is short, the information for one particular tweet is limit. If we only extract the tweets which contain the keywords for the symptoms of foodborne diseases, the number of tweets may be too few. And due to the short content of one particular tweet, it is difficult to detect the foodborne disease event only with the tweets extracted by the keywords for the symptoms of the foodborne diseases. Moreover, except for the tweet containing the keywords for the symptoms of foodborne diseases, the tweets in its context may contain other important information for the foodborne diseases, such as the location where the foodborne diseases are caused. Assuming the tweets for a user is a tweet list $S = \{T_1, T_2, \ldots, T_k, \ldots, T_n\}$, where T_k is the tweet which contains the keywords for the foodborne disease symptoms and T_i is posted earlier than T_j if $i < j$. For a user whose tweet is selected according to the keywords, we can simply get more tweets by extracting the m tweets which are before and after the tweet which contain the keywords. The candidate tweet set is denoted as formulas (1).

$$C = \{T_{k-i+1}, T_{k-i+2}, \ldots, T_k, \ldots, T_{k+j}\}, 0 < i < k, k < j < m \quad (1)$$

We call this method as the fixed context window selecting. The fixed tweet window $[T_p, T_q]$ which contains the p tweets before T_k and the q tweets after T_k. The algorithm is shown in Algorithm 1.

Algorithm 1 Fixed context window selecting.

Input: A tweet list for a user $S = \{T_1, T_2, \ldots, T_k, \ldots, T_n\}$; The tweet T_k which contains the keywords for the foodborne disease symptoms; The upper bound P for the fixed window and the lower bound Q for the fixed window.

Output: The candidate tweet set C

```
Initialize T = T_k, C = ∅
Push T_k into C
for i=1 to P do
  Push T_{k-i} into C
end for
Update T = T_k
for j=1 to Q do
  Push T_{k+j} into C
end for
return C
```

However, Algorithm 1 does not consider the semantic similarity between the tweets. The extracted tweets may contain no relevant information with the foodborne disease but introduce too much noise.

To solve the above problems, we design an algorithm to dynamically choose more relevant tweets in the context to get a larger candidate corpus. We carry on a tokenization for the tweets and const d in formulas (2).

$$Sim(T_i, T_j) = \cos(v_i, v_j) \quad (2)$$

The similarity between two tweets gets higher while the value of cosine is larger. We compute a dynamic window in the context for a particular tweet T_k which contains the keywords for the foodborne disease symptoms. The algorithm is shown as Algorithm 2.

For a tweet T_i, we compute the similarity between T_k and T_i. If the similarity is greater than the threshold U, the tweet T_i will be added into the candidate tweet set C. We find the similar tweets with T_k before and after it. The threshold U is first computed as the average of the similarities between all the tweets and T_k. The decreasing rate η is a constant rate. Our method takes account of the semantic similarity between the tweets and makes sure that the tweets in the candidate set are relevant to the foodborne disease. At the same time, less noisy data are introduced.

After the candidate tweet set is built, we try to extract the key-phrases for the tweets. The easiest method is based on TF/IDF. But it only considers the statistical properties of the phrases. The relationships between the phrases are not considered and the phrases with low frequency will be ignored. Therefore, the TF/IDF method is not suitable for the Weibo data, which are short for a particular tweet and the phrases vary frequently. In this paper, we use TextRank, which is a graph based key-phrase extraction algorithm [26], to extract the key-phrases for the tweets. The basic idea of TextRank is similar to that of PageRank algorithm in the field of information retrieval. It divides the text into several segments and builds the graph model for them. The voting schema is used to rank the phrases of the text and the key-phrases are extracted according to the rank.

We manually label some tweets which are indeed related with foodborne disease. And then we use the extracted key-phrases for the tweets together with the keywords for the symptoms of the foodborne disease to train a SVM classifier to detect the foodborne disease event.

Algorithm 2 Dynamic context window calculation.

Input: A tweet list for a user $S = \{T_1, T_2, \ldots, T_k, \ldots, T_n\}$; The tweet T_k which contains the keywords for the foodborne disease symptoms; The threshold U for the similarity between tweets; The decreasing rate for the similarity measure η; The upper bound P for the dynamic window and the lower bound Q for the dynamic window.
Output: The candidate tweet set C

```
Initialize T = T_k, C = ∅
Push T_k into C
for i=1 to P do
  if Sim(T, T_{k-i}) > U then
    Push T_{k-i} into C
    Update T = T + T_{k-i},Update U = U * η
  else
    break;
  end if
end for
Update T = T_k
for j=1 to Q do
  if Sim(T, T_{k+j}) > U then
    Push T_{k+j} into C
    Update T = T + T_{k+j},Update U = U * η
  else
    break;
  end if
end for
return C
```

3.3. Location determination for foodborne disease event

When the foodborne disease events are detected, we hope to determine the location of the events. Then we could refer the locations or the restaurants which are related with the foodborne diseases. We can then associate the results with the restaurant recommendations.

To find the restaurants which are related with the foodborne diseases, the geo-location for the foodborne disease event should be determined. There may be location description in the registration information for the Weibo users. But this location is the zone where the user lives, not exactly the location of the foodborne disease event occurs. There are also GPS data for some mobile users, but the data are also sparse. On the other hand, we observe that lots of the tweets related with foodborne disease contain the information of the restaurant or the name of the food. And some users also refer to the location when they post a tweet. We utilize this information with the aid of other data to find the restaurants related to the foodborne disease.

We define the restaurant and food information in the tweet as information A. If the tweet contains the restaurant, we will get the location of the restaurant from the website of dianping (https://www.dianping.com/), which has the information for the restaurants and their detailed locations. If the tweet contains the name of the food, we could find all the restaurants with this food and their location from Baidu API. The tweets may also contain the geo-location when they are posted. We define this kind of information as information B. We will use the administrative location data on tcmap (http://www.tcmap.com.cn) to get the detailed location for information B. The location in the registration information of the Weibo users are defined as information C. Note that information A and B may be missed, but information C is usually available. We then design an algorithm which utilizes the information A, B and C to get the location L of the foodborne disease events. The location L is determined as in formulas (3).

$$\begin{cases} L = \{A_i | minDist(A_i, B_j), A_i \in A, B_j \in B\}, & A \neq \emptyset, B \neq \emptyset \\ L = \{A_i | A_i \in C, A_i \in A\}, & A \neq \emptyset, C \neq \emptyset, B = \emptyset \\ L = \{B_j | B_j \in C, B_j \in B\}, & B \neq \emptyset, C \neq \emptyset, A = \emptyset \end{cases} \tag{3}$$

When A and B are both available, we compute the nearest two points $A_i \in A$ and $B_j \in B$ and A_i is the desired location. The algorithm for location determination when A and B are both available is shown in Algorithm 3.

Algorithm 3 Location determination(1).

Input: Three kinds of location information A, B and C.
Output: The location for foodborne disease event L.

```
Initialize L = ∅, D = Max
if A ≠ ∅ and B ≠ ∅ then
  for A_i ∈ A do
    if DISTANCE(A_i, B_j) < D then
      D = DISTANCE(A_i, B_j)
      L = A_i
    end if
  end for
end if
return L
```

When A and C are available and B is empty, the locations in A which are also in C are the desired locations. The algorithm for location determination in this case is shown in Algorithm 4.

Algorithm 4 Location determination(2).

Require: Three kinds of location information A, B and C.
Ensure: The location for foodborne disease event L.

```
Initialize L = ∅
if A ≠ ∅ and B = ∅ then
  if A_i ∈ C then
    L = A_i
  end if
end if
return L
```

When B and C are available and A is empty, the locations in B which are also in C are the desired locations.

After the locations for the foodborne disease events are determined, we find all the restaurants which are related with the foodborne disease events. We give a score for each restaurant and the score is lower for a restaurant related with the foodborne diseases. Then we insert the foodborne disease related information as a factor in the restaurant recommendation algorithm.

4. Experiments

In the experiments, we extract the tweets from 933,313 users in Beijing, China which contain the 31 keywords for the symptoms of foodborne diseases. The tweets are posted between August 2014 and October 2014. We use Algorithm 1 to get the candidate tweet set and set $m = 200$. That means we extract the 200 tweets before and after the tweet containing the keywords for the symptoms as the tweet set. There are totally about 80 million tweets in the set.

We use the ten features in Table 2 to construct a SVM classifier to filter the tweets.

After data filtering, we get about 31% of the tweets in the candidate tweet set.

We randomly select parts of the data and manually label some key-phrases which are related with foodborne diseases. We use

Table 2
The features for data filtering.

#Followings	#Followers	Length of personal description
#All tweets	#Average retweeting	#Average recommendation
#Average comments	Average length of the tweets	Time of posting
#Average links in tweets		

Table 3
Comparison of algorithms for key-phrase extraction with fixed context window.

Method	Total number	Correct number	Accurate rate(%)
TF/IDF	2000	423	21.15
TextRank	2000	644	32.2

Table 4
Comparison of algorithms for key-phrase extraction with dynamic context window.

Method	Total number	Correct number	Accurate rate(%)
TF/IDF	2000	478	23.9
TextRank	2000	706	35.3

the TF/IDF and TextRank to extract the key-phrases respectively. If the extracted phrases are 80% correct, we say that the extracted phrases are key-phrases. When we use the fixed context window algorithm to select the tweet set, the accurate rate for the TF/IDF and TextRank for key-phrase extraction is shown in Table 3.

From Table 3 we can see that for TF/IDF method, there are 423 tweets in 2000 tweets which can get correct key-phrases. The accurate rate is 21.15%. But for TextRank algorithm, 644 in 2000 tweets can get correct key-phrases. The accurate rate is 32.2%. TextRank can get higher accurate rate for key-phrase extraction for algorithm 1 which selects a fixed number of tweets before and after the tweet T_k which is foodborne disease related.

We construct the word vectors for the tweets using word2vec and then use Algorithm 2 to dynamically choose the candidate tweet set. We also compute the accurate rate for TF/IDF and TextRank in this case when dynamic context window is calculated to generate the tweet set. The result is shown in Table 4.

From Table 4 we can see that for TF/IDF method, there are 478 tweets in 2000 tweets which can get correct key-phrases. The accurate rate is 23.9%. But for TextRank algorithm, 706 tweets in 2000 tweets can get correct key-phrases. The accurate rate is 35.3%. TextRank can also get higher accurate rate than TF/IDF when dynamic context window is selected for tweet set. From Tables 3 and 4, we can see that when we use TextRank algorithm to extract key-phrases for tweets and use dynamic context window calculation algorithm for tweet set selection, we can get higher accurate rate for the foodborne disease event detection. When we use dynamic context window calculation algorithm for tweet set selection, the topics for the candidate tweets will be more similar, and the noise will be less. So the dynamic context window calculation algorithm with TextRank will give better results and is more suitable for foodborne disease event detection.

We also design algorithms for location determination. We utilize the content of the tweets, the data on dianping, tcmap and Baidu API to determine the location of the foodborne disease events. We make statistics to the information *A, B, C* mentioned in Section 3.3. There are 13% of all the tweets which contain both *A* and *B*, 19% containing *A* and *C* but not *B*, and 16% containing *B* and *C* but not *A*. For the case where *A, B* and *C* are all available, we select part of the tweets and manually label the geo-location of the tweets. Then we calculate the location for the tweets using our algorithm. We compare the results of our algorithm and the manually correct result and compute the accurate rate for our location determination algorithm. The result is shown in Table 5.

Table 5
The accurate rate for location determination.

Total number	Correct number	Accurate rate(%)
500	332	66.4
1000	647	64.7
1500	1009	67.3
2000	1280	64.0

From Table 5 we can see that our location determination algorithm can get about 65% accurate rate.

From all the above experiment results, we see that using TextRank algorithm for key-phrases extraction and dynamic context window calculation algorithm for tweet set selection can get better results than the baseline method of TF/IDF and fixed context window selection. After the data preprocessing, we use the dynamic context window calculation algorithm to extend the data source. This makes sure the higher accurate rate.

5. Conclusion

In this paper, we show how to use the social media data to extract valuable information about foodborne diseases. Algorithms are designed to get the foodborne disease related tweets and detect the foodborne disease events. The algorithm is novel and efficient. We use not only the mobile data, but also all the Weibo data. We design a method to identify the location for the foodborne disease event. The method is novel. This is a new try to determine the location of the event for non-mobile users. The experiments show that our methods are effective. However, the accurate rate for the event detection is not high due to the sparsity and continuously changed topics of Weibo. In the future work, we will try to improve the detection algorithm and use the algorithm to assist the restaurant recommendation.

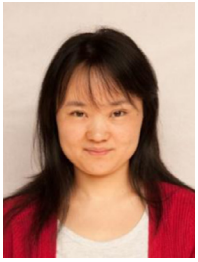
Acknowledgments

This work was supported by The National Key Research and Development Plan under Grant No. 2016YFB1000605 and 2016YFB0501901, the National Natural Science Foundation of China under Grant No. 61402435,41371386,91224006, The Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA06010307), Special Research Funding of National Health and Family Planning Commission of China (No. 201302005), and the Knowledge Innovation Program of Chinese Academy of Sciences under Grant No. CNIC_QN_1507.

References

[1] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734–749.
[2] L. Sharma, A. Gera, A survey of recommendation system: research challenges, Int. J. Eng. Trends Technol. 4 (5) (2013) 1989–1992.
[3] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer (8) (2009) 30–37.
[4] X. Su, T.M. Khoshgoftaar, A survey of collaborative filtering techniques, Adv. Artif. Intell. 2009 (2009) 4.
[5] Y. Shi, M. Larson, A. Hanjalic, Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges, ACM Comput. Surv. 47 (1) (2014) 3.

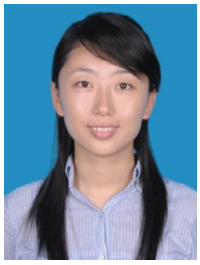
[6] H. Xie, Q. Li, X. Mao, X. Li, Y. Cai, Y. Rao, Community-aware user profile enrichment in folksonomy, Neural Netw. 58 (2014) 111–121.
[7] H. Xie, Q. Li, X. Mao, X. Li, Y. Cai, Q. Zheng, Mining latent user community for tag-based and content-based search in social media, Comput. J. 57 (9) (2014) 1415–1430.
[8] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with twitter: what 140 characters reveal about political sentiment., ICWSM 10 (2010) 178–185.
[9] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 851–860.
[10] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: detecting influenza epidemics using twitter, in: Proceedings of the Conference on Empirical methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1568–1576.
[11] A. Signorini, A.M. Segre, P.M. Polgreen, The use of twitter to track levels of disease activity and public concern in the US during the influenza a H1N1 pandemic, PloS ONE 6 (5) (2011) e19467.
[12] A. Culotta, Detecting influenza outbreaks by analyzing twitter messages, arXiv:1007.4748 (2010).
[13] J. Gomide, A. Veloso, W. Meira Jr, V. Almeida, F. Benevenuto, F. Ferraz, M. Teixeira, Dengue surveillance based on a computational model of spatio-temporal locality of twitter, in: Proceedings of the 3rd International Web Science Conference, ACM, 2011, p. 3.
[14] Centers for Disease Control and Prevention (CDC), CDC estimates of foodborne illness in the United States, Retrieved March 23 2011.
[15] R.W. Newkirk, J.B. Bender, C.W. Hedberg, The potential capability of social media as a component of food safety and food terrorism surveillance systems, Foodborne Pathog. Dis. 9 (2) (2012) 120–124.
[16] J.K. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen, J. Bhatt, S. Brown, Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013–2014, Morb. Mortal. Wkly. Rep. 63 (32) (2014) 681–685.
[17] H. Xie, D. Zou, R.Y. Lau, F.L. Wang, T.-L. Wong, Generating incidental word-learning tasks via topic-based and load-based profiles, IEEE MultiMedia 23 (1) (2016) 60–70.
[18] H. Xie, X. Li, T. Wang, L. Chen, K. Li, F.L. Wang, Y. Cai, Q. Li, H. Min, Personalized search for social media via dominating verbal context, Neurocomputing 172 (2016) 27–37.
[19] L. Yu, X. Sun, Z. Huang, Robust spatial-temporal deep model for multimedia event detection, Neurocomputing 213 (2016) 48–53.
[20] Y. Bin, Y. Yang, F. Shen, X. Xu, Combining multi-representation for multimedia event detection using co-training, Neurocomputing 217 (2016) 11–18.
[21] X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, J. Xia, Event detection and popularity prediction in microblogging, Neurocomputing 149 (2015) 1469–1480.
[22] A. Sadilek, S. Brennan, H. Kautz, V. Silenzio, nemesis: Which restaurants should you avoid today? in: First AAAI Conference on Human Computation and Crowdsourcing, 2013.
[23] A. Sadilek, H. Kautz, L. DiPrete, B. Labus, E. Portman, J. Teitel, V. Silenzio, Deploying nemesis: preventing foodborne illness by data mining social media, The AAAI Conference on Artificial Intelligence (AAAI) (2016) 3982–3990.
[24] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations (ICLR), Scottsdale, Arizona, USA, 2013.
[25] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[26] R. Mihalcea, P. Tarau, Textrank: bringing order into texts, in: Proceedings of the Empirical Methods of Natural Language Processing (EMNLP), 2004, pp. 404–411.



Wenjuan CUI is an Assistant Professor in the Computer Network Information Center, Chinese Academy of Sciences. She received her BSc (2009) from Shandong University and Ph.D. degree (2013) from City University of Hong Kong. Her research interests include data mining, recommender systems, semantic analysis and Bioinformatics. She has published papers in several journals and conferences.

Pengfei Wang is a student of Chinese Academy of Sciences. He is pursuing his Ph. D. degree. His research interest is recommender system and machine learning.

Yi Du received his BSc from Shandong University and Ph.D. degree from Institute of Software, Chinese Academy of Sciences. He is currently an assistant professor in the Computer Network Information Center, Chinese Academy of Sciences. His research interest is visual analytics.



Xin Chen received her Ph.D. degree from Institute of Software, Chinese Academy of Sciences. She is currently an assistant professor in the Computer Network Information Center, Chinese Academy of Sciences. Her research interest is computer graphics.

Danhuai Guo is an associate professor in the Computer Network Information Center, Chinese Academy of Sciences. His research interest is geoinformatics, data mining and visualization.

Jianhui Li is a professor in the Computer Network Information Center, Chinese Academy of Sciences. His research interest is massive data processing, cloud computing.

Yuanchun Zhou is a professor in the Computer Network Information Center, Chinese Academy of Sciences. His research interest is data mining, cloud computing and massive data processing.