Pers Ubiquit Comput (2015) 19:1125–1132
DOI 10.1007/s00779-015-0877-5

ORIGINAL ARTICLE

Chinese social media analysis for disease surveillance

Xiaohui Cui[1] · Nanhai Yang[1] · Zhibo Wang[1,2] · Cheng Hu[1] · Weiping Zhu[1] · Hanjie Li[1] · Yujie Ji[1] · Cheng Liu[3]

Received: 2 January 2015 / Accepted: 1 May 2015 / Published online: 11 September 2015

Abstract It is reported that there are hundreds of thousands of deaths caused by seasonal flu all around the world every year. More other diseases such as chickenpox, malaria, etc. are also serious threats to people's physical and mental health. There are 250,000–500,000 deaths every year around the world. Therefore proper techniques for disease surveillance are highly demanded. Recently, social media analysis is regarded as an efficient way to achieve this goal, which is feasible since growing number of people have been posting their health information on social media such as blogs, personal websites, etc. Previous work on social media analysis mainly focused on English materials but hardly considered Chinese materials, which hinders the application of such technique to Chinese people. In this paper, we proposed a new method of Chinese social media analysis for disease surveillance. More specifically, we compared different kinds of methods in the process of classification and then proposed a new way to process Chinese text data. The Chinese Sina micro-blog data collected from September to December 2013 are used to validate the effectiveness of the proposed method. The results show that a high classification precision of 87.49 % in average has been obtained. Comparing with the data from the authority, Chinese National Influenza Center, we can predict the outbreak time of flu 5 days earlier.

✉ Xiaohui Cui
xcui@whu.edu.cn

[1] International School of Software, Wuhan University, Wuhan 430079, China

[2] Software College, East China Institute of Technology, Nanchang 330013, China

[3] Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Keywords Social media · Chinese · SVMLIGHT · Classification · Prediction · Flu

1 Introduction

With the popularity and development of Internet, new social media such as blogs, personal websites, instant messages, and etc. have been greatly changing people's life. These social media propel the spread of social news, public opinion and personal daily information in human's society, playing an important role in current information dissemination. According to the survey of iResearch (as shown in Fig. 1), time that people spend on social media reaches 4.6 h every week by the end of September 2009 [1], and it will keep increasing in the future [2].

Since people spend so much time on social media, it is worthy to utilize them to uncover and collect various kinds of health information. There are many researchers engaging in the analysis of English social media for disease surveillance or related works. In 2006, a website called "who is sick" is developed for people to post their sickness information [3]. In 2009, Ginsber [4] predicted flu 1–2 weeks earlier than Centers for Disease Control (CDC) through analyzing the log files of Google search. After that, more researchers predicted the outbreak time of diseases by using internet data [4–6]. In 2010, Lampos [7] analyzed Twitter, wrote a tool that could automatically track flu, then compared with the same period in Health Protection Agency (HPA), and obtained the correct results. Ficeifeld et al. [8] collected users' health information through an application installed on mobile phones and then detected diseases. Many scientists have been doing various experiments which are using support vector machine (SVM) and Naïve Bayes (NB) or other technologies to classify data

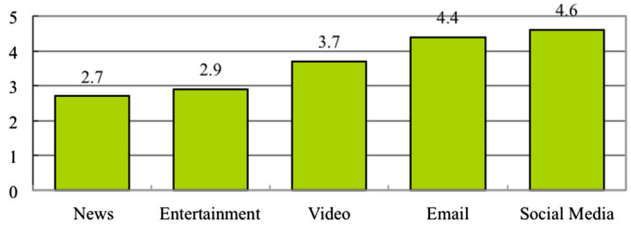

Fig. 1 Global Internet users' online time assignments in 2010

into different areas. Sadilek [9, 10] analyzed the contents of Twitter of the users and their friends, and then used them to predict the users' body health status. Kaundal [11] used machine learning technique in disease forecasting. However, few researches relating to Chinese social media have been finished. Jin [12] used Flickr for prediction and forecast. Zhengyan [13] classified short text into different kinds of categories (sports, news etc.) using K-nearest neighbor (KNN). Yang [14] proposed a method of automatic detection of rumor Sina micro-blogs. These works are not for disease surveillance. To the best of our knowledge, there are no researches on disease surveillance through mining Chinese social media.

This paper aims to predicate people's health status in a region of Beijing based on Sina micro-blog, a famous social medium in China. We collect the related data and propose an effective classification method for such purpose. As a result, we can classify the micro-blog data with 89.77 % precision and 88.76 % recall in average.

This work is an important step toward predicting disease based on Chinese social media. It explores the process of classification of short text and compares different kinds of method in the process of classification to get a better result. It also provides foundation for researches on predicting disease by analyzing social media information.

The rest of the paper is organized as follows: Sect. 2 presents the procedure of this research. Section 3 describes how to get the data and the character of this data. Section 4 discusses our simulation environment and describes the method used in this research procedure and the experiment result. Finally, Sect. 5 concludes the paper.

2 Research procedure

We mainly work out our prediction according to our classification results. After we have got the classification results, we analyze them in a statistical way and present with line charts. We then compare the charts with the line chart whose data come from weekly reports on the official website of National Influenza Center. After the credibility of the classification is proved, we can successfully finish our disease surveillance work.

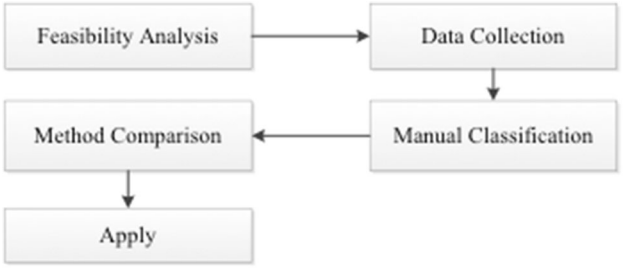

Fig. 2 Research procedure

Our research flow is shown in Fig. 2. We begin with feasibility analysis and find no researches have been done in disease surveillance on Chinese Social Media and the amount of data we can get from blogs can be enormous which brings much difficulty to our research. Then we start to collect data from Sina and have the data cleaned. The biggest challenge during this step is natural language processing. After this procedure, we obtain micro-blogs without urls, '@' symbols, emoji expressions or some other special symbols. And then we select sample data randomly and classify them manually. We then build upon previous work on classification of text messages (K-means, KNN, SVM), to get a better result with a relatively better classifier and apply those methods to our classification. Since some of the algorithms, such as SVM, contain too many parameters, we have to repeat the experiment for many times with different combination of different parameters to get a conclusion with high accuracy.

3 The data

Micro-blog is a kind of blog service through which people can post their messages with no more than 140 characters. It enables users to express their thoughts briefly and encourages frequent information updates. In China, there are mainly four such kind of micro-blogs: Sina micro-blog (http://weibo.com/), Tencent micro-blog (http://t.qq.com/), NetEase micro-blog (http://t.163.com/) and Sohu micro-blog (http://t.sohu.com/). According to the analysis of google trend from 2010 (Fig. 3), Sina micro-blog is the most popular Chinese social media, so our analysis and evaluation are based on the data obtained from Sina micro-blog.

Sina provides convenient API for obtaining data, such as 'Public Micro-blog', 'Location Based Micro-blog', 'Location Nearby Micro-blog', and etc. 'Public Micro-blog' provides all the micro-blogs without location information; 'Location Based Micro-blog' provides geotag micro-blog in specific place, and it cannot cover most of the micro-blog; 'Location Nearby Micro-blogs' provides geotag

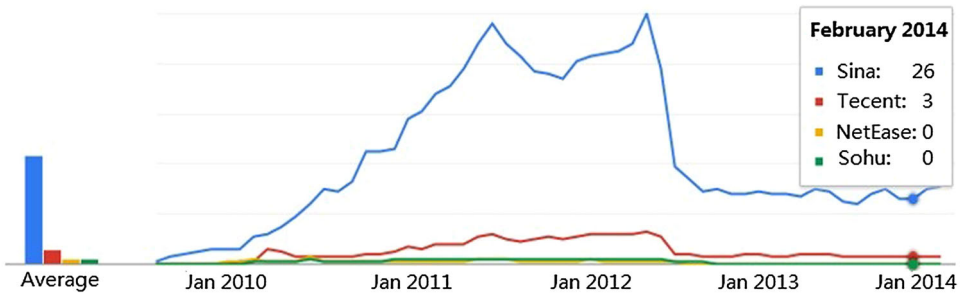

Fig. 3 The different micro-blogs' using trend in China

Table 1 Parameters of 'location nearby micro-blog'

Parameter	Meaning
Lat	Latitude
Long	Longitude
Range	Radius of search range
StartTime	Time start to obtain data, expressed as UNIX timestamp
EndTime	Time end to obtain data, expressed as UNIX timestamp

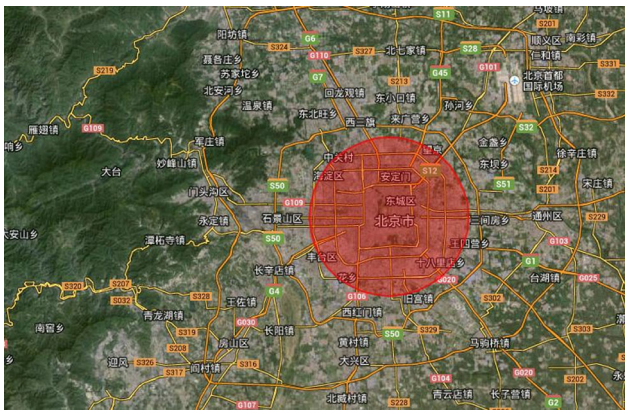

Fig. 4 Places and scopes to obtain data

micro-blog, and it can cover the micro-blogs over the area we choose.

We choose 'location nearby micro-blog' (Its main parameters for the APIs are listed in Table 1) because we want to get all the micro-blogs in a specific area with location information [15]. We choose a circle area in Beijing as shown in Fig. 4.

Using 'location nearby micro-blog' API and JavaScript, we collect Sina micro-blogs in that circle area (longitude: 116.39750833333, latitude: 39.908647222222, range 11,120) from September 2013 to December 2013. There are 3505110 pieces of micro-blog in total which includes 951,299 pieces of micro-blog in September, 900,337 pieces in October, 861,590 pieces of micro-blog in November, and 791,884 pieces in December. Our dataset relates to

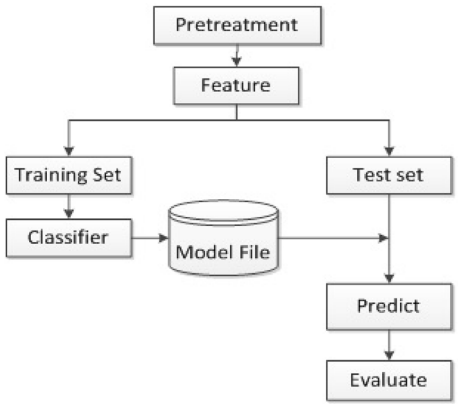

Fig. 5 Classification procedures

374411 people, so 4.4 pieces of micro-blogs for per person in average. [16] In that about a million pieces of data need to be checked, our experiment uses a pre-treatment function to filter data [17].

We select 5000 pieces of status randomly and classify them manually into two categories: one is 'sick micro-blog' which indicates that the author is sick; the other one is 'not sick micro-blog' which indicates that the author is not sick. Among these micro-blogs, we get 285 'sick micro-blog'. Then we select 285 'not sick micro-blog' as training data and test data.

4 Approach

In this section, we follow the steps illustrated in Fig. 5 to conduct the prediction. Firstly, we conduct some preprocessing to eliminate the noise information and then use text model to express the content of every pieces of micro-blog. Secondly, we train the classifier using training data and record the results into model file. After that, we predicate the flu based on the training model and the test data. Finally, we evaluate our proposed method and report the results.

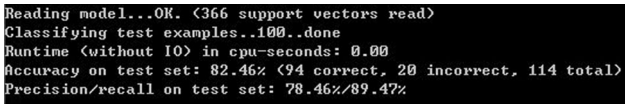

Fig. 6 Result of the classification using character as textual feature

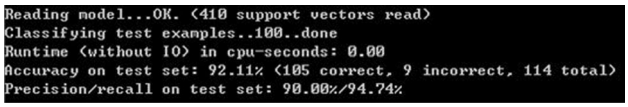

Fig. 7 Result of the classification using word as textual feature

4.1 Textual feature

One important different of text processing between English and Chinese is the feature of extracting text. Although there are many Chinese word segmentation systems which can provide word segmentation with a high accuracy, it is not easy to determine whether using word or character as textual features.

This paper uses the same dataset and the same classifier to compare these two kinds of features. According to our result shown in Figs. 6 and 7, using word as features can get a higher precision and recall, which reaches 0.90 precision and 0.947 recall compared with 0.785 precision and 0.895 recall by using character as features. So we use word as features, for example: a micro-blog 'I am sick' is represented by following feature vector:

(我感冒了) = (我,感冒,了)
Means: (I am sick) = (I, am, sick)

4.2 Word weighting

Since word is the advanced language used by human, human can gain the information from the text, but computers cannot. So it is necessary to convert the text message into the message computers can understand, and we need to pre compute the weight of different word. Word weighting method is another problem which should be considered when we start to classify micro-blogs. There are four kinds of word weighting method: Boolean weighting, term frequency weighting (TF), inverted document frequency weighting (IDF) and term frequency-inverted document frequency weighting (TFIDF). Boolean weighting does not consider the importance of each word, and term frequency weighting does not consider the entire corpus, so this paper compare IDF with TFIDF, the accuracy of TFIDF (Fig. 8) is 92.11 % which is much higher than IDF (accuracy = 64.91 %) (Fig. 9), so this paper uses TFIDF as a method of word weighting [18–20].

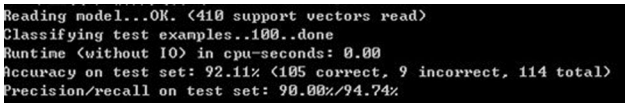

Fig. 8 Result of the classification using TFIDF as word weighting

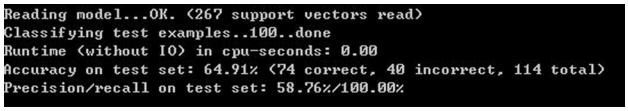

Fig. 9 Result of the classification using IDF as word weighting

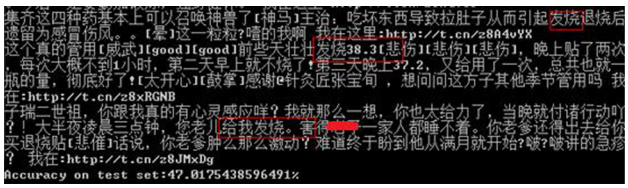

Fig. 10 Result of K-means classification

4.3 Classifier

For the classification, machine learning is widely used in this area. Machine learning can be mainly classified into two categories: Supervised learning and unsupervised learning. SVM and K-means are two of the main methods in the machine learning area. And these algorithms are improved by many intelligent computer scientists like Deng et al. [21]. In this paper we select SVM as an example of supervised learning and K-means as an example of unsupervised learning to compare their performance. We uses 570 classified data to compare their performance for our problem.

As a result (shown in Fig. 10), a class does not belong to sick contains more sick micro-blog, the classification accuracy rate is low. Because of the text for micro-blog is colloquial, and if its category does not clear, the use of K-means unsupervised machine learning cannot achieve good results. K-means belongs to unsupervised learning, clustering center is unable to control, unless has obvious category relationship, K-means cannot obtain the good effect.

As shown in Table 2. we get the accuracy of 89.77 % for SVM and 70.44 % for K-means. We think the reason for that is we can provide specific labels for SVM in this problem which facilitates the classification and has a better

Table 2 The performance of K-means and SVM

Algorithm	Evaluation (%)	
	K-means	SVM
Precision	70.44	89.77
Recall	49.98	88.76
F1-measure	58.47	89.21

Table 3 KNN classification confusion matrix

Manual	Classifier	
	Sick	Not sick
Sick	17	40
Not sick	2	55

performance. Therefore we adopt SVM for our classification.

We then use K-nearest neighbors (KNN) as a classifier (as shown in Table 3) to compare with SVM.

K-nearest neighbors aims at finding k-nearest class:

$$p(d_{\text{new}}, c_i) = \sum_{k=1}^{m} \text{Similarity}(d_k, d_{\text{new}}) \times y(d_k, dc_i)$$
$$(d_k, dc_i) = \begin{cases} 1 & d_k \text{ belongs to } c_i \\ 0 & d_k \text{ not belongs to } c_i \end{cases} \quad (1)$$

d_{new} stands for a new document need to be classified. d_k stands for the kth document of the corpus.

Support vector machine aims at finding a hyper plane to classify samples:

$$\text{svm}(x) = \text{sgn}\left\{\sum_{i=1}^{N} a_i y_i K(x_i \cdot x) + b\right\} \quad (2)$$

where sgn(.) stands for sign function, and k(.) stands for the kernel function of SVM, a_i is determined by slack variables, y_i stands for the label of x_i, x stands for input text, b is determined by penalty factor [22–25].

Using KNN [26], we achieve 63.15 % precision, and the result suggests that KNN is also better than K-means, but is not as well as SVM which achieves 90.00 % precision.

What's worse, KNN has a lower efficiency than SVM when classifying the big data. If the number of micro-blog needed to be classified grows from 1000 to 100,000, the time consumed by KNN to finish this task raises from 9.8 to 524.67 s. However, the time consumed by SVM is always less than 1 s (as shown in Fig. 11).

4.4 Experiment

According to the experiment results above, we decide to use words as features, TFIDF as word weighting method, SVM as classifier, and we use the SVM_LIGHT as a convenient tool for our experiments. This SVM is robust

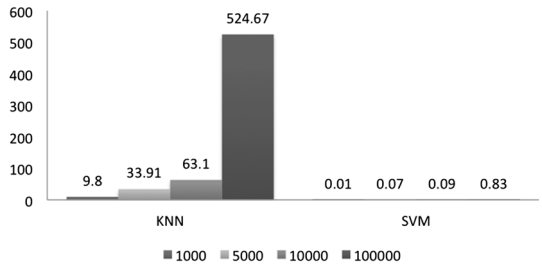

Fig. 11 Time consumed by KNN and SVMLIGHT

and includes algorithm for approximately training large transductive SVMs for big data set.

In text categorization system, it generally divides the corpus into two parts: the training set and the test set. The training set is made up of many pre-classified documents and is using for learning the category attributes. The test set is using for evaluating the classifier by assigning a category to each unclassified article in it. As the training data and test data, we choose.

To prove this classifier is stable and reliable, this paper uses K-Fold cross validation for verification, which randomly divides data into K parts, then take one part as a test set, the remaining $K-1$ parts as a training set, In this experiment, we use fivefold for this test (as shown in Table 4).

From this experiment, we finally achieve 89.77 % precision in average, 88.76 % recall in average and 89.21 % F1 measure in average. These prove that using our classification model can distinguish between 'sick micro-blog' and 'not sick micro-blog'. So we use this model to our dataset from September 2013 to December 2013 (as shown in Table 5; Fig. 12).

According to the positive rates of influenza detection on the official website of National Influenza Center, which show the percentage of influenza like cases in the total number of visits to a hospital, we have got a line chart, as shown in Fig. 13.

Compared with Figs. 12 and 13, these two charts have some similar points. Firstly, from September to October, both charts' dependent variables are all rising. Secondly, those two charts are rising evidently and approach to the top in December. These two charts both predict that the state of flu is urgent in December. As a result, Figs. 12 and 13 have some similarities, which means the credibility of our experiment.

However, we can see that these two charts have some differences. There are some reasons to illustrate this problem shown as following.

Firstly, the source of data in each map is different. For data shown in Fig. 12, it is collected from people who

Table 4 5-Fold classification experiment

Experiment	Evaluation (%)					Average (%)
	1-kfold	2-kfold	3-kfold	4-kfold	5-kfold	
Precision	90.00	91.23	92.31	86.21	89.09	89.77
Recall	94.70	91.23	84.21	87.72	85.96	88.76
F1 measure	92.29	91.23	88.07	86.96	87.50	89.21

Table 5 Classification between September and December 2013

	Status no.	Sick status no.	Ratio (%)
September	935,646	63,284	6.763669
October	885,436	60,874	6.875031
November	848,685	57,296	6.75115
December	780,890	55,826	7.149022

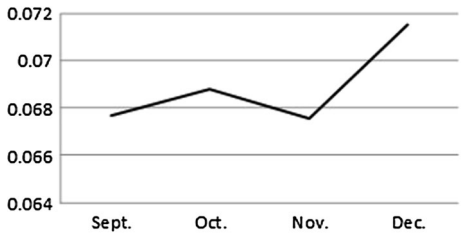

Fig. 12 Classification result between September and December 2013

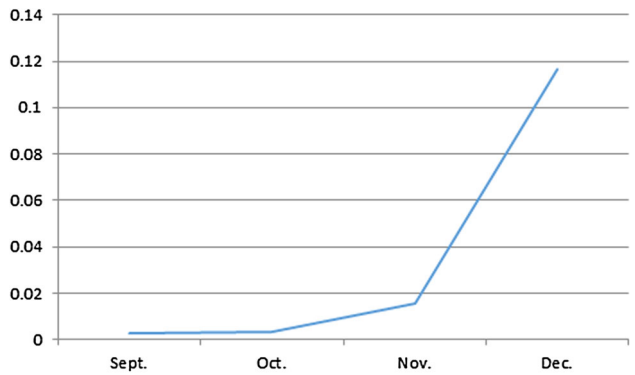

Fig. 13 CNIC report between September and December 2013

publish messages on "Sina micro-blog", and the region that messages came from is around Beijing. However, data shown in Fig. 13, collected by CNIC, is from people who are diagnosed at the hospital, and the region covers Northern China which is larger than Beijing. As a result, the ratios are different.

Secondly, the difference has some relation to the climate. From this two charts, we can get a conclusion that the difference occurs between October and November. When it comes to November, weather is getting colder. As a result,

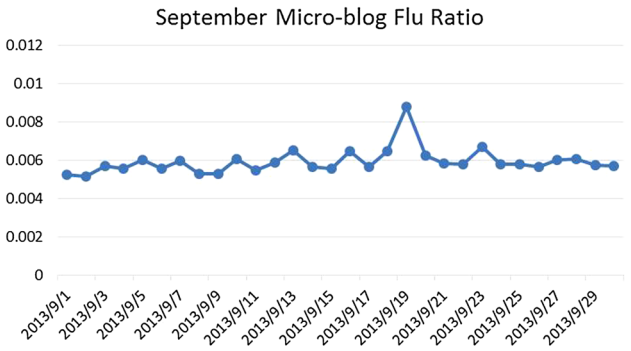

Fig. 14 Classification result in September 2013

the number of flu-patients who publish messages on the "micro-blog" will decrease. Also, because flu is hard to be cured at this time, if people get flu at this time, people are more likely to go to the hospital for treating instead of staying at home and eating some pills. So, according to the data got from micro-blog, the rate of getting flu is smaller. However, in December, people in Beijing start to use heating installation. As a result, compared with November, patients are more likely to publish micro-blog in a warmer condition.

Then this paper uses the same classification model, SVM, to classify the micro-blogs day by day in the 4 months, and gets the result as shown in Figs. 14, 15, 16 and 17.

As shown in Fig. 14, the ratio of 'sick micro-blog' rises from September 17, 2013, and peaks as 0.877 % on September 19, 2013. Compared with the data from China Nation Influenza Center (CNIC), we predict outbreak time of flu 3 days earlier.

As shown in Fig. 15, the ratio of 'sick micro-blog' rises from October 26, 2013, and peaks as 0.722 % on October 31, 2013. Compared with the data from China Nation Influenza Center (CNIC), we predict outbreak time of flu 3 days earlier.

As shown in Fig. 16, the ratio of 'sick micro-blog' rises from November 8, 2013, and peaks as 0.9 % on November 11, 2013. Compared with the data from China Nation Influenza Center (CNIC), we predict outbreak time of flu 6 days earlier.

As shown in Fig. 17, the ratio of 'sick micro-blog' rises from December 15, 2013, and peaks as 5.00 % on December 17, 2013. Compared with the data from China

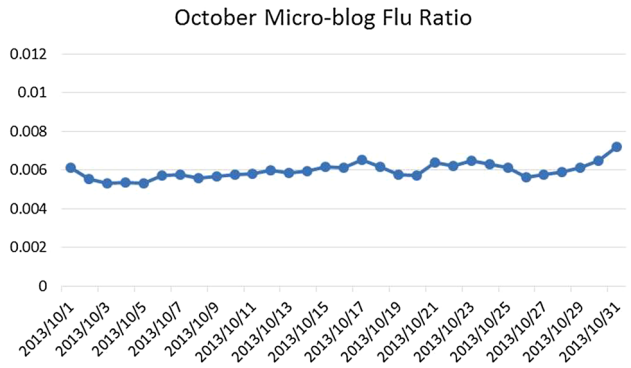

Fig. 15 Classification result in October 2013

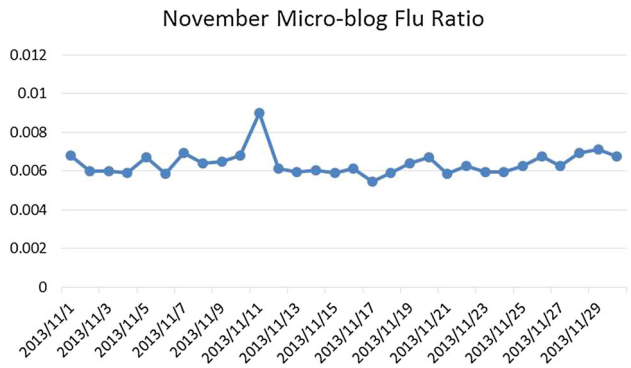

Fig. 16 Classification result in November 2013

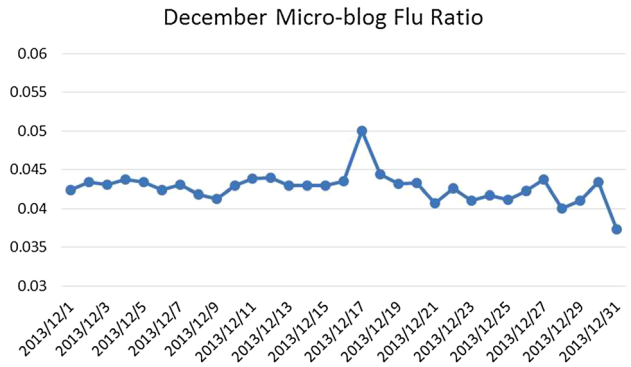

Fig. 17 Classification result in December 2013

Nation Influenza Center (CNIC), we predict outbreak time of flu 5 days earlier. CNIC's data suggest that seasonal influenza achieves a little higher level from December 15, 2013, and rises to a much higher level in this month.

5 Conclusion

In this paper, we investigate how to predict the trend of diseases in the real world based on Chinese social media. We combine social media data with spatio-temporal data and successfully predict outbreak time of flu 5 days earlier than Chinese National Influenza Center. We also think that our method for processing Chinese social media data can be used in other related fields for Chinese big data analysis.

In the future, we need to take more considerations on spatio-temporal data and investigate the flu's influence on individuals for a period time. Since most users of Sina micro-blogs are the generation born in 80s and 90s, we also need to obtain more data from other ages to get more comprehensive results.

Acknowledgments This research is supported in part by National Nature Science Foundation of China No. 61440054, Fundamental Research Funds for the Central Universities of China No. 216-274213, and Nature Science Foundation of Hubei, China No. 2014CFA048. Outstanding Academic Talents Startup Funds of Wuhan University, No. 216-410100003 and 216-410100004.

References

1. IResearch (2010) In 2010 the global Internet users spend most of their time in social media. http://service.iresearch.cn/others//20101129/128573.shtml
2. Infographic (2012) The growing impact of social media. http://www.sociallyawareblog.com/2012/11/21/time-americans-spend-per-month-on-social-media-sites/
3. Collier N, Son NT, Nguyen NM (2011) OMG u got flu? Analysis of shared health messages for bio-surveillance. J. Biomed Semant 2(S–5):S9
4. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2008) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014
5. Mangold WG, Faulds DJ (2009) Social media: the new hybrid element of the promotion mix. Bus Horiz 52(4):357–365
6. Kamel Boulos MN, Sanfilippo AP, Corley CD, Wheeler S (2010) Social web mining and exploitation for serious applications. Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. Comput Methods Programs Biomed 100(1):16–23
7. Lampos V, De Bie T, Cristianini N (2010) Flu detector-tracking epidemics on twitter. In: European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD 2010), Barcelona, Spain, pp 599–602
8. Freifeld CC, Chunara R, Mekaru SR, Chan EH, Kass-Hout T, Iacucci AA, Brownstein JS (2010) Participatory epidemiology: use of mobile phones for community-based health reporting. PLoS Med 7(12):e1000376
9. Sadilek A, Kautz HA, Silenzio (2012a) Predicting disease transmission from geo-tagged micro-blog data. In: Twenty-sixth AAAI conference on artificial intelligence
10. Sadilek A, Kautz H, Silenzio V (2012b) Dublin: modeling spread of disease from social interactions. In: Proceedings of sixth AAAI international conference on weblogs and social media (ICWSM)
11. Kaundal R, Kapoor AS, Raghava GP (2006) Machine learning techniques in disease forecasting: a case study on rice blast prediction. BMC Bioinform 7(1):485
12. Jin X, Gallagher A, Cao L, Luo J, Han J (2010) The wisdom of social multimedia: using flickr for prediction and forecast. In: Proceedings of the international conference on multimedia. ACM, pp 1235–1244

13. Zheng-yan C (2010) Short message classification of microblogging based on semantic. Mod Comput 8:006
14. Yang F, Liu Y, Yu X, Yang M (2012) Automatic detection of rumor on sina weibo. In: Proceedings of the ACM SIGKDD workshop on mining data semantics. ACM, p 13
15. Bao M, Yang N, Zhou L, Lao Y, Zhang Y, Tian Y (2013) The spatial analysis of weibo check-in data–the case study of wuhan. In: Geo-informatics in resource management and sustainable ecosystem. Springer, Berlin, pp 480–491
16. Sun Y, Yan H, Lu C, Bie R, Zhou Z (2014) Constructing the web of events from raw data in the web of things. Mob Inf Syst 10(1):105–125
17. Ritchie M, Charlish A, Woodbridge K, Stove A (2011) Use of the Kullback–Leibler divergence in estimating clutter distributions. In: 2011 IEEE on radar conference (RADAR). IEEE, pp 751–756
18. Amati G, Van Rijsbergen CJ (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans Inf Syst (TOIS) 20(4):357–389
19. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523
20. Liu J, Li B, Zhang W-S (2012) Feature extraction using maximum variance sparse mapping. Neural Comput Appl 21(8): 1827–1833
21. Deng S, Xu Y, Li L, Li X, He Y (2013) A feature-selection algorithm based on support vector machine-multiclass for hyperspectral visible spectral analysis. J Food Eng 119(1): 159–166
22. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
23. Joachims T (1999) Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A (eds) Advances in kernel methods—support vector learning. MIT-Press, pp 41–56
24. Chang C-C, Lin C-J (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27
25. Yang N, Li S, Liu J, Bian F (2014) Sensitivity of support vector machine classification to various training features. TELKOMNIKA Indones J Electr Eng 12(1):286–291
26. Han E-HS, Karypis G, Kumar V (2001) Text categorization using weight adjusted k-nearest neighbor classification. Springer, Berlin