International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.91-100
http://dx.doi.org/10.14257/ijdta.2016.9.5.09

A Method for Tracking Flu Trends through Weibo

Yang Li and Changjun Hu

School of Computer & Communication Engineering, University of Science & Technology Beijing, Beijing, China
liyang@ustb.edu.cn, hu.cj.mail@gamil.com

Abstract

Real-time monitoring the spread of disease and taking a rapid response is necessary. The traditional public health report is accurate, but requires a lot of manpower and resources. The main drawback is the lag of time. Notifiable infectious diseases report generally lags behind medical diagnosis about 4-5 weeks. In this paper, social network data are used to detect disease and track its rapidly changing trends. We take flu data in Sina weibo as an example and analyze flu related weibos from temporal and spatial dimensions. Compared with the previous work, most studies filter out the non-infection weibo noises directly, however, the noises have close association with flu activities. We are not simply discard these data but use them to capture the public nuanced attitude changes. Flu related weibos are divided into four categories which represent four states of public concern. The four states, gradually upgrading from concern about news to anxiety of illness, help to capture the public nuanced attitude changes toward flu trends. Flu weibos concern distributed map and influenza activities curve are drawn to show the analyze result. Multiple classification systems' accuracy are investigated. The proposed method twice iterative classification makes the system accuracy up to 89.50%.

***Keywords**: Weibo, Flu, Classify, Nuanced attitude change*

1. Introduction

Disease surveillance is through long-term, continuous collection, collation, analysis of disease data to discover the dynamic distribution (information, time and geographic) and influential factors (natural factors, social factors). Disease surveillance can reduce the impact of disease outbreak and pandemic.

The People's Republic of China National Health and Family Planning Commission (NHFPC) collects previous months' disease data from hospital and disease prevention institutions at all levels. Although this method can ensure the accuracy and authenticity, its main drawback is the 4-5 weeks' time lag between the time of medical diagnosis and the time when the data becomes available.

Ginsberg *et al*. proposed a method to detect flu related activities by Google search engine [1]. This method studied users' search habits, and found a fixed pattern in all the results to distinguish the real flu infections from interests. The estimated results had a high degree of consistency with the suspected cases of flu data released by U.S. CDC. Without sufficient context information, some surge search queries caused by discussions on traditional media are not easy to filter. They may lead to completely unrelated search peaks to an incidence of a disease. Besides the data is not suitable to make further subdivide. There exist another two problems in search queries data, location and privacy protection. Firstly, search queries can't get accurate location information by parsing IP address. Secondly, user searching results involve their privacy, which are only owned by a few companies, not available reproducibility and follow-on research.

Sina weibo is a social network platform for broadcasting real-time brief information. Up to the first half year 2013, Sina weibo has 536 million registered users and processes about 130 million weibos per day, in which more than 60% used handle devices to login and publish messages making it convey more immediacy than other platforms. In whole year of 2013, total number of sixty million flu-related weibos were published. User's privacy weibos do not be retrieved and read by general public. Each weibo contains 140 words of brief content and metadata with a semi-structure, such as time and location information. Weibo is becoming a new promising platform that could be used to monitor flu activities.

To capture different states of public concern, flu related weibos are divided into four stages, which are beginning to care about the news, taking precautions, anxious about illness and finally infected. For this purpose, we designed and tested a variety of classifiers to process and separate weibos to obtain the best results.

This paper is organized as follows. Section 2 introduces related works and Section 3 describes the study method which mainly consists of two parts. The first part describes whether flu activities can be detected by Sina weibo. Flu weibo concern map and the comparison of influenza weibo's data and NHFPC flu data curves are drawn in this part. The second part describes the steps how to classify flu related weibos into four categories of public concern states, and the results are discussed. Finally, the conclusion and directions for future work are given in Section 4.

2. Related Works

Social network is playing an increasingly prominent role in monitoring the real world events. The United States Geological Survey (USGS) uses Tweet Earthquake Dispatch (TED) system to monitor seismic activities within the global scope [2]. The TED system, using public Twitter data, can detect earthquakes in anywhere from 30 seconds to two minutes. It has a great impact on monitoring seismic activities in remote areas. Seismometers take as long as 20 minutes to confirm earthquakes.

Andranik Tumasjan *et al*. (2010) analyze over 100,000 tweets containing a reference to parties or politicians before the German federal election 2009 [3]. The result shows that tweets can be used as a reflection of political sentiment and a predictor of the election result. Online messages on Twitter validly mirror offline political sentiment.

Several twitter disease surveillance have been proposed in previous studies. Alessio Signorini *et al*. (2011) introduce a surveillance system which can show the 500 most recent flu related tweets on U.S. map [4]. They examine flu's impact on people's lives and design ILI estimation model to predict weekly ILI with simple SVM classifier to filter noises. Further studies lead by Alex Lamb revealed the relation between infection tweets and past/present tense, self/others to separate infection from fear [5].

3. Methods

3.1 Real Time Flu Surveillance

For real-time tracking public interests and concerns of influenza and monitoring influenza activities, we continue gathering and mining a large amount of publically-available weibos from Sina weibo platform. Sina weibo is one of the China's most widely used microblogging platform. However, Sina does not provide weibo retrieval application programmer's interface. We design a weibo crawler by using the search function on web page. Surveillance system stores 140 million flu weibos containing keyword "flu" since January 1, 2010. Moreover, our main interest was to monitor influenza related traffic within China, we excluded all weibos tagged as originating outside China, and any weibos not written in Chinese.

Weibo contains semi-structured metadata (time-stamp, geolocation, original). Its geolocation is user self-declaration. Compared with the national data published by NHFPC and the provincial data published by Google, the accuracy of geographic location of weibos can reach municipal level. Because the volume of posts on Sina weibo varies over time as well as across geographic regions, usage weighted percentage of user volumes whose weibos mentioned “flu” to represent the level of concern in corresponding time interval and geographic region. Different from Alessio Signorini *et al*. using the fraction of the total tweets, adapt the number of users because the increasing of flu weibo number do not mean the level rising of public concerns. Make the maximum and minimum cities concern as the interest threshold and divide threshold values into nine. Level of flu concerns is differentiated by colors for easy comparison. The map in Figure 1 is an output of the geographical analysis and shows the public concerns by municipal.

Map provides a direct way to monitor the flu attention of different cities, which can be used to analyze the spread and evolution of influenza. What’s more, as it shows public concerns of a region in real-time, hospitals could decide whether they need to reserve more drugs accordingly.

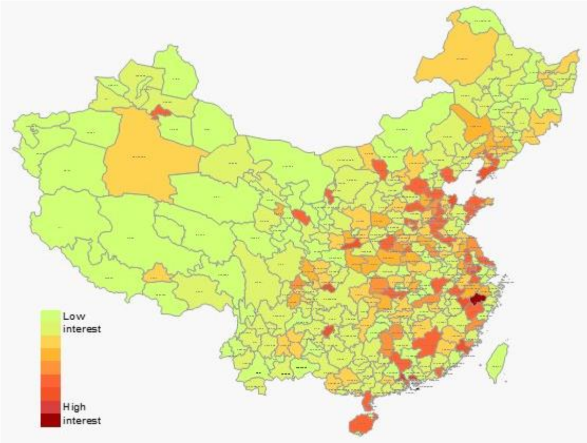


Figure 1. China’s Flu Weibo Concern Distributed Map

In order to find out whether the flu weibos mining method can reflect the real situation, Figure 2 shows the comparison curves between the data published by NHFPC (orange line) per month and the flu related weibos (blue line) from January 2010 to December 2013.

In Figure 2, the main vertical axis is NHFPC flu data published per month, side vertical axis stands for public concerns. We can see that the users’ flu attention have some relevance to NHFPC published data. The data released by NHFPC have one month lag time, while method based flu weibos data is real-time and has diminish time interval. But the data trends are not matched very well. The reason may be the existence of non-flu infection weibos, which made the curve couldn’t fully reflect the real flu situation.

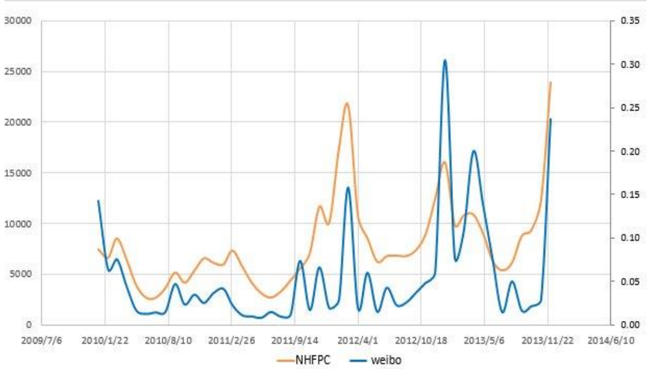


Figure 2. The Comparison of Influenza Weibos Data and NHFPC Flu Data

3.2 Tracking Nuanced Attitude Trend Change

Influenza weibos contain plenty of unreal flu infection activities, which have interference on monitoring the flu trend. Compared with the search engine way, weibo has a wealth of contextual information. Natural language processing is used to analyze weibo's context to identify whether it is infection activity. Most of non-flu infection weibos can be summarized as the following three types shown in Table 1.

Table 1. The Main Three Types of Influenza Weibo Noise

News	[Zhejiang add two cases of human infection with H7N9 cases]Zhejiang Provincial Health and Family Planning Commission briefing today, the province added two cases of human infection with the H7N9 avian influenza. Up to now, since the start of the year, Zhejiang Province, has a total 87 cases of H7N9 confirmed cases of avian influenza.
Prevention	[Four principles must know to prevent influenza] Pediatrician remind parents should pay attention to various aspects of children's basic necessities when the seasonal change, the most important thing is to master the four principles against influenza. Principle One: Control the ambient temperature; II: Timing ventilation; three: Stick to outdoor activities; four: rational drug of choice. You do it?
Anxious	I suddenly felt a sore throat, do not be the flu. I do not want a shot I do not want to spend money. Sleep sleep sleep, go to work tomorrow ...

Sometimes news discussion on traditional media will cause a surge in flu related weibos, which does not mean a significant increasing in the amount of influenza infections. *e.g.* April 2013, Jiangsu Province Health Department report that Radix can prevent H7N9 avian influenza. This month the flu related weibos increased by 77%, while the actual number of infections only increased by 0.77%. In this paper, raw weibo data searched by keyword "flu" are classified into four categories. Class A stands for influenza infection activities, class B for related news, class C for precautions, class D for anxious. At class B flu begins draw public attention, gradually upgraded to fear and panic at class D. The other weibos generally consists of the word like "潮流感" which takes a very small proportion of the flu weibos, so it can be ignored when design the classifier.

393 marked weibos are selected from January 2013 to March 2013 in flu related weibos. ICTCLAS [6] is used in weibos word segmentation. The class of news and precautions begin with "[" and "]" more than 87%, so we can simply filter out other punctuation during word segmentation, finally we get 9824 tokens and 3499 words.

3.2.1 Automatically Select Stop Words

Stop words can decrease the vector space and improve the efficiency of the text classifier effectively. Appropriate choosing the set of stop words may have a very small negative impact on the classification results. Catarina Silva verified that appropriate stop words can decrease the feature space vocabularies, especially useful in improving the accuracy of text classification based on support vector machine [7]. Stop words set automatically selected is more targeted compared with common used stop words set. And the meaningless words will be deleted from the feature words and thus we could obtain a better classification results.

Let D = $\{d_1, d_2 \dots d_n\}$ be a set of weibos which contain keyword “flu” and W = $\{w_1, w_2 \dots w_r\}$ is the set of feature items of word segmentation. The average amount of information is used to evaluate the actual implications of each term “w” in the set, based on the Shannon entropy theory [8].

$$W(w) = 1 + \frac{1}{\ln(n)} \sum_{i=1}^{n} P_i(w) \ln[P_i(w)] \tag{1}$$

$P_i(w)$ is the probability of term “w” occurrence in weibo “i”. The resulting list can be ordered by ascending entropy to reveal the words that have a greater probability of being noise words. We randomly select 200 weibos per month from January 2010 to December 2013, 9800 word segments in total. After word segmentation, we finally get 269,223 tokens and 17,656 words. 500 words whose entropy less than -0.1 are chosen as the stop words set S = $\{s_1, s_2 \dots s_k\}$.And then, 393 marked weibos dataset D1 are filtered by S, got 3053 terms remained.

3.2.2 Select the Feature Terms

In order to obtain better classification results, three different feature evaluation functions are used for scoring feature terms w to find the most representative characteristic of text category.

1) Information gain

A key measure of information gain is how much information feature terms can bring to the classification system [9]. The more information feature term brings, the more important it is. For a given term w, the difference between the prior entropy of text and posterior entropy is the information gain it brought to the system. The information gain of term “w” is:

$$\begin{aligned} \mathrm{IG(W)} &= H(C) - H(C \mid W) \\ &= -\sum_{j-1}^{m} P(C_j)\log_2 P(C_j) + P(w)\sum_{j-1}^{m} p(C_j \mid w)\log_2 P(C_j \mid w) \\ &\quad + P(\overline{w})\sum_{j-1}^{m} p(C_j \mid \overline{w})\log_2 p(C_j \mid \overline{w}) \end{aligned} \tag{2}$$

C_j ranges from 1 to 4.

2)Chi-Square Statistic

Chi-square statistic (χ^2) quantify the importance of the feature terms by estimating the correlation between terms and classes [10]. The terms should be selected with the highest correlation.

$$\chi^2(c, w) = \frac{n(P(c,w)P(\bar{c},\overline{w}) - P(c,\overline{w})P(\bar{c},w))^2}{P(c)P(w)P(\bar{c})P(\overline{w})} \tag{3}$$

Taking the maximum value of class A to class D as the chi-square statistic result of the feature item “w”.

3) Document frequency

Document frequency is the easiest method to select the feature item, which is defined as the frequency of feature item “w” appearing in the set D [11]. Set the minimum and maximum threshold, and calculate the document frequency of each feature term. If the term’s frequency is greater than the maximum threshold or less than the minimum threshold, this feature item would be deleted, otherwise be retained.

The average number of tokens of flu infection weibos is 14.04, while other three classes is 33.11. With the clear distinction, the byte stream length is defined as one of the features to describe the weibos. The news category are usually released by Sina verified accounts called "big V", such as government agencies, the media, scholars and celebrities.

And then those weibos will be forward by ordinary users in a large quantity. Because the influenza infection weibos usually released to express the illness of themselves or their close relatives, only the original weibo can be marked as class A when judging flu activities. And whether a weibo is released by Sina verify account or original account is taken as a feature item. The results of the selected features are shown in Table 2.

Table 2. Feature Items Extracted by Different Methods

Methods	Feature Items
Information Gain	[,], disease, infection, avian flu, diagnosis, prevention, attention, death, reports, epidemic, use, type, not, institutions, methods, express, city, aged, vitamins, inform, ill, until, citizen, public health, already, children, diet
χ^2 test	disease, avian influenza, diagnosis, prevention, death, infection, methods, epidemic, reporting, use, enter, [, city, express, attention, supplement, diet, vitamins, agency, until, notification, disease, citizen,], insist
Document Frequency	[,], Colds, prevention, attention, season, infection, health, body, influenza virus, people, vaccines, anti, weather, drink, hospitals, effective, avian flu, water, use, enhance, immunity, heat, say, hand, health, up

3.2.3 Calculate the Feature Weights

TFIDF is the most widely used weight calculation algorithm in text processing field [12]:

$$\mathrm{T(w)} = f_i(w) \times \log(\frac{n}{n_k} + l) \tag{4}$$

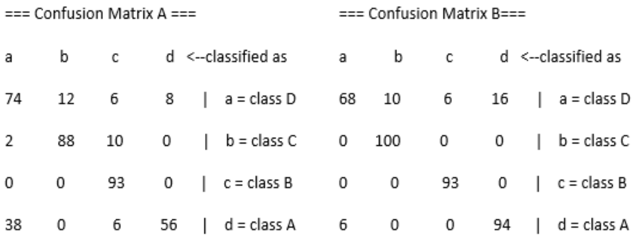
In this expression, $f_i(w)$ is the number of times that term w occurs in a weibo, n_k stands the number of weibos containing the term. Taking the impact of a weibo's length, we normal the weight to [0, 1]. And "l" takes the empirical value 0.01.

$$\mathrm{T(w)} = \frac{f_i(w)\times\log\left(\frac{n}{n_k}+0.01\right)}{\sqrt{\sum_1^n (f_i^2(w)+log^2(\frac{n}{n_k}+0.01))}} \tag{5}$$

3.2.4 Twice iterative Classification

The open source data mining platform WEKA [13] is used to train microblogging classifier which is used for target classification. WEKA provides classifier based on different algorithms such as Navie Bayes algorithm, KNN algorithm, neural networks algorithm, SVM algorithm [14] and their improved algorithms. A variety of classification algorithms are investigated. When weibos are divided into four categories directly, the highest classification accuracy is 76.24%. Its results are not satisfactory. Analyze the confusion matrix, shown in Figure 3 confusion matrix A. Class A and class D has many similarities from the content. There are 38 class A instances are wrongly classified into class D. The primary purpose of the system is to track flu infection activities. This type error is the most difficult to tolerate. To make sure class A can be correctly classified we propose a twice iterative classification method.

Temporarily mark all class C and class D instances in dataset D1 as class B. Select the feature terms. Calculate the feature weights and trained classifier to classify D1. Instance of 91.16% are correctly classified. Only three instances of class A are wrong assigned to class B. The data belonging to class B at first classification is marked as D2. Recovery the label of class B, class C, class D in D2 and label misclassified class A as class D. Reselect feature terms of D2. Calculate weights and twice iteration classify D2. The process is shown in Figure 4.The second classification accuracy is 93.65%. After twice iterative classification, the confusion matrix shown in Figure 3 confusion matrix B. Classification system accuracy increases to 89.50%.



```
=== Confusion Matrix A ===                    === Confusion Matrix B===

a     b     c     d   <--classified as        a     b     c     d   <--classified as

74    12    6     8   |  a = class D          68    10    6     16  |  a = class D

2     88    10    0   |  b = class C          0     100   0     0   |  b = class C

0     0     93    0   |  c = class B          0     0     93    0   |  c = class B

38    0     6     56  |  d = class A          6     0     0     94  |  d = class A
```

Figure 3. WEKA Direct Classification Confusion Matrix A and Twice Iteration Classification Confusion Matrix

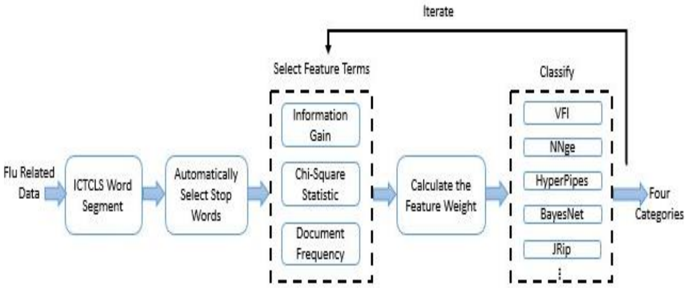


Figure 4. Twice Iterative Classification Process

Multiple classification algorithm are tested to find a suitable flu weibos classification system. Table 3 compares the results of the various classification systems.

Table 3. TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area of each System

System	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
IG-VFI-VFI	0.854	0.038	0.856	0.854	0.854	0.924
χ^2-VFI-VFI	0.895	0.031	0.905	0.905	0.905	0.977
DF-VFI-VFI	0.777	0.036	0.847	0.777	0.81	0.943
IG-NNge-HyperPipes	0.793	0.043	0.828	0.793	0.81	0.882
χ^2-NNge-HyperPipes	0.875	0.032	0.844	0.835	0.839	0.892
DF-NNge-HyperPipes	0.814	0.05	0.82	0.814	0.817	0.886
IG-BayesNet-JPip	0.854	0.038	0.855	0.854	0.854	0.958
χ^2-BayesNet-JRip	0.809	0.05	0.807	0.809	0.808	0.938
DF-BayesNet-JRip	0.81	0.05	0.815	0.81	0.809	0.936

Comparison of three characteristic evaluation function, chi-square statistic classification accuracy is better than information gain and document frequency. Chi-square selects local terms. However, information gain and document frequency can only select the global terms. All weibos collected from January 2010 are classified to monitor influenza related events. The result curves show that the classification accuracy is not stable. Due to public concerns change over development of time and events, the classifier is difficult to fit weibos of all months. Following Google's strategy, each season we update classification system to improve its performance. Results are shown in Figure 5. The blue line represents the proportion of inflection weibos to total weibos. To get a better visual effect in Figure 5, we expand the data amount 30 times. The green line shows the NHFPC's measured influenza activities. The proportion of flu inflection, news, precautions and anxious weibos to all flu related weibos are represented by blue, purple, green and orange histograms.

Even after filtration, the number of influenza infections is still much larger than NHFPC measured. There are three possible reasons: first of all, weibos cover a wider range than NHFPC data collection agency. Everyone can publish weibos wherever they could access to the internet by PC or the mobile phones. Secondly, the system cannot guarantee completely correct classification. A certain number of non-infection flu weibos are divided into class A. Thirdly, the authenticity of each flu weibo also cannot be promised. So in Figure 5, we shrink the number of flu weibos by several orders of magnitude. We can see that after filtration, the flu weibos trend is more correlation with NHFPC data than before, and it can represent the real flu trends basically. In addition, we can see the nuanced changes of four categories and help capture the tendency of specific events.

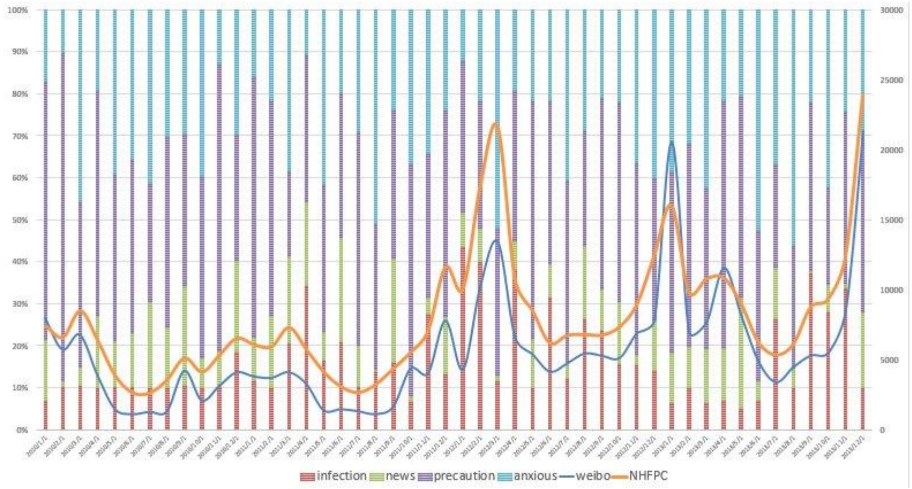


Figure 5. Comparing with the Filtered Data and NHFPC

4. Conclusion & Future Work

Experiments show that mining and analyzing the flu related weibos to track flu activity is available. We build a real-time flu surveillance system via using weibos data which can diminish the time interval of monitor to one hour. Flu weibos concern distributed map and influenza activities curve show the people clearly who use weibos and get flu distributed in space, timeline and quantity. Meanwhile, they also can be used to study how flu or other diseases transmission and evolution. In order to capture public concern states, we divide flu weibos into four categories with attention to nuanced differences between flu weibos instead of two classes which always adopted in previous studies. Twice iterative classification method we proposed can enhance the classification accuracy from 76.24% to 89.50%. For future work, we plan not only to classify a larger number of messages, but also make weibo data models to forecast trends and outbreaks of flu.

Acknowledgements

This work was supported by the R&D Infrastructure and Facility Development Program of China (Grant No. 2005DKA32800), the 2012 Ladder Plan Project of Beijing Key Laboratory of Knowledge Engineering for Materials Science (Grant No. Z121101002812005), the Key Science–Technology Plan of the National Twelfth Five-Year-Plan Project of China (Grant No. 2011BAK08B04), the Fundamental Research Funds for the Central Universities (No. FRFTP-12-079A) and the National Key Basic Research and Development Program (973 Program) of China (Grant No. 2013CB329601).

References

[1] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, "Detecting influenza epidemics using search engine query data", Nature, vol. 457, no. 7232, **(2009)**, pp.1012-1014.

[2] E. Paul, C. Daniel and R. Michelle Guy, "Twitter earthquake detection: earthquake monitoring in a social world", Annals of Geophysics, vol. 54, no. 6, **(2011)**, pp. 708-715.

[3] A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe, "Redicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", ICWSM, vol. 10, **(2010)**, pp. 178-185.

[4] A. Signorini, A. M. Segre and 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, vol. 6, no. 5, **(2010)**, e19467.

[5] A. Lamb, M. J. Paul and M. Dredze, "Separating fact from fear: Tracking flu infections on twitter", In Proceedings of NAACL-HLT, **(2013)**, pp. 789-795.

[6] H. P. Zhang, H. K. Yu, D. Y. Xiong and Q. Liu, "HHMM-based Chinese lexical analyzer ICTCLAS", In Proceedings of the second SIGHAN workshop on Chinese language processing, Association for Computational Linguistics, vol. 17, **(2003)**, pp. 184-187.

[7] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization", In Neural Networks, Proceedings of the International Joint Conference on, vol. 3, **(2003)**, pp. 1661-1666.

[8] M. Dash and H. Liu, "Feature selection for classification", Intelligent data analysis, vol. 1, no. 3, **(1997)**, pp. 131-156.

[9] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization", Information processing & management, vol. 42, no. 1, **(2006)**, pp. 155-165.

[10] Y. T. Chen and M. C. Chen, "Using chi-square statistics to measure similarities for text categorization", Expert Systems with Applications, vol. 38, no. 4, **(2010)**, pp. 3085-3090.

[11] N. Azam and J. Yao, 'Comparison of term frequency and document frequency based feature selection metrics in text categorization", Expert Systems with Applications, vol. 39, no. 5, **(2012)**, pp. 4760-4768.

[12] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", Carnegie-Mellon University Pittsburgh Pa Department of Computer Science, CMU-CS, **(1996)**, pp. 96-118.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringe, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter, vol. 11, no. 1, **(2009)**, pp. 10-18.

[14] T. Joachims, "Learning to classify text using support vector machines: Methods, theory and algorithms", Kluwer Academic Publishers, vol. 29, no. 4, **(2003)**, pp. 655-664.

Authors

Yang Li, received the Ph.D. degree from University of Science and Technology Beijing, China, in 2010, where she is also received her BSc degree. Her research interests include data integration, semantic web, cloud computing and software engineering.

Changjun Hu, He received the Ph.D. degree from Peking University, Beijing, China, in 2001. He is currently a Professor at the School of Information Engineering at the University of Science and Technology Beijing, China. His main research interests include parallel computing, parallel compilation technology, parallel software engineering, network storage system, data engineering and software engineering.