Detect and Analyze Flu Outlier Events via Social Network

Quanquan Fu[1], Changjun Hu[1], Wenwen Xu[1], Xiao He[1], and Tieshan Zhang[2]

[1] School of Computer & Communication Engineering,
University of Science & Technology Beijing, Beijing, China
[2] Information Office, China-Japan Friendship Hospital, Beijing, China
fuquanquan06@sina.com, huchangjun@ies.ustb.edu.cn,
{xuwenwenustb,tie_shan}@163.com, hexiao83@gmail.com

Abstract. The popularity of social networks provides a new way for constant surveillance of unusual events related to a certain disease. Some researchers have begun to use twitter to estimate the situation of public health, as well as predict disease trends. However, previous studies usually focused on the infection data but not the data judged as non-infection, which was usually filtered directly in their studies. We believe that the non-infection data is also essential for monitoring disease activity, because of their inherently subtle connections. Firstly, we construct a time series outlier model that can detect flu outlier events of different region in China with high precision and good recall by mining all the flu related data. Secondly, those outlier events are used to find out hot topics by SN-TDT and use the twice iteration classification method which is designed to analyze users' status who published a flu-related weibo. These results could provide science reference for deploying sickness prevention resources, and make recommendation about which place pose a high risk of getting infected.

Keywords: weibo, outlier events, time series, twice iterate classification.

1 Introduction

Sina Weibo is a social network platform for broadcasting real-time brief information. Having wide user groups range makes it a new way to monitor disease activity. By the end of 2013, Sina Weibo has 281 million registered users, and nearly up to 33% in all Chinese internet users. Among them more than 69.7% people use weibo mobile client, which make it has good real-time features than the traditional way reported by disease prevention and control institutions. Superior to other platforms, such as the search engines, each item weibo contains 140 words brief content and metadata with a semi-structure, just like time and location information, by which we are able to get richer information than the number of infections. Recent work has demonstrated that micro-blogging data can be used to track levels of disease activity and public concern (Alessio Signorini 2010) [1], predict flu trends (Harshavardhan Achrekar 2011) [2], fine-grained predict the health of specific people (Adam Sadilek 2012) [3], and detect health conditions (Victor M. Prieto 2014) [4].

W. Han et al. (Eds.): APWeb 2014 Workshops, LNCS 8710, pp. 136–147, 2014.
© Springer International Publishing Switzerland 2014

So far, most of the research is focused on classification for the infection of weibo, ignoring the research on the non-infection. However, more than 80% are non-infection ones in flu related weibos, which contains a wealth of user information. In previous studies, non-infection weibos were filtered directly, only a very small part of the infection weibos were analyzed, which made it difficult to reflect the integrity of the flu event. A different kind of problem one can pose, however, is to combine with infection the non-infection weibos to analyze those complete flu activities.

This paper discusses the detection of abnormal flu related events and analyses the relationship between infection and non-infection weibos. Two conditions should be satisfied in detection of abnormal weibos events. Frist, the outlier detection algorithm must meet the real-time requirement, that is to say, outliers can be found immediately once they happens. Second, outlier detection algorithm should be able to find out both of the patterns outlier and single point outlier.

We combine both reachable neighbor outlier factor and time sequence outlier detection methods to analyze the quantity changes of flu related weibos in each province every day. Outlier patterns and point are detected in the time sequence. Since the attention differences on every event, there would be large changes in weibo quantity, which require that the detection algorithm should have good adaptability to make continuous dynamic adjustment. Then we analyze a certain number of weibos within the time scope of the outlier. The SN-TDT model is adopted based on the short text characteristic of social network. The cluster analysis is made to find out the topic of the outlier event and mining users' focus. In this way, flu related weibos are divided into four stages, which include caring about the news, taking precautions, anxious about illness and finally infection. For this purpose, we designed twice iterative classifier to process and separate weibos.

Analysis and detection of outlier events have practical significance. The research results can provide reference for government grasping public opinions, and correctly guide the dissemination of information, avoid panic due to information asymmetry. According to user status, when the attentions of prevention measures become high in some place, government can increase the deployment of medical resources and promote the correct disease prevention knowledge. What's more, advice could provide to citizen that where is the high risk area.

This paper is organized as follows. Section 2 introduces related works and section 3 describes the study method which mainly consists of three parts. The first part describes outlier detect model. The sliding window-reachable neighbor detection method is introduced in this part. The second part introduces the SN-TDT algorithm. The third part describes the steps how to classify flu related weibos into four categories of public concern states. Section 4 is the experiments and results. Finally, the conclusion and directions for future work are given in section 5.

2 Related Works

A number of studies have been conducted using different forms of social networks to monitoring and prediction hot social events. Like Takeshi Sakaki et al. (2010) [5] investigate the real-time interaction of earthquakes in Twitter. They consider each Twitter user as a sensor. When the user feels the earthquake occurrence, he may make a

tweet to broadcast the fact. False-positive ratio is used to calculate the probability of earthquake occurrence when there are n items of weibo classified as positive. The alarm will be set when the probability exceed the threshold. The author records each user's longitude and latitude marked in weibos. Then Kalman filtering and particle filtering are applied to find the center and the trajectory of the event location.

The abnormal event detection is rarely used in flu detection. The outlier detection technique based on relative density and the outlier detection technique based on time sequence are two common outlier detection technologies. The outlier detection technique based on density believe that normal data is in high density area, while the abnormal point in the low density area. But if the data is in the variable density region, the technique based on relative density will greatly affected. In order to solve this problem, Markus M. Breunig et al. [6] defined LOF (Local Outlier Factor) as the ratio of average density of K nearest neighbor and the density of the data itself, Using LOF as the outlier degree. Zakia Ferdousi et al. [7] use Peer Group Analysis (PGA), which is an unsupervised technique for fraud detection. PGA characterize the expected pattern of behavior around the target sequence in terms of the behavior of similar objects, and then to detect any difference in evolution between the expected pattern and the target.

Classification systems have also been used to filter the original data to obtain better results. Adam Sadilek et al. [3] predicted disease transmission from geo-tagged Micro-Blog data. They worked on SVM classifier to classify health-related text messages and detecting illness-related messages such as flu, sick, headache, stomach etc.

3 Methods

3.1 Detect Flu Related Weibos' Outlier

In order to detect abnormal events, we collect flu weibo data every day continuously. The number of flu weibos is analyzed in the stage of outlier detection. We think that the total number of flu weibo should change smoothly and continuously if there not any outlier happened. The sudden appearance of the point deviation and mode change can be judged as a flu outlier. We collected flu related weibo number every day from March 11th, 2013 to May 31th, as shown in Figure 1.

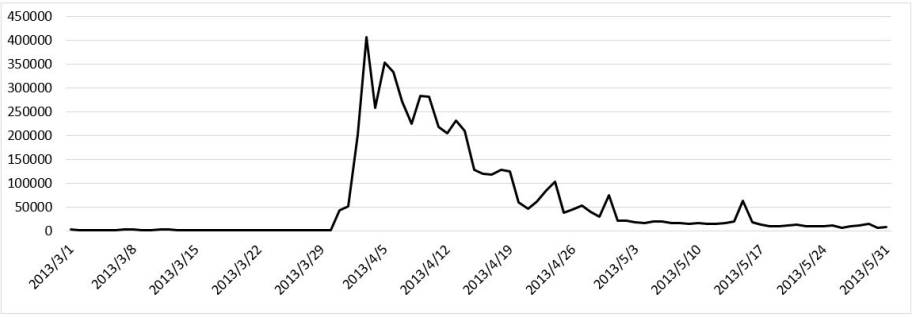

Fig. 1. The quantity of flu related weibos

We can see that the rapid growth of the flu weibo in April in figure 1. It leads to the number of weibo in March cannot be measures by the same magnitude of that in April. This requires outlier detection algorithm should aim at dealing with local data, and have the ability to adapt to new data dynamically. The sooner we discover the outliers, the early corresponding measures could be implemented. All of that make the algorithm should be in real time, the result could be found immediately once the outlier happens.

Take the daily amount of flu weibo data of every provinces and the record time as an ordered set $X= \langle x_1=(t_1,v_1),x_2=(t_2,v_2),\ldots x_n=(t_n,v_n) \rangle$, the recording time is strictly increasing ($i < j \Leftrightarrow t_i < t_j$). The number of flu weibo series drawn a series of discrete points in a daily interval, and a straight line can be got by connecting the two adjacent points, which makes a time series. Sub pattern can be divided by determining the edge points using the slope in analytic geometry. Method for judging whether a day is an edge point is:

Calculate the slope k_1 by the number of weibo x_i in day i and the number of weibo x_{i-1} in day i-1, and slope k_2 by the number of weibo x_{i-1} in day i-1 and the number of weibo x_{i-2} in day i-2. If$|k_1 - k_2| > maxSlope$, it can be judged as the edge point of a sub model p. Use its length, intercept and mean value as characteristics of this sub model. In this way, the pattern density of time series can be converted into the density of data objects in data set D. When pattern length is set to 1, pattern outlier becomes single point outlier.

The density of p is defined as following:

Definition 1: *k-distance* of an object p

For any positive integer k, the k-distance of object p, denoted as k-dist(p),is defined as the distance d(p, o) between p and an object $o \in D$ such that:

(1) For at least k objects $o' \in D\{p\}$,it holds that $d(p, o') \leq d(p, o)$ and

(2) For at most k-1 objects $o' \in D\{p\}$,it holds that $d(p, o') < d(p, o)$.

Definition 2: k-distance neighborhood of an object p:

$Nk(p) = \{q \in D\{p\} \mid d(p, q) \leq \text{k-dist}(p)\}$

Definition 3: k- reachable neighbor of p:

$\forall p \in D$, if $q \in D\{p\}$,there is $p \in Nk(q)$, we call Nk(q) is one of p's k- reachable neighbor about k-dist(p), marked as RNK(p). It is can be seen that a smaller |RNK(p)| stands for a smaller chance the object p in other objects' reachable neighbor. Otherwise, object p is in an intensive position.

Definition 4: the density of an object p:

$$\text{RNrk}(p) = \frac{1}{|N_k(p)|}\sum_{o \in N_p(p)} \frac{|RN_k(p)|}{|RN_o(p)|}$$

RNrk (p) is the average ratio of the k- reachable neighbor of object p and others which reflect the local density of object p.

Definition 5 the k-reachable neighbor outlier factor of an object p:

$$\text{RNOFk}(p) = \max\{1\text{-RNrk}(p), 0\};$$

The bigger RNOF, the more likely p is an outlier mode.

Outlier detection of flu activity requires an online, real-time-alarming algorithm. We introduce the sliding fixed size window model. When the number of object is greater

than the window length, each insertion of a new object means the deletion of an old object from the window tail.

Flu outlier event monitoring definition and algorithm are given as following in table 1:

Time series = $\langle v_1, v_2 \ldots v_m \rangle$, edge point =$\{X_{i_1}, X_{i_2} \ldots X_{i_n}\}$. Sub pattern p = L ($X_{i_j}, X_{i_{j+1}}$). Use the sub model to represent the time series = $\langle \mathrm{L}(X_{i_1}, X_{i_2}), \mathrm{L}(X_{i_2}, X_{i_3}), \ldots \mathrm{L}(X_{i_{n-1}}, X_{i_n}) \rangle$

Table 1. flu outlier events detection algorithm

```
while((v_i = Time series data) is valid ){
  if( v_i is the edge point ){
    new pattern p = L(X_{i_j}, X_{i_{j+1}})
    if(time window is full)
      Delete the old object from the tail of the queue,
insert the new object in the header.
    else
      Insert a new object
    for(each object p in window W){
        calculate the k-dist and k-distance
  neighborhood of p
        make every p and k- reachable neighbor of p in
N_k(p) increased by 1.
     }
    for(each object p in window W)
      for(each object in N_k(p)){
        calculate the outlier factor RNOF_k of p
      arrange RNOF_k in ascending order
      if(p ∈ ranking of the k top  && slope of p > 0)
        judge p as outlier
      }
    }
  the next point in the time sequence
  }
```

3.2 Short Text Topic Detection and Tracking

After the outliers in flu time series has been detected, we extract hot topics to make further study about the content of those events. For that purpose, we capture most the weibos in the days which are judged as a sub pattern outlier.

When detecting the outliers, we only give alarms if the slope of an event is greater than 0. The reason lies in that people's attention on an event usually has the properties

of mutual exclusion. When someone pays attention to something in the near period of time, his attention on the other thing will drop. So when the quantity of flu weibos drops sharply, there may be due to other events, affecting the users' attention on flu. Because we can't collect all the information in Weibo platform, neither can we judge what causes the drop, so we do not discuss when the slope of an event is smaller than 0.

The topic detection and track algorithm based on the characteristics of social network, short for SN-TDT, proposed by Liubao Yu is adopted to extract top N topic flu events.

SN-TDT algorithm can be described as follows:

Step 1. Read the new text, if there is no new text, jump to step 7, else to step 2.

Step 2. Mark the feature words of this text.

Step 3. Judging the correlation degree between text and topics, if there are not any existing topic associated with the text, in other words, the correlation degree between the topic and text is lower than a given threshold, jump to step 4, otherwise to step 5.

Step 4. Create a new topic, add the text to the newly created topic, go to step 6.

Step 5. Select the topic most associated with the text, and add the text into the topic.

Step 6. Adjust each attribute of the topic, and jump to step 1.

Step 7. Sort all the topics and the top N are the hot topics.

3.3 Capture Users' Different States

Most of flu related weibos are non-infection. If people are not sick, it is what their mood would be when they send these weibos that is focused in our next research.

Compared with other ways such as search engine, weibo has a wealth of contextual information. Natural language processing is used to analyze users' different states.

Most of non-infection weibos can be summarized as the following three types, news, preventive measures and anxiety. In this paper, raw weibo data searched by keyword "flu" are classified into four categories. Class A for related news, class B for precautions, class C for anxiety and class D stands for influenza infection activities. At class A flu begins to draw public attention, gradually upgraded to fear and panic at class C. 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 designing the classifier. ICTCLAS is used in weibos word segmentation. The most of news and precautions weibos begin with "[" and "]", so we can simply filter out other punctuations during word segmentation.

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

Information gain. A key measure of information gain is how much information feature terms can bring to the classification system [8]. 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:

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

C_j ranges from 1 to 4.

Chi-Square Statistic. Chi-square statistic (χ^2) quantify the importance of the feature terms by estimating the correlation between terms and classes [9]. The terms should be selected with the highest correlation.

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

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

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 [10]. 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.

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 forwarded 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 D when judging flu infections. And whether a weibo is released by Sina verify account or original account is taken as a feature item.

Calculate the Feature Weights. TFIDF is the most widely used weight calculation algorithm in text processing field [11]:

$$\mathrm{T(w)}=f_i(w)\times\log\left(\frac{n}{n_k}+l\right) \tag{3}$$

In this expression, $f_i(w)$ is the number of times that term w occurs in a weibo, n_k stands for 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.

$$\mathrm{T(w)}=\frac{f_i(w)\times\log\left(\frac{n}{n_k}+0.01\right)}{\sqrt{\sum_1^n(f_i^2(w)+log^2(\frac{n}{n_k}+0.01))}} \tag{4}$$

Twice Iterative Classification. The open source data mining platform WEKA [12] 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 [13] and their improved algorithms. A variety of classification algorithms are investigated. When weibos are divided into four categories directly, the results are not satisfactory. To make sure weibos can be correctly classified we propose a twice iterative classification method.

Temporarily mark all class B and class C instances in dataset D1 as class A. Select the feature terms. Calculate the feature weights and trained classifier to classify D1. The data belonging to class A at first classification is marked as D2. Recovery the label of class A, class B, class C in D2 and label misclassified class D as class C. Reselect feature terms of D2. Calculate weights and twice iteration classify D2. The process is shown in Figure 2.

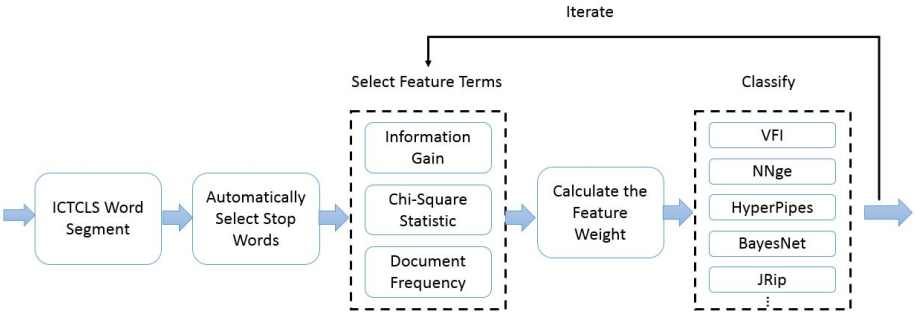

Fig. 2. Twice iterative classification process

4 Experiment and Result

In this section, we evaluate our approach by a number of experiments. The experimental data acquires 34 provinces and cities nationwide daily flu related weibos from March 1, 2013 to May 31, 2013. The sliding window-reachable neighbor algorithm proposed above is tested for anomaly detection with the flu weibos data in Beijing, Shanghai and Guangdong, and when the parameter k=3, the window length = 5 can get a better result. Take outlier threshold $\lambda = 0.5$, the range of outlier factor [0, 0.5]. Because couldn't be determine what happens when weibo data drops, the sub model will be not shown on the diagram if its slope is less than 0.

We can see that using sliding window – reachable neighbor algorithm can basically detect outlier events in Beijing, Shanghai and Guangdong, but at the beginning of the time sequence, the algorithm performance was not very good. In the time series of Beijing form March 3rd to March 5th, flu weibo data increases rapidly, but the detection algorithm fails to find it out. When the data window is full, the algorithm shows a high rate of accuracy and good recall with basically no errors and omissions. The algorithm is designed for outlier patterns in sliding window, therefore the outlier factors are equal within the same sub pattern, and that is to say, we care about both the growth process of outliers in sub models and the peak value.

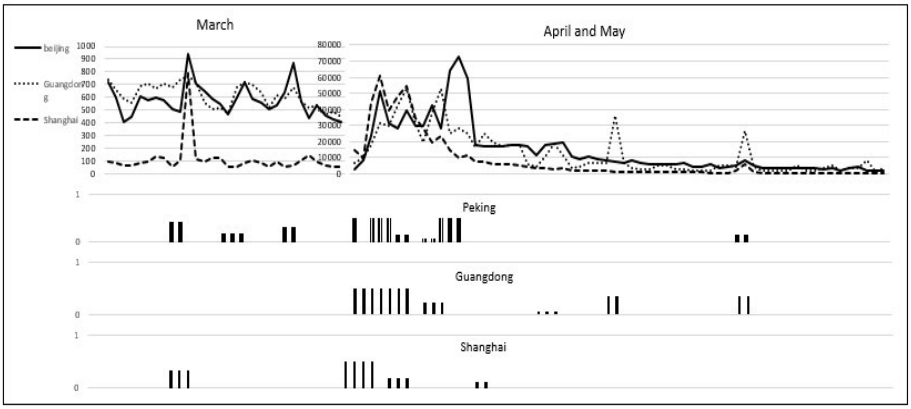

Fig. 3. Experiment of weibo outlier events detection

About 180000 (110M) flu outlier weibos are detected, and Fig. 4 shows the mined hot topics by the SN-TDT algorithm, which includes provincial and topic content.

Fig. 4. SN-TDT hot topic mining results

We can see the reason why flu weibos appears mass exceptions in April is that Shanghai and Anhui confirmed 3 cases of human infection with the H7N9 avian flu virus in March 31st, which is the first cases of H7N9 human infection in world, caused widespread concern. All the outlier events from April to May is about H7N9.

For monitoring flu user status, we use twice iterative classification method to classify flu related weibos of Shanghai in the March 31st. We randomly select 500 item of news, prevention, anxious and infection weibos to label as training set manually from March to May. And randomly select 2000 from flu weibos of

Shanghai labeled as a test set, including 921 news, 364 prevention, 662 anxious, and only 53 infection.

Feature words selected by three different feature evaluation functions are shown in Table 2.

Table 2. Feature items extracted by different methods

Methods	Feature Items
Information Gain	[,], disease, infection, avian flu, H7N9,diagnosis, reports, prevention, methods, attention, express, death, epidemic, use, type, not, institutions, city,
χ^2 test	disease, avian influenza,H7N9, diagnosis, prevention, death, infection, insist, epidemic, notification, report
Document Frequency	[,], H7N9, avian flu, report, attention , infection, virus, body , people, vaccines, anti, effective, drink, hospitals,

Multiple classification algorithm are tested to find a suitable flu weibos classification system. When using WEKA classification directly on the test set, the classification accuracy can be up to 71.4%. System accuracy increases to 79.64% with twice iterative classification. 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
IG-VFI-VFI	0.7686	0.0418	0.7704	0.7686	0.7695
χ^2-VFI-VFI	0.8055	0.0341	0.7964	0.8145	0.8055
DF-VFI-VFI	0.6993	0.0396	0.7623	0.6993	0.7308
IG-NNge-HyperPipes	0.7137	0.0473	0.7452	0.7137	0.7295
χ^2-NNge-HyperPipes	0.7875	0.0352	0.7596	0.7515	0.7556
DF-NNge-HyperPipes	0.7326	0.055	0.738	0.7326	0.7353
IG-BayesNet-JPip	0.7686	0.0418	0.7695	0.7686	0.7691
χ^2-BayesNet-JRip	0.7281	0.055	0.7263	0.7263	0.7272
DF-BayesNet-JRip	0.729	0.055	0.7335	0.729	0.7313

Comparison of three characteristic evaluation functions, 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. The system **χ^2**-VFI-VFI have the best performance.

In order to observe the classification results more intuitive, the map of national influenza attention in March 31st is drawn, and the flu weibos classification results in

Shanghai are shown in Figure 5. As can be seen from the diagram, although there is a high degree of attention on Shanghai, but the actual number of infections is not so much. People pay more attention to the news. The number of Sina weibos classified as infections in March 31st accounted for 3%, about 420 cases.

Even after filtration, the number of influenza infections is still larger than NHFPC measured (99 case in March). 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 D. Thirdly, the authenticity of each flu weibo also cannot be promised.

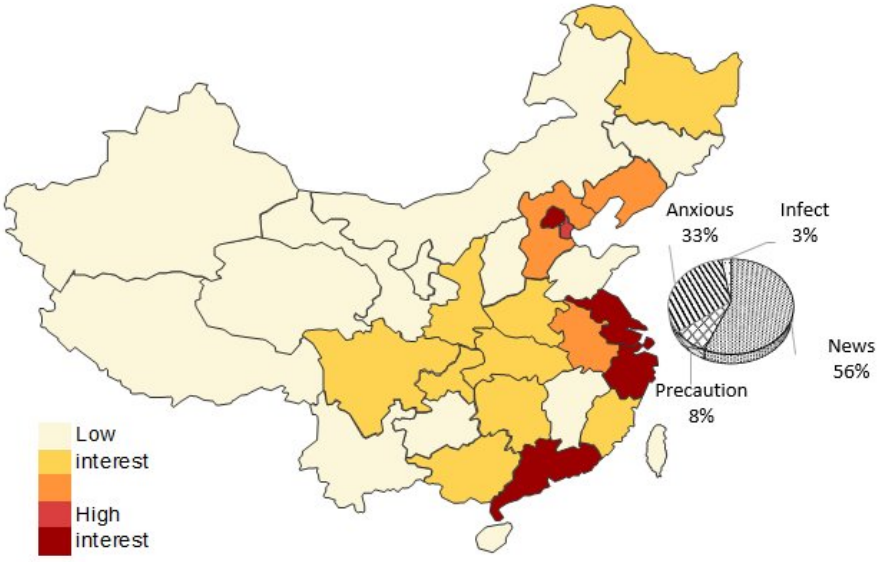

Fig. 5. The map of national flu attention in March 31st distributed in China

5 Conclusions and Future Work

This work focuses on both injection and non-injection flu weibos. By analyzing the number of flu weibos, we can detect outlier events of any province timely and effectively. Not only the point of unexpected events can be detected, outlier patterns caused by high continued attention can also be detected by using the sliding window - density detection method. Some of the detected event has a nationwide influence, taking March 31, 2013 as an example, the first H7N9 diagnosis lead to a surge in the number of flu weibos, but some only caused attentions in the local area. For instance March 24, 2013 the news "Japan research center found that drinking yogurt can effectively reduce the probability suffering from flu" attract users in Beijing attention.

According to the characteristics of social networks, the designed SN-TDT and twice iteration classification method can extract the hot topic and the analyze users' states accurately. In the future, we will study the flu weibos on temporal and spatial, and capture how the disease spread geographically.

References

1. Signorini, A., Segre, A.M., Polgreen, P.M.: The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS One 6(5), e19467 (1946), doi:10.1371/journal.pone.0019467
2. Achrekar, H., Gandhe, A., et al.: Predicting Flu Trends using Twitter Data. In: The First International Workshop on Cyber-Physical Networking Systems, pp. 702–707 (2011)
3. Sadilek, A., Kautz, H.: Vincent Silenzio: Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 136–142 (2012)
4. Prieto, V.M., Matos, S., Alvarez, M., et al.: Twitter: A Good Place to Detect Health Conditions. PLoS One 9(1), e86191 (2014), doi:10.1371/journal.pone.0086191
5. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: 19th International Conference on World Wide Web, pp. 851–860 (2010)
6. Breunig, M.M., Kriegel, H.-P., et al.: LOF: Identifying Density-Based Local Outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (2000)
7. Ferdousi, Z., Maeda, A.: Unsupervised Outlier Detection in Time Series Data. In: The 22nd International Conference on Data Engineering Workshops, 0-7695-2571-7 (2006)
8. Lee, C., Lee, G.G.: Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing & Management 42(1), 155–165 (2006)
9. Chen, Y.T., Chen, M.C.: Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications 38(4), 3085–3090 (2010)
10. Azam, N., Yao, J.: Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications 39(5), 4760–4768 (2012)
11. Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Carnegie-Mellon Univ. Pittsburgh Pa Dept of Computer Science. No. CMU-CS 96-118 (1996)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
13. Joachims, T.: Learning to classify text using support vector machines: Methods, theory and algorithms, p. 205. Kluwer Academic Publishers (2005)