Real Time Early-stage Influenza Detection with Emotion Factors from Sina Microblog

Xiao SUN
School of Computer and Information
Hefei University of Technology
Hefei, Anhui, China
Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine
suntian@gmail.com

Jiaqi YE
School of Computer and Information
Hefei University of Technology
Hefei, Anhui, China
Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine
lane_3000@163.com

Fuji REN
School of Computer and Information
Hefei University of Technology
Hefei, Anhui, China
Faculty of Engineering, University of Tokushima
Tokushima, Japan
ren2fuji@gmail.com

Abstract

Influenza is an acute respiratory illness that occurs every year. Detection of Influenza in its earliest stage would reduce the spread of the illness. Sina microblog is a popular microblogging service, provides perfect sources for flu detection due to its real-time nature and large number of users. In this paper we investigate the real-time flu detection problem and describe a Flu model with emotion factors and sematic information (em-flu model). Experimental results show the robustness and effectiveness of our method and we are hopeful that it would help health organizations in identifying flu outbreak and take timely actions to control.

1 Introduction

Influenza is a highly contagious acute respiratory disease caused by influenza virus. As the highly genetic variation, influenza can cause global epidemic, which not only brought huge dis-asters to people's life and health, but also have significant disruptions to economy. There are about 10-15% of people who get influenza every year and results in up to 50 million illnesses and 500,000 deaths in the world each year. Influenza is a worldwide public health problem and there are no effective measures to control its epidemic at present. The prevalence of influenza in China is one of the most notable problems.

The epidemic of SARS, H1N1 and H5N9 influenza make us realized that people really need to expand surveillance efforts to establish a more sensitive and effective precaution indicator system for infectious disease forecasting. In order to detect influenza epidemic timely and im-prove the ability of early precaution, the research of early forecasting technique is urgently needed.

Nowadays influenza surveillance systems have been established via the European Influenza Surveillance Scheme (EISS) in Europe and the Centre for Disease Control (CDC) in the US to collect data from clinical diagnoses. The research of forecasting methods started relatively late in China and these systems have about two-week delay. The need for efficient sources of data for forecasting have increased due to the Public health authorities' need to forecast at the earliest time to ensure effective treatment. Another surveillance system is Google's flu trends service which is web-based click flu reporting system. Google's flu trend uses the linear model to link the influenza-like illness visits.

Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pages 80–84,
Dublin, Ireland, August 23-29 2014.

Sina Weibo is a Chinese popular microblog service that can potentially provide a good source for early stage flu detection due to its large data scale and real-time features. When flu breaks out, infected users might post related microblog with corresponding emotions in a timely way which can be regarded as indicators or sensors of Influenza. Based on the real-time data of mi-croblog, there has been many applications such as earthquake detection (Sakaki T et al., 2010), public health tracking (Collier N, 2012; Paul M J et al., 2011) and also flu detection (Achrekar H et al., 2011; Culotta A,2010).

The measures of collecting clinical diagnoses and web-based clicks on key word with linear model are quite good but not fair enough. Our research tries to use the big real-time data as re-sources and design a machine learning mode with the emotional factors and sematic information to help find the break point of influenza.

The rest of this paper is organized as follows: In section 2, we describe our Flu model with emotion factors (em-flu model). We describe the preparation of our dataset in Section 3. Exper-imental results are illustrated in Section 4. We conclude this paper in Section 5.

2 Em-flu Model

Existing works on flu prediction suffer the following limitations: Spatial information is seldom considered and sematic or emotion factors are out of consideration. To address this problem, in this paper, we try to introduce an unsupervised approach called Em-flu Markov Network for early stage flu detection. Spatial information are modelled in a four-phase Markov switching model, i.e. non-epidemic phase (NE), rising epidemic phase (RE), stationary epidemic phase (SE) and declining epidemic phase (DE). Our approach assumes microblog users as "sensors" and collective posts containing flu keywords as early indicators. Our algorithm can capture flu outbreaks more promptly and accurately compared with baselines. Based on our proposed algorithm, we create a real-time flu surveillance system. For early stage flu detection, we use a probabilistic graphical Bayesian approach based on Markov Network. The key of the flu detection task is to detect the transition time from non-epidemic phase to epidemic phase.

Basically, our model is based on a segmentation of the series of differences into an epidemic and a non-epidemic phase using a four-stage Markov switching model. Suppose we collect flu related microblog data from N location. For each location $i \in [1,N]$, we segment the data into a time series. $Z_{i,t}$ denotes the phase location i takes on at time t. $Z_{i,t}=0,1,2,3$ correspond to the phase NE, RE SE and DE. $Y_{i,t}$ is the observant variable, which denotes the number of flu related microblog at time t, join in location i. $\Delta Y_{i,t} = (Y_{i,t} - Y_{i,t-1})/ Y_{i,t-1}$. The underlying idea of Markov switching models is to associate each $Y_{i,t}$ with a random variable $Z_{i,t}$ that determines the conditional distribution of $Y_{i,t}$ given $Z_{i,t}$. In our case, each $Z_{i,t}$. is an unobserved random variable that indicates which phase the system is in. Moreover, the unobserved sequence of $Z_{i,t}$. follows a four-stage Markov chain with transition probabilities. For location i, N(i) denotes the subset containing its neighbors. We simplify the model by only considering bordering states in N(i).

We model the spatial information in a unified Markov Network, where the phase for location i at each time is not only dependent upon its previous phase, but its neighbors. In this work, for simplification, we only treat bordering States as neighbors. Since the influence from non-bordering States can be transmitted through bordering ones, such simplification makes sense and experimental results also demonstrate this point. A Generalized Linear Model is used to integrate the spatial information in a unified framework. For location i at time t, the probability that Zi,t takes on value Z is illustrated as follows:

$$P = \Pr(Z_{i,t} \mid Z_{j,t+1}, Z_{i,t-1}) = \frac{\exp(\psi Z_{i,t-1}, Z_{i,t} + \psi Z_{i,t}, Z_{i,t+1} + \sum \Theta Z_{j,t}, Z_{i,t})}{\sum_z \exp(\sum \Theta Z_{j,t}, Z_{i,t} + \psi Z_{i,t-1}, Z_{i,t} + \psi Z_{i,t}, Z_{i,t+1})}, j \in N(i) \qquad (1)$$

Where Ψ and Θ respectively correspond to parameters that control temporal and spatial influence. We give a non-informative Gaussian prior for each element in Ψ and Θ:

$$\Theta_{i,j} \sim N(0, \delta_{i,j}^2) \qquad \psi_{i,j} \sim N(0, \Phi_{i,j}^2) \qquad (2)$$

Next, we describe the characteristics for the dynamics of different phases. Generally speaking, the course of influenza may last a week or two, for a single microblog user, we believe his or her microblog contents will record a series of feelings when user is sick or catching flu. When a person got the flu, he will go through NE, RE, SE, DE phases; the main emotion in these four phases would natu-

rally change by the phase change to another phase. All these individuals' data could be combined into datasheet segmented by time. From the statistics theories, the dynamics for NE, RE, DE and SE can be characterized as Gaussian process:

$$\Pr(\Delta Y_{i,t} \mid z) \sim N(\mathrm{E}_{day(t)}, \delta^2_{day(t)}) \quad (3)$$

Where $\mathrm{E}_{\mathrm{day(t)}}$ corresponds to the average microblog records' number every day, and $\delta^2_{day(t)}$ corresponds to the variance of the records.

3 Data Preparation

We extend our earlier work on Sina microblog data acquisition and developed a crawler to fetch data at regular time intervals. We fetched microblog records containing indicator words shown in Table 1 and collect about 4 million flu-related microblog starting from January 2013 to January 2014. Location details can be obtained from the profile page. We select tweets whose location are in China and discard those ones with meaningless locations.

Indicator words	止咳药(pectoral),输液(transfusion),伤风(cold),流涕(running nose),流感(flu),咳嗽(cough),抗生素(antibiotic),喉咙疼(Sore throat),感冒(influenza),发烧(fever),发高烧(high fever),鼻涕(snot)

Table 1: Indicator seed words set for data collection

Not all microblog containing indicator keywords indicate that the user is infected. Meanwhile the indicator words list may not be perfect, so the indicator words list needs to expand from the data we have and the dataset needs to be processed before be used for our task.

The words in Table 1 will be used as seed words to find the initial dataset and then computing vector in the dataset to find other keyword which can be the representations of seed words. In this way, words list could be expanded and adapt the changes of cyber word. The necessity of filtering in real-time task has been demonstrated in many existing works (Aramaki E et al., 2011; Sakaki T et al., 2010).To filter out these bias tweets, we first prepared manually labeled training data, which was comprised of 3000 microblog records containing key words. We manually annotate them as positive examples and negative ones.

We built a classifier based on support vector machine. We use SVMlight with a polynomial kernel, and employ the following simple text-based features.

Feature A: Collocation features, representing words of the query word within a window size of three.

Feature B: unigrams, denoting the presence or absence of the terms from the dataset.

Performances for different combinations of features are illustrated at Table 2. We observe that A+B is much better than A or B. So in our following experiments, microblog are selected according to a classifier based on feather A+B.

Features	Accuracy	Precision	Recall
A	84.21%	82.31%	89.40%
B	85.10%	84.92%	87.00%
A+B	87.40%	88.75%	**89.64%**

Table 2: Result of different combinations of features for filtering

We briefly demonstrate the relatedness between microblog data and CNIC (Chinese National Influenza Center) surveillance weekly report data, which would support the claim that microblog data can be used for the flu detection task. We observe that performing svm filtering and microblog selection would definitely make microblog data more correlated with real world CNIC data.

For these flu-related microblog records, we generate another microblog web crawler to deal with every record. For every record's user, we use this tool to backup user's microblog content and cut records by a window of time with one week before and after the flu-related microblog record which we had captured. Then the emotional SVM is established to help get the trend of these series of microblog records.

4 Experiments and Data Analysis

The main goal of our task is to help raise an alarm at those moments when there is a high probability that the flu breaks out. In real time situations, for each time, available data only comes from the previous days, and there is no known information about what will happen in the following days or week. By adding the data day by day, we calculate the posterior probability for transiting to epidemic states based on previous observed data. The sum over parameter $Z_{i,t-1}$ and $Z_{j,t}$ makes it infeasible to calculate. We use Gibbs Sampling by first sampling $Z_{i,t-1}$ and $Z_{j,t}$ first and then attain the value of $Z_{i,t}$ given Zi,t-1,Zi,t-1,…:

$$Z_{i,t} = \arg\max P(Z_{i,t} = z \mid Z_{j,t}, Z_{i,t-1}, ..., Y_{j,t}, Y_{i,t-1}, ...) \quad (3)$$

Figure 1 shows the global distribution of DE, SE and RE in the year of 2013. The left hand side figure corresponds to number of flu-related microblog records overtime. Purple symbols denote the phase of RE, red symbols denote the phase of SE and white symbols denote the phase of DE.

Figure 2 shows the result of searching key words like influenza on Baidu Index platform. Compared to Figure 1 seems our influenza curve matches well. The interesting thing we observe from figure 1 is that if the percentage of RE > 0.5, there is strong possibility to convince the flu alarm is coming.

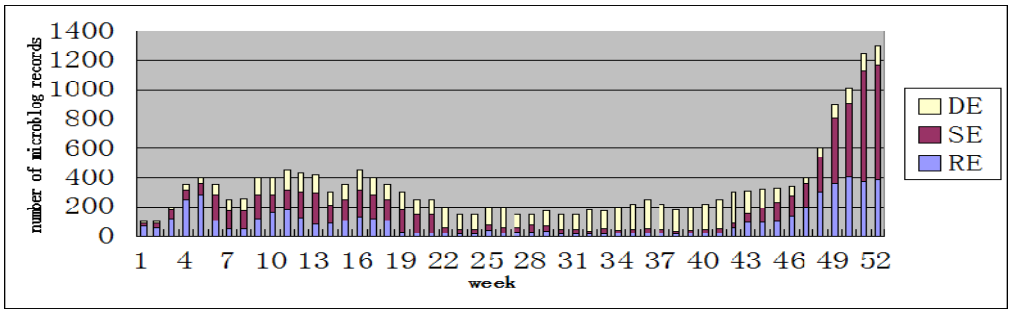


Figure 1: Predictions of the year 2013

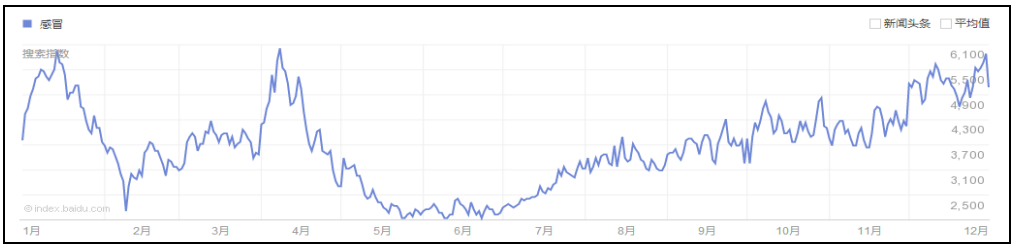


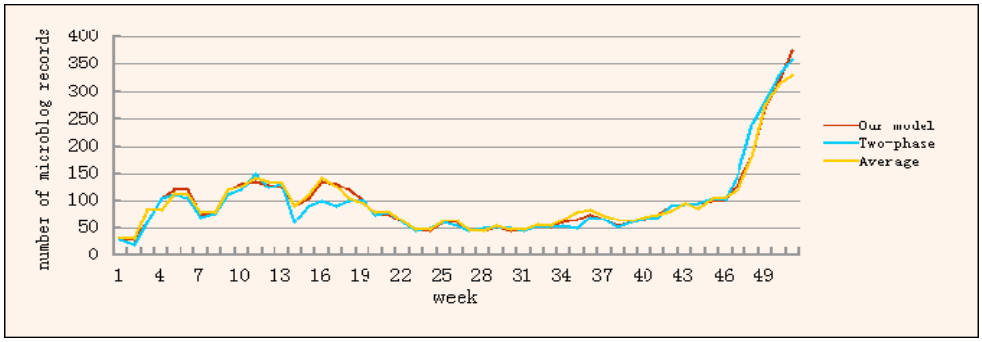
Figure 2: Searching Resutl on Baidu Index platform

For comparison, we employ the following baseline in this paper:

Average: Uses the averager frequency of micrblog records containing keywords based on previous years as the threshold.

Two-Phase: A simple version of our approach but using a simple two-phase in Markove network.

We only report partial experimental results for one province. As we can see from figure 3, our model can best fit the actual microblog data and semms stable. The other two measures also represent the actual truth but not stable enough.



Figre 3: Prediction of Anhui province of the year 2013

5 Conclusions

In this paper, we introduced an unsupervised Bayesian model based on Markov Network based on four phases and microblog emotional factors are appended in the model to help detect early stage flu detection on Sina Microblog. We test our model on real time datasets for multiple applications and experiments results demonstrate the effectiveness of our model. We are hopeful that our approach would help to facilitate timely action by those who want to decrease the number of unnecessary illnesses and deaths. At present, the method also has a few shortcomings; we will continually develop it for further research and exploration.

ACKNOWLEDGMENT

The work is supported by National Natural Sci-ence Funds for Distinguished Young Scholar(No.61203315) and 863 National Advanced Tech-nology Research Program of China (NO. 2012AA011103), and also supported by the Funding Project for AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, HeFei University of Technology.

Reference

Sakaki T, Okazaki M, Matsuo Y. 2010. Earthquake shakes Twitter users: real-time event detection by so-cial sensors[C]//Proceedings of the 19th international conference on World wide web. ACM, 851-860.

Collier N. 2012. Uncovering text mining: A survey of current work on web-based epidemic intelli-gence[J]. Global public health, 7(7): 731-749.

Paul M J, Dredze M. You are what you Tweet: Analyzing Twitter for public health[C]//ICWSM. 2011.

Achrekar H, Gandhe A, Lazarus R, et al. 2011. Predicting flu trends using twitter data[C]//Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. IEEE, 702-707.

Culotta A. 2010. Towards detecting influenza epidemics by analyzing Twitter messag-es[C]//Proceedings of the first workshop on social media analytics. ACM, 115-122.

Aramaki E, Maskawa S, Morita M. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Asso-ciation for Computational Linguistics, 1568-1576.

Lamb A, Paul M J, Dredze M. 2013. Separating fact from fear: Tracking flu infections on twitter[C]//Proceedings of NAACL-HLT.789-795.

Sakaki T, Okazaki M, Matsuo Y. 2010. Earthquake shakes Twitter users: real-time event detection by so-cial sensors[C]//Proceedings of the 19th international conference on World wide web. ACM, 851-860.

Achrekar H. 2012. ONLINE SOCIAL NETWORK FLU TRACKER A NOVEL SENSORY APPROACH TO PREDICT FLU TRENDS[D]. University of Massachusetts,

Aschwanden C. 2004.Spatial Simulation Model for Infectious Viral Diseases with Focus on SARS and the Common Flu[C]//HICSS.