Predicting Popularity of Microblogs in Emerging Disease Event

Jiaqi Liu[1], Zhidong Cao[1(✉)], and Daniel Zeng[1,2]

[1] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{Jiaqi.liu, Zhidong.cao, dajun.zeng}@ia.ac.cn
[2] University of Arizona, Tucson, AZ, USA

Abstract. During emerging disease outbreaks, massive information are disseminated through social network. In China, Sina microblog system as the biggest social network provide a novel way to monitoring the development of emerging disease and public awareness. However, only a small percentage of microblogs could wide spread. Therefore, predict popularity of microblogs timely are meaningful for emergency management. In this paper, a Judgment method for popularity level prediction of microblog is proposed and the temporal pattern between cases number and repost number is verified. Repost number is considered to measure the impact of microblogs. To predict the popularity of microblogs, Granger causality test was used to verify the temporal correlation pattern between development of disease and public concern while an Judgment method based on five classical classification models were proposed. Through analyses, case number of emerging disease are Granger causality of the popularity level of microblogs and the regression model got the best result when lag was three. By Judgment method, more than 86 % microblogs can be classified correctly. The proposed Judgment method based on user, microblog and emerging disease information could analysis the popularity level of microblogs speedily and accurately. This is important and meaningful for monitoring the development of future public health event.

Keywords: Microblogs · Popularity prediction · Granger causality · Classification

1 Introduction

During emerging disease outbreaks, massive information are disseminated not only through online announcements by government agencies but also through lots informal channels [1]. Massive freely available Web-based sources of information make digital disease surveillance possible [2, 3]. Systems based on scanning news media (e.g., GPHIN [4], MedISys [5]), modeling search query (e.g., Google [6], Baidu [7]), and monitoring real-time social media (e.g., Twitter [8], Sina Microblog [9]) were created. In particular, Sina Microblog can be invaluable source of real time individual data in China [10] and millions of users are willing to spread health-related information [11, 12]. However, not all health related messages posted in the microblog system are important that only few of them could attracted public attention and became popular

Y. Chen et al. (Eds.): WAIM 2014, LNCS 8597, pp. 3–13, 2014.
DOI: 10.1007/978-3-319-11538-2_1

through the network [13]. Therefore, concerning and tracking all microblogs would consume a large number of resources and might obtain no achievement. In other words, predict popular microblogs become an urgent problems for situation forecast [14].

Several popularity prediction researches in social media has already obtained some interesting findings. The basic information of user and microblog has strong impact on its spreading scale [15]. More specifically, large repost number tweets would spread quickly in users and exert big influence on users [16] as well as the impact of the temporal [17–19] and spatial [20, 21] characteristics of microblogs are also investigated comprehensively [22]. In addition, some previous research indicated that machine learning methods (e.g., Bayesian Network, Support Vector Machine) have strong classification ability to identify the microblog with high popularity through some basic microblog features [23, 24] and the posting number of tweets and cases number of emerging disease are highly correlated in time [25]. In this paper, the repost number we used is the most common way to evaluate the popularity level of microblogs in information diffusion area [14].

As we consider the popularity level prediction of microblogs about emerging disease event is classification problem, the main work of this paper is proposing a Judgment method by integrating results of five classic classification algorithms. Emerging disease event of human infection with influenza A(H7N9) provided a golden opportunity to test our Judgment Method. Similar with previous studies, several quantitatively features of users and related microblogs are considered in our method. In addition, whether and how the disease situation will influence the popularity of the related microblogs is also an interesting question which is not been studied. In this paper, Granger causality test was used to verify the temporal correlation between the popularity level of microblogs and the development of emerging disease. Based on these analyses, some basic information about the emerging disease were also used in our method. Through analyses, the proposed Judgment method offer an new way to predict the popularity level of microblogs about emerging disease and have important implications for the future digital disease surveillance.

The rest of this paper is organized as follows. Section 2 present the methods we used. Section 3 describes the experimental results. Finally, Sect. 4 concludes the paper and outlines the directions for future work.

2 Methods

2.1 Data Collection and Datasets

H7N9 Case Number. Human infection with influenza A(H7N9) was notified by National Health and Family Planning Commission of the People's Republic of China(NHFPC) on Mach 31, 2013 [26]. It is the first time that human infection with this type of avian virus has been identified. 130 cases were laboratory-confirmed in 8 provinces and 2 municipalities up to May, 14 [27]. After official announcement of three human H7N9 cases on March 31, detailed disease information was daily updated on NHFPC's website till April 24 and then H7N9 related health information was weekly reported. We manually collected these detail message of each human H7N9 cases. Only location, date and number were used in this paper and no personal information is involved (see Fig. 1).

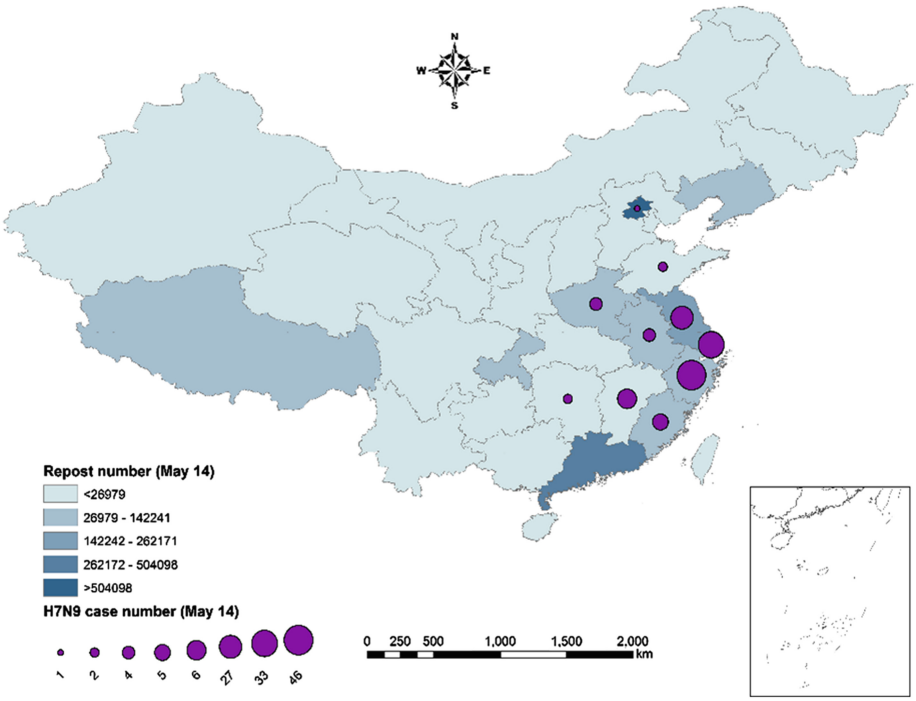

Fig. 1. The geographic distribution of repost and case

Microblogging Data. Sina microblogs (Weibo), which has more than 500 million users and 46.2 million daily active users, is the biggest microblog system ever used in China [28]. After H7N9 event happened, lots of related microblogs were posted and reposted. Hot word index(HWI) (Fig. 2) is calculated according to massive daily microblog data by Sina microblogging system [29]. It reflects the mentioned number of a keyword, includes original and repost, in the whole microblog system. We developed an Internet Epidemic Surveillance Platform(IWSP) which continuously collected textual information from Weibo and other websites or BBS related to public health. Since Weibo has lots of users and a large number of microblogs were created over time [30], we cannot get all of these messages. This platform pay more attention to Internet celebrities, opinion leaders, news agencies and user or account which related to public health. We filter out microblogs which contain keyword or hashtag "H7N9". Unlike Twitter, Weibo count number of repost and comment for each microblog, we can easily evaluate the influence scope of each original microblog. Hence, we excluded repost microblogs. Between March 31 and May 14, 2013, we archived 146,684 original microblogs from 105,721 users. In addition, we updated users' information, such as followers, friends and statuses number and registered province, in August 7.

2.2 Temporal Correlation Analysis

After H7N9 event happened, lots of people concerned about the development of the emerging disease and some of them posted or reposted related microblogs. To verify

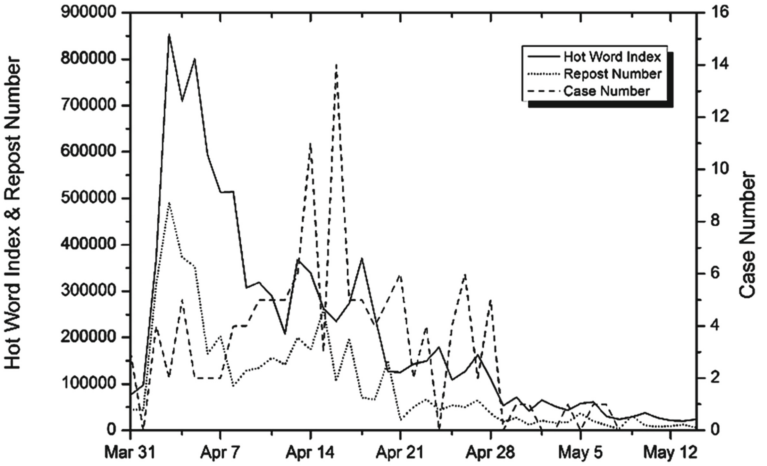

Fig. 2. The time series of hot index, repost number and case number

whether these two aspects have causal relationship, especially temporal correlation pattern, Granger causality test were used. The Granger causality test [31] is a statistical hypothesis test for determining whether one time series is useful in forecasting the another. When values in series X provide statistically significant information about the value or the future values in the other series Y, the first is said to be the Granger causality of the second series.

$$y_t = \sum_{q}^{i=1} \alpha_i x_{t-i} + \sum_{q}^{j=1} \beta_j y_{t-i} + u_t \tag{1}$$

In addition, the stationary of two time series is a preconditions for Granger causality test, which means that the joint probability distribution of the series does not change when shifted in time, otherwise spurious regression problems could occur. Thus, unit root test need to be done on both series in advance. The model for unit root test in this paper is Augmented Dickey-Fuller(ADF) test with zero lag length. Due to different trend of the two time series, the ADF test for case number series is just include intercept while trend and intercept are both considered for reposting number series.

2.3 Popularity Level Prediction Model

Predict popularity level of microblogs is a typical classification problems and several models have been proposed. Five classical classification models in WEKA [32] were used in our experiments, which is logistic regression model, J48 decision tree model, Bayes net classification model, support vector machines and adaptive boosting model. Logistic regression model is a type of probabilistic statistical classification model which usually use one or more continuous variables to make the prediction and success to handle the threshold question [33]. Decision tree model uses an attribute selection

measure to determine which attribute has high distinguish [34]. Bayes net is based on Bayes' theorem of posterior probability [35]. Sequential minimization algorithm(SMO) is an important application of support vector machine which have yielded excellent generalization performance on lots problems [36]. And adaptive boosting model, as voting classification algorithm, has been shown successful in improving the accuracy of certain classifiers [37]. However, all these classification models have their advantages and disadvantages. Hence, we proposed the Judgment method to integrate all results from the five models by regarding the majority result as the final popularity level. Several common evaluation index, such as accuracy, precision, recall and F-value, were used to measure the classification result.

Table 1. Microblog features

Features	Description	Type
Follower	# of users who follows the author of the microblog	Numerical
Following	# of users the author is following	Numerical
Status	# of microblogs posted by the author since the creation of the account	Numerical
Location	The reported location or register location	Nominal
Lag of Days	# of days after March 31 when the microblog posted	Numerical
Local Lag of Days	# of days after the province been infected when the microblog posted	Numerical
Case	# of humans infected by H7N9 avian influenza	Numerical
URL	Whether the microblog contain URL	Nominal
Tag	Whether the microblog tag other user with '@'	Nominal

After deciding the classification models and the judgment method we use, the input and output of these models need to be determined. The features we selected to make the prediction are shown in Table 1. The first four features are user's basic information, which we can direct get through Sina Application Programming Interface(API). The province is nominal variable while others are numerical. The following two numerical features are time information about microblog posted. Microblog posted in different day got varying concerns and local news agencies also play different role after the area got first attacked. H7N9 case number is also considered as numerical. The last two features are two typical and easy accessible nominal features. Microblog which contain URL [17] or tag [18] other user would be marked. For the popularity level, we separated microblogs into two categories according to whether it had been reposted more than or equal to ten, which is a balance between event description and resource utilization. 6.77 % of the microblogs were marked as high reposted microblog (positive tuples) as this criteria. Since class-imbalanced dataset will significant influence the result of most classify model, we used under sampling process to decrease the number of negative tuples. We kept all 9932 microblogs which reposted more than or equal to ten, and then random chose microblogs in negative tuples to constitute a training set with 20000 microblogs.

3 Results

3.1 Temporal Correlation Analysis

The results of ADF unit root test on case and repost number series are shown in Table 2. Both results reject the null hypothesis ($P_{case} < 0.001$, $P_{RTNum} < 0.01$), which means these series don't have unit root. The coefficients of the regression model are all statistical significant ($P < 0.01$). The Akaike information criterion and Schwarz criterion is small enough while Durbin-Watson statistic indicated autocorrelation. According to these results, case and repost number series are both stationary time series.

Table 2. ADF unit root test result

CASE				
	Coefficient	Std. Error	t-Statistic	Prob.
ADF test statistic	-	-	-4.119420	0.0023
Variable: CASE(-1)	-0.586431	0.142358	-4.119420	0.0002
Variable: C	1.664457	0.594751	2.798575	0.0077
Statistic	Value	Statistic		Value
R-squared	0.287769	Akaike info criterion		4.933848
F-statistic	16.96962	Schwarz criterion		5.014948
Prob(F-statistic)	0.000174	Durbin-Watson stat		2.445098
Repost Number				
	Coefficient	Std. Error	t-Statistic	Prob.
ADF test statistic	-	-	-4.334661	0.0067
Variable: RTNum(-1)	-0.547483	0.126304	-4.334661	0.0001
Variable: C	143347.4	36778.24	3.897613	0.0004
Variable:@TREND(1)	-3910.022	1110.155	-3.522051	0.0011
Statistic	Value	Statistic		Value
R-squared	0.318543	Akaike info criterion		25.08405
F-statistic	9.582602	Schwarz criterion		25.20570
Prob(F-statistic)	0.000385	Durbin-Watson stat		1.984863

Table 3 shows that the case series is Granger causality of the repost series. With lag number of days varying from one to five, the results of statistical tests are all significant in different level. The regression model got the best result when lag was three which means the sum of first three lags of case series can predict the most similar results with repost series.

From the results obtained so far, several conclusions can be made. First, new case number will affect users' attention of health event while case series is causality of repost series. Second, the effect of new case number would last for a few days while user would swing their attention on the health event base on the changing trend of the emerging disease. Third, the repost behavior have hysteresis that some users does not log in frequently and might repost the microblog a few days later. Overall, new case number will affect people's attention and the reposting scale of public health event to a

certain extent. However, as long as no burst point of the event, people's attention will gradually decay after reaching the peak. The decreasing trend will not fundamentally changed even sporadic small-scale events.

Table 3. Granger causality result

Lags	Observations	F-Statistic	Prob.
1	44	2.63601	0.1121
2	43	3.51495	0.0398
3	42	4.79999	0.0066
4	41	2.74538	0.0453
5	40	2.14205	0.0885

3.2 Popularity Level Prediction Model

When finished modeling, we tested our models on training dataset first. The results of five classical classification models are shown in Fig. 3. All models achieved good results. Since the training dataset is balanced data, we assigned the β of F-value equal to one which means same weight to precision and recall was given. The F_1-value of models are within the scope of 0.72 (SMO) to 0.90 (J48 decision tree). Moreover, J48 decision tree model obtained the best result in all four measures. In other words, the result indicated that the attributes we selected are effective and information gain ratio showed good performances. Logistic model and adaptive boosting models are good at precision rate (both 0.88) which implies that the high popularity level microblogs they choice out is more likely to be real high popularity level microblogs. In addition, Bayes net and adaptive boosting model got relative high recall (both 0.82) and accuracy rate (Bayes 0.84, Adaboost 0.86) that most of the high popularity level microblogs were found. Majority rules were used to combine the results and the accuracy of Judgment method result is 0.86 while the F_1-value is 0.85. Even if this result is not the highest, all four measures of Judgment method were 3 % higher than the average. Overall, most of the high popularity level microblogs are identified in balanced training dataset.

Then, we test our prediction model on the whole dataset. As class-imbalanced dataset it is, the recall rate is the most important measure for models [38]. Therefore, we assigned $\beta = 2$ which weights recall twice as much as precision. The results of models testing on whole dataset are shown in Fig. 4. The highest F_2-value is 0.69 (J48 decision tree) while the Judgment method got 0.65. For the recall rate, J48 decision tree model obtained the highest value (0.89) while Bayes net, adaptive boosting and Judgment method also got acceptable value above 0.80. The majority of important microblogs were recalled. The Judgment method, with the second highest F_2-value (0.65), achieved the highest precision rate (0.37). Because the dataset is highly imbalanced, even though the precision is relative low, lots of monitoring costs are still save. The density of important microblog was improved from 6.77 % to more than one-third. All models performed very well on accuracy within a range from 0.86 (Bayes net) to 0.89 (Judgment method). Most of the massive low popularity level microblogs are identified and filtered out.

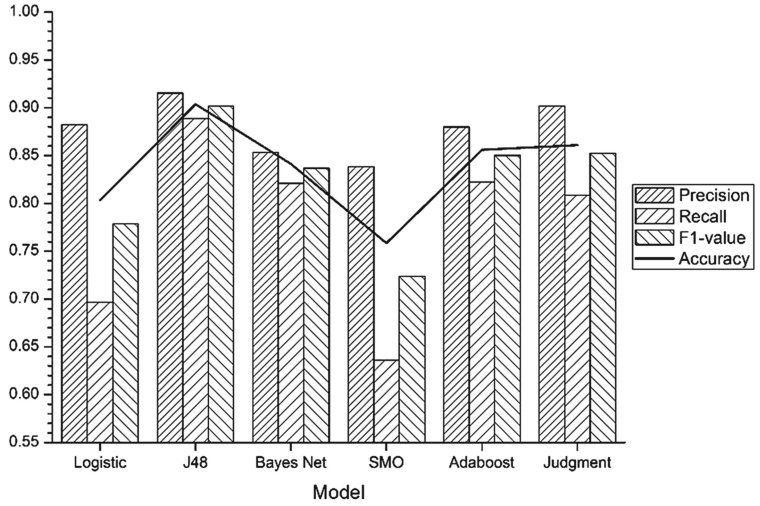

Fig. 3. Classification results on training dataset

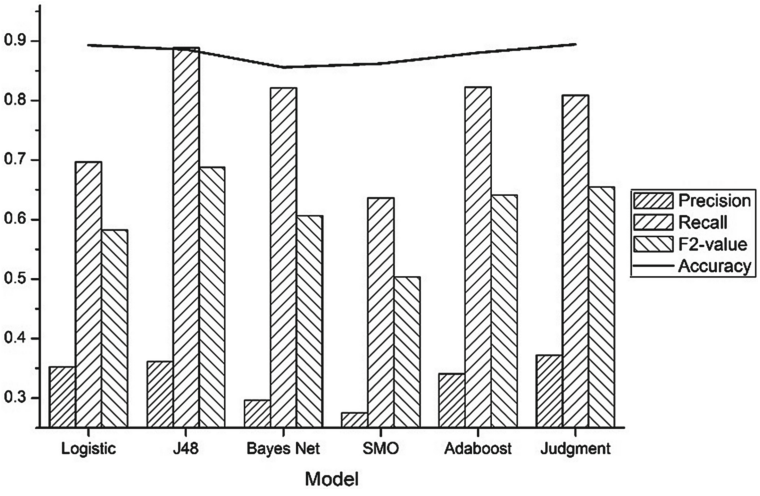

Fig. 4. Classification results on whole dataset

The results obtained by our proposed model proved that the potential popularity of microblogs can be identified exactly through some features of microblogs and disease information. Judgment can be made as soon as the microblog crawled by our platform. Microblogs with high popularity level would be preserved and used for further analysis immediately while low popularity level microblogs would be discarded to save resources.

4 Discussion

With the development of social network and digital disease surveillance, a novel monitoring approach on public health have been proposed and the further intervention would more tangible. The results of our study indicate that the new case number of emerging infectious diseases is the Granger causality of the repost number of microblogs and the new case number will significant affect the influence of microblogs. Therefore, more attention would pay when large number of emerging disease occur to avoid false information or rumor occupy public opinion. Either health related guided approach or warning messages about the emerging disease can widely spread. Moreover, since the sum of first three lags of case series is most predictable for predicting daily microblogs level through the analyses, the control measures should have continuity in order to get enduring influence.

In addition, potential high popularity microblogs have big impact on public awareness on emerging disease event and should be pay attention at any time. The proposed popularity level prediction model which based on the user, microblog and emerging disease information gets high accuracy and F-value on predicting popularity level of microblogs. Using this model, the influence of microblogs can be identified quickly at the time it posted and only potential high popularity level microblogs would be preserved and kept watching. In this way, not only lots storage space and monitoring cost would be saved, but the effect and trend of emerging disease event on Internet would be detected timely. The policymakers could also make judgments based on monitoring results of classification while some interventions could be implemented on both online and offline before the event causing uncontrollable consequences. Several Internet monitoring measures, such as highlight, recommend or delete, can make a hug difference.

Although we got some promising results, there are still several weaknesses to this study. Firstly, the dataset we used is a subset of whole H7N9 related data in Sina microblog system. Although we had tested the representation of our data, some further results might achieve with more comprehensive data. Secondly, the relationships between users might have some impacts on their repost behavior. Once someone in a small group with closely contacted comment on emerging infectious disease event, others have more possibilities to join the discussion. However, these network features were not consider in our research this time since they would significantly increase the complexity of the model. Thirdly, limitation also exist in content analysis that some burst events or words might influence on the popularity level of microblogs, but we didn't include these factors since we think that their impact had already contained in statistical results and wouldn't increase the accuracy visibly. These weaknesses will be our important direction for future research. Although limitations exist, this research has already verified the temporal correlation between emerging disease case number and repost number of microblogs. The proposed prediction model has sufficiently high accuracy and will be deployed in our IWSP platform as soon as possible.

Acknowledgments. This study was funded by National Natural Science Foundation of China (Nos.90924302, 91224008, 91024030, 91324007) and Important National Science & Technology Specific Projects (Nos.2012ZX10004801, 2013ZX10004218).

References

1. Brownstein, J.S., Freifeld, C.C., Mado, L.C.: Digital disease detection-harnessing the web for public health surveillance. New Engl. J. Med. **360**, 2153–2157 (2009)
2. Wilson, K., Brownstein, J.S.: Early detection of disease outbreaks using the internet. Can. Med. Assoc. J. **180**, 829–831 (2009)
3. Lampos, V., De Bie, T., Cristianini, N.: Flu detector - tracking epidemics on twitter. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part III. LNCS, vol. 6323, pp. 599–602. Springer, Heidelberg (2010)
4. The global public health intelligence network (GPHIN). http://www.who.int/csr/alertresponse/epidemicintelligence/en/
5. The medical information system (medisys). http://medusa.jrc.it/medisys/
6. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., et al.: Detecting influenza epidemics using search engine query data. Nature **457**, 1012–1014 (2008)
7. Yuan, Q., Nsoesie, E.O., Lv, B., Peng, G., Chunara, R., et al.: Monitoring influenza epidemics in china with search query from baidu. PLoS ONE **8**, e64323 (2013)
8. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: Bad news travel fast: A content-based analysis of interestingness on twitter. In: Proceedings of the 3rd International Web Science Conference (2011)
9. Salathé, M., Freifeld, C.C., Mekaru, S.R., Tomasulo, A.F., Brownstein, J.S.: Influenza A (H7N9) and the importance of digital epidemiology. New Engl. J. Med. **369**, 401–404 (2013)
10. Yang, J., Counts, S.: Comparing information diffusion structure in weblogs and microblogs. In: ICWSM (2010)
11. Fernandez-Luque, L., Karlsen, R., Bonander, J.: Review of extracting information from the social web for health personalization. J. Med. Internet Res. **13**, e15 (2011)
12. Kostkova, P.: A roadmap to integrated digital public health surveillance: the vision and the challenges. In: Proceedings of the 22nd International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, pp. 687–694 (2013)
13. Yang, J., Counts, S.: Predicting the speed, scale, and range of information diffusion in twitter. ICWSM **10**, 355–358 (2010)
14. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 57–58. ACM (2011)
15. Jenders, M., Kasneci, G., Naumann, F.: Analyzing and predicting viral tweets. In: Proceedings of the 22nd International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, pp. 657–664 (2013)
16. Starbird, K., Palen, L.: (how) will the revolution be retweeted? information diffusion and the 2011 egyptian uprising. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 7–16. ACM (2012)
17. Chew, C., Eysenbach, G.: Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak. PLoS ONE **5**, e14118 (2010)
18. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the Fourth ACM INTERNATIONAL Conference on Web Search and Data Mining, pp. 177–186. ACM (2011)
19. Lee, C., Kwak, H., Park, H., Moon, S.: Finding influentials based on the temporal order of information adoption in twitter. In: Proceedings of the 19th International Conference on World Wide Web, pp. 1137–1138. ACM (2010)

20. Hay, S.I., George, D.B., Moyes, C.L., Brownstein, J.S.: Big data opportunities for global infectious disease surveillance. PLoS medicine **10**, e1001413 (2013)
21. Paul, M.J., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: ICWSM (2011)
22. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600. ACM (2010)
23. Signorini, A., Segre, A.M., Polgreen, P.M.: The use of twitter to track levels of disease activity and public concern in the us during the influenza A H1N1 pandemic. PLoS ONE **6**, e19467 (2011)
24. Collier, N., Son, N.T., Ngoc, M.N.T.: Omg u got u? analysis of shared health messages for bio-surveillance. In: Semantic Mining in Biomedicine (2010)
25. Chunara, R., Andrews, J.R., Brownstein, J.S., et al.: Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. Am. J. Trop. Med. Hyg. **86**, 39 (2012)
26. Human infection with influenza A(H7N9) virus in china. http://www.who.int
27. Human infection with influenza A(H7N9) virus in china. http://www.chinapop.gov.cn/yjb/s3578/201305/67d505cd37eb4a419f17518bdbe05b54.shtml
28. Guo, Z., Li, Z., Tu, H.: Sina microblog: an information-driven online social network. In: 2011 International Conference on Cyberworlds (CW), pp. 160–167. IEEE (2011)
29. The hot index of hotword in sina microblog system. http://data.weibo.com/index/hotword
30. The registered users of sina microblog system exceeded 300 million and more than 100 million microblogs a day. http://news.xinhuanet.com/tech/2012-02/29/c_122769084.htm
31. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econom. J. Econom. Soc. **37**, 424–438 (1969)
32. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., et al.: The weka data mining software: an update. ACM SIGKDD Explor. Newsl. **11**, 10–18 (2009)
33. Bishop, C.M., et al.: Pattern Recognition and Machine Learning, vol. 1. Springer, New York (2006)
34. Lim, T.S., Loh, W.Y., Shih, Y.S.: A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach. Learn. **40**, 203–228 (2000)
35. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan kaufmann, San Francisco (2006)
36. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to platt's SMO algorithm for SVM classifier design. Neural Comput. **13**, 637–649 (2001)
37. Margineantu, D.D., Dietterich, T.G.: Pruning adaptive boosting. In: ICML. Citeseer, volume 97, pp. 211–218 (1997)
38. Barboza, P., Vaillant, L., Mawudeku, A., Nelson, N.P., Hartley, D.M., et al.: Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. PLoS ONE **8**, e57252 (2013)