Using multi-source web data for epidemic surveillance: A case study of the 2009 Influenza A (H1N1) pandemic in Beijing

Yuan Luo[1], Daniel Zeng[1,2], Zhidong Cao[1], Xiaolong Zheng[1], Youzhong Wang[1], Quanyi Wang[3], Huimin Zhao[4]

[1] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China

[2] MIS Department, University of Arizona, Tucson, Arizona 85721, USA

[3] Beijing Center for Disease Control and Prevention, Beijing 100013, China

[4] Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, USA

Abstract—**Timely and effective surveillance is critical for the prevention and control of epidemics. However, due to technical challenges and shortage of human resources, comprehensive and timely data collection required for effective surveillance, especially collection of data about sudden epidemic outbreaks, is still very difficult. In this paper, we propose the use of multi-source web data for epidemic surveillance. We use the 2009 Influenza A (H1N1) pandemic in Beijing as a case study to demonstrate the utility of our proposed approach. Experiments using data from the Beijing Center for Disease Control and Prevention (CDC) and several search engines show encouraging results. This case study also has direct practical values in the real setting.**

I. Introduction

Epidemics constitute an important source of safety threats to public health. As a new type of influenza, Influenza A (H1N1) swept the globe soon after its outbreak in Mexico in April 2009. As of November 2009, it had spread to 206 countries and regions worldwide, causing over 6,250 deaths [1]. It reached most cities in China, including Beijing, and severely impacted on public health, societal stability, and economic development.

Timely and effective surveillance is critical for epidemic prevention and control. Prompt data collection and trend discovery plays a vital role in epidemic surveillance and provides fundamental support to decision-making for epidemic prevention and control. Traditional surveillance mechanisms, characterized by the use of large systems and passive data collection, are insufficient for comprehensive, timely, and effective data collection, especially in case of sudden large-scale epidemics. When Influenza A (H1N1) was spreading to many parts of the world and causing large-scale pandemic in the population, the World Health Organization (WHO), recognizing the difficulty in prompt data collection, stopped requiring each country to report case-by-case details after the first confirmed case has been reported [2]. Besides relying on traditional data collection and analysis methods, a variety of other data sources, such as over-the-counter pharmaceutical sales data [3], have been tracked and analyzed for the purpose of epidemic surveillance. However, such data collection and analysis methods suffer from similar difficulties, due to technical challenges and shortage of human resources, as traditional methods do.

With the growing popularity of the Internet, now hundreds of millions of users search for information on health or diseases over the Internet every day. In China, by the end of June 2009, the number of search engine users had reached 235 million [4]. Several researchers have used web data in epidemic studies. Hulth et al. [5] monitored influenza outbreak by calculating the number of queries with influenza-related keywords on Swiss medical web sites. Researchers at Yahoo studied the relationships among the positive rate of disease, mortality, and the number of queries whose keywords include *flu* or *influenza* [6]. Researchers at Google record and analyze influenza-related search query

terms and report their results through Google Insights of Search, leading to detection of the outbreak of a flu one week ahead of the U.S. Center for Disease Control and Prevention [7]. Their research has been published in the journal *Nature* and aroused great attention in the field of public health. Google is collaborating with the centers for disease control and prevention of a number of countries, hoping to apply this research to establish an accurate and reliable mechanism for forecasting and monitoring influenza disease outbreaks [8].

In light of the initial success of previous research utilizing a single web data source, we further propose the use of multi-source web data to improve epidemic surveillance in this paper. We have tested the utility of our proposed approach through a case study of the 2009 Influenza A (H1N1) pandemic in Beijing. We collected and compared open-source information from multiple Internet search engines for analyzing and predicting the trend of the pandemic. Our experiments show that the approach is promising in alleviating the difficulty of data collection due to the fast spread of the pandemic and improving forecasting. This approach has direct practical values in the real setting and can effectively complement the existing mechanisms used by relevant authorities to improve decision-making and better control the outbreak.

II. Research Methods

Fig. 1 outlines our analysis procedure. We describe the steps in this section and discuss the results in the next section.

A. Data collection and preprocessing

We collected Influenza A (H1N1)-related data from the Beijing Center for Disease Control and Prevention (CDC), as well as open-access web sources. We used two kinds of web data. One consists of web queries submitted to two search engines, Google and Baidu, which are used by 32.7% and 92.9% of search engine users in China, respectively [9]. The other consists of online news available from news.google.com. We crawled all the open source data in the web query repositories of Google Trends

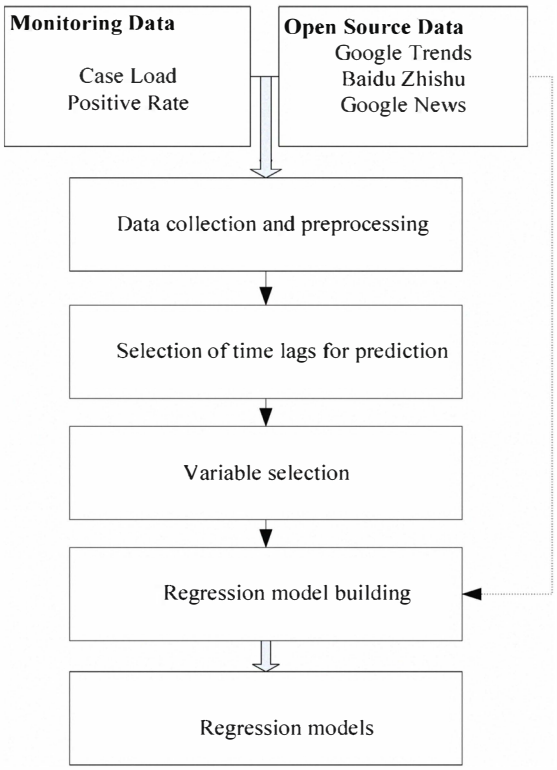


Fig.1. Data collection and analysis for A H1N1 influenza surveillance

(trends.google.com) , Baidu Zhishu (zhishu.baidu.com), and Google News (news.google.com). These repositories all contain time and location information. We focused our data collection to the time period from August 23, 2009 to January 2, 2010 in Beijing. Table I summarizes the data we collected.

Google Trends analyzes a portion of Google web searches to compute the ratio of the number of searches for each query term in a particular week to the total number of searches in a certain period, i.e., the trend of term t in weak w is

$$\text{trend}\,(t,w) = c * \frac{\text{Number of searches with term } t \text{ in week } w}{\text{Total number of searches in some period}}$$

This value may change over time as the denominator is changing. However, the ratio between the trends of two weeks is invariable. We therefore normalized a series of trends $G = (g_1, g_2, \ldots)$ by the first value g_1 to get an invariable series X_1, i.e., $x_{1i} = \frac{g_i}{g_1}$.

Baidu Zhishu shows how many searches have done for each term every day. Advanced news search in Google News can return the number of articles related to a keyword added to Google News in a week. We can get data related to the Beijing region using advanced search tools of both search

engines. To standardize the unit of all the data, we selected week as the unit of time interval.

The data from Beijing CDC include case load and positive rate. The case load is the number of laboratory-confirmed Influenza A (H1N1) cases. The positive rate (true-positive rate) is the ability of a test to correctly identify people with a certain condition. In other words, the positive rate is the ratio of the confirmed case load to the suspected case load.

TABLE I

SUMMARY OF RESEARCH DATA

Data Type	Explanation	Time Interval	Source
Open Source	Data about queries with the keyword "甲流" (A H1N1) in Beijing	Week	Google Trends
	Data about queries with the keyword "甲流" (A H1N1) in Beijing	Day	Baidu Zhishu
	Data about online AH1N1-related news	Week	Google News
Monitoring Data	Case load of A H1N1 in Beijing	Week	Beijing CDC
	Positive rate of A H1N1 in Beijing	Week	Beijing CDC

B. *Time series analysis*

We performed time series analysis to study the relationship between the Beijing CDC data (case load and positive rate) and the data from the three web sources. Related work at Yahoo! and Google showed that their methods using web search data can reveal trends about a week earlier than the U.S. CDC. We first determined an appropriate time lag (i.e., number of weeks) t_0 of prediction for a web source through correlation. Let $(X, Y_{t_0}) = ((x_{1+t_0}, y_1), (x_{2+t_0}, y_2), \ldots, (x_{n+t_0}, y_n))$ be a pair of time series, where X is the time series from a web source and Y_{t_0} is a time series from CDC. We calculated the correlation between the two time series

$$\rho(X, Y_{t_0}) = \frac{\mathrm{Cov}(X, Y_{t_0})}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y_{t_0})}}$$

. The t_0 that gave the maximum correlation was chosen for subsequent analysis.

C. *Stepwise regression*

We then selected the influencing factors (i.e., trends from web sources) for predicting the epidemic situation (characterized by case load and positive rate) about Influenza A (H1N1) through stepwise regression analysis. Stepwise regression is one of several iterative variable-selection procedures. At each step, it first determines whether any of the variables that are already included in the model can be removed (the initial model includes all variables). If none of the variables can be removed, it then determines whether a not-yet-included variable can be added. We used the stepwise regression procedure in the statistical tool R and selected AIC as the criterion [10].

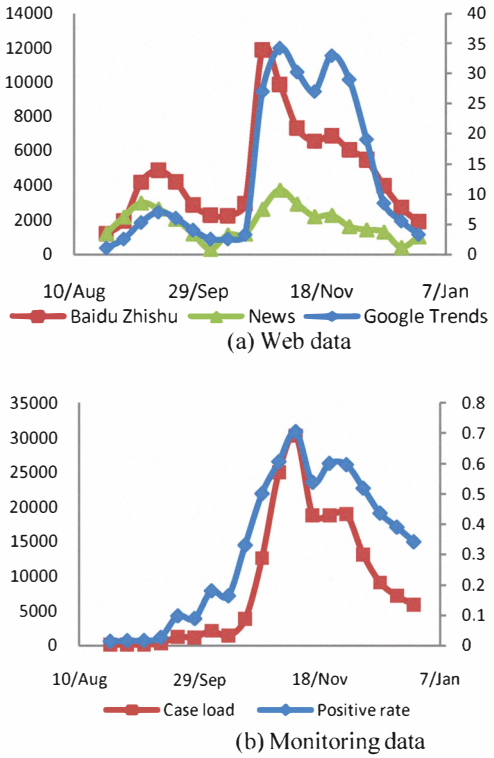
The final regression models are in the form of

$$Y_i(t) = \beta_0 + \beta_1 \times X_1(t - t_{10}) + \beta_2 \times X_2(t - t_{20}) + \beta_3 \times X_3(t - t_{30}) + \varepsilon, \quad i = 1, 2,$$

where Y_1 is the case load, Y_2 is the positive rate, X_1, X_2, and X_3 are the time series from Google Trends, Baidu Zhishu, and Google News, respectively, and $t_{n0}(n = 1, 2, 3)$ is the time lag (measured in number of weeks) for web source *n*. Note that some of the independent variables may not appear in both models.

III. RESULTS

A. *Preliminary data analysis*



(a) Web data

(b) Monitoring data

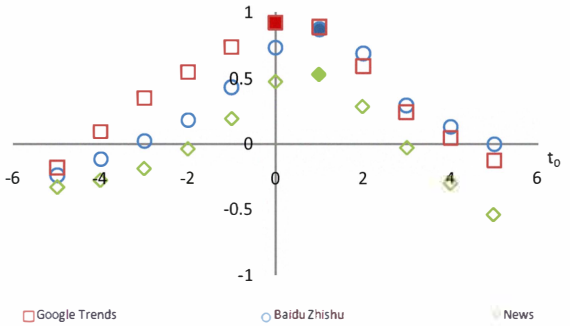
Fig.2 Contrast of trends.

There is a close relationship between the number of people searching for Influenza A (H1N1) related keyword and the number of people who have Influenza A (H1N1) symptoms. According to a local custom in Beijing, we selected the

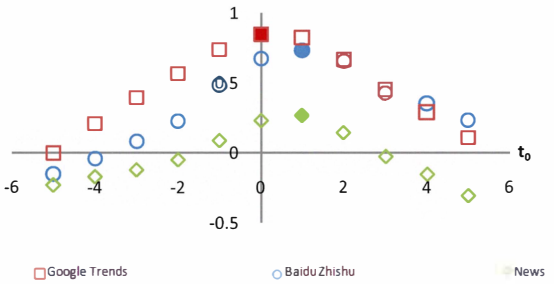
Chinese word “甲流” (pronounced jialiu) as the influenza-related keyword for search queries.

Fig.2 shows an overview of the data. Fig. 2 (a) shows that the trend of the search data form Google Trends is very similar to that from Baidu, and the news data also exhibit similar overall trend as the search data do. This implies that news is somewhat reflective of the concern of the public. Fig. 2(b) shows that case load and positive rate are somewhat correlated.

B. Time series analysis



（a）Correlation between case load and web data



（b）Correlation between positive rate and web data

Fig.3. Correlation between monitoring data and web data over time. (Each solid marker corresponds with the time lag t_0 with the maximum correlation for a time series.)

TABLE II

TIME SERIES- ANALYSIS RESULTS

Pair of time series	ρ_{max}	Selected time lag t_0
Case load, Google Trends	0.92294	0
Case load, Baidu Zhishu	0.87458	1
Case load, Google News	0.52531	1
Positive rate, Google Trends	0.84791	0
Positive rate, Baidu Zhishu	0.73541	1
Positive rate, Google News	0.26581	1

Fig.3 shows the correlation between a monitoring time series and a web data time series under varying time lags. The time lag that has the largest correlation and is selected for subsequent analysis is summarized in table II. It appears that the best time lag for Baidu Zhishu and Google News is one week and that for Google Trends is zero week. The time series based on Google Trends is most highly correlated with the monitoring data. The correlation between the time series based on Baidu Zhishu and that of the monitoring data is also relatively high. The correlation between Google News and the monitoring data is not high. These findings seem to imply that the media has relatively weak relationship with the pandemic but the behaviors of search engine users are influenced by the pandemic to some extent.

C. Stepwise regression

Based on the selected time lags, we performed stepwise regression. The results are presented in tables III and IV.

The variables selected for predicting case load are “google” and “news”, while all three variables are selected for predicting positive rate. This shows that using multiple variables is better than using a single variable and therefore our proposed multi-source approach is advantageous over one that relies on a single source. However, adding more variables is not always better and thus variable selection is necessary.

Fig. 4 contrasts the case load and positive rate predicted by the regression models to the actual values. It appears that the models fit the monitored epidemic curves very closely indeed.

TABLE III

STEPWISE REGRESSION RESULT FOR CASE LOAD

R-SQUARED = 0.82, ADJUSTED R-SQUARED = 0.78, P = 1.77E-05

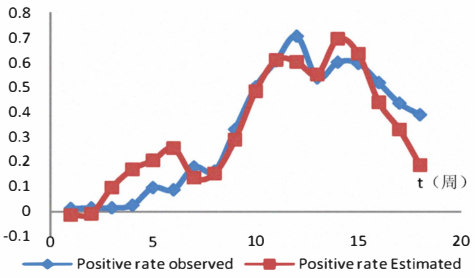
	Df	Sum of Sq	RSS	AIC
<none>	-	-	2.62E+08	303
news	1	61811931	3.24E+08	305
google	1	4.15E+08	6.62E+08	320

TABLE IV

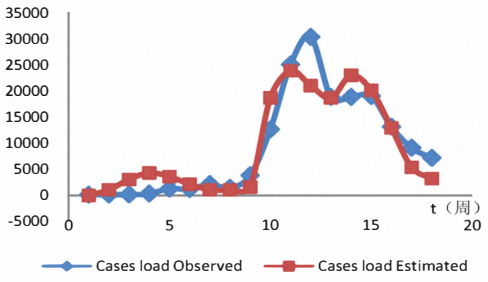
STEPWISE REGRESSION RESULT FOR POSITIVE RATE

R-SQUARED = 0.82, ADJUSTED R-SQUARED = 0.80, P = 2.67E-06

	Df	Sum of Sq	RSS	AIC
<none>	-	-	0.148	-78.447
baidu	1	0.038	0.186	-76.339
news	1	0.134	0.281	-68.854
google	1	0.18	0.327	-66.125



(a) Positive rate



(b) Case load

Fig.4. Contrast of predicted (in red) versus observed (in blue) case load and positive rate of the 2009 Influenza A (H1N1) in Beijing.

IV. Conclusion

In this study, we proposed a multi-source approach for epidemic surveillance and used the Influenza A (H1N1) pandemic in Beijing as a case study to test the utility of the approach. Our results show that the approach is advantageous over one that relies on a single data source.

The name of the disease itself, Influenza A (H1N1), is one of the most relevant search keywords, and the news media has done a lot of publicity even prior to its outbreak in China. Therefore, in our case study, we only extracted the keyword “甲流” (pronounced jialiu), which is most representative of and in accordance with the users' search behaviors, to conduct our analysis. Naturally, not everyone who searches with Influenza A (H1N1) related keywords is a suspected or confirmed patient. However, when massive data about the flu-related web queries are gathered, collective group behaviours may be revealed while the influence of irrelative web queries is diluted. A good example to illustrate the point is the trend about the keyword “地图” (Map) at Baidu (See Fig. 5). The peak always occurs on Fridays (unless there are important holidays) while the lowest point always occurs on weekends. To an individual, the query for map may be a random behaviour. However, the query performed by millions of users is reflective of collective group behaviour.

While our proposed multi-source approach for surveillance is still at an early phase of development, it is convenient, easy, and accessible. It can be used in conjunction with traditional surveillance systems to improve the timeliness and effectiveness of epidemic surveillance.

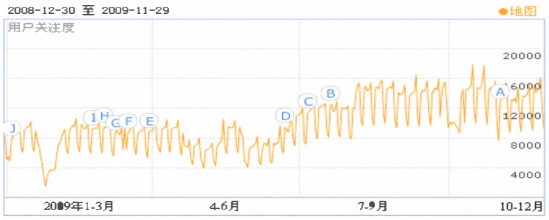


Fig. 5 Trend about the keyword “地图” (Map) at Baidu.

Acknowledgments

The authors would like to thank Nie cheng, Xiarong Li, Huiqian Li, and Fen Xia for valuable discussions and suggestions. This work was supported by the National Natural Science Foundation of China (Grants No. 60875049, 40901219, and 90924302), the Chinese Ministry of Health (Grant No. 2009ZX10004-315100), the Chinese Academy of Sciences (Grants No. 2F07C01 and 2F09N06), and the China Postdoctoral Science Fund (Grant No. 20080440559).

References

[1] WHO. *Pandemic (H1N1) 2009, Ukraine - update 1*. 2009 [cited 2009 11-3]; Available from: http://www.who.int/csr/don/2009_11_03/en/index.html.

[2] Lipsitch, M., et al., *How to maintain surveillance for novel influenza A H1N1 when there are too many cases to count.* The Lancet, 2009.

[3] Magruder, S., *Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease.* Johns Hopkins APL Tech. Digest, 2003. **24**: p. 349–353.

[4] CNNIC, *2009 Chinese search engine netizens behavior research report(Sep. 2009)*. 2009.

[5] Hulth, A., G. Rydevik, and A. Linde, *Web queries as a source for syndromic surveillance.* PLoS ONE, 2009. **4**(2).

[6] Polgreen, P., et al., *Using internet searches for influenza surveillance.* Clinical Infectious Diseases, 2008. **47**(11): p. 1443-1448.

[7] Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. **457**(7232): p. 1012-1014.

[8] Google.org. *Flu trends: How does this work?*; Available from: http://www.google.org/flutrends/intl/en_us/about/how.html.

[9] Hongzhi, A. and G. Lan, *Fast stepwise procedures of selection of variables by using AIC and BIC criteria.* Acta Mathematicae Applicatae Sinica, 1989. **5**: p. 60-67.