Monitoring Influenza Epidemics in China with Search Query from Baidu

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1371/journal.pone.0064323

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:11708653

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Monitoring Influenza Epidemics in China with Search Query from Baidu

Qingyu Yuan\(^1,2,9\), Elaine O. Nsoesie\(^2,3,4,9\), Benfu Lv\(^1\), Geng Peng\(^1\), Rumi Chunara\(^2,3\), John S. Brownstein\(^2,3,5\)

1 Management School, University of Chinese Academy of Sciences, Beijing, China, 2 Children’s Hospital Informatics Program, Division of Emergency Medicine, Boston Children’s Hospital, Boston, Massachusetts, United States of America, 3 Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, United States of America, 4 Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Virginia, United States of America, 5 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Abstract

Several approaches have been proposed for near real-time detection and prediction of the spread of influenza. These include search query data for influenza-related terms, which has been explored as a tool for augmenting traditional surveillance methods. In this paper, we present a method that uses Internet search query data from Baidu to model and monitor influenza activity in China. The objectives of the study are to present a comprehensive technique for: (i) keyword selection, (ii) keyword filtering, (iii) index composition and (iv) modeling and detection of influenza activity in China. Sequential time-series for the selected composite keyword index is significantly correlated with Chinese influenza case data. In addition, one-month-ahead prediction of influenza cases for the first eight months of 2012 has a mean absolute percent error less than 11%. To our knowledge, this is the first study on the use of search query data from Baidu in conjunction with this approach for estimation of influenza activity in China.


Editor: Benjamin J. Cowling, University of Hong Kong, Hong Kong

Received December 26, 2012; Accepted April 13, 2013; Published May 30, 2013

Copyright: © 2013 Yuan et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by research grants from the National Library of Medicine, the National Institutes of Health (5R01LM010812-03 and S508LM008776-03), the National Natural Science Foundation of China under Grant 71202115, 71172199, 71203218, Foundation of Dean of Graduate University of Chinese Academy of Sciences under Grant Y151010Y00, and Postdoctoral Science Foundation under Grant 2011M500422. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Qingyu.Yuan@childrens.harvard.edu

These authors contributed equally to this work.

Introduction

Seasonal influenza epidemics result in an estimated three to five million cases of severe illness and 250,000 to 500,000 deaths worldwide each year [1]. In order to prepare for the next severe pandemic and better control seasonal influenza epidemics, researchers have proposed several approaches to achieve near real-time surveillance of the emergence and spread of influenza. Some novel approaches for rapid disease outbreak detection and surveillance include online surveillance systems utilizing informal sources such as news reports [2], social media data [3–16], and search query data [17–20]. The idea of using search query data for detecting outbreaks was first introduced in 2006 [[17]. Ginsberg et al [18] later discussed how monitoring search queries on Google could be used to detect influenza outbreaks in the United States. Several studies followed, which pointed to the effectiveness and limitations of detecting influenza epidemics using search query data [19], [20]. Although there are limitations, such as the lack of Internet access in some regions of the world and the noise of irrelevant information, Internet search query data is being explored as a low-cost approach to estimating disease activity in near real-time.

Besides influenza surveillance, search query data has also been widely used for research in fields such as, economics and finance. In the same year as the Ginsberg’s publication [18], several studies investigated the usefulness of Google searches for forecasting unemployment in various countries [21–25]. Several papers also used search query data to predict consumption [26], [27], house pricing and sales [28], and travel and consumer confidence [27]. Though studies using web search query data have achieved good results in empirical practice, the field is still young and rapidly developing, with room for discussion and improvement.

We introduce a novel method for estimating influenza activity using search query data from Baidu. Data on Internet searches are available on a daily basis, while routine surveillance data from China’s Ministry of Health (MOH) are typically reported with a one to two-weeks lag. The objective is therefore to estimate present influenza activity based on previously observed laboratory surveillance data plus timely search query data before official reports from China’s MOH. Beyond the use of search query data in a new geographic region and the use of a different search engine, this study is an improvement on other research in this area in that, the keyword selection and composition approach presented is more economical in terms of computational resources and cost compared to the original method by Ginsberg et al [18]. Unlike the United States, in China alternative search engines such as Baidu are more widely used than Google. The market share of Google in China is less than 20%, while that for Baidu is more
than 80% [29]. The wide use of Baidu in China makes it a more representative search query source for this analysis. Several methods have been proposed for detecting and predicting trends of influenza epidemics in China [30–32]. However, most of these techniques solely use influenza-like-illness (ILI) or influenza case data. In this study, we use a combination of influenza case counts and real-time search query for modeling and detection of current influenza activity. Improving methods for surveillance, modeling, detection and prediction of influenza epidemics in China is extremely important. Two of the three pandemics of the 20th century are thought to have started in China [38], [39]. In addition, the severe acute respiratory syndrome (SARS) of 2002 had its origins in the Guangdong Province of China. Therefore, refining approaches for rapid detection of outbreaks of influenza and other respiratory illnesses in China should benefit global public health.

### Approach

Given data on influenza activity from an official source, the approach in this paper can be summarized as follows: (i) search for keywords or terms which might be related to influenza; (ii) process keywords by eliminating those unrelated to influenza epidemics, those with an interrupted time-series representing search query volume and those not correlated to the influenza epidemic curve; (iii) define weights and composite search index, and (iv) fit regression model using selected keyword index to influenza case data. Whereby, the fitted model uses both the influenza case data and the search index.

### Methods

#### Data Sources

**Official case counts.** The counts shown in Table 1 reflect monthly aggregated influenza case counts from March 2009 to August 2012 for China. The data is publicly available on China’s Ministry of Health (MOH) site [http://www.moh.gov.cn/] and typically released 1–2 week after the end of each month. A network of physicians report laboratory confirmed cases to the MOH on a daily basis. However the data is only released to the public at a monthly resolution. The data is solely laboratory confirmed influenza cases and does not include ILI cases. Furthermore, during the 2009 H1N1 pandemic, infections resulting from the new influenza strain were reported separately from cases resulting from circulating seasonal influenza strains in China [40]. The data in this study is solely for seasonal influenza. No ethics committee approval is required to obtain the data since it is publicly available. In addition, only count data is presented, no personal information is revealed, thereby maintaining confidentiality.

**Search query data from baidu.** Baidu’s database [http://index.baidu.com/] contains logs of online search query volume submitted from June 2006. However, since the influenza case count data is available from March 2009, we use Baidu’s data from March 2009 to August 2012. Unlike the case data from the Ministry of Health, Baidu’s search query data is available on a daily basis. The data is therefore converted to monthly counts for analysis. User confidentiality is also maintained, since only the combined term frequency data is available. In addition, Baidu releases search query volume for the entire country.

#### Keyword Selection and Filtering

Different keywords have different search frequency and can therefore produce diverse modeling outcomes. So keywords are carefully selected to reflect terms most likely associated with influenza epidemics. Note, observations from previous studies such as Ginsberg et al [18], have indicated that more keywords do not necessarily assure better model fit. The marginal contribution of adding terms to a “saturated” model is limited, but costly. Ginsberg et al [18] only selected 45 significant keywords from 50 million. The method of exhaustion employed by Ginsberg et al [18] is computationally expensive and not easily reproducible by researchers with limited resources [27]. In some cases, researchers have solely relied on keywords recommended by Google [23], [24], [26]. Keywords recommended by search engines tend to be comprehensive, but not always relevant to the subject. Therefore, further analysis is required to extract keywords, which are most pertinent to the study.

Keywords used in this study are obtained from the following Chinese website: [http://tool.chinaz.com/baidu/words.aspx](http://tool.chinaz.com/baidu/words.aspx) (hereafter referred to as keyword tool). Keywords suggested by the keyword tool include recommendations from Baidu, and others mined using semantic correlation analysis from portal websites, blogs, and online reports. “Flu” (疾病” in Chinese) is the core keyword in this study. Upon entering “流感” into the keyword tool, we obtain 94 related keywords (Table 2). Although recommended by the keyword tool, some of the 94 keywords are not related to influenza epidemics in China. We therefore filter the keywords as follows: (i) the selected keywords should represent factors that might influence the influenza epidemic. (ii) The search query data for each keyword should be represented as a sequential

---

**Table 1. Influenza case data from China’s MOH.**

<table>
<thead>
<tr>
<th>Month</th>
<th>ICD*</th>
<th>Month</th>
<th>ICD</th>
<th>Month</th>
<th>ICD</th>
<th>Month</th>
<th>ICD</th>
<th>Month</th>
<th>ICD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009-03</td>
<td>8015</td>
<td>2009-12</td>
<td>29777</td>
<td>2010-09</td>
<td>5114</td>
<td>2011-06</td>
<td>3065</td>
<td>2012-03</td>
<td>21625</td>
</tr>
<tr>
<td>2009-04</td>
<td>6794</td>
<td>2010-01</td>
<td>10415</td>
<td>2010-10</td>
<td>4121</td>
<td>2011-07</td>
<td>2654</td>
<td>2012-04</td>
<td>10707</td>
</tr>
<tr>
<td>2009-05</td>
<td>7769</td>
<td>2010-02</td>
<td>6595</td>
<td>2010-11</td>
<td>5323</td>
<td>2011-08</td>
<td>3243</td>
<td>2012-05</td>
<td>8520</td>
</tr>
<tr>
<td>2009-06</td>
<td>7999</td>
<td>2010-03</td>
<td>8488</td>
<td>2010-12</td>
<td>6529</td>
<td>2011-09</td>
<td>4360</td>
<td>2012-06</td>
<td>6195</td>
</tr>
<tr>
<td>2009-07</td>
<td>7791</td>
<td>2010-04</td>
<td>6357</td>
<td>2011-01</td>
<td>6072</td>
<td>2011-10</td>
<td>5525</td>
<td>2012-07</td>
<td>6738</td>
</tr>
<tr>
<td>2009-08</td>
<td>14548</td>
<td>2010-05</td>
<td>3865</td>
<td>2011-02</td>
<td>5930</td>
<td>2011-11</td>
<td>7055</td>
<td>2012-08</td>
<td>6793</td>
</tr>
<tr>
<td>2009-09</td>
<td>43596</td>
<td>2010-06</td>
<td>2642</td>
<td>2011-03</td>
<td>7299</td>
<td>2011-12</td>
<td>11631</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009-10</td>
<td>25132</td>
<td>2010-07</td>
<td>2627</td>
<td>2011-04</td>
<td>5727</td>
<td>2012-01</td>
<td>10046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009-11</td>
<td>43018</td>
<td>2010-08</td>
<td>3588</td>
<td>2011-05</td>
<td>4130</td>
<td>2012-02</td>
<td>17421</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*ICD is the abbreviation for influenza case data.*

doi:10.1371/journal.pone.0064323.t001
<table>
<thead>
<tr>
<th>Table 2. Complete keyword list.</th>
</tr>
</thead>
<tbody>
<tr>
<td>流感(flu)</td>
</tr>
<tr>
<td>流感疫苗(influenza vaccine)</td>
</tr>
<tr>
<td>甲型h1n1流感(type a h1n1 flu)</td>
</tr>
<tr>
<td>山东流感(Shandong flu)</td>
</tr>
<tr>
<td>流感的预防措施 (prevention measures of influenza)</td>
</tr>
<tr>
<td>流感的预防 (prevention of flu)</td>
</tr>
<tr>
<td>流感症状 (influenza symptom)</td>
</tr>
<tr>
<td>流感病毒 (influenza virus)</td>
</tr>
<tr>
<td>流感大额下载 (qq influenza game)</td>
</tr>
<tr>
<td>流感的症状 (the influenza symptom)</td>
</tr>
<tr>
<td>流感防治知识 (the knowledge of influenza prevention)</td>
</tr>
<tr>
<td>流感疫苗 (love flu virus)</td>
</tr>
<tr>
<td>山东新流感 (Shandong swine flu)</td>
</tr>
</tbody>
</table>

Note: Web users use Chinese characters to search in Baidu. Keywords in English are listed to show the corresponding translation of each Chinese character. The keywords in bold are excluded at filtering step (i). The keywords in italics are excluded at filtering step (ii) and keywords in bold and italics are excluded at filtering step (iii).

doi:10.1371/journal.pone.0064323.t002
time series with a daily, weekly or monthly resolution. (iii) Lastly, the time series of selected keywords should have a maximum cross-correlation coefficient of at least 0.4 with the influenza case data.

Keywords that remain after the filtering analysis are considered for inclusion in the composite search index. The goal of search index composition is to build the most correlative and stable indicator for the influenza case data based on the available information. The search index is composed in two steps. First, we define synthetic weights for each of the keywords. Next, we combine the weighted time series for the keywords.

Search Index Composition

We consider two approaches for defining synthetic weights: the method of systematic assessment and the strength of the correlation coefficient. The method of systematic assessment [34], [35] involves rating the selected indicator according to the principle of prior evaluation and defining the ratings as weights. The method is comprehensive but highly subjective. Alternatively, the correlation coefficient between the influenza epidemic curve and the keyword frequency curve can be used to represent the weight [18], [33]. This approach is usually combined with Analytic Hierarchy Process (AHP) [36] for better performance. However, solely using the correlation coefficient without adjustments appears to be sufficient for this study.

The search index is defined as: \[ \text{index}_j = \sum_{i=1}^{I} \omega_i x_i^j, \] where \( \omega_i \) is the weight of the \( i^{th} \) keyword and \( x_i^j \) represents the sequence after alignment. Although the definition of the composite index allows for alignment, it is not required for combining the time series in this study since maximum correlations are observed at lag 0. The final set of keywords is selected using the following model:

\[ y = \beta_0 + \beta_1 \text{index}_j + \epsilon \tag{1} \]

In (1), \( \text{index}_j \) represents the search index for \( j \) keywords, \( y \) denotes influenza case counts, \( \beta_0, \beta_1, \epsilon \) denote the intercept, coefficient and error term respectively.

Using a stepwise approach generally used in the selection of variables in a multiple regression framework, keywords are

Table 3. Keywords in composite index.

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>English</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>流感预防</td>
<td>(prevent influenza)</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>流感的症状</td>
<td>(the influenza symptom)</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>甲型流感疫苗</td>
<td>(type a influenza vaccine)</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>流感症状</td>
<td>(flu symptom)</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0064323.t003
selected based on their contribution to the model’s goodness of fit. Partial F test is used to evaluate the goodness of fit after adding data for each keyword to the index. A significant F-statistics implies that the keyword should be added to the composite index, and vice versa. The search index is defined based on the model with the best goodness of fit statistics.

The initial model is based on the keyword with the highest correlation with the influenza case data. In this case, “流感预警” (prevent influenza) has the highest correlation at 0.93 at lag 0. Keywords are then added sequentially based on the correlation coefficient and the partial F test is examined for improved fit. The process is repeated until the goodness of fit can no longer be improved.

Model

As stated, the objective of this paper is to present a method for faster detection of influenza activity in China using search query data. China’s MOH typically releases monthly influenza case data 1–2 weeks into the next month. We therefore aim to provide estimates of case data before the MOH data is publicly available.

The most significant correlations between the composite index and the case data are observed at lag 0 (P = 0.959) and lag 1 (P = 0.658). Correlations at lags 2 and 3 are 0.491 and 0.227 respectively. We therefore fit the following model:

\[
\text{ICD}[t] = \beta_0 + \text{ICD}[t-1] + \beta_1 \times \text{index}[t] + \beta_2 \times \text{index}[t-1] + \varepsilon
\]

(2)
Table 5. Predicted values and error.

<table>
<thead>
<tr>
<th>Month</th>
<th>Actual value</th>
<th>Predicted value</th>
<th>Absolute error</th>
<th>Percent absolute error</th>
</tr>
</thead>
<tbody>
<tr>
<td>01–2012</td>
<td>10046</td>
<td>10230</td>
<td>184</td>
<td>1.8%</td>
</tr>
<tr>
<td>02–2012</td>
<td>17421</td>
<td>14578</td>
<td>2843</td>
<td>16.3%</td>
</tr>
<tr>
<td>03–2012</td>
<td>21625</td>
<td>18429</td>
<td>3196</td>
<td>14.8%</td>
</tr>
<tr>
<td>04–2012</td>
<td>10707</td>
<td>11785</td>
<td>1078</td>
<td>10.1%</td>
</tr>
<tr>
<td>05–2012</td>
<td>8520</td>
<td>8618</td>
<td>98</td>
<td>1.2%</td>
</tr>
<tr>
<td>06–2012</td>
<td>6195</td>
<td>6621</td>
<td>426</td>
<td>6.9%</td>
</tr>
<tr>
<td>07–2012</td>
<td>6738</td>
<td>5240</td>
<td>1498</td>
<td>22.2%</td>
</tr>
<tr>
<td>08–2012</td>
<td>6793</td>
<td>5983</td>
<td>810</td>
<td>11.9%</td>
</tr>
</tbody>
</table>

ICD represents influenza case data, \( \beta_0, \beta_1, \beta_2 \) are the coefficients, index is the composite search index and e is the error term. The model estimates ICD at time t based on ICD at time t-1 and the composite search index at time t and t-1. For example, case counts for February 2012 are estimated at the end of February based on the composite search index for February and January, and the case count for January. We also examine the residuals to evaluate the adequacy of the model.

The influenza case data is divided into a fitting and validation set. Data from March 2009 to December 2011 is used for model fitting, while data from January 2012 to August 2012 is used for validation. We also consider models with second and third order lags. Models are evaluated based on R-squared, AIC and significance of the coefficients. Studies have suggested that solely using an extrapolation of the influenza activity curve for predictions usually results in a higher error rate [32], [33]. The analysis is performed using the Eviews software.

Results

Based on the filtering analysis, 14 out of the 94 keywords are not related to influenza epidemics, 20 keywords do not have sequential time series due to low search volume and only 40 keywords are significantly correlated to the case data (see Table 2). With the stepwise approach, only 8 of the 40 keywords are used in the composite search index (see Table 3). The estimated cross-correlation coefficient between the search index and influenza case data is 0.96 at lag 0 (Figure 1). Influenza epidemics are observed in the spring and winter as expected. Note that the search index clearly captures the peaks and troughs of the influenza time series curve, thereby making it a good indicator for influenza activity in China.

The coefficients \( \beta_0, \beta_1, \beta_2 \) for model (2) are 0.56 (\( P = 0.001 \)), 0.25 (\( P < 0.001 \)) and -0.14 (\( P = 0.004 \)) respectively. Note the model’s R-squared is 0.95 and the AIC is 18.50. In addition, the Durbin-Watson test statistic is 1.89 suggesting that autocorrelation is not an issue (see Table 4). The null hypothesis of the Durbin-Watson test is that the autocorrelation parameter is zero.

The model is validated by predicting influenza cases one month at a time, from January 2012 to August 2012. The results are listed in Figure 2 and Table 5. The mean absolute percent error of prediction for the consecutive eight months is 10.6% (see Table 5). We also consider models with second order lags and third order lags but neither of their statistical results are better than that of model [2] (see Tables S1 and S2).

Discussion

We develop a comprehensive method for pre-processing Internet search data for modeling and detecting influenza epidemics in China. The combined keyword index is significantly correlated to the case data and mean absolute percent error of predicting 2012 monthly influenza cases is less than 11% based on one-step predictions for eight months. Although the monthly search query data and influenza case data are almost synchronous, the search query data can still be used in detecting influenza cases because of the time delay of official reports.

This study contributes to the pool of novel sources of data, such as web-based data, used as early indicators for disease outbreaks. To our knowledge, this is the first study utilizing Baidu search query data in conjunction with this approach for estimating influenza activity in China. Baidu has a significantly higher market share than Google in China, thereby making it a better search query source for this study. The proposed approach is not meant to replace actual estimates of influenza cases, rather it is an indicator of influenza activity, which is freely available in near real-time. This is especially relevant for a country such as China, which has been coined the “epicenter of influenza” [39] by some.

However, there are several limitations to using search query data. Although the selected keywords perform well at capturing the temporal trend of the epidemic curve, there is no guarantee that this would be consistent in future dates. Individual behavior is constantly changing and different factors influence keywords queried by individuals. Another limitation is the unavailability of Internet access in rural regions. The China Internet Network Information Center (CNNIC) currently estimates Internet penetration in China at 39.9%. Surveillance using web-query data depends on adequate Internet access. In addition, not all searches on influenza-related terms are necessarily linked to influenza morbidity. Search queries can be a result of panic during a novel respiratory outbreak, coverage of influenza-related deaths in the media, fear or curiosity. Using several years of data in modeling should hopefully mitigate occurrences of panic induced searches since the weight of various keywords is likely to deviate from one influenza season to another. Furthermore, correlation does not imply causation, which suggests that predictions made using such novel data sources should be carefully evaluated.

Limitations also exist in the data used in this study. Influenza-like illness data might be a better indicator of influenza activity since influenza cases are not always confirmed and case data might underestimate the true burden of the disease. However, China’s Ministry of Health only releases influenza case data for the entire country. In addition, there are likely to be major differences in
timing and duration of epidemics from province to province. Analysis at the province level would therefore be more beneficial. Unfortunately, both the case data and search query volume are only available for the entire country. Though, the model can be easily extended to detect influenza activity at a province level.

Although limitations exist, having more methods and resources geared towards infectious disease surveillance provides a step towards rapid detection and control of emerging and re-emerging outbreaks. Public health scientists and epidemiologists could use observations from such approaches as an indicator for further investigations. These tools are freely available in near real-time and can be especially valuable in regions where official reports of case counts are delayed.

Supporting Information

Table S1 Results for model with second order lag included.

(OCX)

Table S2 Results for model with second and third order lags included.

(OCX)

Author Contributions

Conceived and designed the experiments: JSB QY. Analyzed the data: QY. Contributed reagents/materials/analysis tools: GP BL. Wrote the paper: QY EON. Developed and evaluated the model: QY EON. Edited and revised the manuscript: JSB QY EON RC.

References


