Journal of Data Analysis and Information Processing, 2018, 6, 79-92
http://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

# The Study on China's Flu Prediction Model Based on Web Search Data

**Yan Bu[1], Jinhong Bai[1], Zhuo Chen[1,2], Mingjing Guo[1*], Fan Yang[1]**

[1]School of Economy and Management, China University of Geosciences, Wuhan, China
[2]School of Business, Central South University, Changsha, China
Email: *guomingjing@cug.edu.cn

**How to cite this paper:** Bu, Y., Bai, J.H., Chen, Z., Guo, M.J. and Yang, F. (2018) The Study on China's Flu Prediction Model Based on Web Search Data. *Journal of Data Analysis and Information Processing*, **6**, 79-92.
https://doi.org/10.4236/jdaip.2018.63006

**Received:** June 12, 2018
**Accepted:** July 1, 2018
**Published:** July 4, 2018



## Abstract

Influenza is a kind of infectious disease, which spreads quickly and widely. The outbreak of influenza has brought huge losses to society. In this paper, four major categories of flu keywords, "prevention phase", "symptom phase", "treatment phase", and "commonly-used phrase" were set. Python web crawler was used to obtain relevant influenza data from the National Influenza Center's influenza surveillance weekly report and Baidu Index. The establishment of support vector regression (SVR), least absolute shrinkage and selection operator (LASSO), convolutional neural networks (CNN) prediction models through machine learning, took into account the seasonal characteristics of the influenza, also established the time series model (ARMA). The results show that, it is feasible to predict influenza based on web search data. Machine learning shows a certain forecast effect in the prediction of influenza based on web search data. In the future, it will have certain reference value in influenza prediction. The ARMA(3,0) model predicts better results and has greater generalization. Finally, the lack of research in this paper and future research directions are given.

## Keywords

Data Mining, Web Search, Machine Learning, Baidu Index, Influenza Prediction

## 1. Introduction

Influenza, referred to as the flu, is an acute respiratory infectious disease caused by influenza virus that cannot be completely controlled until now [1]. According to the WHO (World Health Organization) study of seasonal influenza, seasonal influenza causes about 3 to 5 million serious diseases each year, resulting in approximately 250,000 to 500,000 deaths [2]. From the Spanish flu (H1N1) in 1918,

the Asian flu (H2N2) in 1957, the Hong Kong flu (H3N2) in 1968, and the Russian flu (H1N1) in 1977 to April 2009, the outbreak of H1N1 has caused a huge loss of human society for every outbreak of flu [3] [4] [5]. For all countries in the world, the prevention and control of influenza has always been a serious problem.

First of all, in order to control the spread of influenza virus and reduce the losses caused by influenza, it is necessary to use reasonable methods to predict the trend of influenza activity. However, the influenza virus has the characteristics of strong infectiousness, rapid propagation, wide spread, and antigen variability [6], which brings great difficulties to prevention and monitoring. As a result, researchers in various countries are focusing more on improving the timeliness of forecasting the flu epidemic. Second, the use of more timely and accurate data sources is the main means of improving timeliness. In order to obtain influenza case data, most national influenza surveillance agencies generally conduct surveys on suspected influenza cases in hospitals. However, this method requires the collection of national influenza case data. There are complex data processing processes, heavy workload, and monitoring data lag about influenza development and other issues. Finally, in order to obtain more data on the flu cases, flu monitoring agencies used data such as telephone consultations on influenza, sales of flu-type non-prescription drugs, and page views of relevant websites to predict the incidence of influenza [7]. To a certain extent, it improves the accuracy and timeliness of short-term forecasting.

Nowadays, search engines are increasingly becoming the main method for people to obtain information. Web search data has become an ideal data source for influenza surveillance. In the United States, about 90 million adults annually search the Internet for health information such as disease and medicine [8]. Compared with other data sources, web search data has a stronger tendency and immediacy, and search keywords can directly reflects the intent of the inquirer, and the search data can be collected in a timely manner to maintain complete synchronization with the development of the flu epidemic. In addition, the search data has a wider range of survey populations. It can show the attention of all Internet users in a certain area to the flu, and the data is closer to the true whole. Using web search data to monitor epidemic disease is a faster, more accurate and low-cost way. It can be used as an auxiliary measure of traditional investigation methods to provide early warning of disease and is important for the prevention and control of infectious diseases in China and beyond.

## 2. Literature Review

Influenza has caused great difficulties in prevention and monitoring due to its rapid mutation rate. Therefore, the most important task in influenza epidemic surveillance research is to improve the timeliness of predictions. The use of more immediate and accurate data sources is the main reason for improving timeliness [7]. In the era of big data, web search data has become an ideal data source

for influenza surveillance. The flu monitoring application based on web search data mainly includes the following aspects.

### 2.1. Using Search Engines for Influenza Surveillance

In 2008, Polgreen *et al.* [9] used the Web search data for the first time. They used the search volume of influenza-related search terms on the Yahoo! search engine in the United States to verify the correlation between search volume and influenza mortality. Jeremy G *et al.* [10] published a flu trend monitoring research based on Google search data in Nature, which laid the theoretical foundation for the Google Flu Trends (GFT) launched by Google later [11]. GFT is an online flu trend online warning system based on its own search data released by Google. It provides flu trend predictions in 28 countries around the world. After the GFT was released, it was applied to influenza surveillance activities in different countries. In addition to Google's search engine, search data can also be obtained through other methods, such as China's Baidu Index, Weibo Micro Index, etc. Q. Yuan *et al.* [12] studied the relationship between search terms and flu trends through Baidu Index and fitted a multiple regression monitoring model. Lu Li *et al.* [13] compared and analyzed the role of Baidu Index and Sina Weibo micro index in the monitoring of influenza in China and found that the Baidu Index was more relevant to the flu epidemic. Search engine-based influenza surveillance estimates the incidence of influenza due to the search frequency of keywords alone. This can easily lead to over-sensitivity of the model, causing "overestimation" of the epidemic, as well as seasonal and geographical impacts. After it is still insufficient, it needs to improve.

### 2.2. Using Social Networks for Influenza Surveillance

The prediction of events through social networks is a hot topic of big data research. In foreign countries, there are many researchers who use the social platform Twitter to do data analysis, including flu trend monitoring. Nigel Collier *et al.* [14] used SVM algorithm to analyze the epidemic situation by studying user behavior information in the information posted by users on Twitter, and compared the results with that of the CDC (United States Centers for Disease Control and Prevention), and found that it had a very strong relationship with that. Lampos, V. *et al.* [15] observed and tracked the Twitter information published by users in the UK's most popular 49 regions. Using the flu keyword weighted filtering method, it was found that the flu episode showed strong linear correlation with the HPA's influenza-like illness (ILI) data. Similar examples of flu predictions based on social platforms are numerous. Chen *et al.* [16] used Facebook, micro-blog, and Instagram as research data to filter textual data for flu symptoms keywords, to obtain suspected influenza users, and to associate GPS information on Instagram to geographically monitor the flu. Also as a social media in recent years, Weibo has been popular among Chinese citizens. At present, there are many researchers who are doing data mining based on Weibo, such as: anal-

ysis of social relations based on Weibo, public opinion analysis based on Weibo, and outbreak analysis based on Weibo [17]. However, the data did not make significant progress in the study of seasonal influenza surveillance based on Weibo.

### 2.3. Using Existing Disease Surveillance Platforms for Influenza Surveillance

At present, the most representative foreign influenza surveillance platform is Flu Near You. Flu Near You is a flu monitoring and visualization system that can be intuitively displayed on maps. It is also participatory for the general public. Users can submit the relevant information about flu symptoms every week. These data for researchers better understand the spread of the flu, while ordinary citizens can also watch the surrounding communities where they live and the spread of national flu [18]. In China, Baidu and the Chinese Center for Disease Control and Prevention launched its disease prediction platform. The Baidu Disease Forecasting Platform provides an online map tool to show people how active certain diseases are in each region, and to make predictions about disease changes in the past 30 days and the next seven days.

Nowadays, there is no standardized flu prediction model in China and there are not many researches on the use of Web search data to study flu prediction models. This study establishes some prediction models by using Python to crawl relevant flu data together with machine learning. Considering the seasonality of influenza, a time series model has also been established, which has certain reference value for the monitoring and prevention of influenza.

## 3. Sources of Data

### 3.1. Influenza-Like Illness

A major indicator of influenza surveillance at home and abroad is the proportion of influenza-like illness (ILI). It refers to fever (body temperature ≥ 38˚C) and cough in all outpatient clinics at sentinel hospitals. Sore throat is one of the cases of acute respiratory infection [19]. The flu epidemic data used in this paper comes from the weekly influenza surveillance report (http://www.cnic.org.cn/) published by the China National Influenza Center website. The sample period is from the 16th week of 2016 (2016/16, starting on April 25, 2016) to the 16$^{th}$ week of 2018 (2018/16, April 23, 2018). The data collected in this paper are mainly the proportion of influenza-like cases in the country (The proportion of flu-like cases is the total number of ILI patients divided by the number of outpatients, expressed as ILI%).

### 3.2. Web Data

In order to obtain more time-sensitive data, this paper uses Python to write a crawler program and uses Baidu's webpage data as crawling objects. Selected more primitive search terms such as "flu vaccine", "cold", "flu treatment",

"flu medicine" and "H7N9", and summarized the keywords selected by other related studies and keyword recommendations of search engines. Expand the number of words for each keyword type, sum up all valid keywords, and form an initial vocabulary. As shown in Table 1, the data were crawled according to the initial vocabulary.

## 4. Model Introduction

### 4.1. Feature Selection

At the beginning of the model establishment of data mining and machine learning algorithms, in order to minimize the problem of model deviation due to the lack of important variables, we usually choose as many independent variables as possible. However, during the actual modeling process, it is usually necessary to find the subset of independent variables that have the ability to interpret the response variables to improve the model's ability to interpret and predict. This process is called feature selection.

Principal component analysis (PCA) is a method of dimension reduction for unsupervised learning. It requires only eigenvalue decomposition to compress and denoise the data. Therefore, this paper uses PCA algorithm to extract features of influenza keywords.

Algorithm flow: Input: n-dimensional sample set $D=\left(x^{(1)},x^{(2)},\cdots,x^{(m)}\right)$, to reduce dimension to $n'$ dimension. Output: Sample set $D'$ after dimension reduction.

1) Centralize all samples: $x^{(i)}=x^{(i)}-\frac{1}{m}\sum_{j=1}^{m}x^{(j)}$,

Table 1. Key words and extended primaries.

| Classification | Key words | Extended words |
|---|---|---|
| Prevention stage | Flu vaccine | Flu vaccine side effects, Influenza vaccine necessary to fight it, Bird flu vaccine |
| | Flu prevention | How to prevent flu, Prevent influenza, Influenza outbreak, Prevent influenza A |
| Stage of symptoms | Cold | Gastro-intestinal flu, Influenza, Cold symptoms, Viral influenza |
| | Respiratory infection | Upper respiratory tract infection, Nasal congestion, Cough, Bronchitis, Sore throat, Runny nose, Rhinitis, Pharyngitis |
| | Fever | Fever, High fever, Headache, Dizziness, Fatigue, Fever, Chills |
| Treatment stage | Flu treatment | A stream of treatment, What medicine to eat flu |
| | Cold medicine | Baijiahei, Contac, Tylenol, Gankang, Amoxicillin, Cough, Antipyretics, Cephalosporins, Oseltamivir |
| Commonly used words | H1N1 | H1N1 flu, H7N9, H7N9 flu |
| | Influenza-A | Influenza A, Type A H1N1 flu, Type A flu symptoms, What is Influenza A |
| | Influenza | Flu virus, Swine flu, Bird flu |

2) Calculate the sample's covariance matrix $XX^{\mathrm{T}}$,

3) Perform eigenvalue decomposition on the matrix $XX^{\mathrm{T}}$, and take out the eigenvector corresponding to the largest $n'$ eigenvalue $\left(w_1, w_2, \cdots, w_{n'}\right)$, After all the eigenvectors are normalized, they form a matrix of eigenvector $W$,

4) Transform each sample $x^{(i)}$ in the sample set into a new sample $z^{(i)} = W^{\mathrm{T}} x^{(i)}$,

5) Get output sample set $D' = \left(z^{(1)}, z^{(2)}, \cdots, z^{(m)}\right)$.

## 4.2. Model Introduction

### 4.2.1. Support Vector Regression

Provided that the training sample is $\left(x_i, y_i\right)$ and $\left(i = 1, 2, \cdots, l\right)$, the simplest support vector regression (SVR) uses a linear function $f\left(x, \omega\right) = \left(\omega \cdot x\right) + b$ to model the sample points Together, where $w$ and $b$ are the normal vector and the offset of the linear regression function respectively. Assume that all training data are fitted with a linear function without errors under $\varepsilon$. Solve the following optimization problem:

$$\min_{\alpha} \Phi\left(w\right) = \frac{1}{2}\left\|w\right\|^2 \tag{1}$$

$$\text{s.t.} \begin{cases} \left(w \cdot x_i + b\right) - y_i \le \varepsilon, i = 1, 2, \cdots, l \\ y_i - \left(w \cdot x_i + b\right) \le \varepsilon, i = 1, 2, \cdots, l \end{cases} \tag{2}$$

When we cannot fully satisfy the above two-condition constraint, we introduce the slack variables $\xi_i$, $\xi_i^*$ and the penalty parameter $C$ to "soften" the same as the linear inseparable support vector classification. The original optimization problem becomes:

$$\min_{a, \xi_i, \xi_i^*, b} \frac{1}{2}\left\|w\right\|^2 + C * \frac{1}{l}\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \tag{3}$$

$$\text{s.t.} \begin{cases} \left(w \cdot x_i + b\right) - y_i \le \varepsilon + \xi_i, i = 1, 2, \cdots, l \\ y_i - \left(w \cdot x_i + b\right) \le \varepsilon + \xi_i^*, i = 1, 2, \cdots, l \\ \xi_i \ge 0, \xi_i^* \ge 0, i = 1, 2, \cdots, l \end{cases} \tag{4}$$

To solve the problem, you can get the normal vector and the regression function of the regression function:

$$w = \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right)x_i \tag{5}$$

$$f\left(x\right) = \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right)\left(x_i \cdot x\right) + b \tag{6}$$

Here, $\left(x_i \cdot x\right)$ is the inner product of the vector $x_i$ and the vector $x$.

### 4.2.2. Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO), also known as linear regression L1 regularity, is a kind of compression estimation. It obtains a refined model by constructing a penalty function, making it compress some coefficients and setting some coefficients to zero. Therefore, the advantage of subset shrinkage is preserved, which is a kind of biased estimation of multiple colinearity

data. The objective function is:

$$J(w) = \min_{m} \left\{ \frac{1}{2N} \left\| X^{\mathrm{T}} w - y \right\|_2^2 + \alpha \|w\|_1 \right\} \tag{7}$$

Among them, $y$ is the proportion of influenza-like cases, $X$ is the independent variable that affects influenza cases, $N$ is the number of data groups, $\alpha$ = 0.001, and $w$ is the regression coefficient of the influenza model.

#### 4.2.3. Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a deep neural network model containing convolutional layers. It has become a hot topic in the field of speech analysis and image recognition. Since CNN's feature detection layer learns through training data, when CNN is used, explicit feature extraction is avoided, and learning is implicitly performed from training data. Furthermore, because the neuron weights on the same feature map are the same, the network can learn in parallel. Therefore, this paper selected CNN to establish influenza prediction model.

Several important levels of convolutional neural networks:

1) Convolution layer: Each neuron is seen as a filter, which calculates the local data. Take a data window, this data window slides continuously until all samples are covered.

2) Pooled layer: The pooled layer is sandwiched between successive convolution layers to compress the amount of data and parameters and reduce overfitting.

3) Excitation layer: The excitation layer has an excitation function that performs non-linear mapping of the convolutional output.

4) Fully connected layer: In the fully connected layer, all neurons between the two layers have the right to reconnect. Usually the fully connected layer is at the tail of the convolutional neural network because the amount of information at the tail does not begin to be as large.

In this paper, CNN is divided into six layers: input layer, first convolution layer, pooled layer, second convolution layer, fully connected layer, and output layer. Here, the convolutional layer excitation layer adds the excitation function ReLU to each convolution process. In addition, the droupout layer was also added to the fully connected layer, and the inactivation ratio was 0.3, which means that 70% of the neurons were retained and the overfitting phenomenon was reduced.

Enter a size of 1*16 for each training matrix. Before the first convolutional layer, change the matrix size to 4*4 and use a convolution kernel of 2*2*32. The horizontal step is 1 and the vertical step is 1, the result is 4*4*64. Enter the pooling layer to get a 2*2*32 matrix. The function used by the pooling layer is MaxPool. Then enter the next layer of convolution layer, enter 2*2*32, use the convolution kernel as 2*2*64, get 2*2*64, horizontal step is 1, vertical step is 1. Finally enter the fully connected layer, the learning efficiency is 0.01, finding

the best value of the mean-square error (MSE) function by using the stochastic gradient method, the results obtained before reduce the dimension, stretched into a 512*1 matrix, and set the deactivation rate. The output to the output layer completes a training. The CNN training was completed after 500 training steps.

#### 4.2.4. Time Series Model

Taking into account the seasonal characteristics of influenza, this article considers the establishment of a time series model. The time series modeling refers to the model established by using only its past values and random disturbance terms. Its general form is:

$$Y_t = F\left(Y_{t-1}, Y_{t-2}, \cdots, u_t\right) \tag{8}$$

At present, there are two types of time-series models. One is the ARMA (Auto Regression Moving Average) model, which is an autoregressive moving average model; the other is the ARIMA (Auto Regression Integrated Moving Average) model, which is an autoregression integral moving average model. The ARMA model is suitable for stationary time series data, and the ARIMA model is suitable for non-stationary time series data.

## 5. Results Analysis

A total of 47 indicators were crawled in this study from the 16th week of 2016 (started on April 25, 2016) to the 16th week of 2018 (April 23, 2018). Firstly, after PCA dimensionality reduction, there are 16 main components remaining, and the 16 main components after dimensionality reduction are included in SVR, LASSO and CNN for modeling respectively.

In this paper, a total of 105 sets of data were randomly selected from the 105 sets of data to perform tests on 10 groups. SVR, LASSO and CNN were all using the same 10 groups for testing, and the remaining 95 groups were trained.

The fitting results of the SVR, LASSO and CNN models are shown in **Figure 1**. The training results (TR) of the three models fitted with the trends of the flu.

In the SVR model, the ploynomial kernel was used for the kernel function, C = 9.1896, gamma = 0.0474, training RMSE (Root Mean Square Error) = 0.1027, and test RMSE = 6.4906.

The LASSO model uses the penalty function L1, $\alpha = 0.001$, the training RMSE = 3.9954, and the test RMSE = 2.2268.

The learning efficiency of the CNN model is 0.01. In order to prevent over-fitting, the penalty function increases the Dropout layer. Some neurons are randomly deactivated at a ratio of 0.7. The training RMSE = 1.8670 and the test RMSE = 9.6885.

Due to the seasonal features of influenza, the time series model was considered in this paper. Since the time series model requires consistency and completeness of time series data, the first 95 groups were used as training data and the last 10 groups were taken as Test Data. The unit root test results show ADF =

−3.6991, p = 0.0041, indicating that the time series is a stationary time series and can be modeled with time series. The AIC rule of ARMA model is used to determine the order, and the minimum AIC value p = 3 and q = 0 are calculated. The ARMA(3,0) model is selected. The result of ARMA(3,0) fitting is shown in Figure 2. The training RMSE = 1.7123 and the test RMSE = 1.4333.

From the training and predictive results of the SVR, LASSO, CNN and ARMA models, it is feasible to predict the proportion of influenza-like illnesses through the Web search data. Each model shows a certain predictive result, as shown in Figure 3 and Figure 4. Figure 5 shows the accumulation absolute error of the SVR, LASSO and CNN models (SVR-AE, LASSO-AE, CNN-AE). The LASSO model has the smallest absolute error. At the same time point (2016/52, 2017/10, 2017/30, 2018/8) almost all of the three models exhibited relatively large absolute errors. Explain that the three models have poor predictability for certain periods

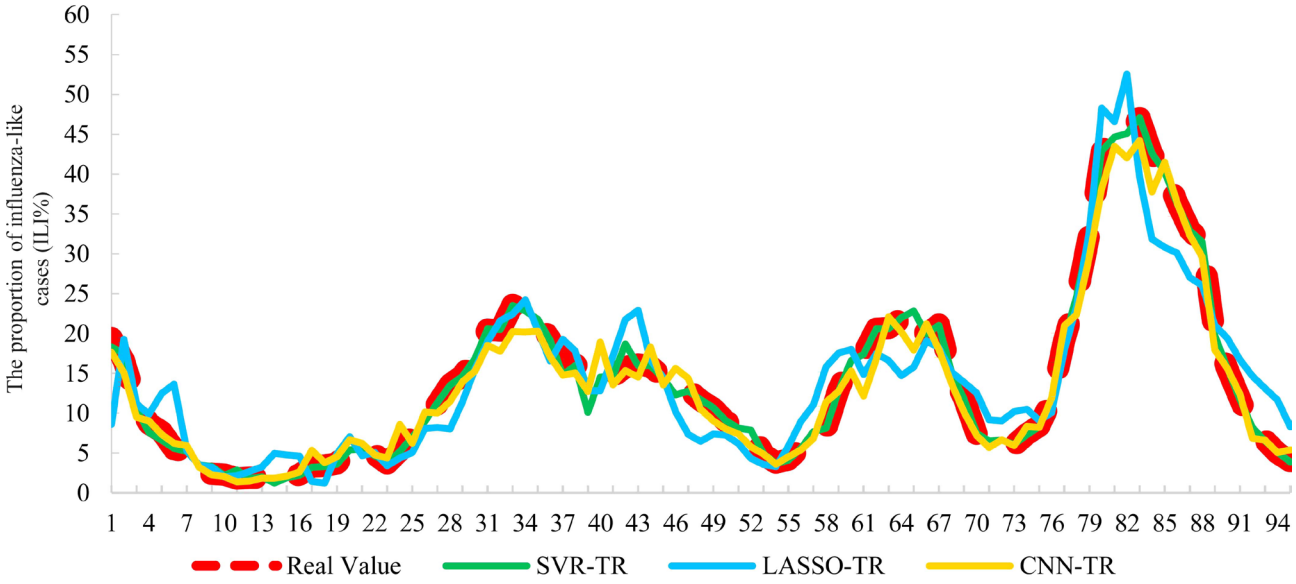


**Figure 1.** Comparison of the fitting results of SVR, LASSO and CNN models.

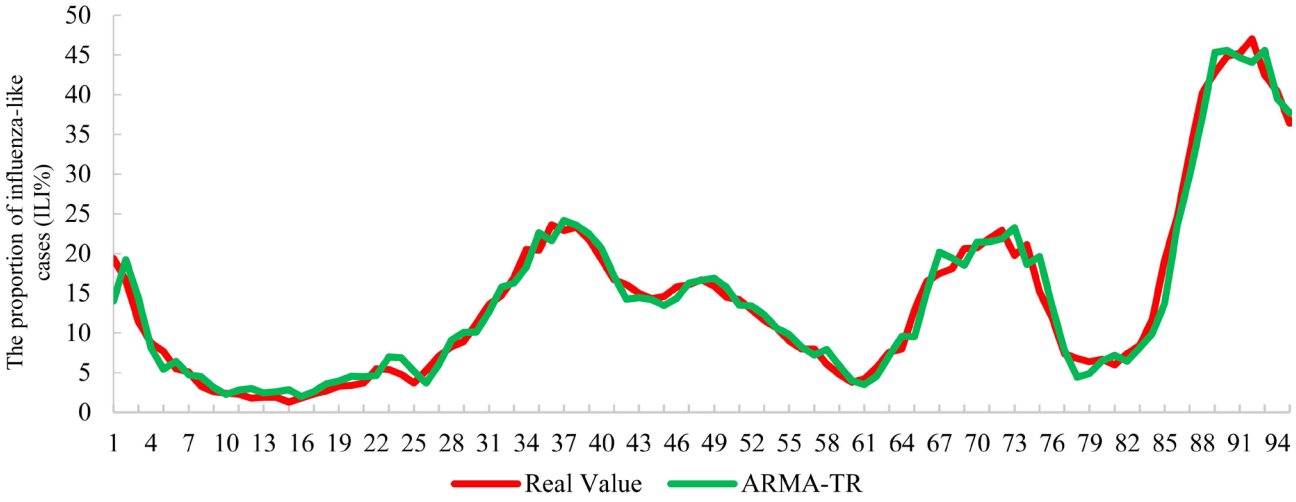


**Figure 2.** The fitting result of ARMA(3,0) model.

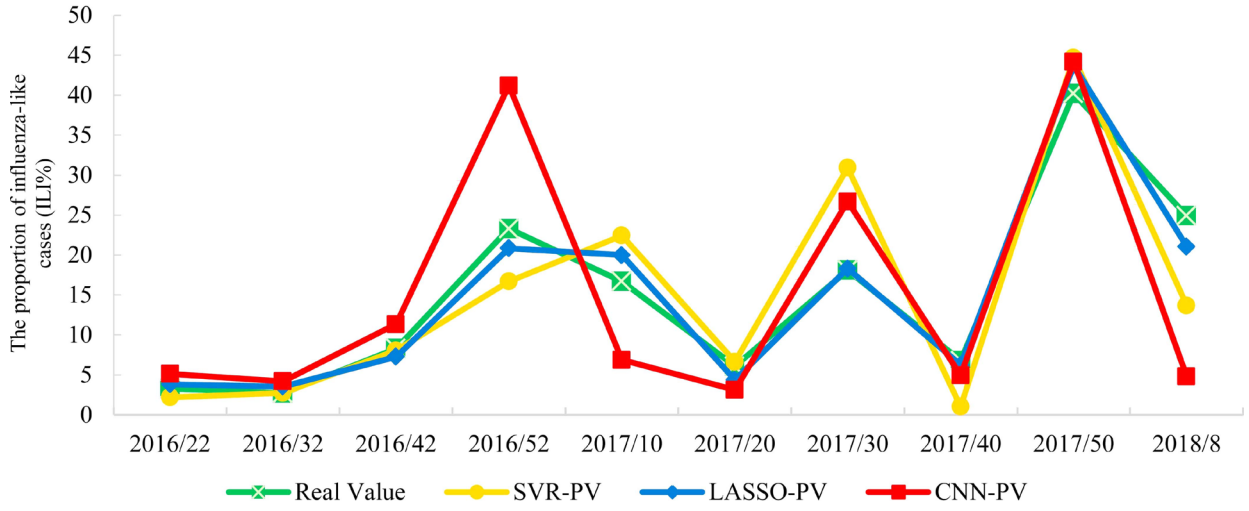


**Figure 3.** Comparison of the prediction results of SVR, LASSO and CNN models.

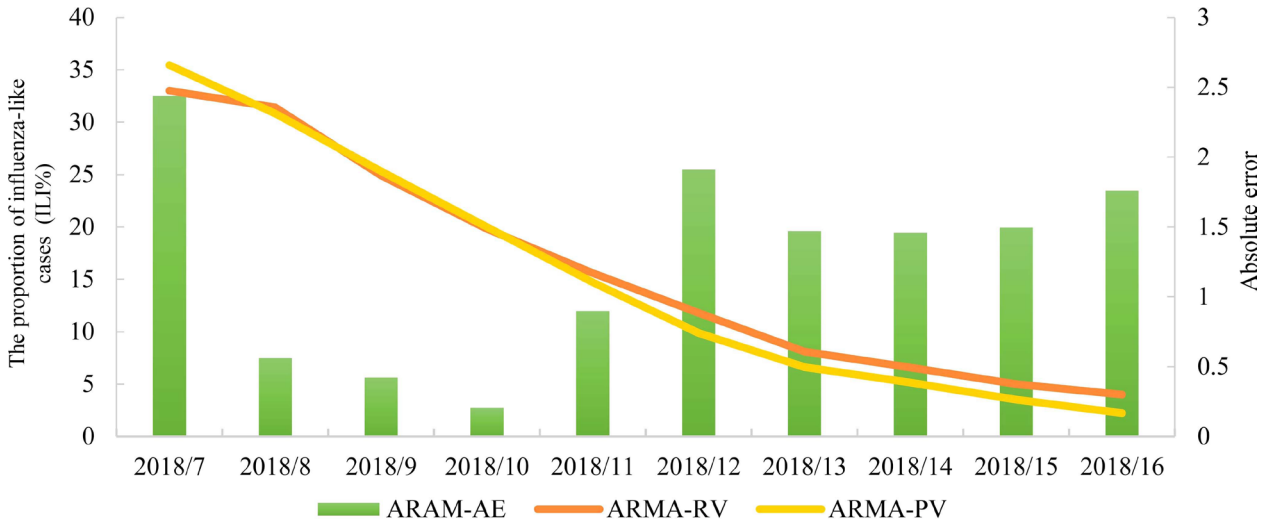


**Figure 4.** The prediction result and absolute error of ARMA(3,0) model.

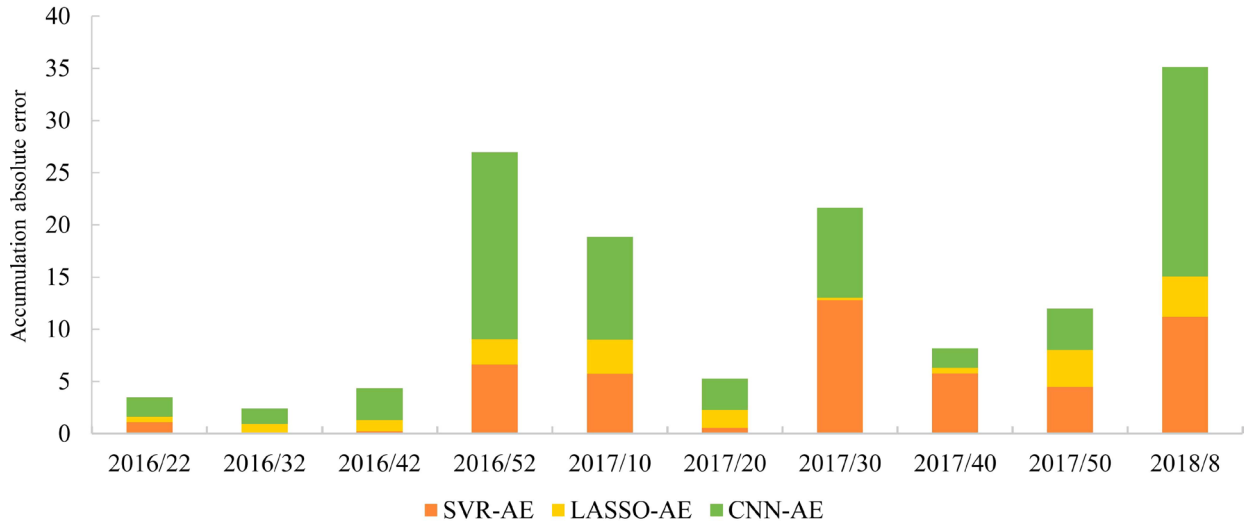


**Figure 5.** Accumulated absolute error of the prediction results of SVR, LASSO and CNN models.

of influenza. The absolute error of ARMA(3,0) is smaller and the error range is (0, 2.5).

From the training RMSE of the model (in Table 2): LASSO > CNN > ARMA(3,0) > SVR, from the perspective of the test RMSE of the model: CNN > SVR > LASSO > ARMA(3,0). By comparison, the ARMA(3,0) model predicts better results and has greater generalization. This reflects the preference for time-series models in predicting the number of influenza cases. The LASSO model also shows a good prediction effect. SVR model performance is poor. The CNN model has the worst prediction effect, which may be due to the small amount of data, resulting in unsatisfactory learning results.

## 6. Conclusions and Prospects

### 6.1. Conclusions

The use of web search data to predict flu epidemics is a popular research in developed countries in recent years. Baidu Index was selected as the web search data source, and the feature extraction of influenza keywords was performed through PCA algorithm. Four influenza prediction models were established and compared to explore the use of web search data to assist in the application of influenza surveillance. The conclusions are as follows:

1) It is feasible to predict the proportion of influenza-like cases by web search data.

2) Machine learning shows a certain predictive effect in the prediction of influenza based on web search data, and it has certain reference value in the future of influenza prediction.

Table 2. Training and prediction results based on SVR, LASSO, CNN and ARMA models.

| Date | Real value (RV) | Predictive value (PV) | | | Date | ARMA | |
|---|---|---|---|---|---|---|---|
| | | SVR-PV | LASSO-PV | CNN-PV | | ARMA-RV | ARMA-PV |
| 2016/22 | 3.3 | 2.190572 | 3.82588 | 5.128867 | 2018/7 | 33 | 35.43721 |
| 2016/32 | 2.7 | 2.759205 | 3.552288 | 4.206986 | 2018/8 | 31.4 | 30.83961 |
| 2016/42 | 8.3 | 8.047085 | 7.270786 | 11.35356 | 2018/9 | 24.9 | 25.3212 |
| 2016/52 | 23.3 | 16.69593 | 20.84384 | 41.21413 | 2018/10 | 19.8 | 20.00432 |
| 2017/10 | 16.7 | 22.42617 | 19.99406 | 6.872914 | 2018/11 | 15.6 | 14.70278 |
| 2017/20 | 6.1 | 6.652845 | 4.370603 | 3.13537 | 2018/12 | 11.8 | 9.889074 |
| 2017/30 | 18.1 | 30.91959 | 18.31482 | 26.71198 | 2018/13 | 8.1 | 6.629661 |
| 2017/40 | 6.8 | 1.038435 | 6.23335 | 4.948258 | 2018/14 | 6.6 | 5.14302 |
| 2017/50 | 40.2 | 44.69814 | 43.72965 | 44.19198 | 2018/15 | 5 | 3.50494 |
| 2018/8 | 24.9 | 13.69753 | 21.04979 | 4.802393 | 2018/16 | 4 | 2.24089 |
| | Training RMSE | 0.1027 | 3.9954 | 1.8670 | | | 1.7123 |
| | Test RMSE | 6.4906 | 2.2268 | 9.6885 | | | 1.4333 |

3) The ARMA(3,0) model has a better predictive result and is more generalized. It also reflects that seasonal characteristics should be taken into account when predicting the proportion of influenza-like cases.

### 6.2. Prospects

The outbreak and epidemic of influenza are affected by a variety of factors, including meteorological factors, virus activity intensity, and air pollution, as well as the combined effects of various factors such as the level of antibody in the population and behavioral patterns. In this study, we only studied flu prediction models by using web search data and influenza history data. Although the use of web search data for influenza surveillance has improved real-time performance, there is still a lack of accuracy, especially at the peak season of the flu season.

Future study directions for this topic include:

1) From the aspect of data sources, on the one hand, we can consider integrating the original search data of multiple search engines to reflect the search behavior of Internet users as fully as possible. In addition, we can obtain interactive behaviors through social networks, professional medical information portals, etc. and browsing behaviors to get more information on influenza concerns; on the other hand, we can collect other metrics that reflect the outbreak and epidemic of flu as a part of the predictive model input.

2) With regard to the scope of research, the scope of the study can be narrowed down to the scope of cities and counties. Based on a regional influenza prediction study, the impact of regional differences can be filtered out, and meteorological factors and other measurement indicators can be introduced more easily.

3) In the aspect of model optimization, more forecasting models can be used for weighted combinatorial optimization, and other better combinatorial optimization methods can also be used. The next optimization goal is to improve the early warning capability and achieve prediction in advance for a period of time.

4) For predictive visualization, some data visualization software can be combined to display the predictive analysis results by using charts and other methods. Displaying the real-time changes of various indicators can help users quickly obtain relevant information and respond quickly.

## Acknowledgements

This project was supported by the Fundamental Research funds for Central Universities, China University of Geosciences (Wuhan) (1810491T09) and Laboratory Research Funds, China University of Geosciences (Wuhan) (SKJ2018240).

## References

[1] Huang, L.R., Yuan, L., Li, R.C., *et al.* (2011) Research on Safety and Immunogenicity of Domestic Influenza Virus Split Vaccine. *Fifth National Symposium on Immunodiagnosis and Vaccine*, Yinchuan, 1 August 2011, 298-301. (In Chinese)

[2] WHO (2014) Seasonal Influenza. http://www.who.int/mediacentre/factsheets/fs211/en/

[3] Brady, R.C. (2010) Influenza. *Adolescent Medicine State of the Art Reviews*, **21**, 236-250.

[4] Shimao, T. (2009) Spanish Flu Related Data. *Kekkaku*, **84**, 685-689.

[5] Centers for Disease Control and Prevention (CDC) (2010) Influenza Activity-United States and Worldwide, June 13-September 25, 2010. *Morbidity and Mortality Weekly Report*, **59**, 1270-1273.

[6] Dijk, A.V., Aramini, J., Edge, G. and Moore, K.M. (2009) Real-Time Surveillance for Respiratory Disease Outbreaks, Ontario, Canada. *Emerging Infectious Diseases*, **15**, 799-801. https://doi.org/10.3201/eid1505.081174

[7] Li, X.T., Liu, F., Dong, J.C.H., *et al.* (2013) Chinese Influenza Surveillance Based on Internet Search Data. *Systems Engineering Theory and Practice*, **33**, 3028-3034. (In Chinese)

[8] FOXS (2016) Online Health Search 2006. http://www.Pewinternet.Org/2006/10/29/online-health-search-2006/

[9] Polgreen, P., Chen, Y.D. and Nelson, F. (2008) Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, **47**, 1443-1448. https://doi.org/10.1086/593098

[10] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009) Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, **457**, 1012. https://doi.org/10.1038/nature07634

[11] Wang, R.J. (2016) Comparison and Optimization of Influenza Alerting Models Based on Internet Search Data. Doctoral Dissertation, Nankai University, Tianjing. (In Chinese)

[12] Yuan, Q., Nsoesie, E.O., Lv, B., *et al.* (2013) Monitoring Influenza Epidemics in China with Search Query from Baidu. *PLoS ONE*, **8**, e64323. https://doi.org/10.1371/journal.pone.0064323

[13] Lu, L., Zou, Y.Q., Peng, Y.S., *et al.* (2016) Comparative Analysis of Baidu Index and Micro Index in Chinese Influenza Surveillance. *Computer Applied Research*, **33**, 392-395. (In Chinese)

[14] Collier, N. (2011) Omg u Got Flu? Analysis of Shared Health Messages for Bio-Surveillance. *Journal of Biomedical Semantics*, **2**, S9. https://doi.org/10.1186/2041-1480-2-S5-S9

[15] Lampos, V., Bie, T.D. and Cristianini, N. (2010) Flu Detector-Tracking Epidemics on Twitter. In: *European Conference on Machine Learning and Knowledge Discovery in Databases*, Vol. 6323, Springer-Verlag, Berlin, 599-602. https://doi.org/10.1007/978-3-642-15939-8_42

[16] Xie, Y., Chen, Z., Cheng, Y., Zhang, K., Agrawal, A., Liao, W.K., *et al.* (2013) Detecting and Tracking Disease Outbreaks by Mining Social Media Data. *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, Beijing, 3-9 August 2013, 2958-2960.

[17] Huang, J., Zhao, H. and Zhang, J. (2013) Detecting Flu Transmission by Social Sensor in China. *IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, Beijing, 20-23 August 2013, 1242-1247. https://doi.org/10.1109/GreenCom-iThings-CPSCom.2013.216

[18] Lu, L. (2015) Prediction of Chinese Flu Trends Based on Internet Data. Doctoral

Dissertation, Hunan University, Changsha. (In Chinese)

[19] Chinese Center for Disease Control and Prevention (2005) National Implementation Plan for Influenza/Human Bird Flu Surveillance. (In Chinese)