Journal of Big Data

RESEARCH Open Access

Search-engine-based surveillance using artificial intelligence for early detection of coronavirus disease outbreak

Ligui Wang[1†], Yuqi Liu[1†], Hui Chen[1†], Shaofu Qiu[1], Yonghong Liu[3], Mingjuan Yang[1], Xinying Du[1], Zhenjun Li[2*], Rongzhang Hao[4*], Huaiyu Tian[3*] and Hongbin Song[1*]

†Ligui Wang, Yuqi Liu and Hui Chen have contributed equally to the work.

*Correspondence: lizhenjun@icdc.cn; hao@ccmu.edu.cn; tianhuaiyu@gmail.com; hongbinsong@263.net

[1] Chinese PLA Center for Disease Control and Prevention, Beijing 100071, China
[2] State Key Laboratory of Infectious Disease Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China
[3] State Key Laboratory of Remote Sensing Science, Center for Global Change and Public Health, College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China
[4] Department of Toxicology and Sanitary Chemistry, School of Public Health, Capital Medical University, Beijing 100069, China

Abstract

The search-engine-based surveillance methods for the early warning and prediction of infectious diseases cannot achieve search engine keywords automatic filtering and real-time updating, lead to powerless for the early warning of emerging infectious diseases. The aim of this study is to develop an artificial intelligence (AI) method for search-engine-based surveillance to improve the early warning ability for emerging infectious diseases. The 32 keywords (444 million search queries) that may be related to the coronavirus disease (COVID-19) outbreak was collected from December 18, 2019 to February 11, 2020 from Baidu's search engine database. The graph convolution network (GCN) model was used to select search engine keywords automatically, and then, multiple linear regression was performed to explore the relationship between the daily query frequencies of keywords and daily new cases. The GCN model was used to automatically select keywords. The prediction trend of the GCN model was highly consistent with the true curve with a mean absolute error of 81.65. Three keywords including "epidemic", "mask" and "coronavirus" were selected. The selection keywords in the search queries were highly correlated with the daily number of confirmed cases ($r = 0.96$, 0.94, and 0.89; $P < 0.01$). An abnormal initial peak (3.05 times the normal volume) in queries appeared on December 31, 2019, which could have served as an early warning signal for an outbreak. Of particular concern, 17.5% of query volume originated from the Hubei Province, 51.15% of which was from Wuhan City. The coefficients of determination (R^2) of our constructed model were 0.88, 0.88, 0.84, 0.77, 0.77, 0.75, 0.73, and 0.73 for a time lag of 0–7 days, respectively, using selection keywords. The model we constructed was used in the Beijing Xinfadi outbreak as an independent test dataset, which successfully predicted the daily numbers of cases for the following days and detected an early signal during the Beijing Xinfadi outbreak ($R^2 = 0.79$). In this paper search-engine-based surveillance based on the AI method was established for the early detection of the COVID-19 epidemic for the first time. The model achieves automatic filtering and real-time updating of search engine keywords and can effectively detect the early signals of emerging infectious diseases.

Keywords: AI, Early detection, Prediction, Search-engine-based surveillance

Introduction

Coronavirus disease (COVID-19) has spread to > 200 countries, causing approximately 532 million confirmed cases and 6 million deaths as of June 10, 2022 [1], and it is the most serious pandemic affecting the world since the 1918 influenza pandemic. Many researchers proposed classification of Covid-19, various types of Pneumonia and normal X-ray images [2, 3]. The machine learning algorithms were used to automatic Diagnosis of Covid-19 from CXR and CT-Scan Images [4, 5]. The novel coronavirus (COVID-19) was found in Wuhan, China in December 2019 [6, 7]. Early warning of infectious disease outbreaks is essential to prevent epidemics. However, at the beginning of the COVID-19 outbreak, traditional infectious disease surveillance did not deliver timely, early warnings. Current research indicates that strict prevention and timely control measures at an early stage can effectively prevent infectious diseases from becoming large-scale epidemics [8, 9]. Thus, countries should place great importance on the construction of infectious disease surveillance systems to detect early signals of any outbreak. However, traditional infectious disease surveillance relies primarily on laboratory diagnosis, making it time-consuming to deliver timely warnings at an early stage [10].

Because the Internet is widely used worldwide, Internet-based surveillance is considered suitable for the early detection and prediction of epidemics [11]. For example, during the 2014 Ebola outbreak in West Africa, HealthMap identified web news reporting a strange fever in Guinea on March 14, 2014, i.e., 9 days before the release of official case information on the ongoing Ebola outbreak. This was the first early warning signal of the Ebola outbreak [12], thereby proving that online data could effectively serve as early warnings of outbreaks. Moreover, online data sounded an alarm for COVID-19 on December 30, 2019 [13]. Polgreen et al. used data from the Yahoo search engine to fit data on laboratory-diagnosed cases of influenza between March 2004 and May 2008; they found that the two sets of data were consistent and that the online data appeared 1–3 weeks earlier than routine reports [14]. Subsequently, Ginsberg et al. proposed the concept of Google Flu Trends (GFT) in 2009 and constructed a GFT prediction model based on 45 keywords related to influenza and influenza-like symptoms, which generated a warning 1–2 weeks earlier than a report from the influenza surveillance system of the Centers for Disease Control and Prevention (CDC) of the United States [15]. This provides preliminary proof that online surveillance can discover epidemics earlier than manual surveillance. Thus, search engine data have been applied to the prediction of other infectious diseases, such as dengue, AIDS, gonorrhea, hand-foot-and-mouth disease, and COVID-19, producing better prediction results [16–21]. Nevertheless, the keyword selection of the aforementioned model depends on manual selection, and it is not possible to update the keywords in real time. Therefore, it is used for the prediction of known infectious diseases and is powerless for the early warning of emerging infectious diseases.

Here, the Baidu Search Index data was used to study the early warning and prediction of COVID-19. Baidu is currently the largest Chinese information search engine and the second largest worldwide. In China, numerous people use Baidu as their preferred search engine, and its market coverage rate is as high as 89.10% [22]. As of June 2019, the number of Internet users in China reached 854 million, with a penetration rate of 61.2%. Additionally, the number of Chinese search engine users reached 695 million [23].

Hence, this study examined search engine-based surveillance based on the AI method for the early detection of COVID-19 using Baidu's search engine data. The study also investigated the possibility of using the abnormal query frequency as an early warning signal for an epidemic and for predicting the number of new cases.

Methods

Data source

The data on confirmed COVID-19 cases were obtained from the Chinese Center for Disease Control and Prevention's surveillance data released by the National Health Commission of China [24]. The time frame was between December 18, 2019 and February 11, 2020. The Baidu search engine query data for all of China were obtained from the Baidu Search Index. In total, 32 keywords were chosen to describe aspects such as infectious diseases, syndromes, pathogens and potential hosts, protective measures, and incidents. The chosen keywords were divided into three groups according to their specificity to COVID-19. Fifteen keywords were general (existing before SARS): "Epidemic" (D), "Masks" (M), "Influenza" (D), "Avian influenza" (D), "Fever" (S), "Cough" (S), "Fatigue" (S), "Muscle soreness" (S), "Shortness of breath" (S), "Respiratory distress" (S), "Sore throat" (S), "Dry cough" (S), "Unsmooth breathing" (S), "ARDS" (Acute respiratory distress syndrome, D) and "Diarrhea" (S). There were eight specific keywords (existing after SARS): "Bats" (P), "Clustered pneumonia" (I), "Wildlife" (P), "SARS" (D), "Pneumonia with unknown cause" (D, I), "Fever with unknown cause" (S, I), "Atypical pneumonia" (D) and "Coronavirus" (P). The final group contained nine highly specific keywords (existing after COVID-19): "Novel coronavirus pneumonia" (D), "Novel coronavirus protection measures" (D, M), "Novel coronavirus" (P), "Wuhan outbreak" (I), "Wenliang Li" (I), "Wuhan" (I), "Eight-doctors' rumor-spreading incident" (I), "Wuhan Seafood Market" (I) and "Wuhan Huanan Seafood Market" (I) (Additional file 1: Table S1). As an independent test dataset, we also collected the query volume of keywords from the Baidu Search Index between May 30, 2020 and July 30, 2020 during the Beijing Xinfadi outbreak and the numbers of confirmed COVID-19 cases of the Beijing Xinfadi outbreak that were released by the Beijing Center for Disease Prevention and Control) between June 11, 2020 and July 5, 2020.

Selection model construction

Data preprocess

Because the situation varied in different cities in China, we first merged the data to the province by summing the query frequency and daily confirmed cases of cities in the same province. We thought that the degree of development of different provinces would affect people's search habits. To balance the different provinces, we designed a standardized search frequency by dividing the query frequency of each province by the standardized per capita gross domestic product. Subsequently, the daily confirmation case was standardized by dividing the standardized population. The data from January 24 to January 29, 2020 were split into a training set for the graph convolution network (GCN) model, and the remaining data were set as the test set. All the specific query words were removed (including Novel coronavirus, Wuhan, Wenliang Li, Novel coronavirus pneumonia, Wuhan outbreak, SARS, Novel coronavirus pneumonia protection measure,

Novel coronavirus protection measure, Wuhan Huanan seafood market, Wuhan seafood market), and 11 keywords including "epidemic," "masks," "influenza," "avian influenza," "bats," "clustered pneumonia," "wildlife," "SARS," "pneumonia with unknown cause," and "coronavirus" was used to construct the GCN model.

GCN model

Graph Convolution Network (GCN) introduced a convolution component designed for graphs, where vertices are allowed to have a varying number of neighbors unlike fixed grids. The framework of keyword selection, including data preprocessing, dataset split, model construction, and feature analysis, is shown in Fig. 1.

The GCN was designed to learn a standardized daily confirmed case. GraphSAGE with a mean aggregator was used to embed the node and extract the graph features as follows:

$$h_v^k \leftarrow W \cdot MEAN\left(\left\{h_v^{k-1}\right\} \bigcup \left\{h_u^{k-1}, \forall u \in N(v)\right\}\right) \tag{1}$$

where W is a weight matrix, $N(v)$ indicates the node set sampled from the neighborhood of the focus node v, and h_v^{k-1} indicates the embedding of layer $k-1$ of focus node v and is the same as h_u^{k-1}. When $k=0$, h_v^0 was set as node feature x_v in the input graph.

The ReLU function was then used to activate the features. Two such convolution layers were stacked, and the output was averaged to standardize the feature as follows:

$$h_g = \frac{1}{|V|}\sum_{v \in V} h_v \tag{2}$$

where V denotes the node set. The output was finally connected to an MLP layer that contained two full connection layers, halving the input size with the ReLU function. As a regression task, the last full connection layer reduced the output feature size by 1. The mean square error function was calculated as a loss function:

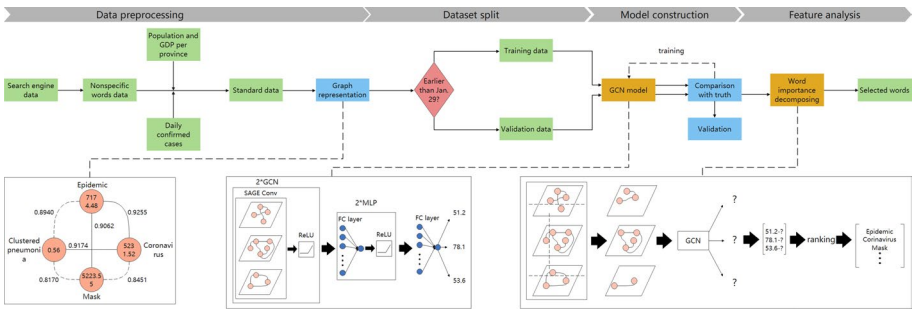

Fig. 1 The framework of keyword selection, containing data preprocessing, dataset split, model construction, and feature analysis. Specific searching words are filtered and standardized by population and the gross domestic product data of each province. Then, the standardized data are represented as graphs with connections between the nodes whose cosine similarity was > 0.9. The graphs are split into the training set (for the train feature learning model) and validation set (only for validating the performance of the feature learning model) corresponding to search engine data up to January 29, 2020. The graph convolution network (GCN) model is used as the feature learning model to learn the relationship between the searching data and epidemic situation. After validation of the GCN model, the importance of the searching words is decomposed by segmenting each node of the graph in turn and evaluating their effect on the result

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{3}$$

The Adam optimizer was used to train the model. The learning rate was set to 0.001, and the batch size was set to 128. The model was trained for 80 epochs to learn the data adequately.

Graph representation

The daily data in each province as a graph was represented. Each query word was represented as a node in the graph, and the corresponding standardized query frequency was set as the node property. We first embedded each query word with the ERNIE 1.0 pretrained language model into a 768 sized vector. The cosine similarity was calculated between each pair of vectors to measure their relevance.

$$\cos\theta_{i,j} = \frac{v_i \cdot v_j}{|v_i| \cdot |v_j|} \tag{4}$$

Then, nodes with a cosine similarity > 0.9 were connected by an edge. Each graph corresponded to a standardized daily confirmed case.

Feature importance

The post-hoc explanation method was used to quantify the feature importance and find frequently used words when the epidemic outbreak occurred. First, we constructed a graph containing nodes corresponding to all query words:

$$G_{total} = \left(Node, Edge\right) \tag{5}$$

where N_w indicates the total index number of searching words.

Each time we selected a word and removed it from the origin graph by deleting the node and cutting the edges connecting it:

$$G_i = \bigcup_{j \neq i} \left(Node_j, Edge_j\right),\ j \in N_w \tag{6}$$

Until all words were selected, we obtained a graph set whose size was equal to the size of the query word set. Then, the graph set was inputted to the model together with the original graph, and each output in the graph set was subtracted from the origin graph output:

$$Imp_i = MLP\left(h_{G_{total}}\right) - MLP(h_{G_i}) \tag{7}$$

With this difference, the importance of all the words was sorted to judge the relevance of each word's query appearance with the epidemic.

Prediction model construction

A multiple linear regression model was used as a quantitative model for the daily numbers of new cases and the daily frequencies of keyword queries (Eq. 8), using the following formula:

$$Y_{(t+j)} = \beta_0 + \beta_1 X_{1(t)} + \beta_2 X_{2(t)} + \beta_3 X_{3(t)}, \ \ j \in \{0, \ldots, 14\} \tag{8}$$

where $Y(t+j)$ represents the data on new cases on day t+j and $X_1(t)$, ..., and $X_3(t)$ represents the query volumes of 3 keywords on day t; these keywords were "coronavirus", "masks, and "epidemic". β_0, ..., β_3 in Eq. (8) represent the coefficients of the variables obtained from the model estimation. The least-squares method was used to optimize the parameters. The coefficient of determination (R^2) was used to evaluate the predictive model.

Results

Automatic selection of query keywords

The GCN model was used to automatically select search engine keywords. We trained the GCN model using the training set and then predicted the situation using a standardized confirmed population in the test set. The predicted trend was highly consistent with the true developing curve, with a mean absolute error of 81.65, which confirmed the performance of our model. Then, the importance of the words was calculated to determine the relevance of each word to the epidemic (Additional file 2: Figure S1A, B). Three keywords including "epidemic," "coronavirus," and "mask" had a positive importance value (47.43, 19.72, and 8.76, respectively), and all the other words were negative.

Correlation analysis of keyword selection

Spearman's correlation analysis was performed on keyword's query frequency and the daily numbers of new cases. The correlation coefficients of the three aforementioned keywords' nationwide query volumes with the daily numbers of new cases were 0.96, 0.94, and 0.89, respectively. The nationwide accumulative query frequencies of the three keywords were highly correlated with the daily numbers of new cases, with a correlation coefficient of 0.96 ($P < 0.01$). In the Hubei province, the correlation coefficient was 0.87 ($P < 0.01$). In Wuhan City, the coefficient was 0.80 ($P < 0.01$) (Fig. 2).

Correlation analysis of search engine retrieval frequency among the cities

The number of keyword queries (per million people) was mainly concentrated in economically developed cities in eastern China. After the first reported case, query volumes increased exponentially ($P < 0.05$). However, there was no significant change in the constituent ratio of query volumes between different cities before and after the first case was reported (Fig. 3A). Furthermore, we analyzed the relationship between the search engine query volume and distance. The results showed that the relevance of the search engine query volume decreased slowly with the increase in the distance between cities; however, when the distance between 2 cities was < 4000 km, the relevance of the search engine query volume and distance between cities still presented a high correlation (Pearson coefficient > 0.90) (Fig. 3B). These results suggest that query frequencies are entirely unrelated to the distance between cities. In addition, there was high consistency in the frequencies of query keywords among the cities.

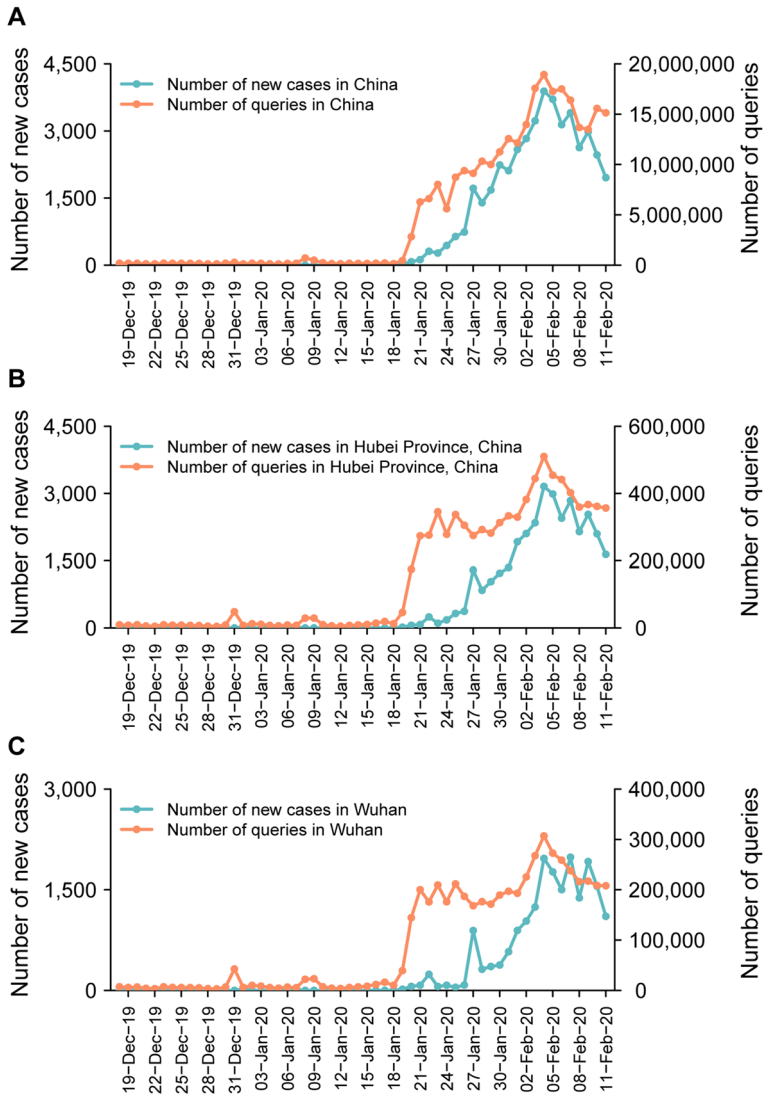

Fig. 2 Correlation between the query volumes and the daily numbers of new cases in China, the Hubei Province, and Wuhan City. (**A**) Correlation of the nationwide query frequencies of three keywords with the daily numbers of new cases. (**B**) In the Hubei Province, the correlation of query frequencies with the daily numbers of new cases. (**C**) In Wuhan City, the correlation of query frequencies with the daily numbers of new cases

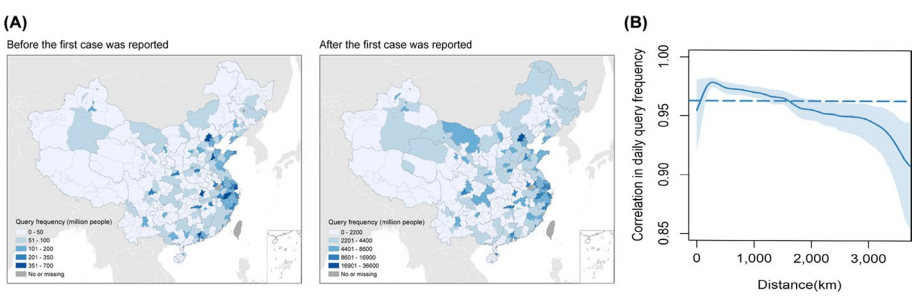

Fig. 3 Relationship between the query frequency and distance between cities. (**A**) Comparison of the query frequency per million population in cities before and after the first case was reported. (**B**) Relationship between the relevance of the search engine query frequency and distance between cities

Detecting early warning signals for COVID-19 based on query volumes

An obvious peak in the query volume (3.05 times the normal volume) was observed on December 31, 2019, based on the volume of the three keywords (Fig. 2A–C). In addition, 17.5% of the nationwide query volume on that day originated from the Hubei Province, 51.5% of which was from Wuhan City. Therefore, the abnormal query volume on that day could have served as an early warning signal for the epidemic. This finding indicates that query frequencies can rapidly reflect the state of public health emergencies, particularly during the early stages of an outbreak.

Further analysis of the query volumes of the 32 keywords on that day compared to the previous day is presented in Additional file 1: Table S1. The query multiple of a keyword was the ratio of its volume on December 31, 2019 to its volume on December 30, 2019. For the 12 keywords, the query multiple was > 2. For three keywords, it exceeded 100: "Wuhan outbreak," 622 times; "pneumonia with unknown cause," 321 times; and "Wuhan Seafood Market," 241 times. All three (100%, 3/3) of those were related to incidents, and 1 of which was also related to infectious disease. Six other keywords with large increases in query volume were "Wuhan Huanan Seafood Market" (93 times), "novel coronavirus" (49 times), "SARS" (46 times), "Wenliang Li" (10 times), "coronavirus" (10 times), and "clustered pneumonia" (which increased from 0 to 8); three (50%, 3/6) of those were related to incidents. For three other keywords, the query volume increased by at least a factor of 2: "epidemic" (4 times), "Wuhan" (2 times), and "masks" (2 times); one (33%, 1/3) of which was related to incidents. For the other 20 keywords, the query volume increased by a factor of < 2; 12 (60%, 12/20) of these keywords were related to symptoms, and only 2 (10%, 2/20) were related to incidents.

Prediction of daily numbers of new cases of COVID-19 using query volumes

Three optimal keywords ("epidemic," "masks," and "coronavirus") was used to predict the daily numbers of new cases and potential future epidemics. To construct a quantitative model of the relationship between keyword query frequencies and the daily numbers of new cases, we performed a correlation analysis on the query volumes and daily numbers of new cases with time lags of 0–14 days. The correlation coefficients were 0.93, 0.93, 0.91, 0.88, 0.85, 0.83, 0.82, 0.80, 0.81, 0.84, 0.81, 0.75, 0.72, 0.68, and 0.60 for a time lag of 0–14 days, respectively; the correlation gradually decreased as the interval between the query and case report increased.

A mathematical model based on the keyword query frequency was constructed to predict the number of new cases with time lags of 0–14 days. Using multiple linear regression on the relationship between the keyword query frequency and the number of reported cases, we obtained the coefficients β_0–β_4. The coefficients of determination R^2 were 0.88, 0.88, 0.84, 0.77, 0.77, 0.75, 0.73, 0.73, 0.76, 0.76 0.72, 0.66, 0.70, 0.72, and 0.63 for a time lag of 0–14 days, respectively (Fig. 4); this indicates that the query-frequency-based model predicted the number of new cases with a 2-day lag, with a coefficient of determination > 0.8. The aforementioned results suggest that this model can use search engine data to accurately predict the number of new cases in the next 2 days, which is earlier than the CDC surveillance data released by the National Health Commission of China (Fig. 4A–H).

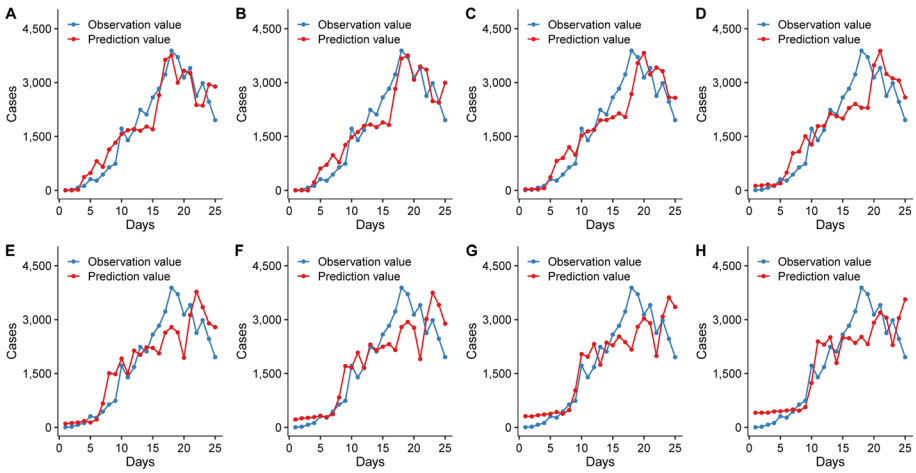

Fig. 4 Using search engine query volumes to predict the numbers of new cases. (**A**)–(**H**) correspond to the prediction of the daily numbers of new cases using query volumes with 0–7-day lags

Verification of the Beijing Xinfadi outbreak

The query volumes of three keywords ("epidemic," "masks," and "coronavirus") and the numbers of confirmed COVID-19 cases during the Beijing Xinfadi outbreak were collected as test data. We found that the query frequencies of the three keywords were highly correlated with the daily numbers of confirmed cases in the Beijing Xinfadi outbreak, with a Pearson correlation coefficient of 0.84 (Fig. 5A). The query volumes of the three keywords increased quickly from 16.77 million on June 11, 2020 to 28.47 million on June 13, 2020 (Fig. 5A), indicating that an early warning signal can be detected using our selected keywords. We predicted the daily number of new cases in the Beijing Xinfadi outbreak using the query volumes of the three chosen keywords ($R^2 = 0.80$) (Fig. 5B).

Discussion

To date, across all studies, search-engine-based surveillance for the early warning and prediction of infectious diseases has been based on statistical methods, and the selection of keywords mainly depends on human experience. Therefore, it cannot achieve automatic filtering and real-time updating of search engine keywords, leading to powerlessness in the early warning of emerging infectious diseases. In this study, we developed an AI method for search-engine-based surveillance that can reduce the number of artificially maintained works and improve the early warning ability of emerging infectious diseases. In the actual work, the keywords used for early warning will be automatically updated according to the retrieval frequency and relevance, and abnormal signals will be quickly identified for early warning of emerging infectious diseases.

The search volume suddenly increased on December 31, 2019, based on our selected keywords, and query volume of keywords specific to COVID-19, such as "Wuhan outbreak," "Wuhan Seafood Market," and "pneumonia with unknown etiology/cause (PUE)," increased sharply (by a factor of 10–600), which probably triggered an increase in the volume of general keywords such as "masks," "avian influenza," and "epidemic," causing an early warning signal. We found that search engine data

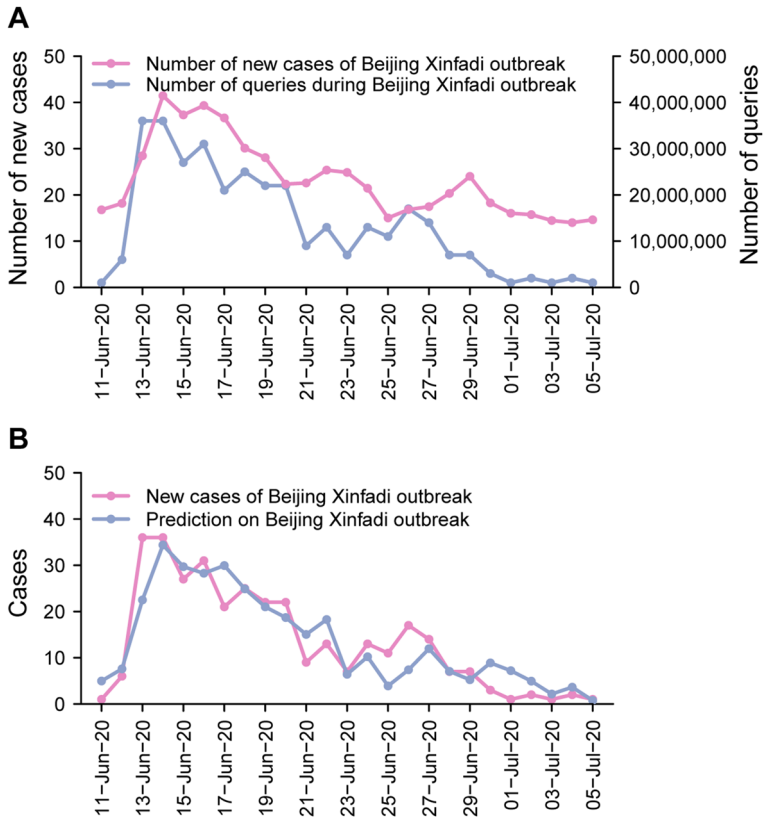

Fig. 5 Query volumes and the daily numbers of new cases during the Beijing Xinfadi outbreak. (**A**) Correlation between the query volumes and the daily numbers of new cases in the Beijing Xinfadi outbreak. (**B**) Prediction of the daily numbers of new cases, based on search engine query volumes, during the Beijing Xinfadi outbreak

revealed an abnormal increase on December 31, 2019 before the officially confirmed epidemic outbreak in Wuhan, which may have been a consequence of online information regarding the discovery of pneumonia cases with unknown causes in Wuhan [25, 26]. On that day, the query volume of the "Wuhan outbreak" had a 622-fold increase, "PUE" had a 321-fold increase, and "Wuhan Seafood Market" had a 241-fold increase. As these messages spread throughout the Internet, the related query volumes soared, and specialist teams from the National Health Commission started their investigations in Wuhan. Our results show that the early warning signal from a search engine usually started from the keywords of special events and then spread rapidly to general keywords.

The constructed model was used for the Beijing Xinfadi outbreak as an independent test dataset, which successfully predicted the daily numbers of cases for the following days and detected an early signal during the Beijing Xinfadi outbreak. Our study demonstrates that our model can detect early warning signals and predict the daily number of new cases accurately and rapidly. The search-engine-based method is faster than traditional infectious disease surveillance systems that rely on laboratory tests or case reports. Therefore, we suggest that in countries with inadequate screening abilities for obtaining precise and timely information on the numbers of cases,

query-volume-based prediction of the potential number of cases would be a powerful tool for estimating the infectious disease trend, particularly at an early stage of an outbreak. In conclusion, our study established search-engine-based surveillance using the AI method for the early detection of the COVID-19 epidemic for the first time. The model achieves automatic filtering and real-time updating of search engine keywords and can effectively detect the early signals of emerging infectious diseases.

Conclusions

The study established search-engine-based surveillance using the AI method for the early detection of the COVID-19 epidemic for the first time. The model achieves automatic filtering and real-time updating of search engine keywords and can effectively detect the early signals of emerging infectious diseases.

Limitations

We only used a COVID-19-related query as an example to show how to choose and optimize search keywords for timely detection and prediction during the early stage of an epidemic. The 32 keywords chosen in this study may not cover all related and sensitive keywords for COVID-19; some important keywords will inevitably be omitted. Some search keywords in this study are not specific to COVID-19 and could be applicable to various other infectious diseases, thereby possibly producing false-positive query results for COVID-19.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-023-00847-9.

Additional file 1. Table S1. Correlation coefficients of the query volumes of 32 keywords with the numbers of new cases of coronavirus disease and their query multiples (increases in query volume) from December 30, 2019 to 31 December 2019.

Additional file 2. Figure S1. Validation result of the graph convolution network model and importance of searching words. (**A**) Summative standardized proportion of coronavirus disease-confirmed cases in populations of each province in the validation set from January 30, 2020 to February 11, 2020. (**B**) Sorted word importance of filtered searching words. "Epidemic," "coronavirus," and "mask" had positive importance, indicating a strong correlation with the spreading progress of the disease.

Acknowledgements
None.

Author contributions
H.S, H.T, Z.L and R.H designed the experiments. L.W, H.C, S.Q, Y.L, M.Y and X.D collected and analyzed data. L.W and H.S wrote the main manuscript text. All authors reviewed the manuscript.

Funding
This work was financially supported by the National Key R&D Program of China [grant number 2021YFC2302004], Beijing Science and Technology Planning Project [grant number Z201100005420010], and National Natural Science Foundation of China [grant number 82073616].

Availability of data and materials
The search engine query data are available from Baidu. Restrictions apply to the availability of these data sets, which were used under license for the current study, and so are not publicly available. These data sets are however available from the authors upon reasonable request and with the respective permission of Baidu.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 2 September 2022 Accepted: 17 October 2023
Published online: 11 November 2023

References

1. WHO. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/. Accessed at 12 June, 2022.
2. Kumar N, Gupta M, Gupta D, Tiwari S. Novel deep transfer learning model for COVID-19 patient detection using X-ray chest images. J Ambient Intell Humaniz Comput. 2023;14(1):469–78.
3. Kaur M, Kumar V, Yadav V, Singh D, Kumar N, Das NN. Metaheuristic-based deep COVID-19 screening model from chest X-ray images. J Healthc Eng. 2021;2021:8829829.
4. Kumar N, Hashmi A, Gupta M, Kundu A. Automatic diagnosis of Covid-19 related pneumonia from CXR and CT-scan images. Eng Technol Appl Sci Res. 2022;12(1):7993–7.
5. Kumar N, Aggarwal D. LEARNING-based focused WEB Crawler. IETE J Res. 2023;69(4):2037–45.
6. Narayan Das N, Kumar N, Kaur M, Kumar V, Singh D. Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays. Ing Rech Biomed. 2022;43(2):114–9.
7. Kumar N, Narayan Das N, Gupta D, Gupta K, Bindra J. Efficient automated disease diagnosis using machine learning models. J Healthc Eng. 2021;2021:9983652.
8. Lu FS, Hattab MW, Clemente CL, Biggerstaff M, Santillana M. Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. Nat Commun. 2019;10(1):147.
9. Nsoesie EO, Kluberg SA, Mekaru SR, Majumder MS, Khan K, Hay SI, Brownstein JS. New digital technologies for the surveillance of infectious diseases at mass gathering events. Clin Microbiol Infect. 2015;21(2):134–40.
10. Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis. 2014;14(2):160–8.
11. Madoff LC, Li A. Web-based surveillance systems for human, animal, and plant diseases. Microbiol Spectrum. 2014;2(1):OH-0015–2012.
12. Milinovich GJ, Magalhães RJS, Hu W. Role of big data in the early detection of Ebola and other emerging infectious diseases. Lancet Glob Health. 2015;3(1):e20–1.
13. Science. Artificial intelligence systems aim to sniff out signs of COVID-19 outbreaks. https://www.sciencemag.org/news/2020/05/artificial-intelligence-systems-aim-sniff-out-signs-covid-19-outbreaks. Accessed at 12 June 2022.
14. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using internet searches for influenza surveillance. Clin Infect Dis. 2008;47(11):1443–8.
15. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012–4.
16. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis. 2014;8(2): e2713.
17. Ling R, Lee J. Disease monitoring and health campaign evaluation using Google search activities for HIV and AIDS, stroke, colorectal cancer, and marijuana use in Canada: a retrospective observational study. JMIR Public Health Surveill. 2016;2(2): e156.
18. Xiao Q, Liu H, Feldman M. Tracking and predicting hand, foot, and mouth disease (HFMD) epidemics in China by Baidu queries. Epidemiol Infect. 2017;145(8):1699–707.
19. Senecal C, Widmer RJ, Lerman LO, Lerman A. Association of search engine queries for chest pain with coronary heart disease epidemiology. JAMA Cardiol. 2018;3(12):1218–21.
20. Ben S, Xin J, Chen S, Jiang Y, Yuan Q, Su L, Christiani DC, Zhang Z, Du M, Wang M. Global internet search trends related to gastrointestinal symptoms predict regional COVID-19 outbreaks. J Infect. 2022;84(1):56–63.
21. Rajan A, Sharaf R, Brown RS, Sharaiha RZ, Lebwohl B, Mahadev S. Association of Search Query Interest in Gastrointestinal Symptoms With COVID-19 Diagnosis in the United States: Infodemiology Study. JMIR Public Health Surveill. 2020;6(3): e19354.
22. Huang S, Liu K, Jiang J. Progress in research of infectious disease surveillance and prediction based on internet search engine. Disease Surveill. 2018;33(11):945–9.
23. The 44th China Statistical Report on Internet Development. http://www.cac.gov.cn/2019-08/30/c_1124938750.htm. Accessed at Mar 15, 2020.
24. National Health Commission of the People's Republic of China. http://www.nhc.gov.cn/. Accessed February 25, 2020.
25. CNTV. http://news.cctv.com/2020/01/09/ARTIwHRH1FDONdbpuIwSucm4200109.shtml. Accessed March 24, 2020. News.
26. Sina. http://finance.sina.com.cn/china/gncj/2020-03-19/doc-iimxyqwa1748367.shtml. Accessed March 19, 2020. News.

Publisher's Note