

# Viewing the Progression of COVID-19 with NewsStand

John Kastner  
kastner@umd.edu  
University of Maryland  
College Park, Maryland 20742

Hong Wei  
hyw@cs.umd.edu  
University of Maryland  
College Park, Maryland 20742

Hanan Samet  
hjs@cs.umd.edu  
University of Maryland  
College Park, Maryland 20742

## ACM Reference format:

John Kastner, Hong Wei, and Hanan Samet. 2016. Viewing the Progression of COVID-19 with NewsStand. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17)*, 4 pages. DOI: 10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

With the continuing spread of COVID-19, it is clearly important to be able to track the progress of the virus over time to be better prepared to anticipate its emergence in new regions. There are application to monitor officially released numbers of cases [1] which are likely to be the most accurate method for tracking the progress of the virus; however, they will not necessarily paint a complete picture. To begin filling any gaps in official reports, we have developed an application<sup>1</sup> usable on desktop and mobile devices that allows users to explore the geographic spread in discussion about the virus through analysis of keyword prevalence in geotagged news articles.

The application we describe here relies heavily on the existing NewsStand system [4]. NewsStand is a well suited basis for this application because it is a spatio-textual search engine. That is, it has the capacity to query a database of news articles in terms of both the raw textual content of the articles and the implicit spatial information encoded by toponyms mentioned in the article. In terms of what this paper requires, textual queries are needed to identify articles that contain keywords of interest to the user while spatial queries are needed to display the geographic distribution of articles that contain the keywords.

## 2 GEOCODING NEWS KEYWORDS

Two steps are required to obtain geocoded news keywords: important keywords must be extracted from news articles, and the news articles must be geocoded by identifying and resolving toponyms to specific geographic coordinates. Once we have identified a set of keywords and a set of geographic locations in an article, we take the cross product of these sets as the set of geocoded keywords. In other words, we create a pair associating every keyword in the article with every location mentioned in the article. Each of the

<sup>1</sup><https://coronaviz.umiacs.io>

This work was sponsored in part by the NSF under Grant iis-18-16889.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference'17, Washington, DC, USA

© 2016 ACM. 978-x-xxxx-xxxx-x/YY/MM...\$15.00

DOI: 10.1145/nnnnnnn.nnnnnnn

keyword location pairs is also associated with the time of publication of the article in order to enable the temporal component of our application.

## 2.1 Keyword Extraction

The extraction of news keywords is handled by the original NewsStand system. NewsStand identifies keywords based on their TF-IDF scores. Words with high TF-IDF scores, words that appear much more frequently in a given document than in other documents in the database, are likely to be important to the document in which the words appear. The word with the highest TF-IDF score for a document is therefore the best keyword. More keywords can be obtained by selecting words with progressively lower scores.

For the purpose of this application, once keywords have been extracted from article text, we examine only those that we believe will be most relevant to the disease we are tracking. For instance, we may specifically look for news articles containing the keyword “coronavirus”. This approach should achieve high precision, since it is unlikely that an article unrelated to the virus will contain this keyword, but it may suffer from low recall.

## 2.2 Geocoding

Geocoding is the process of associating concrete geographic information (i.e. latitude longitude pairs) with a piece of text. Geocoding can be framed as a specific variant of the keyword extraction task in the sense that we must first find keywords that are likely to refer to geographic locations and then decide which of these locations are important enough to associate with the documents [2]. The third step in geocoding which is not required for general keyword extraction is toponym resolution. In toponym resolution, a toponym must be assigned a single latitude longitude pair to be its final location. This is nontrivial because there many ambiguous toponyms that can be used to refer to multiple distinct locations (e.g. Paris, London, etc.) [3].

## 3 APPLICATION INTERFACE

Our application is implemented as an interactive website using HTML, CSS, and JavaScript that communicates with our database through a Ruby On Rails web server. It consists of an interactive map query interface that is used to display our data and animations. In addition, a collection of controls are used to download data from our server and to select how it is displayed on the map. A screenshot of our application is show in Figure 1.

### 3.1 Map Query Interface

Our map query interface is an interactive web map provided by the Leaflet JavaScript Library. Data is rendered onto this map using a technique called marker clustering, which is implemented by an

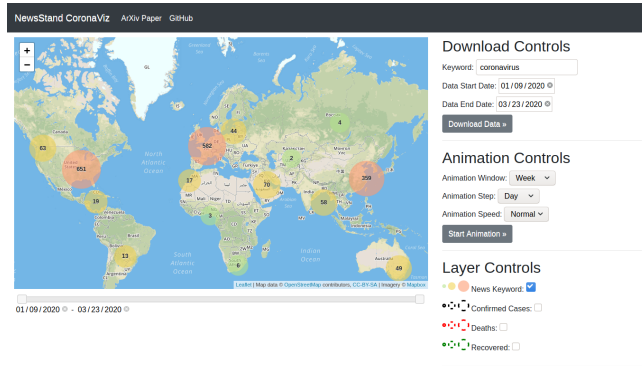


Figure 1: Application Screenshot

extension to Leaflet. Marker clustering allows large numbers of points to be rendered quickly without overloading the user with information. Rather than rendering each point individually, points are clustered into aggregate markers. As the user increases the zoom level of the map, focusing on a smaller area, these aggregate markers are split into two or more new markers that together represent all the points represented by the original marker. This decomposition allows more detail to be displayed when examining a small area without showing too much detail at lower zoom levels.

Each marker represents some collection of points, so the size and color of the markers are scaled to represent the magnitude of this collection. Markers representing a single point are set to have a radius of 40 pixels. From there, the marker's radius scales with the squared logarithm of the number of points ( $r_M = 40 + \log^2|M|$ ). Marker color follows a linear scaling from green for markers representing a single point, up to red for those that have more than 100 points. These functions are chosen to be visually appealing, but there is no other concrete reason for their selection.

The map query interface also gives access to the news articles from which the markers are derived. When a marker that represents more than one physical location is clicked, the zoom level is increased to a point where the constituent points are represented by separate markers, but when a marker that represents a single location is clicked, a pop up listing the articles from within the specified time range that mention the location is opened.

### 3.2 Interface Controls

The first control located below the map interface is a slider bar that can be manipulated to control the range of dates in the data displayed on the map. Moving either knob on the slider causes the map interface to be updated. The exact dates selected are displayed below the slider.

Below the date range slider there are three panels labeled "Download Controls," "Animation Controls," and "Layer Controls."

The download controls are used to retrieve data from our database. First, there is a text entry box labeled "keyword" in which the user types a keyword that they want to visualize. There are then two date entry fields labeled "Data Start Date" and "Data End Date". These control the range of dates for data that is downloaded from our server. Note that this is not necessarily the same as the range

of dates for data that is rendered on the map. Finally, there is a button labeled "Download Data" that, when clicked, sends a request to our server for data within the specified time range matching the provided keyword. This data is rendered on the map query interface, replacing any previous data.

The animation controls deal with using downloaded data to generate time series animations. Most prominently, there is the "Start Animation" button that initiates the animation when clicked. Clicking this button second time will stop an ongoing animation. The properties of this animation are controlled by the next three inputs. The first input controls the size of the range of dates displayed in each frame of the animation. It can be set to hour, day, week, or month. It can also be set to a custom size through use of the "Displayed Data Range" controls discussed above. Next, there is a selection input for the interval advanced between each frame of the animation. This can be set to hour, day, week, or month. The final input field is used to set the speed of the animation to either "slow", "normal", or "fast". These values correspond to a minimum number of milliseconds waited between frames of the animation: 500, 100, and 0 milliseconds respectively. While these are minimum delays, the actual interval may be longer due to time required to compute the subsequent frame.

Finally, the layer controls are used to select between the possible data layers that can be rendered on the map interface. By default, only the news keyword data layer is selected. When the confirmed cases layer is selected, there are marker representing the total number of confirmed cases in a location rendered on the map. Similarly there are layers for recoveries and deaths that contain markers representing these quantities. The data used to generate these layers was originally gathered by Dong et al. [1].

## 4 EXAMPLE USAGE

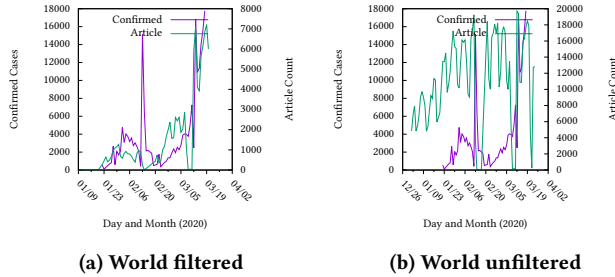
This section provides a walk through demonstrating how our application can be used to obtain an animation to show the geographic change in the discussion of the Novel Corona Virus over time. The instructions here are applicable to other similar uses of our application.

To begin, navigate to [coronaviz.umiacs.io](https://coronaviz.umiacs.io) to access our website. Once there the "Keyword" entry field should be initially populated with "coronavirus". If this is not the case, select the field and type in this term. The next step is selecting the range of dates to be viewed. A reasonable range for visualizing Corona Virus might start in December 2019 and proceed until the current date. To make this selection, click on and select a date for the "Data Start Date" field. "Data End Date" should have been initialized the current date, but, if it is not, a date should also be selected here. Once the keyword and date range have been selected, click "Download Data" to download this data from our server. This may take a noticeable amount of time depending on network speed and the size of the query result. Once the download is complete, data will be rendered on the map.

The next step is to configure the animations controls and initiate an animation sequence. The animation options ("Animation Window" and "Animation Step") are set to week and day respectively by default. These should be reasonable values but, at this point, they can be changed. Clicking the "Start Animation" button starts

Query Area	Correlation coefficient	
	Keyword <i>coronavirus</i>	No filter
World	0.68	-0.05
China	-0.09	-0.08
Hubei Province	-0.11	-0.11
United States	0.71	0.06
Washington state	0.24	0.10

**Table 1: Correlation coefficient between cumulative new articles and confirmed cases at each scale.**



**Figure 2: Filtered and unfiltered data at the global scale.**

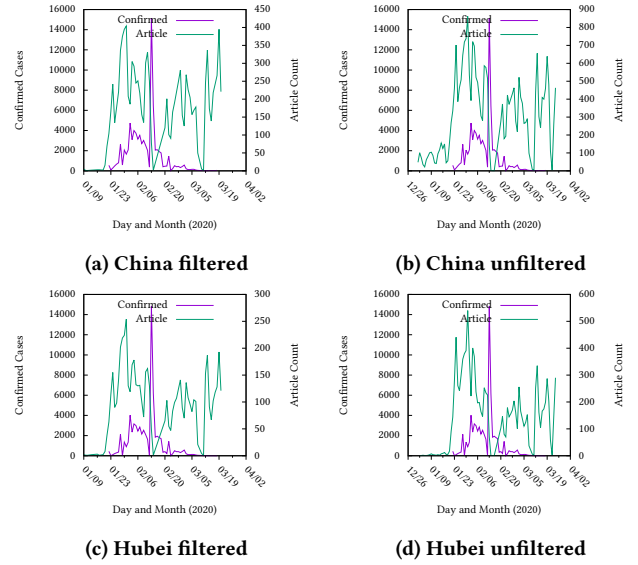
an animation on the map query interface that begins by displaying downloaded data with the earliest publication date and proceeds until the final frame that contains the most recent data. Each frame of the animation shows a period of time defined by “Animation Window” and the start of the frame is advanced by an interval defined by “Animation Step” between each frame. The animation can be terminated early by clicking the button, now labeled “Stop Animation”, a second time.

## 5 EVALUATION

In this section we evaluate the effectiveness of tracking keyword usage in news article to better understand the progression of COVID-19. To do this we compare the confirmed case numbers collected by Dong et al. [1] to the cumulative number of articles recognized by our system. We include geographic information in our analysis by limiting it to confirmed cases and news articles in a specific location. We are able to restrict the analysis in this manner because the dataset of confirmed cases we use already organizes confirmed case numbers by location, and our news article dataset is geocoded as described in Section 2.2. To further investigate the utility of our technique, we perform the comparison at three geographic scales: world, nation, and region. The results of this evaluation are summarized in Table 1 and Figures 2, 3, and 4.<sup>2</sup>

We first ignore specific geographic information extracted from articles and perform this analysis on a global scale. This does not utilize the geographic information we collect, but it is interesting as a point of comparison to other scales. Figure 2a plots the daily number of articles using the keyword “coronavirus” as a time series

<sup>2</sup>In each plot of article counts, the daily number of articles collected is zero from February 14 to February 17, March 9 to March 11, and on March 22. This was caused by problems in our data collection system, so the data points are removed when computing the correlation coefficients in Table 1.

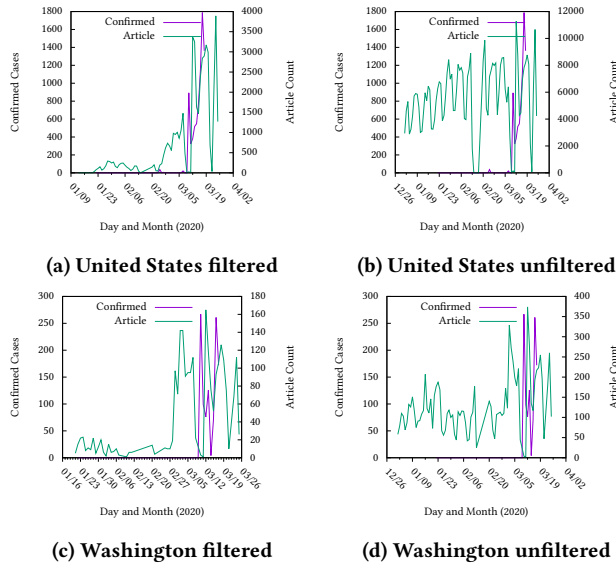


**Figure 3: Filtered and unfiltered data in all of China and Hubei province.**

between January 16, 2020 and March 19, 2020. To demonstrate the utility of news articles in tracking the progress of the virus, this figure also shows a time series for the daily number of new confirmed cases world wide. It is clear that both datasets trend upwards at similar rates. To determine if correlation exists between article counts and case numbers, we calculated the Pearson correlation coefficient between the datasets. We found coefficient to be 0.68, indicating that the datasets are linearly correlated. To investigate the effect of our keyword filter, we performed a similar comparison where we do not filter news articles for any specific terms. Figure 2b show the time series plot for this data. This time, the article counts clearly do no correlate with case numbers. This is realized in a  $-0.05$  correlation coefficient. Since the magnitude of the correlation coefficient is significantly larger when applying the keyword filter, we can see that it is useful in this application.

The plots in Figure 3 are formatted the same as Figure 2, but they focus on smaller geographic scales. In Figure 3a, the plot is restricted to articles geocoded to China and confirmed cases in China. The correlation coefficient between these datasets is  $-0.09$ , indicating that there is no correlation after applying these geographic filters. Figure 3c further restricts the data to the Hubei province of China. The correlation coefficient at this scale is  $-0.11$ , similar to found for all of China. Figures 3b, and 3d show the plots for China and Hubei when the keyword filter is not applied. Since there was no correlation found with the filter applied we do not expect anything to change when it is removed. As expected, the new correlation coefficients are almost unchanged with  $-0.08$  and  $-0.11$  for China and Hubei respectively.

In our final set of evaluations, we shift geographic focus from China to the United States. Figure 4 contains plots for a country level of China analysis of the United States and a more specific analysis of Washington state. In contrast to the country level analysis where no correlation was found, the correlation coefficient



**Figure 4: Filtered and unfiltered data in all of the United States and Washington state.**

found for the United States (0.71) is greater than that found in the global analysis. While the correlation coefficient calculated for the Washington data is much lower (0.24), it is notably higher than the values computed for either China or Hubei. The correlation coefficient for the United States and Washington when not applying the keyword filter (0.06 and 0.10 respectively) is comparable to that for that of the three previous datasets.

In our evaluations we found that the number of news articles using the term "coronavirus" is correlated with the number of new cases of COVID-19 when correlation is computed at a global scale. When restricting the analysis to a country wide scale, we found correlation when using data for the United States, but we were not able to find any correlation in our dataset for China. There is a similar but less pronounced relationship at the region level. While the Washington state dataset does not display particularly strong correlation with the confirmed cases dataset, it is notably higher than the correlation we found for Hubei. A possible explanation for this inconsistency is that many articles discussing COVID-19 will mention China at some point since it is the origin of the virus. To overcome this issue we may be able to develop a technique to determine which term, toponym pairs in a news article should be included in our dataset. Our dataset currently contains an entry for every toponym in an article if the article also contains the query term.

## REFERENCES

- [1] E. Dong, H. Du, and L. Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* (2020).
- [2] M. Lieberman and H. Samet. 2011. Multifaceted toponym recognition for streaming news. In *SIGIR*. 843–852.
- [3] M. Lieberman and H. Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *SIGIR*. 731–740.
- [4] H. Samet, J. Sankaranarayanan, M. Lieberman, M. Adelfio, B. Fruin, J. Lotkowski, D. Panozzo, J. Sperling, and B. Teitler. 2014. Reading news with maps by exploiting spatial synonyms. *Commun. ACM* 57, 10 (2014), 64–77.