

Viewing the Progression of the Novel Corona Virus (COVID-19) with NewsStand

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

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

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

ACM Reference format:

John Kastner, Hong Wei, and Hanan Samet. 2016. Viewing the Progression of the Novel Corona Virus (COVID-19) with NewsStand. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17)*, 3 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. Officially released numbers of cases will likely be the most accurate means by which to track this, but they will not necessarily paint a complete picture. We have developed an application¹ usable on desktop and mobile devices that allows users to explore geographic spread in discussion about the virus through analysis of keyword prevalence in geotaged news articles.

The application we describe here relies heavily on the existing NewsStand system [3]. 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 keyword location pairs is also associated with the time of publication of the article in order to enable the temporal component of our application.

¹<https://coronaviz.umiacs.io>

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 coordinate value 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 [1]. 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 value 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.) [2].

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 shown 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 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

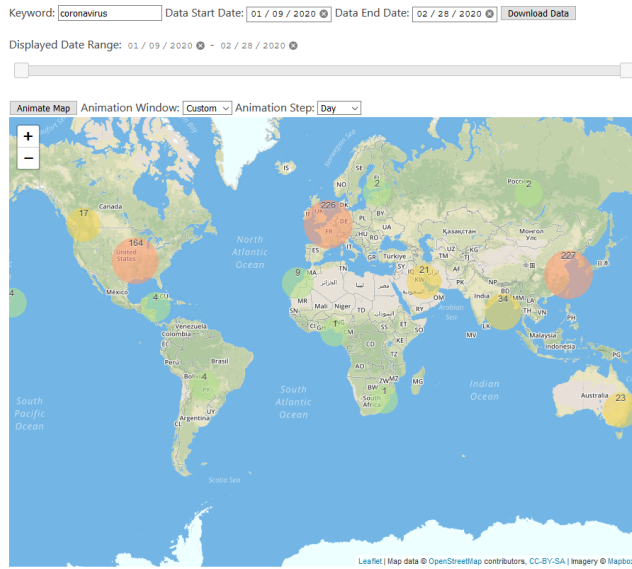


Figure 1: Application Screenshot

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 scalings 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 set of 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 second set of controls select the subset of the downloaded data that is rendered on the map query interface. When any of these controls are used, the data on the map is immediately updated to reflect the new configuration. The range of displayed data can be precisely set using a pair of date entry fields jointly labeled "Displayed Data Range". To set the range in a more intuitive manner, the user can also manipulate the slider bar positioned directly below these fields. Changes made to this slider bar will be shown in the date entry fields. Likewise, changes made to the date entry fields are shown on the slider.

The final set of controls deals with using downloaded data to generate time series animations. Most prominently, there is the "Animate Map" button that initiates the animation when clicked. The properties of this animation are controlled by the next two inputs. The 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. The final input controls the interval advanced between each frame of the animation. This can be set to hour, day, week, or month.

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.umiaccs.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 "Animate Map" button starts 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.

References

- [1] M. Lieberman and H. Samet. 2011. Multifaceted toponym recognition for streaming news. In *SIGIR*. 843f1?852.
- [2] M. Lieberman and H. Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *SIGIR*. 731–740.

- [3] 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.