Visualizing SpatioTemporal Keyword Trends in Online News Articles

John H. Kastner
john.h.kastner@gmail.com
University of Maryland
Computer Science Department
College Park, Maryland

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

ABSTRACT

Online sources of news have steadily supplanted their paper counterparts alongside the growth of the internet. This growth in online news has led to a surplus of data in the form of the text of news articles published online. While an abundance of data is obviously desirable, it can make it difficult for a human to analyze and find trends in the data without assistance. The application demonstrated in the paper aims to aid users in such analysis by building a spatiotextual and spatiotemporal data visualization based on the existing NewsStand architecture. The application is shown to be applicable to tracking the changing geographic prevalence of a disease (e.g., COVID-19) over time.

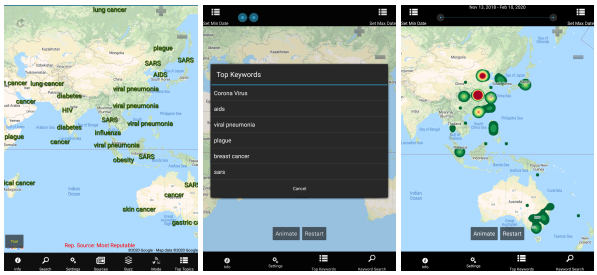
ACM Reference Format:
John H. Kastner and Hanan Samet. 2020. Visualizing SpatioTemporal Keyword Trends in Online News Articles. In *SIGSPATIAL '20: ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 03–06, 2020, Seattle, WA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Internet news outlets have become extremely common in the modern world. Due to their digital nature, they are suited for automated analyses in a way print media can not approach. Making use of this data, however, is not an easy task. This paper aims to aid users in such analysis by building a spatiotemporal data visualization application based on the existing NewsStand[12, 21] system. NewsStand is a well suited basis for this application because it is a spatiotextual 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 the keywords. The temporal information required for this application is obtained from the date recorded when a record is inserted into the database.

SIGSPATIAL '20, November 03–06, 2020, Seattle, WA

ACM ISBN 78-1-4503-6909-1/19/11…$15.00
https://doi.org/10.1145/1122445.1122456



(a) (b) (c)

Figure 1: Keyword and heatmap interfaces. (a) Keyword map, (b) Keyword selection, and (c) Heatmap visualization.

Prior extensions to NewsStand have taken advantage of these abilities for various flavors of keyword detection. In particular, work has been done to identify prominent mentions of brands [1], diseases [8, 9] (also see HealthMap [6]), and crimes [24] by news articles in the NewsStand database. NewsStand has also been ported to mobile devices [20, 23].

The spatiotemporal visualization we present takes advantage of prior work that developed techniques for identifying keywords in NewsStand's database. Keywords found with these methods are paired with spatial data obtained by geocoding the same news articles. A keyword and its location then serve as a data point that can be mapped. In our application, we plot these points on a heatmap, as heatmaps have been established as an effective mechanism for visualising large amounts of geographic information [5]. When temporal information is present in addition to geographic, we can restrict the data plotted on the map to only that which falls within a specific interval of time. By incrementally moving the window forward in time and rendering the resulting heatmaps in order, we can obtain a time series animation showing the change over time of the geographic extent of the use of a keyword. Examining these changes gives users insight into spatiotemporal trends in the use keywords. Screenshots demonstrating our application are shown in Figure 1.

In the rest of this paper we first briefly review related work pertaining to keyword extraction in Section 2 before describing our visualization, the details of its implementation, and our process for obtaining temporally and spatially referenced keywords from news articles in Section 3.1 Section 4 describes our user interface before concluding in Section 5.

2 RELATED WORK

The NewsStand system has previously been used to extract and display information about the prevalence of keywords in online news sources. The original implementation [21] was able to extract keywords from articles based on the TF-IDF scores of words computed for document clustering. Words with high TF-IDF scores occur more frequently in an article relative to their frequency in the entire corpus. As such, these words are generally important to the content of an article and are a good approximation of the article's keywords.

Lan et al. [8, 9] used NewsStand to visualize the progress of potential outbreaks of a disease as measured by the spatial extent of document clusters that mention it. While this work incorporates spatiotemporal analysis into NewsStand, it is limited to tracking disease-related news and only tracks the growth and geographic distribution of a single cluster of documents in relation to a disease. In contrast, our application can track the spread of arbitrary keywords or topics, including diseases, across all documents in the NewsStand database. This extra capability makes the application more applicable than prior work.

Abdelrazek et al. [1] implemented the ability to detect prominent mentions of different brands and companies within news articles. To accomplish this, the authors created two rule based classifiers and trained one supervised machine learning classifier. The goal of all three classifiers was to decide if a brand is mentioned prominently in an article given the name of the brand and the local context in which the brand is mentioned. These classifiers are used as a basis for the brand specific heatmap layer in this paper.

Wajid and Samet [24] developed a classification technique for determining if a news article discusses criminal activity and extracting keywords from the article that are directly related to the crime discussed. To accomplish this the authors first selected articles that contained one of a set of predetermined keywords that are likely indicative of articles about criminal activity (e.g., murder or theft) before classifying each of these article as either primarily about crime or not about crime using a support vector machine (SVM) classifier. This approach is very similar to that of Abdelrazek et al. [1].

In addition to keywords extraction techniques focused on NewsStand, there is large amount of research into keyword extraction from arbitrary documents. Onan et al. [16] provides an overview of some such algorithms.

3 METHODOLOGY

In this section, we first describe the system that is used to extract and geocode keywords from online news sources. Next we describe the heatmap visualization used to display the spatiotemporal news keyword data. Finally, we provide details of the implementation of the visualization. An overview of the described system is given in Figure 2.

3.1 Geocoding News Keywords

Two steps are required to obtain geocoded news keywords. Important keywords must be extracted from news articles, and the news article must be geocoded by identifying and resolving toponyms to specific geographic coordinates. Once we have a set of keywords and a set of locations for an article, we assign to each keyword every location found in the article.

3.1.1 Keywords. Extraction of news keywords is handled primarily by NewsStand and its above extensions [1, 8, 9, 21, 24]. We give a brief overview but consult the original papers for more details.

In the general case, NewsStand identifies keywords based on their TF-IDF scores. Words with high TF-IDF scores are words that appear much more frequently in a given document than in other documents in the database. Such words are likely to be important to the document in which they appear. The word with the highest TF-IDF score for a document is therefore the best keyword, and more keywords can be obtained by selecting words with progressively lower scores.

Keywords in relation to specific topics (e.g., disease, crime, and brands) are identified by the aforementioned extensions to NewsStand. While the implementations for identifying keywords for different topics vary, the principle used is roughly the same. First, words that fall within the scope of the topic must be found in the news article. Such words can be found either by consulting a dictionary of words that are known to be related to the topic, or by training a classifier that takes a word and optionally its context as input and decides if the word is relevant to the topic. This classifier can be obtained using either traditional machine learning techniques or more modern deep learning models. In either case, a dictionary of words related to the topic is used: for consulting directly or for training a classifier. After a word has been identified as related to the topic in question, the second step is to decide if the word is important enough in the news article to be considered a keyword. This can be done using TF-IDF scores as when extracting generic keywords, using raw term frequency, or by training a second classifier to decide if a specific instance of the word is important to the article.

3.1.2 Geocoding. Geocoding is the process of associating concrete geographic information (i.e. latitude longitude coordinate pairs) with a piece of text [10, 11, 13, 18, 19]. Geocoding can be framed as a specific variant of the topic specific keyword extraction task in the sense that you must first find words that are likely to refer to geographic locations and then decide which of these locations are important enough to associate with the documents. The third step in geocoding, which is not required for general keyword extraction, is toponym resolution [11] where a toponym must be assigned to a single latitude longitude pair. This is nontrivial as there are many ambiguous toponyms that can be used to refer to multiple distinct locations (e.g., Paris and London).

3.2 Heatmap Visualization

A geographic heatmap is a cartographic method which assigns individual pixels on the map a color based on the estimated density of data points at their location [15]. We use a heatmap to visualize the results of keyword extraction and geocoding as they have been shown to be effective tools to visualize large amounts of spatial data [5].

In our heatmap, we assign red to areas with the highest density of points, yellow to middle densities, and green to areas with low but non zero densities. Areas without any points are left transparent

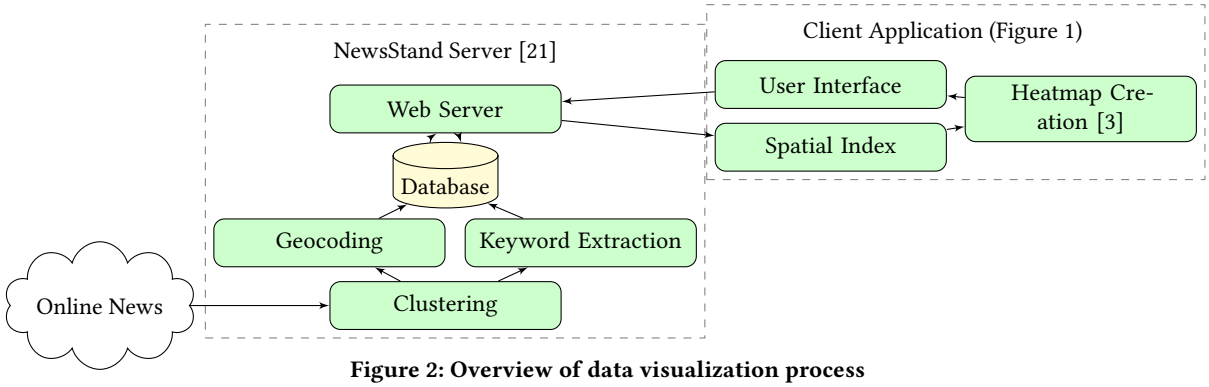


Figure 2: Overview of data visualization process

(a) (b) (c)

Figure 3: Time series of heatmaps for "Corona Virus" on (a) 1/1/20-1/12/20, (b) 1/13/20-1/5/20, and (c) 1/25/20-2/10/20.

on the heatmap layer so that the background map remains visible. To create the images for our heatmaps, we used the `libheatmap` library for the C programming language [3].

Three output heatmaps from our application are shown in Figure 3. Each heatmap shows the result of a query for the disease keyword "corona virus" for a different time period. Note that the geographic extent covered by the colored region in the heatmap and the relative intensity of the different regions change over time. This change corresponds to a change in how much the topic is being discussed and what locations are discussed in the same articles as the topic. When viewed in our application, these three heatmaps, and many more intermediate heatmaps, are rendered as a smooth animation to help users understand this change over time.

3.3 Application Implementation

In order to efficiently handle the large amount of data in the NewsStand database, we needed to apply some optimizations to the construction of our heatmaps. Initially, we attempted to use the heatmap implementation provided by Google in the Android Google Maps package Unfortunately, this implementation was not performing sufficiently well for our application. This led us to use the `libheatmap` library, which was considerably more successful.

When using a map, users often do not want to view the entire world at once. Instead, they can use the pan and zoom functionalities of a digital map to select only the portion of the world they care about. When focused on such a small region, it would be wasteful to compute a heatmap for the entire world only to crop it to a small part of its original size.

We avoid such a scenario by only adding points to the heatmap that are contained in the region of the world that the user is examining. To efficiently select only these points from a very large collection of points, we construct a PR-Quadtree from the data points when they are first received from the server. This index permits efficient retrieval of points within a rectangular region. Since points are often sparse in some areas of the world, this can greatly reduce time to construct heatmaps outside the densest regions.

4 APPLICATION INTERFACE

The user interface for our application consists of two primary components: first, there is a map interface using Google Map tiles as the background layer with our custom heatmap overlay in the foreground; the second component is a collection of interfaces thought which the user is able to indicate the keyword trends they want to investigate. Screenshots displaying these components are shown in Figure 1.

The map interface (Figure 1c) is the primary interface through which a user will interact with our application. The central feature is, of course, the world map with a heatmap overlay. Another feature of note is the slider at the top which allows the user to manually select a period of time to view. There is also a pair of buttons labeled "Animate" and "Reset" overlaid on the bottom of the map. These buttons let the user start an animation sequence for the currently selected keyword and to reset the animation to its default state. Finally, the three buttons below the map give the user access to the keyword selection interfaces.

The keyword selection interfaces (Figure 1b) are used to select what keyword will be used when querying the server. From the map interface, pressing the middle button labeled "Top Keywords" opens an interface that shows a list of keywords for the current keyword mode found in recent news articles. Selecting one of the keywords updates the heatmap to reflect the new selection. In case the desired keyword is not present in the list of top keywords, then the rightmost button labeled "Keyword Search" can be used. This button opens a search window where the user is prompted to type any search term they like. Note that data will not necessarily be available for arbitrary queries. This term will then be used to query the database and construct a heatmap. Finally, the leftmost button labeled "Settings" allows the user to specialize their queries to different domains. Rather than doing a generic keyword query, the user can specialize their query to only search for people, diseases,

brands, or crimes. This specialization takes advantage of the topic specific keyword detection mentioned in Section 3.1.

5 CONCLUSION AND FUTURE WORK

In this paper we presented our application for visualizing spatiotemporal information in online news articles using keyword detection, geocoding, and geographic heatmaps.

To expand this work we plan to incorporate information gathered from sentiment analysis into our visualization. This would allow users to not only see *that* a topic is being talked about but also *how* it is being talked about. However, in order to do this, we will need to develop a technique for assigning sentiments to specific entities rather than to the article as a whole. This is because it is likely that articles will mention one topic in a positive context while also talking about a different topic in a negative context.

While our system is designed to generalize to visualizing trends in arbitrary keywords, it could be specialized to monitor a health crisis such as the COVID-19 pandemic. This could be done with minimal changes to the user interface. Instead, the underlying database could be extended to include data gathered from a variety of sources that are potentially useful for tracking the pandemic. Other work has considered using microblogs such as Twitter for this purpose [7, 14, 17].

We want to improve the performance of heatmap generation. Our application currently runs on a mobile device, so it is restricted in computational power. This means computing a heatmap for each frame in an animation takes a significant amount of time. We mitigate this by storing points in a spatial index and only constructing a heatmap for a limited area but, we are likely obtain larger speedups by exploring quadtree [4, 25] or GPU [26] heatmap construction algorithms. Future work involves incorporation of tabular data [2] as well a merging spatially-adjacent heatmaps via techniques such as connected component labeling [22].

ACKNOWLEDGMENTS

This work was sponsored in part by the NSF under Grants IIS-18-16889 and IIS-20-41415.

REFERENCES

[1] A. Abdelrazek, E. Hand, and H. Samet. 2015. Brands in NewsStand: Spatio-temporal browsing of business news. In *Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, M. Ali, M. Gertz, Y. Huang, M. Renz, and J. Sankaranarayanan (Eds.). Seattle, WA. Article 97.

[2] M. D. Adelfio and H. Samet. 2013. Schema extraction for tabular data on the web. *PVLDB* 6, 6 (2013), 421–432.

[3] L. Beyer. 2017. *lucasb-eyer/libheatmap*. https://github.com/lucasb-eyer/libheatmap

[4] Q. Cai and Y. Zhou. 2016. A quadtree-based hierarchical clustering method for visualizing large point dataset. In *2016 Sixth International Conference on Information Science and Technology (ICIST)*. 372–375.

[5] J. Delort. 2010. Vizualizing Large Spatial Datasets in Interactive Maps. In *International Conference on Advanced Geographic Information Systems, Applications, and Services*. 33–38.

[6] C. Freifeld, K. Mandl, B. Reis, and J. Brownstein. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association* 15, 2 (2008), 150–157.

[7] N. Gramsky and H. Samet. [n.d.]. Seeder finder - identifying additional needles in the Twitter haystack.

[8] R. Lan, M. D. Adelfio, and H. Samet. 2014. Spatio-temporal disease tracking using news articles. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health (HealthGIS 2014)*. Dallas, TX, 31–38.

[9] R. Lan, M. D. Lieberman, and H. Samet. 2012. The picture of health: map-based, collaborative spatio-temporal disease tracking. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health (HealthGIS 2012)*. Redondo Beach, CA, 27–35.

[10] M. D. Lieberman and H. Samet. 2011. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th International Conference on Research and Development in Information Retrieval (SIGIR'11)*. Beijing, China, 843–852.

[11] M. D. Lieberman and H. Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval (SIGIR'12)*. Portland, OR, 731–740.

[12] M. D. Lieberman and H. Samet. 2012. Supporting rapid processing and interactive map-based exploration of streaming news. In *Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, I. Cruz, C. A. Knoblock, P. Kröger, E. Tanin, and P. Widmayer (Eds.). Redondo Beach, CA, 179–188.

[13] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. 2010. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *Proceedings of 6th Workshop on Geographic Information Retrieval*, R. Purves, C. Jones, and P. Clough (Eds.). Zurich, Switzerland. Article 6.

[14] A. Magdy. 2020. Microblogs: a renewable spatio-temporal fortune. *SIGSPATIAL Special* 12, 1 (2020), 41–52.

[15] R. Netek, T. Pour, and R. Slezakova. 2018. Implementation of Heat Maps in Geographical Information System – Exploratory Study on Traffic Accident Data. *Open Geosciences* 10, 1 (2018), 367–384.

[16] A. Onan, S. Korukoğlu, and H. Bulut. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications* 57 (2016), 232–247.

[17] U. Qazi, M. Imran, and F. Ofli. 2020. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special* 12, 1 (2020), 6–15.

[18] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. 2010. Determining the spatial reader scopes of news sources using Local Lexicons. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, A. El Abbadi, D. Agrawal, M. Mokbel, and P. Zhang (Eds.). San Jose, CA, 43–52.

[19] H. Samet. 2014. Using minimaps to enable toponym resolution with an effective 100% rate of recall. In *Proceedings of 8th ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR'14)*, R. Purves and C. Jones (Eds.). Dallas, TX, 9:1–9:8.

[20] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. 2011. Porting a web-based mapping application to a smartphone app. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, D. Agrawal, I. Cruz, C. S. Jensen, E. Ofek, and E. Tanin (Eds.). Chicago, 525–528.

[21] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. 2014. Reading news with maps by exploiting spatial synonyms. *Commun. ACM* 57, 10 (Oct. 2014), 64–77.

[22] H. Samet and M. Tamminen. 1986. An improved approach to connected component labeling of images. In *Proceedings of Computer Vision and Pattern Recognition'86*. Miami Beach, FL, 312–318.

[23] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. 2011. Adapting a map query interface for a gesturing touch screen interface. In *Proceedings of the Twentieth International Word Wide Web Conference (Companion Volume)*, S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar (Eds.). Hyderabad, India, 257–260.

[24] F. Wajid and H. Samet. 2016. CrimeStand: Spatial tracking of criminal activity. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, M. Ali, S. Newsam, S. Ravada, M. Renz, and G. Trajcevski (Eds.). Burlingame, CA. Article 81.

[25] K. Yuan, X. Cheng, Z. Gui, F. Li, and H. Wu. 2019. A quad-tree-based fast and adaptive Kernel Density Estimation algorithm for heat-map generation. *IJGIS* 33, 12 (2019), 2455–2476.

[26] G. Zhang, A. Zhu, and Q. Huang. 2017. A GPU-accelerated adaptive kernel density estimation approach for efficient point pattern analysis on spatial big data. *IJGIS* 31, 10 (2017), 2068–2097.