Spatio-Temporal Disease Tracking Using News Articles*

Rongjian Lan† Marco D. Adelfio Hanan Samet
Center for Automation Research, Institute for Advanced Computer Studies,
Department of Computer Science, University of Maryland
College Park, MD 20742
{rjlan, marco, hjs}@cs.umd.edu

ABSTRACT

Geographical Information Systems have been increasingly used to aid the prompt detection, tracking, and analysis of disease outbreaks. Web content which is full of health-related data also serves as a useful resource for disease outbreak analysis. News posts often report the initial outbreak of diseases and contain valuable information that aids in ascertaining the time and location of the disease outbreak. The locations mentioned in the news posts are specified textually rather than geometrically thereby requiring the use of geotagging methods to detect them and to map the textual specification to the corresponding actual geometric specification. The NewsStand system which aggregates news posts by topic and location while providing a map query interface to them is enhanced to enable disease tracking and analysis by geotagging disease-related web news posts. Besides the powerful functionalities of NewsStand for news exploration, enhancements of NewsStand with respect to the analysis of temporal information are described which include a well-designed time slider, a heatmap-based visualization tool for displaying disease distribution, and intuitive spatio-temporal querying methods. Future improvements to NewsStand are also discussed.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Performance

Keywords

Disease tracking, geotagging, GIS, spatio-temporal

*This work was supported in part by the National Science Foundation under Grants IIS-10-18475, IIS-12-19023, and IIS-13-20791.

†Currently at Google

HealthGIS'14, November 04-07 2014, Dallas/Fort Worth, TX, USA

 978-1-4503-3136-4/14/11...$15.00
http://dx.doi.org/10.1145/2676629.2676637

1. INTRODUCTION

Post disease analysis is increasingly becoming an important part of disease control and prevention in our globally connected world. Organizations such as the US Centers for Disease Control and Prevention and the World Health Organization are investing more money and recruiting more experts on monitoring and analyzing infectious diseases around the globe partially through the application of techniques from big data and geographic information systems. In addition, the increasing prevalence of volunteered geographic information (VGI) [14] in social media such as blogs and tweets can likewise be leveraged to effect greater tracking ability. In fact, disease outbreaks have a strong geographical nature especially as they spread and become epidemics. This is particularly true in the case of environmental conditions and cultural behavioral characteristics which can benefit from the development and deployment of robust health-oriented geographic information systems (GIS). A number of Web-based services for disease tracking have been developed including *ProMED-mail* [16], henceforth referred to as *ProMED*, which is an online alert system intended to quickly disseminate news of the latest outbreaks to medical professionals, as well as to laymen, throughout the world.

Our approach is different from these existing systems in that we focus on the spatial-temporal aspect of the disease outbreaks. We have designed a system to retrieve geographic information from the raw text in web posts that correspond to the output of RSS news feeds, something which is not well supported in existing systems. We achieve this through a process termed *geotagging*. This means that we must recognize terms that correspond to names of locations (termed *toponyms*) in the raw text (termed *toponym recognition*) [24], as well as disambiguate between different interpretations (termed *toponym resolution*) [23]. The result is that, with the aid of a gazetteer, each toponym is assigned its correct spatial interpretation in the form of lat/long values. Toponym recognition is challenging because toponyms are dual use in that they often also serve as the names of people and organizations, and hence we need to use natural language clues. At times, solving both the toponym recognition and resolution problems requires external knowledge such as making use what we term *local lexicons* [27, 34] which help us disambiguate between constructs such as "Paris Hilton"' which can be both a person and a location. Toponym resolution is hard because many locations have the same name (e.g., any of over 60 cities around the world named "Paris") [20, 22, 23, 26, 33]. Such problems are

common in news posts where writers uses names without adequate qualifications. Nevertheless, we are generally capable of handling the ambiguity by using such information as the serving scope of the publication in which the ambiguous reference occurs. Assuming that we can detect all references to geographic locations (e.g., 100% recall for toponym recognition), we can achieve effective 100% recall for the disambiguation step by providing the users all the possible different locations that have the same name provided that some news post is associated with them [36].

In previous work we developed the STEWARD [30] system for geotagging and retrieval of documents from the hidden web and used it for tracking infectious diseases [29]. We later extended it to allow spatio-temporal querying through the use of a time slider [19] for ProMED disease reports. In this paper we turn our attention to news articles. We embed our system in the NewsStand [25, 38, 40, 41, 46] built by our group that monitors RSS feeds from thousands of online news sources and retrieves articles within minutes of publication. It then extracts geographic content from articles using a custom-built geotagger, and groups articles into story clusters using a fast online clustering algorithm.

Newsstand's map-based Web interface also allows for powerful spatio-textual retrieval and display. In particular, it can be used for feature-based queries [7] (e.g., spatial data mining) as well as location-based queries (see also the related QUILT system [39, 44] and the SAND Browser [11, 37]).

In this paper, we introduce the innovative feature of time slider to NewsStand and show how it is integrated into the existing system to handle disease tracking based on information in news postings. This slider allows users to intuitively and easily query news documents based on both time and locations and supports an auto-play function to display the evolution of the news coverage and intensity. These changes in the news coverage across time serves as key evidence for the occurrence of diseases. After a careful exploration of visualization techniques, we adopt the Heatmap technique. We feel that it is the best visualization technique for our need to visualize huge amounts of location-based information which is the most intuitive one for users to perceive quickly and avoid the problem of marker overlaps. With all of these improvements, the NewsStand system is able to help in the analysis of tracking of disease-related news articles in a quick and easy fashion.

The rest of this paper is organized as follows. Section 2 discusses related research work. Section 3.1 introduces our approach for the processing of news postings where we also discuss the geotagging power of NewsStand. Section 4 describes our user interface, including the minimap, the time slider along with its implementation details, and the heatmap visualization. Section 5 contains concluding remarks.

2. RELATED WORK

Disease tracking and monitoring is one of many applications of spatio-temporal visualization. Prior work includes many efforts to generate visualizations of generic spatio-temporal documents, including research on spatio-temporal document profiles described by Strötgen et al. [45]. There, the authors describe a method of extracting event/location pairs from arbitrary text documents, and then provide a map query interface for visualizing the event flow on a map. For example, their system can take a Wikipedia article about an explorer's travels as input and generate a map visualization with a trajectory diagram connecting extracted events in sequence. Another approach is used in the Visits system of Thudt et al. [47], which visualizes paths using a hybrid map/timeline that shows multiple small maps for additional details in areas with many data points. Such techniques works well for tracking explorer paths or personal location histories, but are less appropriate for disease report data where connections between sequential reports is not as important as their spatial distribution over time (for more details on visualizing itineraries, see [1, 3]).

In contrast to visualization systems that attempt to flatten the time dimension and show the events in a static view, our system aims to emphasize the temporal dimension and distribution of data by utilizing an animated heatmap. Heatmaps are common tools for visualizing the spatial distribution of various datasets, with or without a temporal component. An example of a system that uses heatmaps for spatio-temporal data is that of Maciejewski et al. [31], which focuses on the identification of hotspots where many events are clustered. The interface allows selecting data from individual days or constrained geographic regions upon which to generate a heatmap. The variable nature of disease reports and their rates of spread, results in our interface deviating from this model by allowing the specification of a variable-sized temporal window along with an animated mode to emphasize the temporal dimension. Other work in this area by Schulze-Wollgast et al. [43] works to visualize disease spread within Germany using choropleth maps and small multiples of stream graphs. Still another technique for visually indicating the density of events on a map is used in the GeoTemCo system of Janicke et al. [18], which clusters proximate event markers into non-overlapping, proportionally sized marker bubbles to avoid visual clutter. Another system in this domain is HealthMap [10, 13], which gleans information from ProMED [16], GeoSentinel[1], Eurosurveillance[2], Google News[3], Moreover[4], and many other sources.

Additionally, for mapping volunteered geographic information, which includes textual specifications of locations, a geotagging module is required. The NewsStand system makes use of a robust geotagging framework (described in Section 3.1), which extends a rich body of geotagging research. The Web-a-Where system of Amitay et al. [6] uses a variety of context features to interpret geographic references in documents and demonstrates the need for incorporating such features in order to geotag text accurately. Unlike many health mapping systems that rely on machine-specified location information or straightforward geotagging procedures, NewsStand incorporates many context features, including the presence of proximity, sibling, and prominence clues [2, 5, 27], to increase the accuracy of the extracted location references. News Rover [21] is another system, similar to NewsStand, that does aggregation of news videos, rather

[1] `http://geosentinel.org/`
[2] `http://www.eurosurveillance.org/`
[3] `http://news.google.com/`
[4] `http://moreover.com/`

than articles, for exploration and querying.

3. NEWSSTAND DATA PIPELINE

In this section, the process of geotagging the news articles and grouping articles into news clusters are discussed in greater detail.

3.1 Geotagging

After a new article has been introduced to the system, NewsStand must locate and extract the geographic content from the article. This process, described earlier as geotagging, unifies the explicit textual article content with the implicit geography, and enables spatial exploration of the news. NewsStand's geotagging module includes four stages, each of which is a member of the general pipeline.

The first stage of processing deals with extracting the interesting phrases that are most likely to be references to geographic locations and other entities, given the surrounding context. These phrases are collectively called the article's entity feature vector (EFV). We use a statistical Natural Language Processing (NLP) method for Named-Entity Recognition (NER) [50] tagging. NER's goal is to identify phrases from the article that correspond to various entity classes, such as PERSON, ORGANIZATION, and LOCATION. Those phrases tagged as LOCATION are most likely to be locations.

In the second stage of gazetteer record assignment, NewsStand uses a gazetteer, or database of geographic locations, to find those geographic features in the entity feature vector that are names of actual locations. NewsStand uses a gazetteer based on the GeoNames database [49], which is a comprehensive collection of geographic data from over 100 sources, including the GEOnet Names Server (GNS) [32] and Geographic Names Information System (GNIS) [48]. NewsStand's gazetteer contains over 6.5 million geographic locations, and over 8 million location names from around the world (including duplicate names for the same location). The gazetteer contains the latitude and longitude for each record, as well as additional information useful for geotagging, such as alternate names in various languages. The population is also stored for records corresponding to populated places or regions. The gazetteer also stores hierarchical information for each location, including the country and administrative subdivisions that contain the location.

The next and third stage is geographic name disambiguation, where multiple heuristic filters attempt to resolve ambiguous references by selecting the most likely set of assignments for each reference, based on how a human would read the article. These filters rely on our initial assumption that the locations mentioned in the article give evidence to each other, in terms of geographic distance, document distance, and hierarchical containment.

In the fourth and last stage, the geotagger distinguishes between those georeferences that are important to the article, and those that are mentioned only tangentially, by ranking the georeferences by relevance to the article's geographic focus. One basic measure of relevance is the frequency of occurrence throughout the body text. In addition, we found that in a typical news article with a strong geographic focus, important georeferences are mentioned early in the text or in the title. We therefore settled on a weighted frequency ranking that tries to balance these two factors by computing a linearly decreasing weighting of the georeference frequency, with occurrences of a georeference g closer to the beginning of the article giving more weight to g's ranking.

3.2 Online Clustering

We use a clustering algorithm based on vector space model of documents [35] to group together all news articles that describe the same news event into groups of articles named news clusters. A news event is defined in terms of both story content and story lifetime. Articles in the same cluster should share much of the same important keywords, and should have temporally proximate dates of publication. Time is an essential part of grouping news articles, since two articles may contain similar keywords but describe vastly different news events. For example, two stories about separate attempted assassinations in Iraq may share many keywords, but should be placed in separate clusters if one story was breaking news and the other was several days old.

After the news articles are grouped into clusters, we compute the geographic focus for each cluster based on the geotagged information from each documents. This geographic focus will be used in the web user interface display. We choose the geographic focus among all the mentioned locations in the articles by finding the mostly mentioned locations, except those locations that are inconsistently mentioned. We detect the inconsistently mentioned locations using a weighted voting approach.Suppose a news story about College Park in Maryland contains articles mentioning "College Park, MD", with College Park placed in Maryland, and other articles mentioning just "College Park", but placed in Georgia. Because the first set of articles contains qualified "College Park" entities, they cast stronger votes for placing College Park in Maryland, and aggregating votes will likewise place College Park in Maryland. The Georgia interpretation of College Park is thus removed as a candidate for the cluster focus. Once we have resolved inconsistencies in entity interpretations, we compute the cluster focus of a cluster C by collecting the most frequently mentioned locations in articles in C.

4. WEB INTERFACE

In this section we describe our user interface, including the minimap and the time slider as well as the implementation details of the interactive temporal querying capability, and the design principles behind the heatmap visualization. The system can be sampled by going to `http://newsstand.umiacs.umd.edu` and running it in "Time Mode".

4.1 Visual Elements

1 shows the main user interface of NewsStand, which consists of a map visualization with markers representing news articles, and various filters and controls to help users find the information that they seek. Each marker in the map represents a collection of articles that contain the keyword and which were published as news and associated with the marker's location. Filters are provided on the top of the map for users to filter news articles based on their categories, including general, sports, entertainment, health, business,

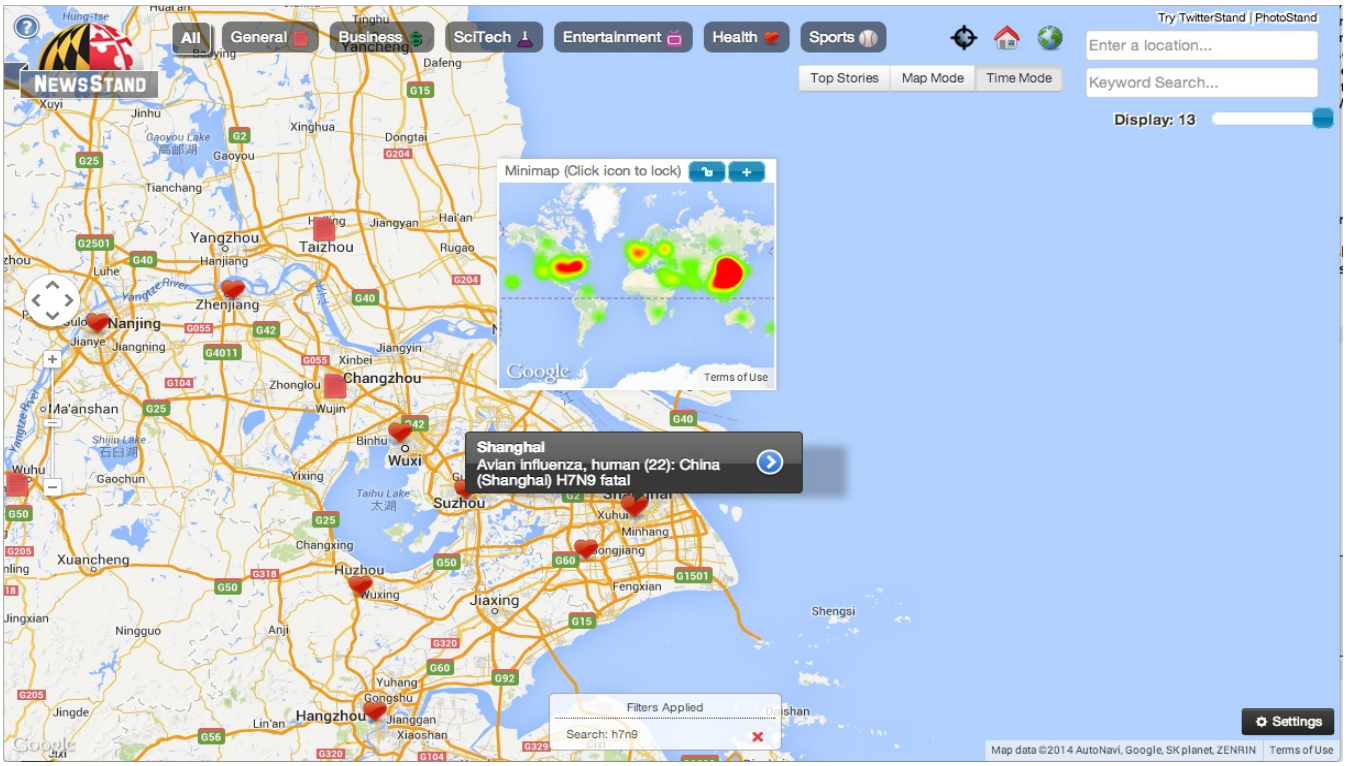

Figure 1: NewsStand's map-based Web interface in Time mode when querying the term "H7N9" and hovering over Shanghai. A minimap contains ball symbols at all the locations mentioned in all of the related news articles over a given period of time that are associated with Shanghai as a result of geotagging the articles.

and science. On the top right are query input boxes where users can specify the keyword or location constraints on the news articles in which they are interested. Around the same area, there are three buttons for switching between different modes of NewsStand. The Top Stories mode allows users to easily view the hottest news articles and is analogous to feature-based queries or spatial data mining [7] and Map mode, the normal and default browsing mode, which allows users to browse the news by location and is analogous to location-based queries. The third mode, Time mode and illustrated in the figure, enables the temporal querying functionalities and returns news articles from the entire news database involving a specific keyword over a user-specified time period.

In Time mode, each marker represents the entire history of published news articles that refer to the specific location and also contain the keyword in the text of the news articles. To help users explore these news articles, they can hover over the marker to bring up an info bubble containing the headline of the most popular or important news article over time that mentioned the keyword and location. Adjacent to the headline info bubble is a minimap with a heatmap visualization showing all the locations mentioned in these articles. We chose the heatmap visualization technique due to its intuitiveness and its ability to facilitate the visualization of large amounts of data points without having to worry about overlap. In this example, the heatmap is color coded into a green-to-red gradient with green representing a smaller amount of data and red representing a large amount of data. Here we can see that overall the news at Shanghai mentioned itself quite often as well as Beijing. To delve further and examine each specific news article, users can click on the arrow button on the extreme right side of the headline info bubble to bring up a list of news articles. The list of articles can be sorted by time of publication as well as by the relevance of the content to the query. This is similar to how Google ranks their search results. Clicking on one of the articles in the list causes the display of the detailed content of that article where NewsStand highlights the matched keywords and location terms so that users can easily verify the relevance of the search result to the query.

Besides the above features for exploring news articles over a period of time, the enhanced features enable users to perform temporal querying over the data. In this case, users simply click the Plus button on the top-right of the minimap window, which cause the minimap to be enlarged to a more appropriate size so that more details can be revealed while the ability to drag the minimap over the main map ensures that the features of the main NewsStand user interface are not completely blocked. The top of the enlarged minimap contains a time slider that enables users to easily specify the temporal component of their query. Users move the slider along the timeline to control the temporal range for which the query results are displayed in the minimap window. Users can drag either of the end points of the slider to adjust the temporal range, which immediately changes the heatmap display in accordance with the new temporal range. For example, in the first figure of 2, the time slider

is focused on the temporal range of March, 2012 to June, 2012, with the heatmap only representing the news within that range. The time slider consists of two components: the top header area and the slider area. The top header area of the time slider contains a number of controls, including a button to adjust the play speed of the auto-play function, a label to show the end points of the time range in greater detail, and a label to show the ratio between the displayed data points and the overall data points. The slider area is the conventional time slider with an auto-play function that automatically moves forward the slider's temporal range to simulate the effect of the lapse of time. Forward and back buttons are also provided for precise control of the slider's temporal range. The functionality of the time slider is achieved by customizing the public API of ArcGIS [12].

The above features enable an intuitive and compelling display of the data's evolution over time, which allows the discovery of temporal trends and outliers. 2 illustrates how these features help in discovering the outbreak and trends of a recent disease, the H7N9 avian flu in Shanghai, China. Recall that in March, 2013, the news about outbreaks of the H7N9 flu around Shanghai burst explosively as the first case of such disease was discovered and the subsequent spread of it. Media coverage spread all over the world with the concentration on the Shanghai and Beijing areas. Later on in the second half of the year, the news about H7N9 was quietened as the disease was being overcome. However, another small burst of media coverage around Beijing and Shanghai marks the second wave of the avian flu in early 2014. Although these examples are simplistic in scope, they serve as a good illustration of the power of temporal querying coupled with NewsStand's geotagging capabilities.

4.2 Temporal Querying

In this section, we provide more details of the implementation of temporal querying for our user interface. Temporal querying functionalities are fully developed within the client side of NewsStand so as to decouple the front and back ends in the updates of temporal queries so as to achieve real-time query adjustment. The temporal querying at a specific location happens when the users hover over the marker in the time mode. The specific keyword and location are passed to the backend of NewsStand where the execution of the database query is conducted. Since our existing database contains several years' worth of news data, such querying over the whole database with filtering is not trivial. To speed up the execution of such a query (i.e., to make it tolerable), we created indexes on the words in the content of the news articles. After the queried data is returned to the front end client, the original minimap is populated with the heatmap visualization of all the data, thereby providing the users an overview of the history and distribution of it all over the world. After bringing up the enlarged minimap with the time slider, users specify the temporal parameters by adjusting the controls of the time slider as described in Section 4.1, thereby controlling which news articles are involved in the heatmap visualization. In other words, the temporal querying component of the front end functions as a post-processing filter of the data.

The fast adjustment of the temporal component of the query depends on the efficient implementation of the time slider

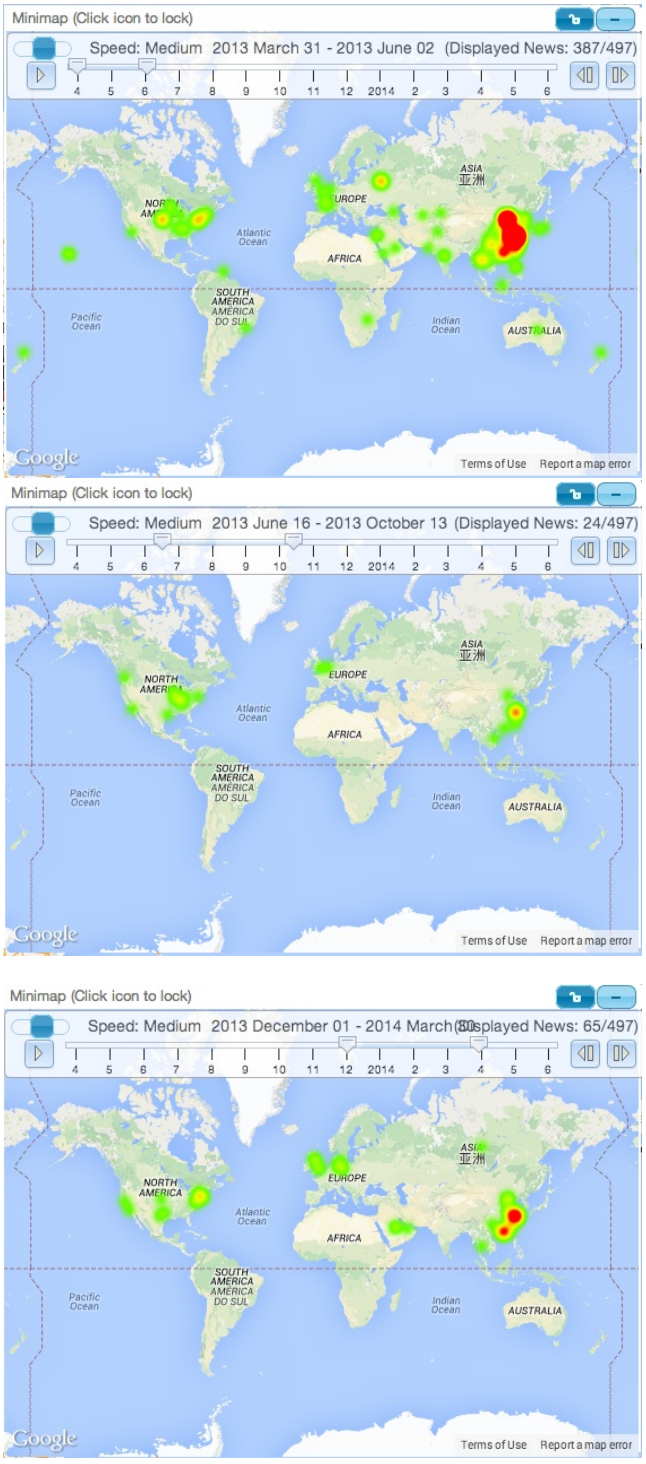

Figure 2: The outbreak of the recurrence of H7N9 avian flu around Beijing and Shanghai China. The visualized data corresponds to geotagged news articles.

function. We managed to solve this problem by first sorting the returned articles based on their time, which is very fast in our use case of hundreds to thousands of articles. With the sorted list of articles, the subsequent querying reduces to binary search of articles over time, which is very efficient. The use of Google Map's visualization API of heatmap also made updating of the visualization very fast.

5. CONCLUDING REMARKS

NewsStand's geotagging capability over news articles along with the enhanced time slider functionality enable the fast and intuitive exploration of disease outbreaks. The news articles data are geotagged by NewsStand data to reveal their geographical location while the time slider enable the spatial temporal querying of the data. The heatmap visualization is used in the display of a large amount of location data over the world which overcomes the problem of overlapping markers and reduces the burden of cognitive perception of the user because of its intuitive color coding. The flexible temporal querying enabled by the time slider and its auto-play function along with the fast implementation of the updates of the visualization also make the analysis of the disease outbreak easy and fast.

However, the current prototype is still in its infancy and requires further customization and improvement. The current database for the data storage is based on a PostgreSQL database, which is not optimized for keyword matching and temporal querying on the data. Alternative databases such as Solr NoSql database would be more suited for use with NewsStand because the database is used mainly for storage and querying purposes instead of the relational querying heavily used in an SQL database. On the front end side, the current temporal querying only works on the data history for the articles related to a specific location, limiting the presentation of the disease trend across the world. With an enhanced backend database, the presentation of the overall situation of the disease outbreak around the world can be achieved, giving the users a higher level understanding of the history of the disease outbreak.

The same ideas can also be adapted to tweets that mention diseases. In this case we would adapt the TwitterStand system [42] which is an extension of NewsStand for tweets. The key difficulty here is that the tweets are limited to 140 characters and thus neither the location information or the disease information is guaranteed to be present in the tweets. This is, in part, because not all tweets have a GPS capability. Moreover, we want to know the location being tweeted about rather than the location at which the tweet was generated. Thus we look for urls to news posts and derive the locations and diseases from their text. In this case, we must also pay attention to who are good and reliable tweeters [15]. It is also useful to detect whether the tweet refers to a past, ongoing, or future event [17]. Spreadsheets are another source of disease mentions accompanied by mentions of locations which could also be visualized in the same manner when the spreadsheet contains a temporal attribute. Of course, we may not have the schema available in which case techniques such as those in [4, 28] that infer the schema as well as WebTables [8, 9] by Google would be useful.

References

[1] M. D. Adelfio and H. Samet. Automated itinerary visualization. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Y. Huang, M. Gertz, J. C. Krumm, J. Sankaranarayanan, and M. Schneider, eds., Dallas, TX, November 2014.

[2] M. D. Adelfio and H. Samet. GeoWhiz: Using common categories for toponym resolution. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, C. A. Knoblock, P. Kröger, J. C. Krumm, M. Schneider, and P. Widmayer, eds., pages 542–545, Orlando, FL, November 2013.

[3] M. D. Adelfio and H. Samet. Itinerary retrieval: Travelers, like traveling salesmen, prefer efficient routes. In *Proceedings of 8th ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR'14)*, R. Purves and C. Jones, eds., Dallas, TX, November 2014.

[4] M. D. Adelfio and H. Samet. Schema extraction for tabular data on the web. *PVLDB*, 6(6):421–432, April 2013. Also *Proceedings of the 39th International Conference on Very Large Data Bases (VLDB)*.

[5] M. D. Adelfio and H. Samet. Structured toponym resolution using combined hierarchical place categories. In *Proceedings of 7th ACM SIGSPATIAL Workshop on Geographic Information Retrieval (GIR'13)*, R. Purves and C. Jones, eds., pages 49–56, Orlando, FL, November 2013.

[6] E. Amitay and N. Har'El and R. Sivan and A. Soffer. Web-a-Where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 273–280, Sheffield, UK, July 2004.

[7] W. G. Aref and H. Samet. Efficient processing of window queries in the pyramid data structure. In *PODS'90: Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 265–272, Nashville, TN, April 1990.

[8] M. J. Cafarella, A. Y. Halevy, and N. Khoussainova. Data integration for the relational web. In *PVLDB*, 2(1):1090–1101, August 2009. Also *Proceedings of the 35th International Conference on Very Large Data Bases (VLDB)*.

[9] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. WebTables: Exploring the power of tables on the web. In *PVLDB*, 1(1):538–549, August 2008. Also *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB)*.

[10] Children's Hospital Boston. HealthMap | Global Health, Local Knowledge, September 2012. URL `http://www.healthmap.org`.

[11] C. Esperança and H. Samet. Experience with SAND/Tcl: a scripting tool for spatial databases. *Journal of Visual Languages and Computing*, 13(2):229–255, April 2002.

[12] Esri. ArcGIS Online, Sept. 2012. URL `http://www.arcgis.com`.

[13] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *JAMIA: Journal of the American Medical Informatics Association*, 15(2):150–157, March 2008.

[14] M. F. Goodchild. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221, August 2007.

[15] N. Gramsky and H. Samet. Seeder finder - identifying additional needles in the Twitter haystack. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN'13)*, A. Pozdnukhov, ed., pages 44–53, Orlando, FL, November 2013.

[16] International Society for Infectious Diseases. ProMED-mail, Sept. 2012. URL `http://www.promedmail.org`.

[17] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN'11)*, Y. Zheng and M. F. Mokbel, eds., pages 25–32, Chicago, November 2011.

[18] S. Jänicke, C. Heine, and G. Scheuermann, Gerik GeoTemCo: Comparative visualization of geospatial-temporal data with clutter removal based on dynamic Delaunay triangulations. *Computer Vision, Imaging and Computer Graphics. Theory and Application Communications in Computer and Information Science*, 359:160-175, 2013.

[19] R. Lan, M. D. Lieberman, and H. Samet. 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)*, pages 27–35, Redondo Beach, CA, November 2012.

[20] J. L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, University of Edinburgh, Edinburgh, Scotland, UK, 2007.

[21] H. Li, B. Jou, J. G. Ellis, D. Morozoff, and S. F. Chang. News Rover: Exploring topical structures and serendipity in heterogeneous multimedia news. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 449–450, Barcelona, Spain, October 2013.

[22] H. Li, R. K. Srihari, C. Niu, and W. Li. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 39–44, Edmonton, Canada, May 2003.

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

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

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

[26] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. 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., article 6, Zurich, Switzerland, February 2010.

[27] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE'10: Proceedings of the 26th International Conference on Data Engineering*, pages 201–212, Long Beach, CA, March 2010.

[28] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Spatio-textual spreadsheets: Geotagging via spatial coherence. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, D. Agrawal, W. G. Aref, C.-T. Lu, M. F. Mokbel, P. Scheuermann, C. Shahabi, and O. Wolfson, eds., pages 524–527, Seattle, WA, November 2009.

[29] M. D. Lieberman, J. Sankaranarayanan, H. Samet, and J. Sperling. Augmenting spatio-textual search with an infectious disease ontology. In *Proceedings of the Workshop on Information Integration Methods, Architectures, and Systems (IIMAS08) (ICDE Workshops 2008)*, pages 266–269, Cancun, Mexico, April 2008.

[30] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD: architecture of a spatio-textual search engine. In *Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems*, H. Samet, M. Schneider, and C. Shahabi, eds., pages 186–193, Seattle, WA, November 2007.

[31] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. Grannis, D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2)205–220, March/April 2010.

[32] National Geospatial-Intelligence Agency. NGA: GNS Home, Sept. 2012. URL `http://earth-info.nga.mil`.

[33] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7):717–745, August 2007.

[34] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. 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., pages 43–52, San Jose, CA, November 2010.

[35] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, November 1975.

[36] H. Samet. 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, November 2014.

[37] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *Communications of the ACM*, 46(1):63–66, January 2003.

[38] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. 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., pages 525–528, Chicago, November 2011.

[39] H. Samet, A. Rosenfeld, C. A. Shaffer, and R. E. Webber. A geographic information system using quadtrees. *Pattern Recognition*, 17(6):647–656, November/December 1984.

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

[41] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. 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., pages 257–260, Hyderabad, India, March-April 2011.

[42] J. Sankaranarayanan, H. Samet, B. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, D. Agrawal, W. G. Aref, C.-T. Lu, M. F. Mokbel, P. Scheuermann, C. Shahabi, and O. Wolfson, eds., pages 42–51, Seattle, WA, November 2009.

[43] P. Schulze-Wollgast, H. Schumann, and C. Tominski. Visual analysis of human health data. In *Proceedings of the 14th International Resource Management Association International Conference*, pages 580–583, Philadelphia, May 2003

[44] C. A. Shaffer, H. Samet, and R. C. Nelson. QUILT: a geographic information system based on quadtrees. *International Journal of Geographical Information Systems*, 4(2):103–131, April–June 1990.

[45] Jannik Strötgen, Michael Gertz, Pavel Popov Extraction and exploration of spatio-temporal information in documents. In *Proceedings of 6th Workshop on Geographic Information Retrieval*, R. Purves, C. Jones, and P. Clough, eds., article 16, Zurich, Switzerland, February 2010.

[46] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, W. G. Aref, M. F. Mokbel, H. Samet, M. Schneider, C. Shahabi, and O. Wolfson, eds., pages 144–153, Irvine, CA, November 2008.

[47] A. Thudt, D. Baur, and S. Carpendale. Visits: A spatiotemporal visualization of location histories. In *Proceedings of the Eurographics Conference on Visualization*, M. Hlawitschka and T. Weinkauf, eds. 79–83, Leipzig, Germany, June 2013.

[48] U.S. Geological Survey. U.S. Board on Geographic Names (BGN), Sept. 2012. URL `http://geonames.usgs.gov`.

[49] M. Wick and B. Vatant. The geonames geographical database [online, cited 24 Jun 2008]. Available from World Wide Web: http://geonames.org/.

[50] G. Zhou and J. Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480, Philadelphia, PA, July 2002.