Representing Dynamic Phenomena Based on Spatiotemporal Information Extracted from Web Documents

Kathleen Stewart Hornsby and Wei Wang

Department of Geography, The University of Iowa, Iowa City, IA 52240 USA
Email: {kathleen-stewart, wei-wang-1}@uiowa.edu

1. Introduction

With the continued development of the Internet, text information is widely available in the form of Web articles, news reports, blogs, Twitter feeds, etc. Geographic information is commonly referenced in these Web documents. In addition, many of these documents describe dynamic occurrences, e.g., the movement of natural phenomena, such as severe storms, as well as human activities, for example, the movement of relief supplies in the wake of a natural disaster, or the movement of troops. In this work we present our research on a set of methods for extracting spatiotemporal information about dynamic events from Web documents. This work combines principles from the field of geographic information retrieval (GIR) with ongoing work in the field of temporal GIS where there is an interest in modeling geographic dynamics. The motivation for this research is to automatically represent the spatiotemporal characterizations of the events described in Web documents in a dynamic mapping environment. In this way, it will be possible to map the neighborhoods affected by moving wildfires, the local extent of flooding in a region, or the progress of relief deployment after a major disaster, such as an earthquake or hurricane directly from descriptions given in Web news stories, reports, or blogs.

2. Background

This paper presents our ongoing research on methods for automatically extracting spatial and temporal references from Web texts and combining this information to afford a spatiotemporal perspective of the dynamics described in the articles. Most GIR systems are based on detecting spatial references in text. Numerous systems have been developed based on GIR techniques including, for example, *GIPSY*, one of the first geo-referenced information processing systems that utilized a gazetteer for matching spatial terms (Woodruff and Plaunt 1994), *MetaCarta*, a system for displaying spatial references in a set of documents on a map in order to visualize the locations for each document in the set (Kornai 2005), *STEWARD*, a system for extracting, querying, and visualizing textual references to geographic locations in unstructured text documents (Lieberman et al. 2007), *NewsStand*, another GIR system that detects geographic-related information from RSS feeds using a custom-built geotagger (Teitler et al. 2008), and more recently *TwitterStand*, for automatically obtaining breaking news from tweets posted by Twitter users, and providing a map interface for reading this news (Sankaranarayanan et al. 2009).

In the field of GIScience, researchers are also growing increasingly interested in incorporating GIR techniques into their studies. For example, methods are being investigated for extracting route directions automatically from text (Klippel et al. 2008), and research involving oral histories and biographies combines narrative

analysis with a GIS-based time-geographic framework for interactive interpretation, analysis, and visualization (Kwan and Ding 2008). In other recent work, humanitarian terms are extracted from articles describing the long-term humanitarian crisis situation in the Sudan and visualized using an earth-based visualization (Tomaszewski 2008).

3. Extracting Spatiotemporal Information from Web Documents

For the work described in this paper, the open source software, GATE (General Architecture for Text Engineering, http://gate.ac.uk/), is used as the primary tool for term extraction. For this study, we are particularly interested in how spatial and temporal references available in text work together to inform us about the dynamics of events. We have been working on an approach that goes beyond the more common extraction of spatial terms, focusing on spatiotemporal information extraction.

The extraction of spatiotemporal information is accomplished in three main steps. The first step involves the parsing and annotation of spatial and temporal terms based on the use of gazetteers for text matching. The default spatial gazetteer for GATE contains general world references plus mostly UK-based places, and so an additional gazetteer using data from ESRI Streetmap and the US Gazetteer Files for the 2000 census has been created to support the extraction of places for articles describing events in the US. GATE also provides basic annotation support for temporal references including common temporal data types, such as day, week, month, and year, and we have extended this by adding a new temporal gazetteer with additional temporal types to provide more refined temporal annotation capabilities.

The second step of the extraction process relates to the sorting of the results of the previous annotation step. Sorting the annotated results based on queries to GATE orders the annotated terms in a systematic way most useful for subsequent spatiotemporal analysis. Annotated results can be sorted according to the order they appear in the document or in other orders, for example, a chronological order of any grouping of location entities (e.g., <*city, state*> pairs) that are extracted. The final step is to export the sorted results in HTML. In this way, the data is ready for geocoding, and mapping and visualization of the extracted spatiotemporal information in an earth-based visualization environment, such as ArcGlobe.

4. Combining Spatiotemporal Information to Model Dynamics

This research is based on exploring new ways to capture spatiotemporal details to inform users about the dynamics described in a document set. If we analyze a document with respect to its spatial details, we are able to obtain basic information about where the dynamics were located. For example, a set of Web news reports on tropical storm Claudette describes the path of this storm that ravaged the southern US, especially Florida, during August 2009. Articles from that period have been obtained from http://news.google.com/archivesearch and processed for their spatial content. Each document includes a header giving the published date of the article and this basic temporal information is also extracted using GATE, allowing locations retrieved from the text to be represented according to different document dates as shown with one document for August 17th (Figure 1).

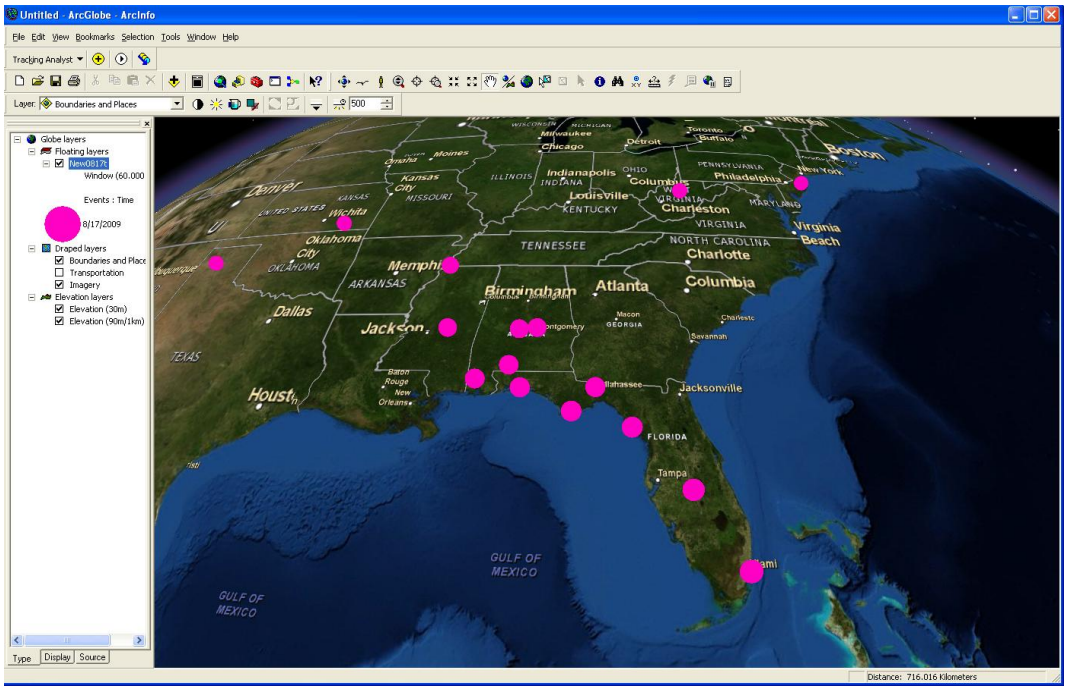

Figure 1. Locations referenced in a news report about tropical storm Claudette on August 17, 2009.

However, documents also often include additional temporal detail that spatial information extraction alone does not reveal, and that enriches our understanding of the storm's dynamics. For this work, we investigate five possible cases for how spatial and temporal information can occur in a document. These cases are a systematic analysis of the types of linguistic forms that might exist in documents, ranging from, for example, scenarios where there is no temporal information available in a sentence, only spatial details, to cases where sentences contain one spatial term along with multiple temporal expressions, or cases that are more complex, for example, sentences that contain multiple spatial *and* multiple temporal expressions, among others. Developing an approach that combines spatial and temporal text references together is more complex requiring, for example, information about the linguistic context of each sentence and methods for disambiguating what happens where and when. In this work, we present our approach for combining space-time textual information to capture the dynamics of events and represent these dynamics using ArcGlobe. Figure 2 shows the case where temporal information about tropical storm Claudette for August 17th is extracted and combined with spatial information, showing that some of the locations identified in Figure 1 are actually associated with dates different to the document date, including locations prior to the 17th, as well as predicted future locations for Claudette.

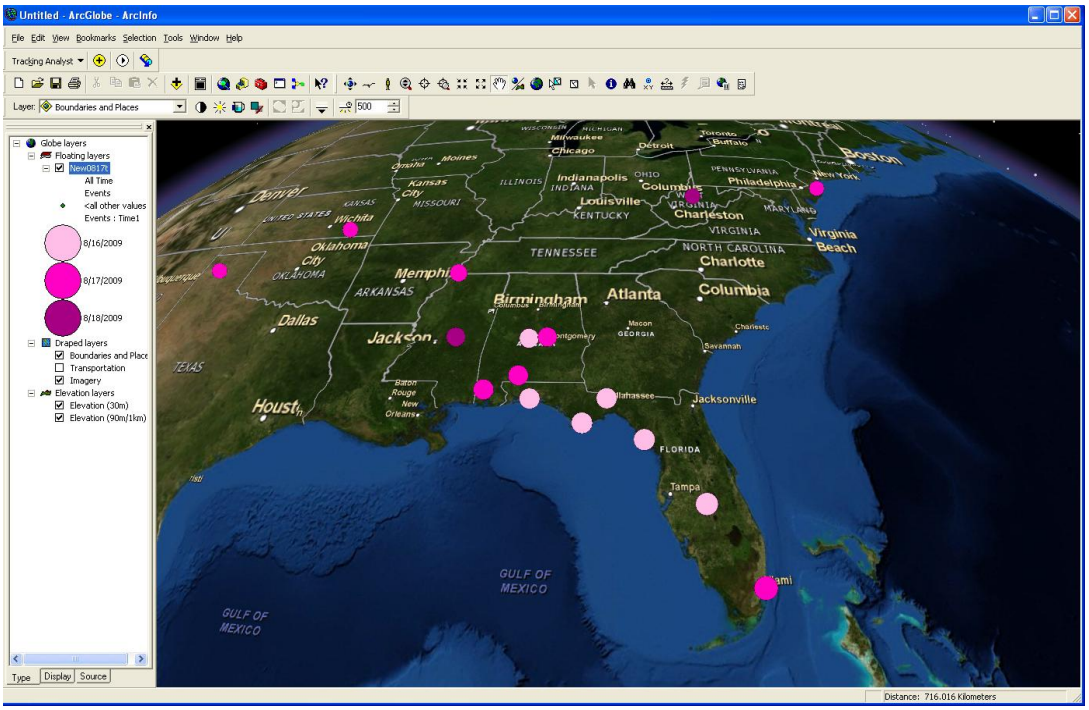

Figure 2. Locations referenced in a news report about tropical storm Claudette on August 17, 2009. Temporal information refines the spatial pattern of locations affected by the storm. Lighter shades signify locations attributed to earlier dates, while dark shades signify likely future locations.

This research uses a number of document sets for analysis, for example, articles on wildfires in California from August - September, 2009**,** flooding in the Midwest March – April, 2009, and we plan to also investigate examples of human activities described in text, for example, the distribution of supplies after the devastating earthquake in Haiti in January 2010. The results show that it is possible to distinguish information about current or past movements *vs.* future movements, and discriminate locations associated with different events. This work can reveal possible spatiotemporal changes and trends that remain otherwise untapped in text document datasets. There are challenges too since different documents often contain varying degrees of spatial and temporal detail, that can lead to uneven descriptions of change. While news articles are helpful for us to develop and test our approach, this work is particularly promising for obtaining information about local dynamics where text-based descriptions may be one of the few available sources of information about events that are happening in a local area. Incorporating temporal information may also contribute to reducing disambiguation of places, still a challenging problem for GIR tasks. Future efforts will investigate how to augment the role of gazetteers with ontologies for spatiotemporal information retrieval since ontologies offer many advantages for reasoning about entities in text especially with respect to generalization and specialization.

Acknowledgements

Kathleen Stewart Hornsby's research is supported in part by grants from the U.S. Department of Defense HM1582-08-2001, HM1582-05-1-2039 and HM1582-08-1-0013.

References

Lieberman M D, Samet H, Sankaranarayanan J and Sperling J, 2007, STEWARD: Architecture of a spatio-textual search engine. *Proceedings of the 15th annual ACM international symposium on advances in geographic information systems*, Seattle, WA, USA, 186–193.

Klippel A, MacEachren A M, Mitra P, Turton I, Zhang X, Jaiswal A, Soon K, Oyler J and Li R, 2008, Geographic analysis of linguistically encoded movement patterns: a contextualized perspective. *Extended abstracts for the 5th international conference GIScience 2008*, Park City, UT, USA, 113–117.

Kornai A, 2005, MetaCarta at GeoCLEF 2005. *The working notes of the CLEF workshop*, Vienna, Austria, 21–23.

Kwan M P and Ding G X, 2008, Geo-narrative geographic information systems for narrative analysis in qualitative and mixed-methods research. *The Professional Geographer*, 60(4): 443 – 465.

Sankaranarayanan J, Samet H, Teitler B E, Lieberman M D, Sperling J, 2009, TwitterStand: News in tweets. *Proceedings of 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, *ACM GIS '09*, Seattle, WA, USA, 45–51.

Teitler B E, Lieberman M D, Panozzo D, Sankaranarayanan J, Samet H, and Sperling J, 2008, NewsStand: a new view on news. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information System*, Irvine, CA, USA, 18.

Tomaszewski B, 2008, Producing geo-historical context from implicit sources: a geovisual analytics approach. *The Cartographic Journal*, 45(3): 165–181.

Woodruff A G and Plaunt C, 1994, GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9): 645–655.