

# COMMUNICATIONS

CACM.ACM.ORG OF THE

# ACM

10/2014 VOL.57 NO.10

## Reading News with Maps by Exploiting Spatial Synonyms

Abstractions for  
Software-Defined  
Networks

Certificate  
Transparency

Disrupting and  
Transforming  
the University

Gradual Evolution

Unconventional  
Computing

DOI:10.1145/262957

**Use this map query interface to search the world, even when not sure what information you seek.**

**BY HANAN SAMET, JAGAN SANKARANARAYANAN, MICHAEL D. LIEBERMAN, MARCO D. ADELFO, BRENDAN C. FRUIN, JACK M. LOTKOWSKI, DANIELE PANOZZO, JON SPERLING, AND BENJAMIN E. TEITLER**

# Reading News with Maps by Exploiting Spatial Synonyms

DO YOU TRAVEL? Do you want to know what is happening in the place and vicinity you are traveling to? Do you want to keep up with the latest news in the place and neighboring vicinity you left, especially if it is where you may have once lived or worked? If you answered yes to any of these questions, then our NewsStand, denoting Spatio-Textual Aggregation of News and Display, and related systems, are for you.

NewsStand<sup>46</sup> is an example application of a general framework for enabling people to search for information with a map-query interface. As such, it is a variant of systems we have been developing for the

past 30 years at the University of Maryland that we call “spatial browsers,” as in Samet et al.<sup>39</sup> and Samet et al.<sup>41</sup> The advantage of the map-query interface is that a map, coupled with the ability to vary the zoom level at which it is viewed, provides inherent granularity to a search process that facilitates approximate search. This capability distinguishes it from prevalent keyword-based conventional search methods that provide a limited facility for approximate searches that are realized primarily by permitting a match through a subset of the keywords. However, users often lack a firm grasp of which keyword to use, and would thus welcome the search to also account for synonyms. For queries to spatially referenced data, termed “spatial queries to spatial data,” the map-query interface is a step in this direction. Consider the action of pointing at a location (such as through the appropriate positioning of a pointing device or gesturing appropriately) and making the interpretation of the precision of this positioning specification dependent on the zoom level. This is equivalent to permitting use of spatial synonyms.

Being able to use spatial synonyms is important, as it enables users to search for data when they are not exactly sure what they seek or what the answer to their query should be. For example, suppose the query seeks a “rock concert in Manhattan.” The presence of “rock concerts” in Harlem, Brooklyn,

## » key insights

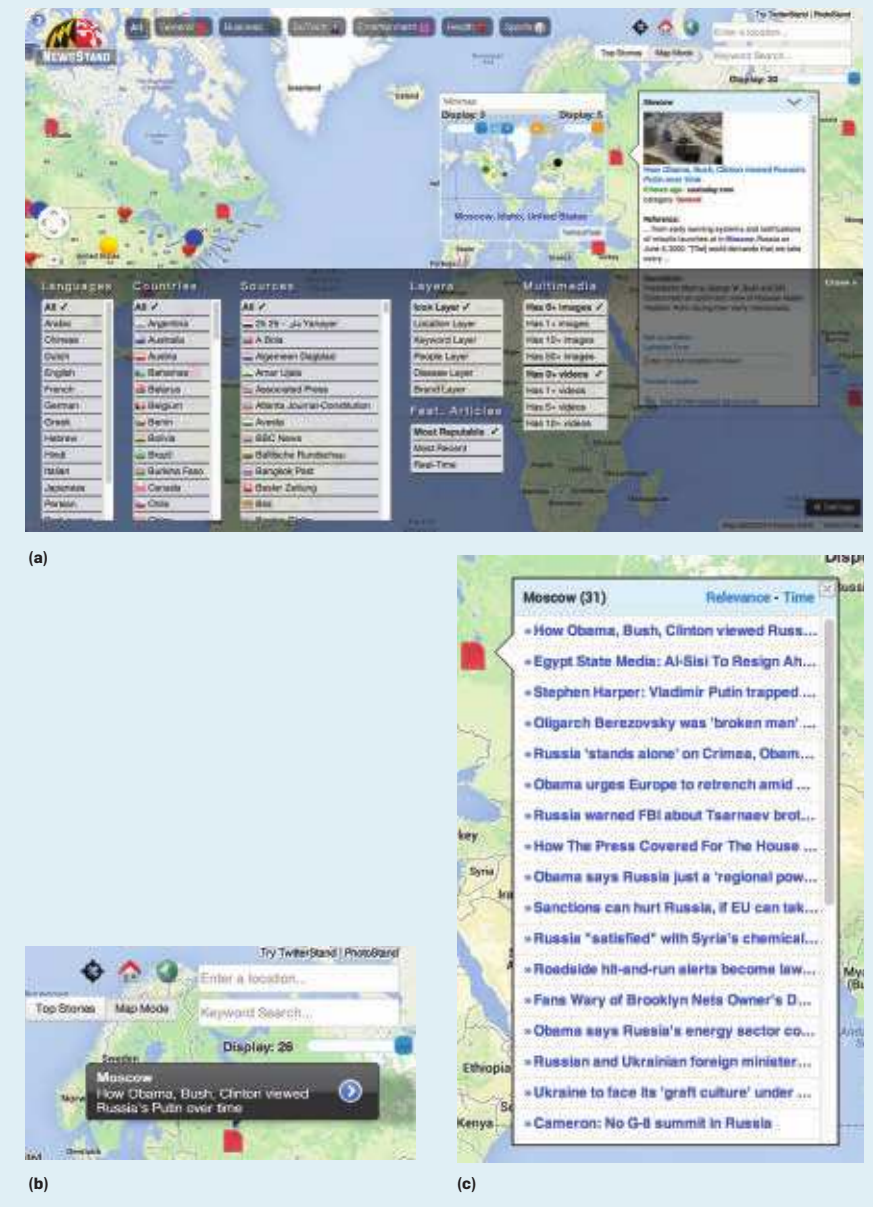
- The NewStand map query interface monitors the output of more than 10,000 RSS news sources within minutes of publication and associates articles with the locations they mention.
- A map coupled with the ability to vary the zoom level at which it is viewed and interpreted provides inherent granularity to the search process, facilitating an approximate search and enabling use of spatial synonyms.
- Textual specification of location is preferable to geometric specification for users of mobile devices but must overcome potential ambiguity.

ILLUSTRATION BY COHERENT IMAGES





**Figure 1. NewsStand Map Mode: (a) Example screenshot for “What is happening at location X on March 26, 2014?”; (b) representative headline in Moscow for the Obama/Putin relationship topic; and (c) representative headlines for topics associated with Moscow.**



or New York City are all good answers when no such events can be found in Manhattan, as they correspond to spatial synonyms: Harlem by virtue of being contained in Manhattan; Brooklyn by virtue of both proximity and being a sibling (both are boroughs of New York City); and New York City by virtue of a containment relationship. Conventional search engines handle spatial queries by dynamically incorporating information gleaned from query-and-click logs, whereby if enough users searching for Manhattan end up clicking on pages associated with Harlem or New York, then over time, the search

engine infers the spatial scope of the documents to be proximate or relevant to New York. More recently, search engines (such as Google’s Knowledge Graph and Microsoft’s Satori) have been using large knowledgebases to understand the spatial focus of keyword search queries, as well as, to a limited extent, the spatial focus of the documents. Notwithstanding such improvements to search engines for understanding locations in documents, the primary utility of the search engines is still based on popularity in the sense that the PageRank algorithm and click logs ensure webpages provided

to the user are ordered by measures that incorporate some aspect of their frequency. In particular, the classic PageRank algorithm uses static data, while click logs correspond to dynamic data. The frequency basis ensures the results are the same as those provided to other users. This property can be characterized as the “democratization of search” in the sense that all users receive equal treatment. A cruder way to describe the resulting effect is that it does not discriminate among users in the sense they all get the same bad (or good) answers. That is, the effect of using the PageRank algorithm and click logs to order results (effectively choosing which results to present to the user) is that if nobody ever looked for some data (or its neighbor in a spatial sense) before or linked to it, then it will never be found and, hence, will never be presented to the user. In some cases, this is fine. However, for synonyms, it has a strongly negative effect on the quality of search results, as it means if nobody linked to similar pages due to their content being equivalent but for the use of the same words, or clicked on a spatial neighbor, then the search engine will never find the similarity. As such, the PageRank algorithm will never be able to find similar pages as it crawls the Web when building an index to the Web pages, and no useful click logs will be found.

NewsStand and related systems we have built at the University of Maryland address the synonym problem for spatial queries. Note that all spatial queries can be broken down into two classes:

*Location-based.* Takes a location  $X$ , traditionally specified using lat/long coordinate values as an argument, and returns a set of features associated with  $X$ ; and

*Feature-based.* Takes a feature  $Y$  as an argument and returns the set of locations with which  $Y$  is associated.

These queries can also be characterized as a pair of functions, with one the inverse of the other. Feature-based queries are also known as “spatial data mining.”<sup>3</sup> Although features are usually properties (also known as attributes) of spatially referenced data (such as crop types, soil types, zones, and speed limits), they and the underlying spatially referenced data domain can be



more broadly interpreted. NewsStand adapts them to the domain of unstructured data consisting of collections of news articles with textually specified locations; the features are the topics. Adapting these concepts results in a location-based query returning all topics and articles mentioning a specific place or region  $X$  and a feature-based query returning all places and regions mentioned in articles about topic  $T$  or just article  $Y$ . Note that NewsStand does not require users to specify  $T$  in advance, in which case the topics are ranked by importance, which can be defined by various criteria, including, but not limited to, the number of articles comprising them. Here is a typical pair of queries: What is happening at location  $X$ ?; and Where is topic  $T$  or article  $Y$  happening?

Their execution is facilitated by building an index on the spatial data,<sup>36</sup> preferably all at once through bulk loading, as in Hjaltason and Samet.<sup>12</sup> An index is relatively easy to construct when the spatial data is specified geometrically and numerically. However, data is not specified this way in NewsStand, as all data is unstructured. In particular, location and feature data are both just collections of words, some of which, in the case of spatial data, can be (but are not required to be) interpreted as the names of locations. That is, spatial data is specified using

text (called “toponyms”) rather than geometry, meaning some ambiguity is involved. This ambiguity has advantages and disadvantages. The advantage is that, from a geometric standpoint, the textual specification captures both the point and spatial extent interpretations of the data, analogous to a polymorphic type in parameter transmission serving as the cornerstone of inheritance in object-oriented programming languages. For example, a city can be geometrically specified by either a point (such as its centroid) or a region corresponding to its boundary, the choice of which depends on the level of zoom with which the query interface is activated. The disadvantage is we are not always sure if a term is a geographic location. For example, does “Jordan” refer to a country, a river, or a surname, as in “Michael Jordan”? The process of answering is called “toponym recognition.”<sup>18</sup> Moreover, if it is a geographic location, then which, if any, of the possibly many instances of geographic locations with the same name is meant. For example, does “London” refer to an instance in the U.K., Ontario, Canada, or one of many others? The process of answering is called “toponym resolution.”<sup>19</sup> Resolving these ambiguities with no errors (or almost none) is one of the main technical challenges we have faced in deploying NewsStand and related systems.

### NewsStand User Interface

NewsStand’s goal is to offer an alternative to the news-reading process and, more important, experience. Users query NewsStand by choosing a region of interest and finding relevant associated topics and articles (experience the NewsStand interface at <http://newsstand.umiacs.umd.edu>). The topics and articles displayed are determined by the location and level of zoom that together dictate the spatial scope of the query, or region of interest. The two ways of interpreting the notion of “region of interest” are in terms of content and of news sources. In the simplest way, there are no predetermined boundaries on the locations of the news sources for the articles being displayed for the region of interest. In the second way, the sources can be limited to a subset of available sources by specifying them explicitly (such as *New York Times* and *Washington Post*), by language, by spatial region that can be specified textually (such as restrict sources to Ireland), or by drawing the region of interest on the NewsStand map (such as a box overlapping Ireland and the U.K.). Users can also constrain the spatial region and news sources; they need not be the same. This is a useful feature, as it enables users to see how one part of the world views events in another part of the world. For example, users may want to see how the Eng-

**Figure 2. NewsStand Top Stories Mode: (a) Example screenshot for “Where is topic  $T$  or article  $Y$  happening on March 26, 2014?”; and (b) subset of images associated with the Obama/Putin relationship topic with duplicates and near-duplicates grayed over.**



lish press views and interprets developments in the Middle East. The result is analogous to sentiment analysis. Other applications include monitoring hot spots for investors, national security, and keeping up with the spread of diseases, as in Lieberman et al.<sup>24</sup>

Figure 1a is a screenshot of NewsStand's output for "What is happening at location *X* on March 26, 2014?" This is NewsStand's "Map Mode." *X* is Africa, Europe, and part of the Americas. The figure includes an excerpt from an article about the Obama/Putin relationship that mentions Moscow. Each icon, or symbol on the map, we call a "marker," represents a set of articles on the same and/or different topics where the main property shared by all the articles is that they mention the corresponding map location. The type of symbol conveys information about the news category (such as general news, business, science and technology, entertainment, health, and sports) spanning most of the article topics associated with the location. The user can select one or more of these categories by toggling the appropriate buttons at the top of the screen.

Figure 1b is an info bubble containing the headline from a representative article on the dominant topic associated with Moscow, or the Obama/Putin relationship. NewsStand obtains these topics by applying a clustering process to all the articles. The info bubble is generated by the user hovering the mouse cursor over Moscow. The hovering action also causes the markers at all other locations on the map associated with this representative article to be replaced by orange balls. In this example, these locations correspond to, in part, the countries involved in, or affected by, the Obama/Putin relationship. Some locations might lie outside the geographic span of the map (such as in North America and the Far East) currently visible in the screenshot.

Including areas of interest beyond the map is achieved through a minimap generated when the user hovers over a marker, along with the headline (not shown here). The action displays the geographic span of the representative article with orange balls at the appropriate locations. The utility of the minimap involves permitting users to see the selected article's geographic fo-



## NewsStand gathers its data by crawling the Web. Its primary sources are thousands of individual news sources worldwide in the form of RSS feeds.



cus, without having to leave their area of interest on the main map, and is independent of the current level of zoom.

Blue balls on both the main map and the minimap indicate other locations with the same name as the one over which the user is currently hovering—here Moscow. Allowing the minimap to include all other locations with the same name may cause the geographic span of the minimap to exceed that of the orange balls. The blue balls enable detecting toponym resolution errors.

A black ball on the minimap marks the location over which the user is currently hovering, or Moscow. Up and down arrows on the minimap allow the user to scroll through the orange and blue balls and output the corresponding location names. Scrolling through the blue balls enables ranking the interpretations of the location name. Green and red balls on the minimap correspond to the current blue and orange balls in the scrolling process. Hovering over an orange ball in the minimap yields the name of the location, while hovering over a blue ball yields both the name of the location and its containing location on the minimap (such as "Moscow, ID, United States"), as all blue balls have the same name.

Figure 1c is an info bubble showing headlines of representative articles for each topic associated with Moscow, the location over which the user hovered most recently. It results from clicking the > symbol in the headline info bubble associated with this location. Clicking on one of the headlines yields the summary info bubble (see Figure 1a), as well as the adjacent corresponding minimap, which is also generated when hovering over a marker. Note that orange balls (but not the blue balls) in the minimap differ as a user scrolls through the headlines of the topics. The summary info bubble also contains links to related images, videos, and other articles. Clicking on the headline in this summary info bubble causes the full text of the article to be displayed and, if it is in a language other than English, an option is available to translate it and/or the headline into English through a translation package (such as Google Translate and Microsoft Translator).

The domain of news sources for the articles from which the representative article is drawn can be restricted by language, geographic region, or country, as well as by specific newspaper. This is done by setting up an appropriate filter using the “settings” button (at the lower-right-hand corner of the screen in Figure 1a) and selecting the appropriate ones, as in the lower grayed half of Figure 1a. Note that users are also able to do a search by location or keyword(s), as well as vary the number of markers to be displayed through a display slider.

Figure 2a is a screenshot of NewsStand’s output for “Where is topic  $T$  or article  $Y$  happening on March 26, 2014? This is NewsStand’s “Top Stories Mode.”  $T$  is one of the topics whose representative headlines are shown in the bottom-left pane ranked using an importance measure. Importance is defined in terms of significance, age, and frequency, though velocity/acceleration of arrival should also be taken into account, as it is a better measure since topics eventually lose their timeliness. The headline displayed is the one that was clicked. It is highlighted (by being grayed) as a result of the user hovering over it, corresponding here to the Obama/Putin relationship topic. Clicking on the headline causes more details (such as an expanded description and the number of related documents, images, and videos) to appear about it, as shown in the top left pane of Figure 2a, along with the means to access them via a subsequent mouse click.

The hovering click in the bottom-left pane of Figure 2a also causes appropriate markers (category symbols) to appear on the map (right pane) at the principal geographic locations associated with the topic. In this example, these locations correspond to, in part, some of the countries involved in, or affected by, the Obama/Putin relationship, including the U.S. and Russia. Hovering the mouse cursor on the map in the right pane causes info bubbles and the associated minimap with the same semantics to appear, as in the “What is happening at  $X$ ?” query in Figure 1. In particular, the orange balls enable the user to differentiate between locations in close proximity (such as London and Wimbledon in the U.K. for a tennis cluster), while the blue balls

capture other instances of geographic locations with the same name (such as “Moscow, PA, United States”).

Users in Map and Top Stories modes can obtain the collection of images and videos associated with each cluster. For images, NewsStand detects duplicates or near duplicates and hides them from view. This is a powerful property, as it uses the words associated with the articles, or their semantics, as the first step in finding similar images, while the duplicates among the similar images can be detected through classical image-similarity methods, including hierarchical color histograms<sup>5</sup> and the Scale-Invariant Feature Transform algorithm, or SIFT.<sup>25</sup> Figure 2b is an example of a subset of such images for the Obama/Putin relationship topic anchored in Moscow.

As outlined earlier, NewsStand’s ultimate goal is to make the map the medium of choice for presenting information with spatial relevance and is thus not restricted to news articles; that is, it can also be applied to search results, images, videos, and tweets. It also enables summarization of news, further exploration, and even knowledge acquisition through discovery of patterns in the news, a direct result of the association of topics or categories with the locations mentioned in constituent articles. For example, queries can be

chained in the sense that an interesting topic might be associated with Paris, France, and the same topic might also be associated with London, U.K., as found through the orange balls. At this point, the user would move the pointing device to London and click to find other related topics mentioning London, as well as other locations to which the user can transition by moving via the map-query interface. This unlimited chaining is possible only in Map Mode, as the queries are location-based, while the queries in Top Stories Mode are topic-based, and the markers on the map are restricted to the locations corresponding to the highest-ranked topics, unless the user does a keyword search.

NewsStand can also compute a cluster disease focus, or the most common term in the cluster corresponding to the name of a disease (such as “Europe on March 26, 2014” in Figure 3). Alternatively, a user can apply the same idea and find the most common term in the cluster corresponding to the name of a person or brand. Finding such a term is achieved by setting the “layers” parameter to “disease,” “people,” or “brand,” respectively.

### Related Work

Comparing NewsStand with existing newsreaders is difficult, as reading the

**Figure 3. NewsStand screenshot showing clusters that mention a disease name for Europe on March 26, 2014; the user is hovering over Valencia, Spain, and the disease is breast cancer. Orange balls in the minimap show all other locations in the world where the relevant cluster mentions breast cancer.**





news with a map is a feature not found in any popular news reader (such as Pulse). News-reading systems (such as Microsoft Bing News, Google News, and Yahoo News) present the news in classical linear fashion with aggregation of different sources for each topic. These providers all include some aspect of locality in terms of aggregation of articles and topics relevant to a user's locality. Aggregation is usually done according to a ZIP or postal code or city-state specification. For example, for ZIP code 20742, topics could mention "College Park, MD." For Google News, this feature seems to be implemented, at least as far as we can tell, by applying Google search with location names as search keys. For example, after determining the user is in ZIP code 20742 (such as by virtue of the user's IP address, absent an alternative specification of the local area), Google News would return the topics mentioning "College Park, MD" or "University of Maryland," as they are known to be associated with this ZIP code. In addition, the resulting list of topics also appears to be based primarily on the location of the news source (usually a newspaper) where the articles comprising the topics are contained, rather than on story content. In these examples, the number of topics displayed is limited, though there is no particular reason for this limitation save for the absence of topics relevant to the user's locality. Note also that in these examples there is no notion of article importance in determining what is shown to the user.

Interestingly, none of the popular news readers use a map to present the articles, though they could all do so with a mashup on their mapping platforms. HealthMap<sup>9</sup> does use a map to present disease outbreaks, where locations are obtained from the dateline of a disease report or metadata from ProMed reports. This use of a map to present disease reports is similar to the "disease layer" in NewsStand (see Figure 3), except that in NewsStand the locations are obtained from the article's actual text. It is also similar to an implementation of our Spatio-Textual Extraction on the Web Aiding the Retrieval of Documents, or STEWARD,<sup>23</sup> system with ProMed reports that can also show disease propagation over time.<sup>16</sup> Note although the mapping

platforms supporting the mashups are able to zoom in, with the exception of NewsStand, none couple zoom with the ability to obtain more articles.

In the past, a number of systems tried to understand geographical locations in news articles and display them, though most are no longer available or accessible. For example, Reuters's NewsMap, the *Washington Post's* TimeSpace, the BBC's LiveStats, and the AP's Mobile News Network tried to associate news articles with a coarse geography based on the wire-service location where the article was filed. An article submitted to the Miami news wire would therefore be listed for all ZIP codes in Miami. Unlike NewsStand, there appears to be no attempt in the AP Mobile News Network to analyze individual articles to determine the main associated location, or geographic focus, or other important locations mentioned in the articles.

It is also useful to compare NewsStand with commercial services for Web search and recommender systems (such as review sites like Yelp and TripAdvisor). The difference is that in these systems, awareness of spatial entities is a result of the explicit population of their databases with spatial information in the form of addresses or GPS, or lat-long, values; hence they can support the exploration of the spatial information. NewsStand has a dual role: discover the spatial information in its input data that is specified textually and usually ambiguously (requiring incorporation of other information, some external to its input data); and exploratory, where the capabilities are similar to those in recommender systems, though there is less emphasis on a map-query interface in the recommender systems.

### NewsStand Architecture

The key elements to understanding news were perhaps best captured in 1902 by Rudyard Kipling in his *Just So Stories*: "I keep six honest serving-men (They taught me all I knew); Their names are What and Where and When And How and Why and Who." NewsStand focuses on the "what" and "where" and to a lesser extent on "when," where "when" is recent. Here, we focus on "what" and later on "where."

NewsStand gathers its data by crawl-

ing the Web. Its primary sources are thousands of individual news sources worldwide in the form of really simple syndication (RSS) feeds; RSS is a widely used XML protocol for online publication, ideal for NewsStand, as it requires only a title, short description, and Web link for each published news item. RSS 2.0 also allows an optional publication date, helping determine the age, or "freshness," of an article. NewsStand currently indexes 10,000 news sources and processes approximately 50,000 news articles per day. It determines the geographic locations mentioned in the article, a process known as geotagging, and tries to determine an article's geographic focus or foci that are the key locations mentioned in it.

NewsStand also aggregates news articles by topic based on content similarity (termed "clustering") so articles concerning the same event are grouped in the same cluster. The main goal of clustering is to automatically group news articles that describe the same news event into sets of news articles, termed "article clusters" (also referred to earlier as "topics" and as "clusters"), such that each cluster contains only the articles encountered in the input seen so far pertaining to a specific topic. As news articles enter this stage, NewsStand assigns them to news clusters, essentially a one-shot process meaning once an article is added to a cluster it remains there forever. NewsStand will never revisit or recluster the article, which is desirable, as articles come into NewsStand at a high-throughput rate, and NewsStand needs a document-clustering system that can process them quickly while still managing to deliver good-quality clustering output. Such a version of the clustering algorithm is characterized as being "online."

Given these requirements, NewsStand uses the leader-follower clustering<sup>7</sup> algorithm that permits online clustering in both the term-vector space using the term frequency-inverse document frequency, or TF-IDF, metric<sup>35</sup> and the temporal dimension. For each cluster, NewsStand maintains a term centroid and time centroid corresponding to the means of all term-feature vectors and publication times of articles in the cluster, respectively. To cluster a new article  $a$ ,



NewsStand checks whether a cluster exists where the distance from its term and time centroids to  $a$  is less than a fixed cutoff distance  $\epsilon$ . If one or more candidate clusters exists,  $a$  is added to the closest such cluster, and the cluster's centroids are updated; otherwise, NewsStand creates a new cluster containing only  $a$ .

NewsStand's online clustering algorithm ranks the clusters based on its notion of "importance," as determined by several factors:

*Number of articles.* The number of articles in the cluster;

*Number of unique news sources in a cluster.* For example, an event in Irvine, CA, is important if carried by multiple news sources, especially if some are geographically distant from Los Angeles (approximately 50 miles from Irvine);

*The cluster's rate of propagation.* Articles about important events are picked up by multiple news sources within a short period of time; and

*Time of addition.* The time at which the most recent addition to the cluster took place, an option exercised by the NewsStand user, precluding consideration of the first three factors.

When clusters are ranked using the first three factors, NewsStand must choose the cluster's representative article, a form of secondary ranking. The nature of this article can be varied by the NewsStand user to be either the most recent article, thereby disregarding the corresponding cluster's importance (the fourth factor), or according

to the cluster's importance, where the choice is between the article from the most reputable source or from the source with the freshest article. Though it is important for NewsStand to show the clusters with the most significant topics in the current viewing window when in Map Mode, simply displaying the highest-ranked topics on the map may not produce a useful display for a wide audience, as these topics tend to be clustered in particular geographic areas. This situation reflects the uneven news coverage of major newspapers, as they tend to focus on these geographic areas. In NewsStand, topic selection is a trade-off between significance and spread. To achieve a balance, NewsStand subdivides the viewing window into a regular grid and requires each grid square contain no more than a maximum number of topics. The topics displayed are selected in decreasing order of significance and age, an approach that ensures a good spread of top topics across the entire map.

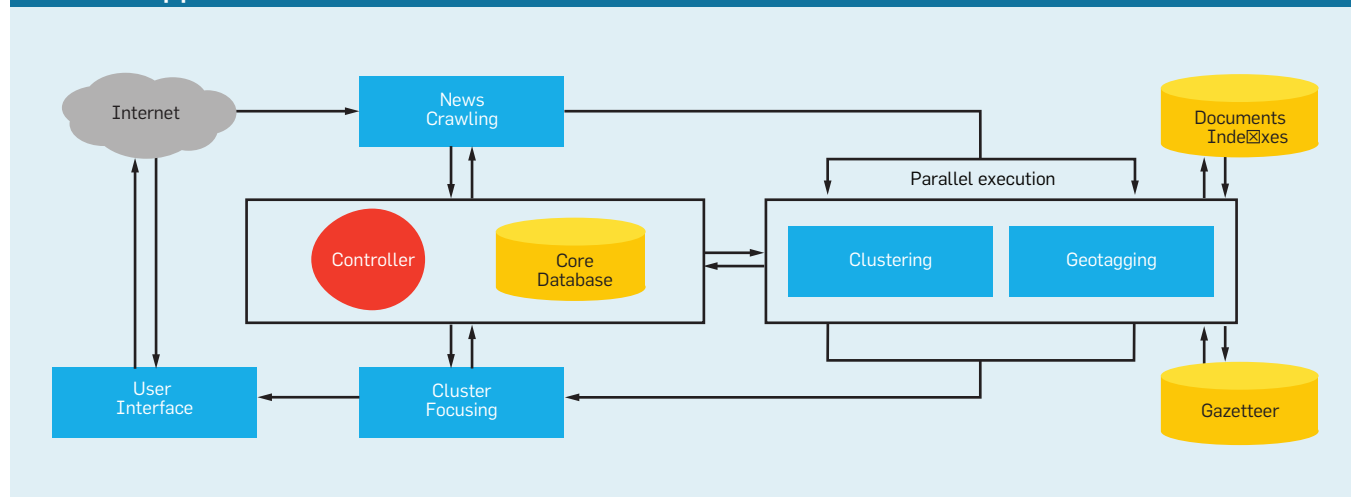
NewsStand also determines the geographic focus or foci associated with the cluster, a determination facilitated through the clustering process vis-à-vis the location feature. NewsStand displays each cluster at the positions of its geographic foci, provided it is one of the most important clusters or its geographic focus is also the focus of one of the most important clusters, where the number of locations is set by manipulating a slider in the upper-right corner of the map. The locations associated

with the most important clusters are thus the ones for which the map contains data. This display is usually done with the aid of symbols corresponding to their news category, as in Figure 1.

However, rather than display the category symbol associated with the cluster, NewsStand can also display the text corresponding to the most prevalent term in the cluster we call the "keyword" by having the user set the appropriate "layers" parameter in Figure 1a. Alternatively, users can also display the actual name of the location that serves as the geographic focus by setting the layers parameter to "location."

Scalability and fast processing of individual articles were the most important criteria in designing NewsStand's architecture<sup>20</sup> (see Figure 4). Additional goals include presenting the latest news as quickly as possible, within minutes of its online publication, and being robust to failure. The NewsStand architecture fulfills these criteria by subdividing its collection and processing into several modules, each able to run independently on separate computing nodes in a distributed computing cluster. The figure outlines how the articles are processed by a sequence of these modules in a computing pipeline. Because each module might execute on a different node, a given article might be processed by several different computing nodes in the system. We also designed the modules in a way that allows for multiple instances of any module to run simultaneously on one or more nodes. News-

**Figure 4. High-level overview of NewsStand's architecture. We designed the system as a pipeline, with individual processing modules working independently. A central control module orchestrates article processing by delegating work to the other modules and tracking articles in the pipeline.**



Stand is thus able to execute as many instances of modules as required to handle the volume of news it receives. Each module receives input and sends output to a PostgreSQL database system that serves as a synchronization point. User actions (such as zoom, pan, and select) in the NewsStand interface are automatically converted into SQL queries that are answered by the PostgreSQL database.

**Geotagging**

NewsStand extracts geographic locations from news articles (termed “geotagging”) and is related to work in geographic information retrieval. Much of the existing work in this area deals with finding the geographic scope of websites and individual documents. In the context of news articles, NewsStand distinguishes among three types of geographic scope:<sup>26</sup>

*Provider.* The publisher’s geographic location;

*Content.* The article or topic content’s geography; and

*Serving.* Based on the reader’s location.

NewsStand relies on article content to determine an article’s geographic scope and also tries to use provider scope, which it knows, and serving scope, which it attempts to learn.

NewsStand extends our earlier work on geotagging in STEWARD<sup>25</sup>

to support spatio-textual queries on documents on the hidden Web. While STEWARD’s technology is applicable for an arbitrary set of documents, NewsStand contains additional modules and features designed specifically for more effective processing of news articles. STEWARD processes each document independent of all other documents, while NewsStand takes advantage of multiple versions and instances of articles about a topic by grouping them, most often from different news sources, into topic clusters that allow for improved geotagging and lets users retrieve related articles easily.

Geotagging consists of two processes: toponym recognition and toponym resolution. Toponym recognition involves geo/non-geo ambiguity, where a given phrase might refer to a geographic location or some other kind of entity (such as deciding whether a mention of “Washington” refers to a location or another entity, like a person’s name). Aliasing is a secondary issue, where multiple names refer to the same geographic location (such as “Los Angeles” and “LA”). Toponym resolution, also known as “geographic name ambiguity,” or polysemy, involves geo/geo ambiguity, where a given name could refer to any of several geographic locations. For example, “Springfield” is the name of many cities in the U.S., including in

Massachusetts and Illinois, where it is the state capital.

**Toponym recognition.** Many different approaches to toponym recognition have been undertaken, though all share certain characteristics. The idea is to extract the “interesting” phrases, or the ones 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,” or EFV. The easiest way to identify the EFV is to look for phrases in the document that exist in a gazetteer or database of geographic names and locations. This approach is utilized by many researchers as their primary search strategy.<sup>2</sup> In particular, Web-a-Where,<sup>2</sup> a system for associating geography with Web pages, uses a small, well-curated gazetteer of approximately 40,000 locations, created by collecting the names of countries and cities with populations greater than 5,000. This size imposes a serious limitation on Web-a-Where’s practical geotagging capabilities, as it precludes it from being able to recognize the lightly populated, usually local, places commonplace in articles from local news sources. Moreover, a small gazetteer means Web-a-Where is more prone to making toponym-recognition errors because it misses out on being aware of geo/non-geo ambiguity afforded by the use of larger gazetteers.

To deal with the geo/non-geo ambiguity inherent in larger gazetteers, researchers, including Martins et al.,<sup>27</sup> Rauch et al.,<sup>33</sup> and Stokes et al.<sup>45</sup> have proposed a variety of heuristics for filtering potentially erroneous toponyms. MetaCarta<sup>33</sup> recognizes spatial cue words (such as “city of”), as well as certain forms of postal addresses and textual representations of geographic coordinates. However, this strategy causes serious problems when geotagging newspaper articles, as the address of a newspaper’s home office is often included in each article. Given MetaCarta’s primary focus on larger prominent locations, these properly formatted address strings play too large a role in its geotagging process, resulting in many geotagging errors.

Other approaches to toponym recognition are rooted in solutions to related problems in natural language process-

**Figure 5. Illustrative local lexicon for readers living in the vicinity of Columbus, OH; note the many local places that share names with more prominent locations elsewhere.**





ing. For example, Named-Entity Recognition (NER)<sup>47</sup> focuses on nouns and noun phrases, aiming to identify noun phrases from an article that correspond to various entity classes (such as PERSON, ORGANIZATION, and LOCATION). Phrases tagged as LOCATION are most likely to be locations and stored as geographic features of the entity feature vector, while ORGANIZATION and PERSON phrases are stored as non-geographic features. NER approaches can be classified roughly as rule-based<sup>18,31</sup> or statistical.<sup>17</sup>

Rule-based solutions feature catalogs of rules listing possible contexts in which toponyms may appear. On the other hand, statistical solutions rely on annotated corpora of documents to train language models using constructs like hidden Markov models (HMMs)<sup>7</sup> and conditional random fields (CRFs).<sup>15</sup> HMMs and CRFs are used widely when annotated corpora are available. NewsStand's toponym-recognition procedure uses the NER tagger of the LingPipe tool kit that was trained on news data from the Message Understanding Conference, or MUC-6, and the well-known Brown corpus.

Note that NER tagging does not preclude use of a gazetteer. Instead, these tagging methods serve as filters or pruning devices for controlling the number of lookups made to the gazetteer. The downside is that if an entity is not identified as a potential location, it will be missed, which happens. NewsStand uses GeoNames (<http://geonames.org/>), an open gazetteer originally assembled from more than 100 gazetteers, including the GEOnet Names Server and Geographic Names Information System. It is maintained by volunteers worldwide and currently contains the names of approximately 8.5 million different geographic locations, of which approximately 5.5 million are unique, with the difference accounting for the need to perform toponym resolution or resolve geo/geo ambiguity. The NewsStand gazetteer contains almost 16.3 million entries due to its need to keep track of the names of each location in multiple languages.

Our experience with the eight million articles most recently processed by NewsStand encountered only approximately 60,000 distinct locations,

though more than 40,000 were subject to geo/geo ambiguity, making toponym resolution critical. The gazetteer also stores the population of populated places or regions, as well as hierarchical information, including the country and administrative subdivisions containing the location, which is useful for recognizing highly local toponyms. Gazetteer lookup is applied to every geographic feature  $f \in EFV$  and the matching locations to form the set  $L(f)$ , where there are as many sets as there are features, or  $|EFV|$ .

**Toponym resolution.** When a toponym is recognized, NewsStand applies a toponym-resolution procedure to resolve the geo/geo ambiguity. The problem of geo/geo ambiguity resolution is related to the more general problem of associating canonical entities with each noun phrase mentioned in a document, which is referred to as “named-entity disambiguation,” or NED. To disambiguate noun phrases, NED resorts to matching noun phrases to a knowledge repository (such as Wikipedia, DBpedia, and Yago). At a high level, noun phrases mentioned in a document are first matched to multiple candidate entities, then disambiguated based on the relatedness of these entities in the knowledge repository. For instance, Milne and Witten<sup>29</sup> used a supervised learning approach using a relatedness measure, where the relatedness between two Wikipedia articles is based on the number of overlapping incoming links. Similarly, Hoffart et al.<sup>13</sup> used “coherence” among the various candidate entities to disambiguate all noun phrases. Some recent efforts have sought to combine NER and NED modules into a Named-Entity Recognition and Disambiguation, or NERD, module<sup>34</sup> that scans a document and outputs entities mentioned in it.

The simplest toponym-resolution strategy is to assign a default sense to each recognized toponym using some prominence measure (such as population). Many researchers, including Amitay et al.,<sup>2</sup> Martins,<sup>27</sup> Purves et al.,<sup>31</sup> Rauch et al.,<sup>33</sup> and Stokes et al.,<sup>45</sup> have done so in combination with other methods. For example, MetaCarta<sup>33</sup> assigns “default senses” in the form of probabilities based on how often each interpretation of a given toponym appears in a pre-collected corpus of geo-

tagged documents. It then alters these probabilities based on other heuristics (such as cue words and occurrence with nearby toponyms). The Spatially aware Information Retrieval on the Internet, or SPIRIT, project<sup>31</sup> uses techniques similar to those in MetaCarta by searching for sentence cues, falling back to a “default sense” for a given geographic reference in the absence of stronger evidence.

Note that using default senses and probabilities based on corpora makes it nearly impossible for the relatively unknown location references in articles (such as any of the more than 2,000 lesser-known instances of “London” around the world) in articles in local newspapers to be selected as correct interpretations, since these smaller places will have appeared in few pre-created corpora of news articles. In contrast, NewsStand uses a concept we call a “local lexicon”<sup>22,32</sup> that is associated with a news source and contains the set of locations in the source's geographic scope. For example, the local lexicon of readers living in “Columbus, OH” includes “Dublin,” “Amsterdam,” “London,” “Delaware,” “Africa,” “Alexandria,” “Baltimore,” and “Bremen” (see Figure 5). Readers outside the Columbus area, lacking these place names in their local lexicons, would likely think first of the more prominent same-name places.

Using the local lexicon is analogous to using a combination of the provider- and serving-scopes interpretation of the geographic scope described earlier. In particular, NewsStand learns its serving scope by forming a corpus of articles for each news source and collecting the geographic locations mentioned in the corpus that are local to it. This approach is based on observing that news articles are written with an assumption of where their reader is located. For example, when the location “Springfield, IL” is mentioned in a newspaper article in Illinois (such as Chicago), the qualifier “Illinois” or “IL” is most likely not used due to the expectation that its readers will make the correct interpretation automatically. On the other hand, an article in the *New York Times* would retain the “Illinois” qualifier when discussing “Springfield” to avoid any possible misunderstanding. Local lexicons are

particularly useful when users zoom in on the map, thereby focusing on relatively small areas where the articles are more local in nature. In this case, knowledge of the provider scope is extremely valuable in overcoming the geo/geo ambiguity.

The local lexicon can also be seen as a “resolving context” for toponym resolution. A related popular strategy<sup>2,27,31,45</sup> for toponym resolution places the resolving context within a hierarchical geographic ontology that involves finding a geographic region in which many of the document’s toponyms can be resolved. For example, Web-a-Where<sup>2</sup> pursues such an approach by searching for several forms of hierarchical evidence in documents, including minimal resolving contexts and containment of adjacent toponyms (such as “College Park, MD”). It identifies a document’s geographic focus through a simple scoring algorithm that takes into account the gazetteer hierarchy, as well as a confidence score, for each location. Ding et al.<sup>6</sup> used a similar approach. MetaCarta<sup>33</sup> and Google Book Search have no notion of a computed geographic focus, and thus require users to determine a focus for themselves. Instead of using content location, Mehler et al.<sup>28</sup> associated documents with provider location, which, at times, is equivalent to using the dateline. Note the central assumption behind finding a minimal resolving context is that the document under consideration has a single geographic focus, useful for resolving toponyms in that focus, but not for resolving distant toponyms mentioned in passing.

Note, too, the local lexicon is just one of many techniques NewsStand uses for toponym resolution, its need manifested by the fact that some features are associated with multiple records, or  $|L(f)| > 1$ . In particular, NewsStand resolves such ambiguous references through heuristic filters that select the most likely set of assignments for each reference, based on how a human would read an article. These filters rely on NewsStand’s initial assumption that locations mentioned in the article give evidence to each other, in terms of geographic distance, document distance,<sup>19</sup> and hierarchical containment. The “object container filter” is one

such filter. It searches for pairs of geographic features  $f_1, f_2 \in EFV$  separated in the article by containment keywords or punctuation symbols (such as “ $f_1$  in  $f_2$ ” or “ $f_1, f_2$ ”). If it finds a pair of locations  $(l_1, l_2)$ , so  $l_1 \in L(f_1)$ ,  $l_2 \in L(f_2)$ , and  $l_1$  is contained in  $l_2$ , then  $f_1$  and  $f_2$  are disambiguated as  $l_1$  and  $l_2$ , respectively. For example, suppose  $f_1 =$  “Brooklyn” and  $f_2 =$  “NYC.” Also, let  $L(f_1) =$  {“Brooklyn, New York City,” “Brooklyn, Shelby County”} and  $L(f_2) =$  {“New York City, New York County,” “North York County, U.K.”}. We now disambiguate  $f_1$  as  $l_1 =$  “Brooklyn, New York City” and  $f_2$  as  $l_2 =$  “New York City, New York County.” This disambiguation is justified by NewsStand’s observation that a pair of features that are textually close in the article, close geographically, and exhibit a hierarchy relationship are unlikely to occur by chance. Another example of this strategy is when a query involves lists of locations, in which case NewsStand tries to use proximity, sibling, and prominence clues to resolve the ambiguity.<sup>1,21</sup>

**Evaluation.** To see how well NewsStand’s geotagging performs, rather than display a news category icon at a location, NewsStand can display the actual name of the location by setting the “layers” parameter to “location” instead of to “icon.” In this way, it can detect wrong geo/geo interpretations (such as placing “Los Angeles” in “Chile” instead of in “California”), as well as wrong classifications of non-geo as geo (such as “George” in “South Africa” instead of “George Anthony” from the 2012 Casey Anthony baby murder trial in “Orlando, FL”) but not vice versa.

Moreover, hovering over the name  $n$  of a location  $l$  (both in the “location” and “icon” layers) causes NewsStand to generate a minimap, as well as markers in the form of blue balls at all other locations  $k$  with the same name  $n$  on both the map and the minimap, such that at least one article cluster is associated with  $k$ . This minimap enables NewsStand to quickly find geotagging errors. Research is under way to use this information to learn better classifiers. The blue balls enable NewsStand to overcome possible toponym resolution errors by providing access to all articles it determines mention a particular location name  $n$  for any

interpretation  $k$  of  $n$  as long as at least one article is associated with interpretation  $k$ , even though  $k$  may not be the correct interpretation, thereby giving the user the final say. By examining all mentions of  $n$  for the correct interpretation subject to NewsStand’s stipulation that at least one article is associated with the interpretation (assuming 100% recall for toponym recognition with lower precision), the result is that NewsStand achieves 100% recall for toponym resolution for the interpretations of a location that are in its gazetteer, with lower precision, though it does not miss any. Note that in some sense NewsStand is ranking its responses, where the highest-ranked response is associated with the queried location on the main map and the lower-ranked responses are associated with the minimap.

Results of Lieberman’s and Samet’s experiments<sup>18</sup> with handcrafted corpuses of articles showed that NewsStand’s toponym recognition<sup>18</sup> and toponym resolution<sup>19</sup> processes outperformed Reuters’s OpenCalais and Yahoo’s Placemaker, which are closed-source commercial products providing public Web APIs that allow for automated geotagging of documents. At one time, the MetaCarta system<sup>33</sup> provided a similar capability by recognizing spatial cue words (such as “city of”), as well as certain forms of postal addresses and textual representations of geographic coordinates in text documents.

### Lessons Learned

Building NewsStand has taught us that the geotagging tasks of toponym recognition and resolution are much more complex than we originally envisioned. For example, NewsStand’s geotagger could use more semantic hints from a document to aid correct geotagging (such as landmarks and rivers). Moreover, geography can be used to improve the clustering of news articles by modifying the TF-IDF framework so terms that are spatial synonyms are merged into one term instead of being treated as separate terms. A primary difficulty involves evaluating NewsStand’s performance on these tasks. Comparing NewsStand with other systems means having to use standardized datasets known as “corpuses.” We performed



this comparison for both components of the geotagging task, with emphasis on recall rather than precision, achieving superior results.<sup>18,19</sup> Nevertheless, this evaluation method involves two shortcomings: the datasets are far too small, and “corpuses are like corpses” in that news and language are constantly changing. The news data can be characterized as streaming data. The evaluation should be conducted more in a spirit of sampling, as in inspection/quality control tasks, something we intend to do in the future.

In a Web browser, NewsStand works well with the mapping API provided by Google Maps to display topics. It has also been adapted to work with Bing Maps and the Google Earth plugin, though the plugin leads to a number of display problems due to the limited number of supported platforms. NewsStand has also been ported to work on devices with a gesturing touchscreen interface (such as smartphones and tablets) for use with Web browsers,<sup>42</sup> albeit with a slightly different user interface, and as an app<sup>38</sup> for the iPhone, Android, and Windows Phone platforms (see Figure 6). NewsStand does not have a “public” API, though much of its functionality and ability to handle different smartphone platforms makes use of its “private” API.

Differences between the browser-

based Web environment and native app environment for mobile devices require changes in user behavior or habits. For example, map-centric applications on the Web function best as single-page applications, meaning external links (such as to news articles in NewsStand) are opened in separate browser tabs to preserve the NewsStand App and its state, which would not occur if the news articles would be opened in the same tab. A concrete example of the undesired ramification of opening the external link in a separate browser tab is that users cannot make use of the “back” button to return to the app and its prior state. Instead, they must explicitly close the newly opened tab, in which case the invoking tab and its state are implicitly restored. Such problems do not arise in the native app environment, which can coordinate fluid transitions among many windows, thereby providing more user-friendly interaction, with the trade-off, in our example, that only one external link to a news article can be opened at a time.

Porting NewsStand to a variety of mobile/smartphone platforms revealed a lack of adherence to classical cartographic principles in the implementations of the underlying mapping APIs. As a result, consistency issues surfaced for some operations (such as zooming

and panning). For example, once the name of a location appears in the map, that name should continue to be present as long as the location remains in the window as the user zooms in further or pans.<sup>40</sup> Curiously, some mapping apps on mobile and smartphone platforms do not enable zooming out so the entire world can be seen on the screen (such as in the Google Maps and Apple Maps mobile/smartphone mapping APIs), thereby requiring further panning to see the rest of the world, though it is present in the “here” Maps API.<sup>40</sup> This phenomenon is especially annoying in NewsStand where users want to see what is happening in the whole world.<sup>40</sup> Minimaps alleviate the problem via, in part, the orange balls showing all other locations mentioned in a particular article highlighted with a headline info bubble.

We had to account for not being able to hover in devices that make use of a gesturing interface when designing the user interface, as it means some features would have to be implemented differently on gesturing-enabled platforms. In particular, hovering enables the user to observe the spatial variability of phenomena being displayed, or expanded, as the pointing device passes over the location. The gesturing interface requires a tap or

Figure 6. NewsStand App screenshots for (a) iPhone, (b) Android, and (c) Windows Phone platforms.




click for such a display action to take place since a continuous motion of a finger over an area is interpreted as a single tap or click, and it is thus difficult to observe spatial variability. On the other hand, the absence of hovering means a transition from map location  $l$  to another map location  $b$  can be made by tapping on  $b$ . In contrast, the hovering needed to transition from  $l$  to  $b$  may lead to certain actions being taken that would destroy the current state of the system.

The design challenge of speedy map labeling involves placement of the minimap in close proximity to the headline bubble and associated info box. It also arises in dynamic display of labels (such as disease names in Figure 3 and keywords and names of people and brands). Our goal is to do so at interactive speeds under panning and zooming, achieving it through techniques developed for dynamic map labeling<sup>30</sup> and incorporated in the PhotoStand system.<sup>37</sup>


### Conclusion

We reviewed the design goals and functionality of the NewsStand system for using a map to read news on the Web, harnessing the power of spatial synonyms. NewsStand demonstrates that extracting geographic content from news articles taps a previously unseen dimension of information that can aid understanding news events across space and time. NEWS can indeed be described as an acronym of north, east, west, south. The increasing popularity of geotagged content on the Internet will enable compelling applications for systems like NewsStand in other knowledge domains. For example, sentiment/content analysis can reveal how the same news story can be interpreted by people in different countries or in different languages and hot-spot analysis based on news, tweets, or other sources of data feeds. Moreover, NewsStand represents a contribution to the emerging field of computational journalism.<sup>8</sup>

Future work includes using a map-query interface to access other media through representative images (such as PhotoStand<sup>37</sup>), videos, and audio clips. We are also working on incorporating other sources of news and information. For example, we have in-




**NewsStand currently indexes 10,000 news sources and processes approximately 50,000 news articles per day.**



corporated Twitter tweets into NewsStand, resulting in the TwitterStand system<sup>44</sup> where the idea is to tap the large volume of news articles to serve as a kind of clustering corpus so very short and information-sparse tweets can be clustered using existing news clusters. An interesting aspect of this method is that the tweets, due to their short length, usually have little or no geographic content but, when clustered, inherit the geographic information associated with the geographic focus of the cluster with which they are associated. The novel result is the focus is now on the geographic regions about which a user is tweeting rather than on the geographic regions from which the user is tweeting (easy to find when the tweeting device has GPS capability). This focus is useful when tweeting about future events,<sup>14</sup> but one must be careful in choosing whose tweets to follow.<sup>11</sup>

### Acknowledgments

This article is based on an earlier paper by Teitler et al.<sup>46</sup> This work was supported in part by the National Science Foundation under Grants IIS-07-13501, IIS-08-12377, CCF-08-30618, IIS-10-18475, IIS-12-19023, and IIS-13-20791, as well as by the Office of Policy Development & Research of the U.S. Department of Housing and Urban Development, Microsoft Research, Google Research, Nvidia, the E.T.S. Walton Visitor Award of the Science Foundation of Ireland, and the National Center for Geocomputation at the National University of Ireland at Maynooth. We also thank Larry Brandt, Jim Gray, Keith Marzullo, and Maria Zeman kova for championing it. 

### References

1. Adelfio, M.D. and Samet, H. Structured toponym resolution using combined hierarchical place categories. In *Proceedings of the Seventh ACM SIGSPATIAL Workshop on Geographic Information Retrieval* (Orlando, FL, Nov. 5). ACM Press, New York, 2013, 49–56.
2. Amitay, E., Har'El, N., Sivan, R., and Soffer, A. Web-where: Geotagging Web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, U.K., July 25–29). ACM Press, New York, 2004, 273–280.
3. Aref, W.G. and Samet, H. Efficient processing of window queries in the pyramid data structure. In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (Nashville, TN, Apr. 2–4). ACM Press, New York, 1990, 265–272.
4. Baldwin, B. and Carpenter, B. *Lingpipe*. <http://alias-i.com/lingpipe/>
5. Chum, O., Philbin, J., Isard, M., and Zisserman, A.



Scalable near-identical image and shot detection. In *Proceedings of the Sixth ACM International Conference on Image and Video Retrieval* (Amsterdam, The Netherlands, July 9–11). ACM Press, New York, 2007, 549–556.

6. Ding, J., Gravano, L., and Shivakumar, N. Computing geospatial scopes of Web resources. *Proceedings of the 2006 International Conference on Very Large Data Bases* (Cairo, Egypt, Sept. 10–14). Morgan Kaufmann, San Francisco, 2000, 545–556.
7. Duda, R.O., Hart, P.E., and Stork, D.G. *Pattern Classification, Second Edition*. Wiley Interscience, New York, 2000.
8. Essa, I. *Computation + Journalism: A Study of Computation and Journalism and How They Impact Each Other*. <http://www.computation-and-journalism.com/>
9. Francis, W.N. A standard corpus of edited present-day American English. *College English* 28(6) (Jan. 1965), 267–273.
10. Freifeld, C.C., Mandl, K.D., Reis, B.Y., and Brownstein, J.S. HeatMap: Global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association* 15, 2 (Mar. 2008), 150–157.
11. Gramsky, N. and Samet, H. Seeder finder: Identifying additional needles in the Twitter haystack. In *Proceedings of the Fifth ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (Orlando, FL, Nov. 5). ACM Press, New York, 2013, 44–53.
12. Hjaltason, G.R. and Samet, H. Speeding up construction of PMR quadtree-based spatial indexes. *Very Large Data Bases Journal* 11, 2 (Oct. 2002), 109–137.
13. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenauf, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Edinburgh, Scotland, July 27–31). Association for Computational Linguistics, Stroudsburg, PA, 2011, 782–792.
14. Jackoway, A., Samet, H., and Sankaranarayanan, J. Identification of live news events using Twitter. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (Chicago, Nov. 1). ACM Press, New York, 2011, 25–32.
15. Lafferty, J.D., McCallum, A., and Peireira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning* (Williamstown, MA, June 28–July 1). Morgan Kaufman, San Francisco, 2001, 282–289.
16. Lan, R., Lieberman, M.D., and Samet, H. The picture of health: Map-based, collaborative spatio-temporal disease tracking. In *Proceedings of the First ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health* (Redondo Beach, CA, Nov. 6). ACM Press, New York, 2012, 27–35.
17. Leidner, J.L. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland, U.K., Oct. 2006. <https://www.era.lib.ed.ac.uk/bitstream/1842/1849/1/leidner-2007-phd.pdf>
18. Lieberman, M.D. and Samet, H. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th International Conference on Research and Development in Information Retrieval* (Beijing, July 24–28). ACM Press, New York, 2011, 843–852.
19. Lieberman, M.D. and Samet, H. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval* (Portland, OR, Aug. 12–16). ACM Press, New York, 2012, 731–740.
20. Lieberman, M.D. and Samet, H. 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* (Redondo Beach, CA, Nov. 7–9). ACM Press, New York, 2012, 179–188.
21. Lieberman, M.D., Samet, H., and Sankaranarayanan, J. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *Proceedings of the Sixth Workshop on Geographic Information Retrieval* (Zürich, Switzerland, Feb. 18–19). ACM Press, New York, 2010.
22. Lieberman, M.D., Samet, H., and Sankaranarayanan, J. Geotagging with local lexicons to build indexes for textually specified spatial data. In *Proceedings of the 2006 IEEE International Conference on Data Engineering* (Long Beach, CA, Mar. 1–6). IEEE Press, 2010, 201–212.
23. Lieberman, M.D., Samet, H., Sankaranarayanan, J., and Sperling, J. STEWARD: Architecture of a spatio-textual search engine. *Proceedings of 15th ACM International Symposium on Advances in Geographic Information Systems* (Seattle, Nov. 7–9). ACM Press, New York, 2007, 186–193.
24. Lieberman, M.D., Sankaranarayanan, J., Samet, H., and Sperling, J. Augmenting spatio-textual search with an infectious disease ontology. *Proceedings of the Workshop on Information Integration Methods, Architectures, and Systems* (Cancun, Mexico, Apr. 11–12). IEEE Computer Society, 2008, 266–269.
25. Lowe, D.G. Object recognition from local scale-invariant features. In *Proceedings of the Seventh International Conference on Computer Vision* (Corfu, Greece, Sept. 20–25). IEEE Computer Society, 1999, 1150–1157.
26. Markowetz, A., Brinkhoff, T., and Seeger, B. Exploiting the Internet as a geospatial database. *Proceedings of the Workshop on Next Generation Geospatial Information* (Cambridge, MA, Oct. 19–21, 2003).
27. Martins, B., Mangunifhas, H., Borbinha, J., and Siabato W. A geo-temporal information extraction service for processing descriptive metadata in digital libraries. *e-Perimeteron* 4, 1 (2009), 25–37.
28. Mehlher, A., Bao, Y., Li, X., Wang, Y., and Skiena, S. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics* 18 (Sept.–Oct. 2006), 765–772.
29. Milne, D. and Witten, I.H. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, CA, Oct. 26–30). ACM Press, New York, 2008, 509–518.
30. Nutanong, S., Adelfio, M.D., and Samet, H. Multiresolution select-distinct queries on large geospatial point sets. *Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Redondo Beach, CA, Nov. 7–9). ACM Press, New York, 2012, 159–168.
31. Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Buchher, B., Finch, D., Fu, G., Jo, H., Syed, A.K., Vaid, S., and Yang, B. The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Systems* 21, 7 (2007), 717–745.
32. Quercini, G., Samet, H., Sankaranarayanan, J., and Lieberman, M.D. 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* (San Jose, CA Nov. 3–5). ACM Press, New York, 2010, 43–52.
33. Rauc, E., Bukatin, M., and Baker, K. A confidence-based framework for disambiguating geospatial terms. In *Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References* (Edmonton, Canada). Association for Computational Linguistics, Stroudsburg, PA, 2003, 50–54.
34. Rizzo, G. and Troncy, R. NERD: A framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France, Apr. 23–27). Association for Computational Linguistics, Stroudsburg, PA, 2012, 73–76.
35. Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513–523.
36. Samet, H. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, San Francisco, 2006.
37. Samet, H., Adelfio, M.D., Fruin, B.C., Lieberman, M.D., and Sankaranarayanan, J. PhotoStand: A map query interface for a database of news photos. *Proceedings of the VLDB Endowment* 6, 12 (Aug. 2013), 1350–1353.
38. Samet, H., Adelfio, M.D., Fruin, B.C., Lieberman, M.D., and Teitler, B.E. Porting a Web-based mapping application to a smartphone app. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Chicago, Nov. 2–4). ACM Press, New York, 2011, 525–528.
39. Samet, H., Alborzi, H., Brabec, F., Esperança, C., Hjaltason, G.R., Morgan, F., and Tanin, E. Use of the SAND spatial browser for digital government applications. *Commun. ACM* 46(6) (Jan. 2003), 63–66.
40. Samet, H., Fruin, B.C., and Nutanong, S. DUKing it out at the smartphone mobile app mapping API corral: Apple, Google, and the competition. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems* (Redondo Beach, CA, Nov. 6). ACM Press, New York, 2012, 41–48.
41. Samet, H., Rosenfeld, A., Shaffer, C.A., and Webber, R.E. A geospatial information system using quadtrees. *Pattern Recognition* 17, 6 (Nov./Dec. 1984), 647–656.
42. Samet, H., Teitler, B.E., Adelfio, M.D., and Lieberman, M.D. Adapting a map query interface for a gesturing touchscreen interface. In *Proceedings of the 20th International World Wide Web Conference* (Hyderabad, India, Mar. 28–Apr. 1). ACM Press, New York, 2011, 257–260.
43. Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M.D., and Sperling, J. TwitterStand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Seattle, Nov. 4–6). ACM Press, New York, 2009, 42–51.
44. Sarma, A.D., Lee, H., Gonzales, H., Madhavan, J., and Halevy, A. Efficient spatial sampling of large geospatial tables. *Proceedings of the ACM SIGMOD Conference* (Scottsdale, AZ, May 20–24). ACM Press, New York, 2012, 193–204.
45. Stokes, N., Li, Y., Moffat, A., and Rong, J. An empirical study of the effects of NLP components on geographic IR performance. *International Journal of Geographical Information Systems* 22, 3 (Mar. 2008), 247–264.
46. Teitler, B., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. NewsStand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Irvine, CA, Nov. 5–7). ACM Press, New York, 2008, 144–153.
47. Zhou, G. and Su, J. Named entity recognition using an HMM-based chunk tagger. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, PA, July 6–12). Association for Computational Linguistics, Stroudsburg, PA, 2002, 473–480.

**Hanan Samet** (hjs@cs.umd.edu) is a distinguished university professor in the Computer Science Department, Center for Automation Research, and Institute for Advanced Computer Studies at the University of Maryland, College Park, MD.

**Jagan Sankaranarayanan** (jsagan@gmail.com) is a member of the research staff at NEC Labs, Cupertino, CA; his research for this article was conducted while he was an assistant research scientist at the Institute for Advanced Computer Studies at the University of Maryland, College Park, MD.

**Michael D. Lieberman** (mike.d.lieberman@gmail.com) is a research physicist scientist at the Johns Hopkins University Applied Physics Laboratory, Laurel, MD; his research for this article was part of his Ph.D. in computer science at the University of Maryland, College Park, MD.

**Marco D. Adelfio** (marco@cs.umd.edu) is a Ph.D. candidate in computer science at the University of Maryland, College Park, MD.

**Brendan C. Fruin** (bcfruin@gmail.com) is a software engineer at Zillow, Seattle, WA; his research for this article was part of his master's in computer science at the University of Maryland, College Park, MD.

**Jack M. Lotkowski** (JackLotkowski@gmail.com) is an undergraduate student at the University of Maryland, College Park, MD.

**Daniele Panozzo** (daniele.panozzo@gmail.com) is a senior researcher at ETH, Zürich, Switzerland; his research for this article was conducted while he was a visiting student at the Institute for Advanced Computer Studies at the University of Maryland, College Park, MD.

**Jon Sperling** (jonxsperling@gmail.com) is a senior researcher in the Office of Policy Development and Research of the U.S. Department of Housing and Urban Development, Washington, D.C.

**Benjamin E. Teitler** (bteitler@cs.umd.edu) did research for this article as part of his master's in computer science at the University of Maryland, College Park, MD.