Reading News with Maps: The Power of Searching with Spatial Synonyms*

Hanan Samet† hjs@cs.umd.edu
Benjamin E. Teitler† bteitler@cs.umd.edu
Michael D. Lieberman† codepoet@cs.umd.edu
Jagan Sankaranarayanan† jagan@cs.umd.edu
Daniele Panozzo† daniele@cs.umd.edu
Jon Sperling‡ Jon.Sperling@hud.gov

ABSTRACT

The NewsStand system is an example application of a general framework that we are developing to enable people to search for information using a map query interface, where the information results from monitoring the output of over 8,000 RSS news sources and is available for retrieval within minutes of publication. The advantage of doing so is that a map, coupled with an ability to vary the zoom level at which it is viewed, provides an inherent granularity to the search process that facilitates an approximate search. This distinguishes it from today's prevalent keyword-based conventional search methods that provide a very limited facility for approximate searches which are realized primarily by permitting a match via use of a subset of the keywords. However, it is often the case that users do not have a firm grasp of which keyword to use, and thus would welcome the capability for the search to also take synonyms into account. In the case of queries to spatially-referenced data, the map query interface is a step in this direction as the act of pointing at a location (e.g., by the appropriate positioning of a pointing device) and making the interpretation of the precision of this positioning specification dependent on the zoom level is equivalent to permitting the use of spatial synonyms. The issues that arise in the design of such a system including the identification of words that correspond to geographic locations are discussed, and examples are provided of the utility of the approach, thereby representing a step forward in the emerging field of computational journalism.

1. INTRODUCTION

Do you travel? Do you want to know what is going on in the town you are traveling to? Do you want to keep up with the latest news in the town you have left, especially when it is your own hometown? If your answer was YES to any of these questions (and who wouldn't :-)?), then NewsStand (denoting Spatio-Textual Aggregation of News and Display), as well as related systems, developed by us, are for you.

NewsStand is an example application of a general framework that we are developing to enable people to search for information using a map query interface. The advantage of doing so is that a map, coupled with an ability to vary the zoom level at which it is viewed, provides an inherent granularity to the search process that facilitates an approximate search. This distinguishes it from today's prevalent keyword-based conventional search methods that provide a very limited facility for approximate searches which are realized primarily by permitting a match via use of a subset of the keywords. However, it is often the case that users do not have a firm grasp of which keyword to use, and thus would welcome the capability for the search to also take synonyms into account. In the case of queries to spatially-referenced data (termed *spatial queries* to *spatial data*), the map query interface is a step in this direction as the act of pointing at a location (e.g., by the appropriate positioning of a pointing device) and making the interpretation of the precision of this positioning specification dependent on the zoom level is equivalent to permitting the use of spatial synonyms.

The ability to use spatial synonyms is extremely important as it enables us to search for data when we are not exactly sure what we are seeking, or what the answer to our query should be. For example, suppose that our query seeks a "Rock Concert in Manhattan". The presence of "Rock Concerts" in Harlem or New York City are good answers when no such events can be found in Manhattan, as they correspond to approximate synonyms: Harlem by virtue of proximity and New York City by virtue of a containment relationship. Conventional search engines that deploy techniques such as the page rank method [5] are good at finding documents containing keywords that we are looking for, but they cannot be easily modified to handle the above query. Moreover, their primary utility is based on grounds of popularity in the sense that the page rank algorithm ensures that the web pages provided to the user as part of the response are ordered by a measure that incorporates some aspect related to their frequency, thereby ensuring that 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 are treated equally. A cruder way of describing the resulting effect is that it doesn't discriminate between users in the sense that they all get the same bad (or good!) answers. In other words, the effect of using the page rank algorithm to order the results (thereby effectively choosing which results to present to the user) is that if nobody ever looked for some data before or linked to it,

*This work was supported in part by the National Science Foundation under Grants IIS-08-12377, CCF-08-30618, and IIS-07-13501, as well as the Office of Policy Development & Research of the Department of Housing and Development, Microsoft Research, Google, 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.

†Department of Computer Science, Center for Automation Research, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

‡HUD Office of Policy Development & Research (PD&R), 451 7th St. SW, Room 8146, Washington, DC 20410, USA.

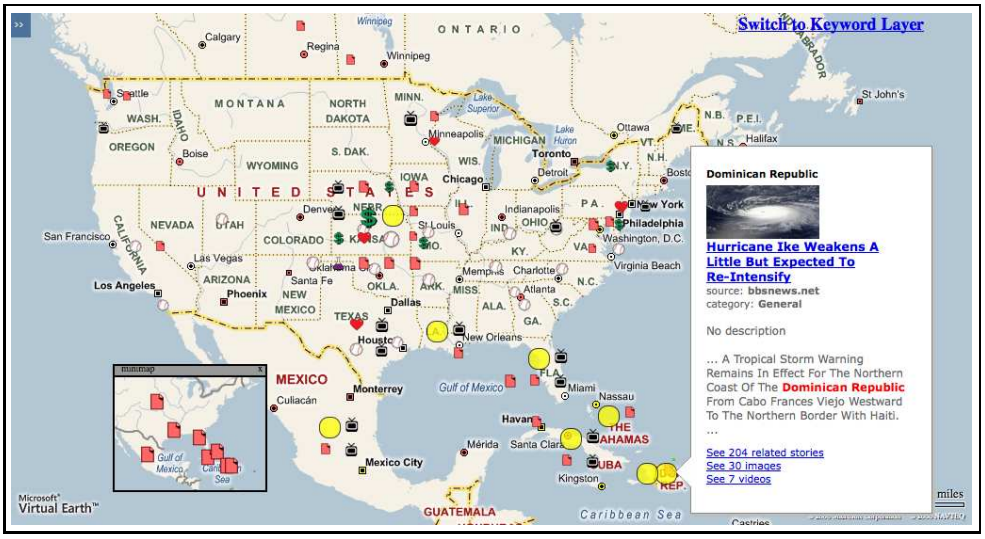

Figure 1: A screenshot of NewsStand's output to the "What is happening at location X?" query, where X is the United States and Gulf of Mexico regions, and showing an article about Hurricane Ike affecting the Caribbean and Gulf of Mexico in September 2008. The highlighted symbols displayed on the large map and the minimap correspond to all locations mentioned in the article. The info bubble shows text tying the article to the highlighted geographic location. Notice that the highlighted symbols correspond to the path traveled by Hurricane Ike. NewsStand's interface is accessible at `http://newsstand.umiacs.umd.edu/`.

then it will never be found and, hence, never presented to the user. In some cases, this is fine. However, in the case of synonyms, this has a strong negative effect on the quality of the search results as it means that if nobody linked to similar pages on account of their content being equivalent but for the use of the same words, then the similarity will never be found by the search engine as the page ranking algorithm will never be able to find the similar pages as it crawls the web when building the index to the web pages.

NewsStand and the related systems that we have built address the synonym problem for spatial queries. The key issue here can be seen by noting that all spatial queries can be broken down into the following two classes:

1. 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.

2. 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 where one function is the inverse of the other. Feature-based queries are also known as spatial data mining [2, 17].

Although features are usually properties (also known as *attributes*) of spatially-referenced data such as crop types, soil types, zones, speed limits, etc., both they and the underlying spatially-referenced data domain can be more broadly interpreted. In particular, NewsStand applies these concepts to the domain of unstructured data consisting of collections of news articles with textually-specified locations and where the features are the topics. In this case, a location-based query returns all topics/articles mentioning a specific place or region X, while a feature-based query returns all places/regions mentioned in articles about topic T, or just article Y. It is important to note that NewsStand does not require T to be known a priori, in which case the topics are ranked by importance, which can be defined by a number of criteria including, but not limited to, the number of articles which comprise them. Thus, a typical pair of queries is:

1. Location-based—"What is happening at location X?"

2. Feature-based—"Where is topic T or article Y happening?"

Figure 1 displays a screenshot of NewsStand's output from the "What is happening at location X query?," where X is the United States and Gulf of Mexico regions. Each symbol, which we term a *marker*, represents a set of articles about a particular topic associated with the corresponding location on the map. The type of the symbol conveys information about the news category in which the topic falls (e.g., news, sports, entertainment, business, science, health, etc.). Hovering the mouse cursor on a topic symbol causes a small info bubble to appear, populated with an overall summary of a representative article on the topic, which, in this case, is an article about Hurricane Ike in September 2008. Clicking on a symbol causes all symbols on the visible map that are also associated with the topic to be highlighted in yellow, which in this case enables us to see the path traveled by the Hurricane thereby pinpointing some of the affected islands in the Caribbean and Gulf of Mexico, as well as the Gulf Coast of the USA. The result is a variant of spatial data mining yielding a form of knowledge discovery. NewsStand also features a smaller map that shows the geographic span of the selected article. This minimap allows users to easily see the selected article's geographic focus, without having to leave their area of interest on the main map, and is independent of the current level of zoom, which may precluded them from being highlighted on the part of the map that is visible.

Figure 2 displays NewsStand's output from the "Where is topic T or article Y happening?, " where T is the visit of President Obama to Copenhagen in October 2009 to lobby the International Olympics Committee on behalf of Chicago. On the left we see a number of topics. Hovering the mouse cursor on a topic (left pane) causes appropriate symbols to appear on the map (right pane) at the principal geographic

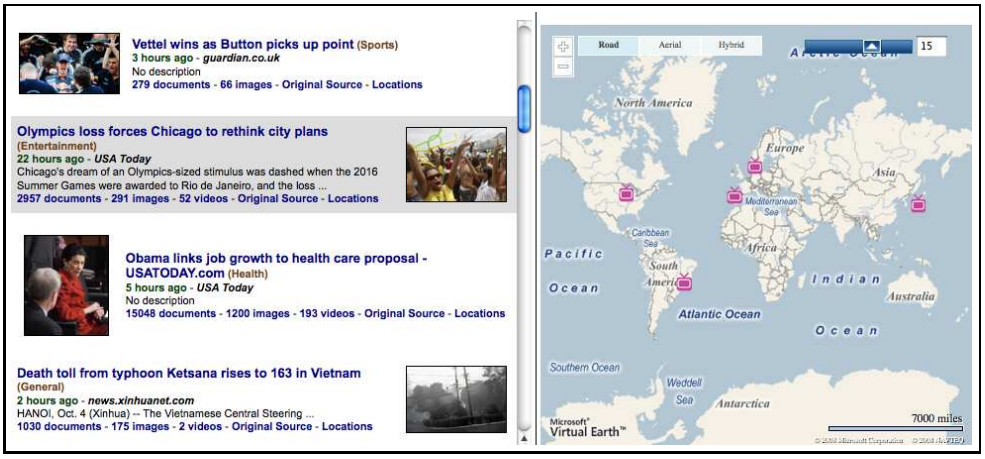

Figure 2: A screenshot of NewsStand's output to the "Where is topic T or article Y happening?, " where T is the visit of President Obama to Copenhagen in October 2009 to lobby the International Olympics Committee on behalf of Chicago. The symbols displayed on the map in the right are positioned at the principal geographic locations associated with this topic. This allows us to easily see the cities under consideration which in addition to Chicago were Madrid, Rio de Janeiro, and Tokyo. The info bubble shows text tying the article to the highlighted geographic location.

locations associated with this topic. This action allows us to easily see the cities under consideration which in addition to Chicago were Madrid, Rio de Janeiro, and Tokyo. Again, the result is a variant of spatial data mining yielding a form of knowledge discovery. Hovering the mouse cursor on the map causes info bubbles to appear as in the "What is happening at X?" query in Figure 1.

The execution of these queries is facilitated by building an index on the spatial data. These indexes are relatively easy to construct when both the spatial and feature data are specified geometrically and numerically, respectively. However, this is not the case in our application as all of the data is unstructured. In particular, both the location and feature data are just collections of words of text that can be (but are not required to be) interpreted as the names of locations in the case of spatial data. In other words, the spatial data is specified using text (called *toponyms*) rather than geometry, which means that there is some ambiguity involved. This ambiguity has both advantages and disadvantages. The advantage of the ambiguity 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 which serves as the cornerstone of inheritance in object-oriented programming languages [40]). For example, geometrically, a city can be 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. On the other hand, the disadvantage of the ambiguity is that we are not always sure if a term is a geographic location or not (e.g., does "Jordan" refer to a country or is it a surname as in "Michael Jordan"?). 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 (e.g., does "London" refer to an instance in the UK, Ontario, Canada, or one of many more others?). Resolving these ambiguities with no errors (or almost none) is one of the main technical challenges in the successful deployment of NewsStand and the related systems.

The rest of this paper is organized as follows. Section 2 presents NewsStand's architecture, as well as positions it in the context of existing alternative approaches to associating news articles with geographic locations. Section 3 discusses *geotagging* (also known as *geoparsing*), the process of determining words in text that correspond to locations, and also outlines how it is done in NewsStand. Section 4 indicates how NewsStand groups similar articles into clusters, thereby resulting in a form of topic detection, in an online manner as the collection of articles is constantly changing. Note that this process must also determine a geographic cluster focus, achieved by a process that takes into account the geographic focis of the individual articles making up each cluster. Section 5 describes the user interface which is the raison d'etre for this work as well as some of the issues that we faced in the display. Concluding remarks are drawn in Section 6.

2. UNDERSTANDING NEWS

The key elements to understanding news have perhaps been best captured by Rudyard Kipling in 1902 who said in *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*.

Given an event, these six "honest serving-men" enable us to determine what happened, why it happened, how it happened, who made it happen, when it happened (with freshness being an important factor in making it news), and, perhaps most importantly, where it happened. Together, the output of these six "honest serving-men forms the cornerstones of a well-written, comprehensible, timely, and relevant news article. The relevancy is in large part a result of the article having a geographic focus, thereby usually emphasizing the "Where" component, and thus reporting events in a certain geographic region. Some related key questions include where are the top stories (i.e., topics) and how do we find them. In particular, we want to know what is happening around the world and to be able to tunnel down (i.e., using zooming) to specific areas such as South Asia, the India-Pakistan border, as well as down to a specific neighborhood such as the one from which the reader hails. However, popular news aggregators such as Google News, Yahoo! News,

Figure 3: A screenshot of NewsStand showing the keywords in news clusters for Europe and the Mediterranean area on a day in 2008. It allows users to gain an overall understanding of the top topics without needing to hover on individual markers. However, doing so will result in the display of a snippet from an article that is a member of the topic cluster.

and Microsoft Bing News have only a rudimentary understanding of the implicit geographic content of news articles, usually based on the address of the publishing news source (e.g., newspaper). Furthermore, these systems group articles by keyword or topic, rather than by geography. On the other hand, the output of NewsStand, instead, can be summarized as using "What" and "When" to identify "Where" and, to a lesser extent in terms of our emphasis, "Who".

NewsStand gathers its data by crawling the web. Its primary source of data are thousands of individual news sources from all over the world in the form of *Really Simple Syndication* (RSS) feeds. RSS is a widely-used XML protocol for online publication and is ideal for NewsStand, as it requires at least a title, short description, and web link for each published news item. RSS 2.0 also allows an optional publication date, which helps determine the age or "freshness" of articles. NewsStand currently indexes 6,500 news sources, and processes about 60,000 news articles per day. It determines the geographic locations mentioned in the article (termed *geotagging*) and also tries to determine the article's geographic focus or foci (i.e., the key locations in the article). In addition, it aggregates news articles by topic based on content similarity (termed *clustering*) so that articles about the same event are grouped into the same cluster. It ranks the clusters based on its notion of "importance", which is determined by factors such as:

1. The number of articles in the cluster.
2. The number of unique news sources in the cluster. For example, an event in Irvine, CA is important if carried by multiple news sources, especially if some of them are geographically distant from Los Angeles (which is approximately 50 miles away from Irvine, CA).
3. The topic's rate of propagation. In particular, articles about important events will be picked up by multiple news sources within a short time period.

NewsStand also tries to determine the cluster geographic focus or foci and associates the cluster with them. The latter is aided by making use of the clustering process vis-a-vis the location feature. Each cluster is displayed at the positions of its geographic foci, which is usually done with the aid of symbols corresponding to its news category as in Figure 1. However, instead of displaying the category symbol associated with the cluster,, we can also display the text corresponding to the most prevalent term in the cluster, called the *keyword* (e.g., Figure 3 which captures the keywords for a day in 2008 for Europe and the Mediterranean area).

Scalability and fast processing of individual articles are the most important criteria in designing NewsStand's architecture, which is shown in Figure 5. Additional goals include presenting the latest news as quickly as possible, within minutes of its online publication, and being robust to failure. These criteria are fulfilled by subdividing NewsStand's collection and processing into several modules, each of which can run independently on separate computing nodes in a distributed computing cluster. From the figure we see that 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. In addition, the modules are designed in such a way that allows for multiple instances of any module to run simultaneously on one or more nodes. NewsStand is therefore able to execute as many instances of modules as required to handle the volume of news that is received. Each module receives input and sends output to a transactional database system that serves as a synchronization point. Using transactions, the database ensures that the overall system state changes atomically and is never internally inconsistent. Furthermore, the database system can be replicated across multiple nodes as necessary to handle increased system load. NewsStand uses the PostgreSQL database package for these purposes.

In addition to individual processing modules, NewsStand makes use of a specially-created master controller module to orchestrate the entire system. The controller module's responsibility is to delegate articles to be processed to the other modules in the system that function as slave nodes. The controller maintains its own collection of database tables that track an article as it moves through the system, as well as the pool of connected slaves. A simple communication protocol allows the master and slaves to send several control messages for assigning work and reporting success or failure. Upon creation, slave modules connect to the master and initiate a handshake that announces the slave's presence and in what role the slave will function. The master then assigns several articles to be processed to the slave and waits for a return message indicating success or failure. If no such response is received after a set time limit, then the master assumes that the slave somehow failed. The master

Figure 4: A screenshot of NewsStand showing the most common terms in each cluster that correspond to the name of the person important to the news topic on a day in October 2009.

then requires the failed slave to resend the handshake before it will delegate additional work to that slave.

NewsStand's goal is to change the news reading process and, most importantly, experience. In particular, users query it by choosing a region of interest and finding topics/articles relevant to it. The topics/articles that are displayed are determined by the location and level of zoom which together dictate the spatial scope (i.e., the region of interest). There are two ways of interpreting the notion of "region of interest". One is in terms of content, while the second is in terms of the news sources. In particular, in the simplest case, there are no predetermined boundaries on the locations of the news sources for the articles that are displayed for the region of interest. In the second case, the sources can be limited to a subset of the available sources (e.g., the New York Times and the Washington Post), or they can also be limited by spatial region, which can be specified textually (e.g., restrict the sources to lie in Ireland), or by drawing the region of interest on the map (e.g., a box overlapping both Ireland and the United Kingdom). Of course, we can also constrain both the content and the news sources, and 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, we may want to see how the English press views/interprets developments in the Middle East. The result is somewhat analogous to sentiment analysis [49]. Some other applications include monitoring hot spots, which is useful for investors, national security, and to keeping up with spread of diseases (e.g., [24]).

As we stated at the outset of the paper, the ultimate goal of NewsStand is to make the map as the medium for the presentation of information that has spatial relevance and thus it is not restricted to news articles—that is, it can also be applied to search results, photos, videos, etc. In addition, NewsStand enables both a summary of the news as well as further exploration and even knowledge acquisition via discovery of patterns in the news, which is a direct result of the association of topics or categories with the locations that are mentioned in their constituent articles. For example, we can also compute a cluster people focus which is the most common term in the cluster which corresponds to a name of a person (e.g., Figure 4 for Europe and the Mediterranean area on a day in October 2009). Similarly, this can be done for diseases where now we have a cluster disease focus which is the most common term in the cluster which corresponds to the name of a disease (e.g., Figure 7 for the US on a day in October 2009).

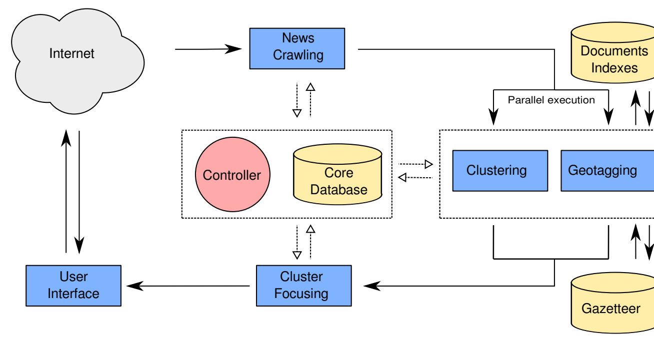

Figure 5: A high level overview diagram of NewsStand's architecture. The system is designed as a pipeline, with individual processing modules working independently. A central control module orchestrates article processing by delegating work to the other modules and tracking articles in the pipeline.

At this point, it is useful to compare NewsStand with existing news readers. News reading systems such as Microsoft Bing News and Google News present the news in the classical linear fashion with aggregation of different sources for each topic. Google NewsMicrosoft Bing News [29], and Yahoo! News [48] all have some aspect of locality in that there exists some aggregation of articles/topics that are relevant to the user's locality. This is usually done according to a ZIP/postal code or city-state specification. For example, these could be the topics that mention "College Park, MD". In the case of Google, this feature seems to be implemented, at least as far as we can tell, by applying Google Search with the location names as the search keys For example, after determining that the user is in ZIP code 20742 (e.g., by virtue of the IP address of the user, absent an alternative specification of the local area), Google local returns the articles that mention "College Park, MD" or the "University of Maryland" as they are known to be associated with this ZIP code. In addition, the resulting list of articles also appears to be based primarily on the location of the news source (usually a newspaper), rather than on story content. In these

Figure 6: Locations mentioned in news articles about the May 2009 swine flu pandemic, obtained by geotagging related news articles. Large red circles indicate high frequency, and small circles are color coded according to recency, with lighter colors indicating the newest mentions.

cases, the number of topics that are displayed is limited to a small number, although there is no particular reason for this save the absence of topics relevant to the user's locality. It is important to note that there is no notion of article importance in determining what is shown to the user.

Recently, individual news services such as Reuters [43] have started to provide a map with each article where geographic locations associated with top articles are highlighted, and users can view the associated articles by clicking on the corresponding location. The locations are determined with the aid of the MetaCarta system (described below), but, unfortunately, the map display is static in the sense that it presents all of the top articles in its collection at just one view (i.e., level of zoom). It is interesting to note that although the mapping platforms that are used do provide the ability to zoom in, the zoom capability is not coupled with an ability to obtain more articles that are commensurate with an increased level of zoom. Moreover, the locations that are associated with the articles are usually the datelines of the article. Another system known as the AP Mobile News Network [42] determines even coarser geography, based on the wire service city where the article was filed. For example, an article submitted to the Maryland news wire would be listed for all postal codes in Maryland. Thus, unlike NewsStand, there does not appear to be an attempt in the AP Mobile News Network to analyze the individual articles to determine what is the main associated location (i.e., the geographic focus) as well as other important locations.

3. GEOTAGGING

NewsStand extracts geographic locations from news articles (termed *geotagging*) which is related to work in *geographic information extraction*. Geotagging is a powerful tool in understanding spatial and temporal properties of news events. For example, Figure 6 illustrates worldwide outbreaks of swine flu in May 2009, obtained by geotagging news articles written about it, which can then be indexed spatially. Large red circles indicate high frequency, and small circles are color coded according to recency, with lighter colors indicating the most recent mentions.

Much of the existing work on geographic information extraction deals with finding the *geographic scope* of websites and individual documents. In the context of news articles, we distinguish between three types of geographic scope [25, 46]:

1. *Provider scope*, the publisher's geographic location;
2. *Content scope*, the article/topic content's geography; and
3. *Serving scope*, based on the readers' location.

NewsStand relies on article content to determine the article's geographic scope, and, as we shall point out later, also tries to make use of the provider scope, which it knows, and the serving scope, which it attempts to learn. Other approaches [7, 8, 11, 27, 51], instead, use the link structure of inbound and outbound links in the article. This solution, also used by search engines, may not be suitable for news articles as well as the *hidden web*, a set of documents intended for internal use in an organization, which typically have few links.

NewsStand extends our earlier work on geotagging in STEWARD (denoting *Spatio-Textual Extraction on the Web Aiding the Retrieval of Documents*) [23] 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. In particular, STEWARD processes each document independently of all other documents, while NewsStand takes advantage of multiple versions and instances of articles about a particular topic by grouping these articles, most often from different news sources, into topic clusters. These clusters allow for improved geotagging, and let users retrieve related articles with ease.

Geotagging consists of two processes: toponym recognition and toponym resolution. The main issue in toponym recognition is *geo/non-geo ambiguity*, where a given phrase might refer to a geographic location, or some other kind of entity—e.g., deciding whether a mention of "Washington" refers to a location or some other entity such as a person's name. A secondary issue is *aliasing*, where multiple names refer to the same geographic location—e.g., "Los Angeles" and "LA"; The main issue in toponym resolution, also known as *geographic name ambiguity* or *polysemy*, is *geo/geo ambiguity*, where a given name might refer to any of several geographic locations—e.g., "Springfield" is the name of many cities in the USA such as in Massachusetts and also the capital of the state of Illinois.

Many different approaches to toponym recognition have been undertaken, although they share similar characteristics. In essence, the idea is to extract the "interesting" phrases, which are the ones 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)* The most common strategy for identifying the *EFV* is simply to look for phrases in the document that exist in a *gazetteer*, or database of geographic names and locations , and many researchers have used this as their primary strategy [1, 30, 35, 44, 45]. In particular, Web-a-Where [1] uses a small, well-curated gazetteer of about 40,000 locations, created by collecting the names of countries and cities with populations greater than 5,000. This small 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 that are commonplace in articles from local news sources. Furthermore, the small gazetteer means that Web-a-Where is more prone to making toponym recognition errors because it misses out on being aware of geo/non-geo ambiguity which is afforded by use of larger gazetteers.

To deal with the geo/non-geo ambiguity inherent in larger gazetteers, researchers [16, 26, 32, 36, 37, 39] have proposed a variety of heuristics for filtering potentially erroneous toponyms. MetaCarta [32] recognizes spatial cue words (e.g.,

Figure 7: A screenshot of NewsStand showing the most common terms in each cluster that correspond to the name of a disease for the US on a day in October 2009. This could be useful in tracking the spread of a disease.

"city of") as well as certain forms of postal addresses and textual representations of geographic coordinates. Unfortunately, this strategy causes serious problems when geotagging newspaper articles, as often the address of the newspaper's home office is included in each article. Given MetaCarta's primary focus on larger, prominent locations, these properly-formatted address strings play an overlarge role in its geotagging process, thereby resulting in many geotagging errors.

Other approaches to toponym recognition are rooted in solutions to related problems in Natural Language Processing (NLP), namely Named-Entity Recognition (NER) [4, 50] and Part-Of-Speech (POS) tagging [18]. NER focuses on the nouns and noun phrases and its goal is to identify phrases from the article that correspond to various entity classes, such as PERSON, ORGANIZATION, and LOCATION, while POS tagging assigns a part-of-speech to each token or word in the article, where nouns are clearly a priority in terms of attention. Those phrases tagged as LOCATION are most likely to be locations and are stored as *geographic features* of the entity feature vector, while ORGANIZATION and PERSON phrases are stored as *non-geographic features*. Regardless of whether NER or POS is used, these approaches can be roughly classified as either rule-based [6, 10, 12, 31, 33, 52] or statistical [21, 23, 41] in nature.

Rule-based solutions feature catalogs of rules that list 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 such as Hidden Markov Models (HMMs) [50] and Conditional Random Fields (CRFs) [20]. They have been used widely where annotated corpora are available. While statistical NER methods can be useful they are more error prone as they provide a finer classification than the POS methods—that is, it identifies proper nouns and also distinguishes between the three types of proper nouns that interest us. Therefore, NewsStand's toponym recognition procedure does not overly rely on any single method; instead, it opts for a hybrid approach (i.e., a combination of rule and statistical-based NER and POS taggers) involving multiple sources of evidence, while making use of the NER tagger of the LingPipe toolkit [3] for some tasks, such as identifying names of people. This tagger was trained on news data from the MUC-6 conference and the well-known Brown corpus [15].

It is important to note that the use of NER and/or POS tagging does not preclude the use of a gazetteer. Instead, these tagging methods just serve as filters or pruning devices to control the number of lookups made to the gazetteer. Of course, the downside is that if an entity is not identified as a potential location, then it will be missed, which does happen. NewsStand uses the GeoNames [47] gazetteer, an open gazetteer originally built from over 100 gazetteers including the GEOnet Names Server (GNS) and Geographic Names Information System (GNIS). It is maintained by volunteers around the world, and currently contains the names of about 6.7 million different geographic locations, where about 6 million of them have different names, with the difference accounting for the need to perform toponym resolution (i.e., resolve geo/geo ambiguity). In fact, the gazetteer contains almost 9.3 different names due to the need to keep track of the names of each location in different languages. It is interesting to note that from our experience with the 8 million articles most recently processed by NewsStand, we observed that only about 60,000 distinct locations were encountered, although over 40,000 were subject to geo/geo ambiguity, thereby clearly making toponym resolution an important task. The gazetteer also stores the population of populated places or regions, as well as hierarchical information including the country and administrative subdivisions that contain 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 form $L(f)$, where there are as many sets as there are features (i.e., $|EFV|$).

Once toponyms have been recognized, a toponym resolution procedure is applied to resolve the *geo/geo ambiguity*. Perhaps the simplest toponym resolution strategy is to assign a default sense to each recognized toponym, using some prominence measure such as population, and many researchers [1, 10, 26, 31, 32, 39, 52] have done so in combination with other methods. MetaCarta [32] assigns "default senses" in the form of probabilities based on how often each interpretation of a given toponym appeared in a pre-collected corpus of geotagged documents. It then alters these probabilities based on other heuristics such as cue words and occurrence with nearby toponyms. The SPIRIT project [31] uses similar techniques to those of MetaCarta by searching for sentence cues, and 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 (e.g., any of the over

Figure 8: An illustration of the local lexicon for readers living in the vicinity of Columbus, Ohio, USA. Notice the many local places that share names with more prominent places elsewhere.

2,000 lesser-known Londons around the world), that so often frequent articles in local newspapers, to be selected as correct interpretations, since these smaller places will have appeared in very few pre-created corpora of news articles. In contrast, NewsStand uses a concept known as a *local lexicon* [22], which is associated with a news source and contains the set of locations in the source's geographic scope. For example, as shown in Figure 8, the local lexicon of readers living in "Columbus, Ohio" includes places such as "Dublin", "Amsterdam", "London", "Delaware", "Africa", "Alexandria", "Baltimore", and "Bremen". In contrast, readers outside the Columbus area, lacking the above places in their local lexicons, would think first of the more prominent places that share their names.

This is analogous to making use of 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 is based on the observation that news articles are written with a knowledge/assumption of where their readers are located. For example, when the location "Springfield, Illinois" is mentioned in a newspaper article in Illinois (e.g., Chicago), the qualifier "Illinois" or "IL" is most likely not used on account of 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" should it be in fact in Illinois in order to avoid a possible misunderstanding. Local lexicons are particularly useful when users zoom in heavily 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 viewed as a "resolving context" for toponym resolution. A related popular [1, 10, 26, 30, 31, 35, 39, 45] strategy for toponym resolution places the resolving context within a hierarchical geographic ontology, which involves finding a geographic region in which many of the document's toponyms can be resolved. Web-a-Where [1] pursues such an approach by searching for several forms of hierarchical evidence in documents, including finding minimal resolving contexts and checking for containment of adjacent toponyms (e.g., "College Park, Maryland"). It identifies a document's geographic focus by using a simple scoring algorithm that takes into account the gazetteer hierarchy as well as a confidence score for each location l, which is the probability that l has been correctly identified. Ding et al. [11] use a similar approach. MetaCarta [32] and Google Book Search have no notion of a computed geographic focus, and thus require users to determine a focus by themselves. Instead of using content location, Mehler et al. [28] associate documents with the provider's location, which, at times, is equivalent to using the dateline. Note that the central assumption behind finding a minimal resolving context is that the document under consideration has a single geographic focus, which will be useful for resolving toponyms in that focus, but will not help in resolving distant toponyms mentioned in passing. Other toponym resolution strategies involve the use of geospatial measures such as minimizing total geographic coverage [21, 37] or minimizing pairwise toponym distance [23].

It is important to note that the local lexicon is just one of a number of techniques used by NewsStand for toponym resolution, whose need is manifested by the fact that some features have multiple records associated with them (i.e. $|L(f)| > 1$). In particular, NewsStand resolves such ambiguous references through the use of heuristic filters that select 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. For example, one such filter, the *object-container* filter, proceeds by searching for pairs of geographic features $f_1, f_2 \in EFV$ that are 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 such that a location $l_1 \in L(f_1)$ is contained in a location $l_2 \in L(f_2)$, then f_1 and f_2 are disambiguated as l_1 and l_2, respectively. This disambiguation is justified by the observation that a pair of features that are textually close in the article, close geographically, and exhibit a hierarchy relationship unlikely to occur by chance.

As we mentioned earlier, the geotagger must also determine the geographic focus of the individual articles. An obvious measure is the frequency of occurrence throughout the body text. We also observed that in a typical news article with a strong geographic component, important georeferences appear early in the text or in the title. Therefore, NewsStand uses 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 start of the article contributing more weight to g's ranking. In addition, NewsStand must also determine the geographic focus, if one exists, of the collection of news articles that forms a topic, rather than just one article, and this is achieved as a byproduct of the clustering algorithm (see Section 4), where the features that correspond to geographic locations are isolated and their cluster center is determined.

4. ONLINE CLUSTERING

The main goal here 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 below as simply *clusters*), such that each cluster should only contain the articles, encountered in the input seen so far, pertaining to a specific topic. This problem, called *topic detection*, is similar to clustering in the document domain. The problem is different from that of clas-

sification, since NewsStand does not the identity of topics beforehand. Furthermore, given that we are interested in detecting new topics, no training set can accurately predict future events. As news articles enter this stage, we assign them to news clusters, which is essentially a one-shot process meaning that once an article is added to a cluster, it remains there forever. We will never revisit or recluster the article, which is desirable as articles are coming into NewsStand at a high throughput rate and we need a document clustering system that can process them quickly, while still managing to give a good quality clustering output. Such a version of the clustering algorithm is characterized as being called *online* and these additional constraints imposed on this problem makes it much harder. In particular, we use a clustering algorithm, called leader-follower clustering [13], which allows for clustering in both by content and by time and modified it sufficiently so that it works in an online fashion as well as becomes resilient to noise.

The online clustering algorithm has a list of active clusters such that, along with each cluster, we associate a list of *feature vectors* (i.e., keywords) and their associated scores. We also store the *time centroid* of each cluster, which is the mean publication times of all the articles forming the cluster. A cluster is marked *inactive* if the time centroid is greater than several days (chosen according to system tuning/capacity considerations), in which case no additional article can be added to the cluster. When an input news article a is obtained, we first represent a by its feature vector representation using TF-IDF. We use a variant of the *cosine similarity measure* [38] for computing the distance between an input article a and a candidate cluster c, defined as follows:

$$\delta(a,c) = \frac{\overrightarrow{TFV_a} \bullet \overrightarrow{TFV_c}}{||\overrightarrow{TFV_a}||\,||\overrightarrow{TFV_c}||}$$

where $\overrightarrow{TFV_a}$, $\overrightarrow{TFV_c}$ are feature vectors of a and c, respectively. Note that a is added to a cluster c, if such a cluster exists, that is closest to a as well as within a distance of ϵ, where ϵ is a pre-specified constant. If no such cluster exists, then a new cluster is started with a as its only member.

To account for the temporal dimension in clustering, we apply a Gaussian attenuator on the cosine distance that favors adding the input articles to those clusters whose time centroids are close to the article's publication time. In particular, the Gaussian parameter takes into account the difference in days between the cluster's time centroid and the article's publication time. Our modified distance formula is

$$\dot{\delta}(a,c) = \delta(a,c) \cdot e^{\frac{-(T_t - T_c)^2}{2(\sigma)^2}}$$

where T_t is a's publication time and T_c is a cluster c's time centroid.

In order to speed up the search for a cluster c that is nearest to a as well as within a distance of ϵ from it, we maintain an inverted index of the cluster centroids. That is, the index stores for each feature f, pointers to all clusters containing f. We use this index to reduce the number of distance computations required for clustering. When a new article a is encountered, we only compute the distances to those clusters that have at least one feature in common with a. As a further optimization, we maintain a list of *active* clusters whose centroids are less than three days old. Only those clusters in the active list are considered as candidates to which a new article may be added. Together, these optimizations enable our algorithm to minimize the number of distance computations necessary for clustering an article.

Just as we computed the geographic focus when geotagging individual articles (see Section 3, we now wish to compute the *cluster geographic focus* of clusters of individual news articles. That is, we wish to decide which locations tagged in the members of an article cluster are relevant to the cluster topic, and which are simply mentioned in passing. The locations determined during the computation of the cluster geographic focus will be used for display on the user interface. Note that even though our clusters were created strictly using term similarity, the clustering ensures that different versions of the same topic are grouped into the same cluster, which should also ensure a grouping of the contained georeferences as well. Therefore, to ensure an accurate computation of the cluster geographic focus, we aggregate the geotagging results for each individual article in the cluster. More specifically, for each location l mentioned in an article in cluster C, we assign l a rank based primarily on how many articles mention l.

This process may be hampered by sporadic location inaccuracies introduced by improperly geotagged articles. Fortunately, we can correct these individual article errors at the cluster level, by using aggregated entity information and geotagging confidence values from the contained articles. If we make a reasonable assumption about article clusters, then we can use specific information discovered when processing each article individually to drastically improve the quality of our cluster geographic focus computation. We assume that if two or more entities found in articles from a particular cluster have the same name, they refer to the same entity. For example, if fifteen of twenty articles in a cluster all mention the entity "Springfield", it is assumed that they all refer to the same "Springfield," whether a person, location, organization, or other entity type. We expect this assumption to hold for article clusters, since we know each article in the cluster is about the same topic—it would be rare for an article to mention two distinct locations with the same name. More commonly, a person or organization mentioned in the article could share a name with a location in the article, but we still expect this case to be rare. We therefore expect that any disagreements among individual articles in a cluster are due to geotagger errors.

Using our assumption, we correct inconsistently-tagged entities (i.e. entities in a cluster that share the same name, but refer to different entities) using weighted voting. Each article in the cluster that mentions an inconsistent entity e casts a vote for its interpretation of e. Those articles with entities tagged with higher confidence cast stronger votes for those entities. For example, several articles may mention "Mr. Springfield", indicating a strong tendency toward an interpretation of "Springfield" as a person's name, so these articles would cast strong votes for their interpretation of "Springfield." On the other hand, an article simply mentioning "Springfield" with no additional qualification, and tagged as a location, would cast a weaker vote for this interpretation. By counting votes we determine that "Springfield" is a person's name, and should thus not be included in the cluster geographic focus.

This concept can be applied to inconsistent locations as well, in that articles can cast votes for their interpretation of location entities. For example, suppose that an article 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. We have found that the above methods generally perform well in extracting the cluster geographic focus, for both large and small cluster sizes.

5. USER INTERFACE AND DISPLAY ISSUES

NewsStand is designed to answer the the two questions "What is happening at X?" and "Where is topic T or article Y happening?" and its user interface has two corresponding modes. Once users have chosen a mode, they interact with NewsStand using *pan* and *zoom* to retrieve additional news articles. As users pan and zoom on the map, the map is constantly updated to retrieve new topics for the viewing window, thus keeping the window filled with topics, regardless of position or zoom level. A slider provides dynamic control over the number of different topics that are presented to the user. A given view of the map attempts to produce a summary of the news topics in the view, providing a mixture of topic significance and geographic spread of the topics. Users interested in a smaller or larger geographic region than the map shows can zoom in or out to retrieve more topics involving that region. NewsStand works best with the mapping API provided by Microsoft Virtual Earth to display topics in a web browser, although it also works with Google Maps and the Google Earth plugin although its use leads to a number of display issues due to limited support for user programming in the API.

Though it is important to show the most significant topics in the current viewing window when in "What is happening at X" mode, simply displaying the top 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 is a manifestation of the uneven news coverage of major newspapers, who tend to focus their publications on these geographic areas. In NewsStand, topic selection is a trade off between topic *significance* and *spread*. To achieve a balance, NewsStand subdivides the viewing window into a regular grid, and requires that each grid square contains no more than a maximum number of topics. As we have seen in Figures 1 and 2, in order to save space, the topics are represented on the display with markers. The topics that are displayed are selected in decreasing order of topic significance and topic age. This approach ensures a good spread of top topics across the entire map. However, a naive implementation may drastically change the appearance of the map with even a small pan request, especially if the geographic locations for many topics lie near borders of grid cells. This can be disorienting for users, who might not expect such large results from small changes. We address this problem by relaxing the restrictions on each grid cell in that we require that a given cell and all its neighbors fulfill the maximum topic requirement. This small change produces a fairly good distribution of topics, as in the above naive algorithm, but adapts better to small pan movements.

One of the main issues in display when in "What is happening at X" mode, is that at times markers may occlude other markers. This is permitted as long as the markers associated with more significant topics are placed above those associated with less significant topics. One exception to this rule is when markers exactly coincide—that is, when several topics involve the same geographic location. Clearly, it is unacceptable to place markers at the exact same coordinates on the map, as users cannot infer that many topics refer to that location. This is often a problem with large cities, as they are part of the geographic focus of many news topics. NewsStand resolves this problem by placing coinciding markers in a spiral, such that the most significant topics lie at the center of the spiral (i.e. the original location), and less significant topics are placed around the center. This allows significant geographic locations to have more of their articles visible, at the expense of accuracy in marker placement. However, due to their regular shape, these spirals are usually easy to identify and do not contribute significantly to user confusion.

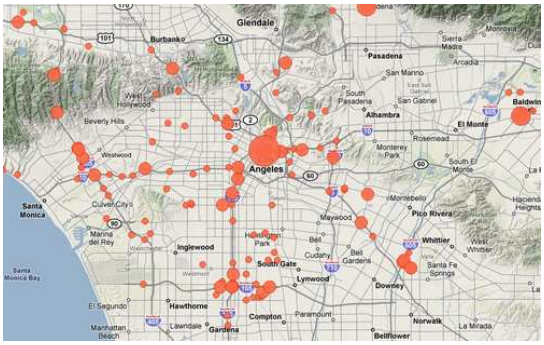

Figure 9: Geotagging of the tweets of accidents in Los Angeles, CA was made possible by incorporating the source location of the user.

6. CONCLUDING REMARKS

The design goals and functionality of the NewsStand system for using a map to read news on the world wide web, thereby harnessing the power of spatial synonyms, were reviewed. NewsStand demonstrates that extracting geographic content from news articles exposes a previously unseen dimension of information that can aid in understanding the news. Indeed, "NEWS" can be succinctly described as an acronym of "North, East, West, South". We believe that the increasing prevalence of geotagged content on the Internet will enable compelling applications for systems like NewsStand in other knowledge domains. Moreover, NewsStand represents a step forward in the emerging field of computational journalism [14, 19].

A number of aspects of NewsStand could benefit from further improvement. For example, NewsStand's geotagger could use more semantic hints from the document to aid in correct geotagging, such as landmarks and rivers. Moreover, geography can be used to improve the clustering of news articles, in addition to terms found in the text. The dynamic display of labels (e.g., keywords, people names, and disease names in Figures 3, 4, and 7, respectively) instead of markers, at interactive speeds under panning and zooming, could be improved by using techniques developed for dynamic map labeling [9]. Furthermore, other media can be placed on the map itself, including representative pictures, videos, and audio clips.

Some directions for future work include processing news articles in languages other than English. Adding this capability to NewsStand would also be useful in reducing its geographic bias towards the areas about which articles are usually written, and where a more uniform coverage (i.e., distributed) of the news is needed. Additional directions include the incorporation of other sources of news and information. For example, recently, we have incorporated Twitter tweets into NewsStand resulting in the creation of a new

system called Twitterstand [34]. The idea is to tap the large volume of news articles to serve as a kind of clustering corpus so that the very short and information-sparse tweets can be clustered using the existing news clusters. The interesting aspect of this method is that the tweets, due to their short length, usually have little or no geographic content. However, once they are clustered, they will inherit the geographic information associated with the geographic cluster focus with which they have become associated. For example, Figure 9 is the result of geotagging tweets of traffic incidents in Los Angeles, CA, which was made possible by incorporating the tweets content with the source location (i.e., Los Angeles) of the user.

7. REFERENCES

[1] E. Amitay, N. Har'El, 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*, pages 273–280, Sheffield, UK, July 2004.

[2] W. G. Aref and H. Samet. Efficient processing of window queries in the pyramid data structure. In *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 265–272, Nashville, TN, Apr. 1990. Also in *Proceedings of the Fifth Brazilian Symposium on Databases*, pages 15–26, Rio de Janeiro, Brazil, April 1990.

[3] B. Baldwin and B. Carpenter. Lingpipe [online]. Available from: `http://alias-i.com/lingpipe/` [cited 14 Nov 2009].

[4] A. Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University, New York, NY, USA, 1999.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International conference on World Wide Web*, pages 107–117, Brisbane, Australia, Apr. 1998.

[6] D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313, Mar. 2008.

[7] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of web pages. In *Proceedings of the Workshop on Web Databases*, pages 91–96, Philadelphia, PA, June 1999.

[8] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In *Proceedings of the ACM SIGMOD Conference*, pages 277–288, Chicago, IL, June 2006.

[9] J. Christensen, J. Marks, and S. Shieber. An empirical study of algorithms for point-feature label placement. *ACM Transactions on Graphics*, 14(3):203–232, July 1995.

[10] P. Clough. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval (GIR'05)*, pages 25–30, Bremen, Germany, Nov. 2005.

[11] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 545–556, Cairo, Egypt, Sept. 2000.

[12] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In A. El Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, and K.-Y. Whang, editors, *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB)*, pages 545–556, Cairo, Egypt, Sept. 2000.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, New York, second edition, 2000.

[14] I. Essa. Computation + journalism: A study of computation and journalism and how they impact each other [online]. Available from: `http://www.computation-and-journalism.com/` [cited 14 Nov 2009].

[15] W. N. Francis. A standard corpus of edited present-day american english. *College English*, 26(4):267–273, 1965.

[16] E. Garbin and I. Mani. Disambiguating toponyms in news. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[17] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2000.

[18] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, Jan. 2000.

[19] K. Kim, S. Oh, J. Lee, and I. Essa. Augmenting aerial earth maps with dynamic information. In *IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2009.

[20] J. D. Lafferty, A. McCallum, and F. C. N. Peireira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*, pages 282–289, Williamstown, MA, USA, June 2001.

[21] 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, Oct. 2006.

[22] M. Lieberman, H. Samet, and J. Sankaranarayanan. Preprocessing issues in constructing indexes for textually-specified spatial data. In *Proceedings of the 26th IEEE International Conference on Data Engineering*, Long Beach, CA, Apr. 2010. To appear.

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

[24] 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, Apr. 2008.

[25] A. Markowetz, T. Brinkhoff, and B. Seeger. Exploiting the internet as a geospatial database. In *Proceedings*

on the Workshop on Next Generation Geospatial Information, Cambridge, MA, Oct. 2003. Online Proceedings.

[26] B. Martins, H. Manguinhas, J. Borbinha, and W. Siabato. A geo-temporal information extraction service for processing descriptive metadata in digital libraries. *e-Perimetron*, 4(1):25–37, 2009.

[27] K. S. McCurley. Geospatial mapping and navigation of the web. In *Proceedings of the 10th International World Wide Web Conference*, pages 221–229, Hong Kong, China, May 2001.

[28] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):765–772, 2006.

[29] Microsoft Corporation. Bing news [online]. Available from: `http://news.bing.com/` [cited 14 Nov 2009].

[30] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best. Geocoding multilingual texts: Recognition, disambiguation, and visualization. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 53–58, Genoa, Italy, May 2006.

[31] 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 Systems*, 21(7):717–745, 2007.

[32] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 50–54, Edmonton, CA, May 2003.

[33] C. Sallaberry, M. Gaio, J. Lesbegueries, and P. Loustau. A semantic approach for geospatial information extraction from unstructured documents. In A. Scharl and K. Tochtermann, editors, *The Geospatial Web*, pages 93–104. Springer, London, England, UK, 2007.

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

[35] F. Schilder, Y. Versley, and C. Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. In *Proceedings of the SIGIR 2004 Workshop on Geographic Information Retrieval (GIR'04)*, Sheffield, UK, July 2004.

[36] M. J. Silva, B. Martins, M. Chaves, and N. Cardoso. Adding geographic scope to web resources. In *Proceedings of the Workshop on Geographic Information Retrieval*, Sheffield, UK, July 2004. Online Proceedings.

[37] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 127–136, Darmstadt, Germany, 2001.

[38] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, pages 1–20, Boston, MA, Aug. 2000.

[39] N. Stokes, Y. Li, A. Moffat, and J. Rong. An empirical study of the effects of NLP components on geographic IR performance. *International Journal of Geographical Information Science*, 22(3):247–264, Mar. 2008.

[40] B. Stroustrup. *The C++ Programming Language*. Addison-Wesley Longman, Reading, MA, third edition, 1997.

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

[42] The Associated Press. Mobile news network [online]. Available from: `http://apnews.com/` [cited 14 Nov 2009].

[43] Thomson Reuters. Reuters news maps [online]. Available from: `http://labs.reuters.com/newsmaps/` [cited 14 Nov 2009].

[44] R. Volz, J. Kleb, and W. Mueller. Towards ontology-based disambiguation of geographical identifiers. In *Proceedings of the WWW 2007 Workshop on I3: Identity, Identifiers, Identification*, Banff, Alberta, Canada, May 2007.

[45] C. Wang, X. Xie, L. Wang, Y. Lu, and W.-Y. Ma. Detecting geographic locations from web resources. In *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 17–24, New York, NY, USA, 2005. ACM.

[46] C. Wang, X. Xie, L. Wang, Y. Lu, and W.-Y. Ma. Web resource geographic location classification and detection. In *Proceedings of the Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pages 1138–1139, Chiba, Japan, May 2005.

[47] M. Wick and B. Vatant. The geonames geographical database [online]. Available from: `http://geonames.org/` [cited 14 Nov 2009].

[48] Yahoo! Corporation. Yahoo! news [online]. Available from: `http://news.yahoo.com/` [cited 14 Nov 2009].

[49] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *IEEE International Conference on Data Mining (ICDM)*, pages 427–434, Nov. 2003.

[50] G. Zhou and J. Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 209–219, Philadelphia, PA, 2001.

[51] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. Hybrid index structures for location-based web search. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 155–162, Bremen, Germany, Oct. 2005.

[52] W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh. On assigning place names to geography related web pages. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362, New York, NY, USA, 2005. ACM.