Reading News with Maps by Exploiting Spatial Synonyms

Hanan Samet*

hjs@cs.umd.edu

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

* Based on Joint Work with Marco D. Adelfio, Brendan C. Fruin, Jack Lotkowski, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Jon Sperling, and Ben E. Teitler
Application

Questions

1. Do you travel?
2. Do you want to know what is going on in the town you are traveling to?
3. Do you want to keep up with the latest news in the town you have left
   - Especially when it is your own hometown?
   - E.g., keep up with the local sports team

Answer: NewsStand

- Enables search with a map query interface instead of by keyword
- Advantage: a map, coupled with ability to vary the zoom level at which it is viewed, provides a granularity to the search and facilitates an approximate search
  - Can do an approximate search with a group of keywords (e.g., synonyms) in the query formulation
  - But users often have no clue as to which keyword to use
  - Would welcome the search automatically taking them into account
- Map query interface is a step in this direction
  - Pointing at a location and making the interpretation of the precision of this positioning dependent on the zoom level is equivalent to using spatial synonyms
Power of Spatial Synonyms

- Enables search for data when not exactly sure of what we are seeking, or what should be the answer to the query.

- Ex: Seek a “Rock Concert in Manhattan”
  - “Rock Concerts” in “Harlem”, “New York City”, or “Brooklyn” are 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
    - “New York City” by virtue of containing Manhattan
    - “Brooklyn” by virtue of proximity and a sibling relationship (neighboring borough)
Conventional Search Engines and Spatial Synonyms

- Use page rank method and good at finding documents containing keywords that we are looking for, but cannot be easily modified to handle spatial proximity query.

- Primary utility is based on popularity in sense of ensuring that web pages in response are ordered by a measure incorporating their frequency of being linked to so results are same as provided to other users.
  - "Democratization of search"
  - All users are treated equally
  - They all get the same bad (or good!) answers.

- Effectively means that if nobody ever looked for some data before or linked to it, then it will never be found and, hence, never presented to users.

- In case of synonyms, if no links to similar pages on account of being equivalent but for the use of the same words, then similarity will never be found by the search engine as the web crawler will never be able to find the similar pages when building the index to the web pages.
Taking Advantage of Spatial Synonyms: Location Specification

- Explicit via geometry (latitude-longitude pairs of numbers)
  - Used in programs and calculations
  - Not in search engines or mobile devices
  - Users don’t know them in this way or used to communicate in this way
- Accustomed to textual specification
  - Easy to communicate on smartphone devices with soft keyboard
  - Can capture verbally by speech recognition (e.g., Siri)
  - Behave like a polymorphic type
    - One size fits all
    - “Los Angeles” can be interpreted as a point or an area and user need not be concerned about it
- Supports use of spatial synonyms
- Drawback is ambiguity
  - Is “London” reference a person or a location? (toponym recognition)
  - If “London” is a location, which of many? (toponym resolution)
Decoding or Forming an E-mail Address
Determining Performance of a Team in a Sports League

Checking scores in Dublin is like checking scores in Dublin.

Use your phone abroad like you do in California with AT&T's international data packages.

Copyright 2014: Hanan Samet
Interpreting Weather Temperature Measurement Unit

Getting the weather in Mexico is like getting the weather in Mexico.

Use your phone abroad like you do in New York with AT&T’s international data packages.

Rethink Possible®

Copyright 2014: Hanan Samet

Reading news with maps – p.8/60
Finding Local Food
33 Different Plymoughs

There are 33 places in the U.S. named Plymouth.

Call it Thanksgiving-induced nostalgia, but we turn around on our pilgrims — forget the nearly three dozen U.S. counties and towns named for one of our test settlements. The Plymouth that stood it all in Massachusetts, appeared in Money magazine’s “Best Places to Live” in 2005, but it wasn’t an easy place to inhabit in 1620, when the colonists who set sail from England’s port of Plymouth to escape the religious rule of King James I made land fall there. About half of them died during the first brutal winter. If only the pilgrims had come to32
Plymouth, California, where fall temps hover around 50 degrees. You know what they say about hindsight...

Copyright 2014: Hanan Samet
Goal: Change News Reading Paradigm

- Use map to read news for all media (e.g., text, photos, tweets, videos)
- Choose place of interest and find topics/articles relevant to it
- Topics/articles determined by location and level of zoom
- No predetermined boundaries on sources of articles
- Application: monitoring hot spots
  1. Investors
  2. National security
  3. Disease monitoring
- One-stop shopping for spatially-oriented news reading
  1. Summarize the news
     - What are the top stories happening?
  2. Explore the news
     - What is happening in Darfur?
  3. Discover patterns in the news
     - How are the Olympics and Darfur related?
- Overall goal: make map medium for presenting all spatially-referenced information
Mapping the News

1. Cluster articles on same topic using TF-IDF and associate clusters with the mentioned locations
   - Same cluster can be associated with many locations

2. As zoom-in, the cluster populations will be smaller as fewer articles refer to the viewing window
   - Location plays a larger role in the clustering algorithm
   - Geotagging errors are less likely to be filtered out

3. Cluster rank vs: cluster spread
   - Don’t want to have empty areas on the map with no articles implying that less important articles are displayed with some regions than others and some important articles are not displayed unless zoom-in
   - As zoom-in and pan want to make sure that once an article has been displayed, it persists until its location is no longer in the viewing window

4. Zoom-In and Pan are expensive as much redrawing
   - Use “Home”, “Local (L)”, and “World (W)” as navigation shortcuts
   - Can use an inset "overview window" to control zoom and pan with little symbolic information that needs to be redrawn
Existing News Readers

1. Popular news aggregators such as Google News, Yahoo! News, 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)

2. Usually a linear presentation format
   - Articles grouped by keyword or topic, rather than by geography

3. Ex: Google News Reader
   - Classifies articles by topic
   - Local news search
     - Aggregates articles by zip code or city, state specification
     - E.g., articles mentioning “College Park, MD”
     - Provides a limited number of articles (about 20 at the moment)
     - Seem to be based on the host of the articles
     - E.g., “LA Times” provides local articles for “Los Angeles, CA”
     - Seems to use Google Search with location names as search keys
     - E.g., articles for ZIP 20742 are those mentioning “College Park, MD” or “University of Maryland”
   - Has no notion of story importance in the grand scheme
   - International versions use international news sources
NewsStand: Spatio-Textual Aggregation of News and Display

1. Crawls the web looking for news sources and feeds
   - Indexing 10,000 news sources
   - About 50,000 news articles per day

2. Aggregate news articles by both content similarity and location
   - Articles about the same event are grouped into clusters

3. Rank clusters by importance which is based on:
   - Number of articles in cluster
   - Number of unique newspapers in cluster
   - Event’s rate of propagation to other newspapers

4. Associate each cluster with its geographic focus or foci

5. Display each cluster at the positions of the geographic foci

6. Other options:
   - Category (e.g., General, Business, SciTech, Entertainment, Health, Sports)
   - Image and video galleries
   - Map stories by disease, brands, people, etc.
   - User-generated news (e.g., Social networks such as Twitter)
NewsStand: Map Mode

NewsStand is at http://newsstand.umiacs.umd.edu/

Query: What is happening at location Y?
NewsStand: Top Stories Mode

NewsStand is at http://newsstand.umiacs.umd.edu/

Query: Where is topic X occurring (spatial data mining)?
Dynamic Disease Tracking Application: Time Mode

E.g., track ebola
Brand Remediation
Port to Mobile Platforms (Apps)

(iOS) (Android) (Windows)
NewsStand’s Architecture

Data flow:
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
5. **Topic Classifier**: Assigns general topics to articles (e.g., “Sports”).
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
5. **Topic Classifier**: Assigns general topics to articles (e.g., “Sports”).
6. **Geotagger**: Finds toponyms and assigns lat/long values to each.
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
5. **Topic Classifier**: Assigns general topics to articles (e.g., “Sports”).
6. **Geotagger**: Finds toponyms and assigns lat/long values to each.
7. **People/Disease Finder**: Finds mentions of people/diseases.
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
5. **Topic Classifier**: Assigns general topics to articles (e.g., “Sports”).
6. **Geotagger**: Finds toponyms and assigns lat/long values to each.
7. **People/Disease Finder**: Finds mentions of people/diseases.
8. **Media Extractor**: Extracts captioned images and videos.
NewsStand’s Architecture

Data flow:

1. **RSS Grabber**: Polls RSS feeds and retrieves URLs to news articles.
2. **Downloader**: Downloads news articles from URLs.
3. **Cleaner**: Extracts article content from source HTML.
4. **Clusterer**: Groups together articles about the same story.
5. **Topic Classifier**: Assigns general topics to articles (e.g., “Sports”).
6. **Geotagger**: Finds toponyms and assigns lat/long values to each.
7. **People/Disease Finder**: Finds mentions of people/diseases.
8. **Media Extractor**: Extracts captioned images and videos.
9. **Web Interface**: Accesses database to retrieve data for display.
Map Query Interface Requires a Geotagger

- Geotagger: processor that converts a textual specification of a location to a geometric one (i.e., latitude-longitude pair)

- Geotagging issues:
  1. Toponym recognition: identify geographical references in text
     - Does “Jefferson” refer to a person or a geographical location?
     - Known as Geo/Non-Geo Ambiguity
  2. Toponym resolution: disambiguate a geographical reference
     - Does “London” mean “London, UK”, “London, Ontario”, or one of 2570 other instances of “London” in our gazetteer?
     - Known as Geo/Geo Ambiguity
  3. Determine spatial focus of a document
     - Is “Singapore” relevant to a news article about “Hurricane Katrina”?
     - Not so, if article appeared in “Singapore Strait Times”
Geo/Non-Geo Ambiguity Example: Obama

Japan's Obama town overjoyed
Wed, Nov 5 2008

By Toshi Maeda

OBAMA, Japan (Reuters) - The sleepy Japanese fishing town of Obama went wild Wednesday as locals gathered to celebrate namesake Barack Obama's victory in the U.S. presidential election.

More than a hundred residents gathered to watch the vote count on television in a public hall in the middle of the day and chanted "Obama, Obama!" as the result was announced on a news program.

Some were clad in hula costumes in honor of Obama's birthplace in Hawaii. Others showed up wearing "I love Obama" T-shirts.

The town has taken advantage of the name - one of many named Obama, or "small beach" in Japanese - to launch products from fish burgers and steamed cakes to chopsticks.

"The next thing we want to do is to go to the White House and dance the hula at Obama's inauguration ceremony," said Tatsuya Sano, 45, who runs a souvenir shop selling locally made Barack Obama souvenirs.

Chikako Shimizu, 35, the leader of an "Obama Girls" hula dance group launched this year, said she was calm while watching the vote count on television because she had no doubt Obama would win.

"I was convinced that he would win. I couldn't be happier," she said.

Obama City residents plan to dance and party more in the evening.
Mayor of Batman sues WB, Nolan

Southeastern city in Turkey fights for name

By ALI JAAFAR

Batman has a new adversary: Batman.

The mayor of an oil-producing city in southeastern Turkey, which has the same name as the Caped Crusader, is suing helmer Christopher Nolan and Warner Bros. for royalties from mega-grosser “The Dark Knight.”

Huseyin Kalkan, the pro-Kurdish Democratic Society Party mayor of Batman, has accused “The Dark Knight” producers of using the city’s name without permission.

“There is only one Batman in the world,” Kalkan said. “The American producers used the name of our city without informing us.”

No one from the town of Batman has explained why it took so many years to take legal action. Batman first appeared as a comicbook character in 1939 and the “Batman” TV series started in 1966. Tim Burton’s first bigscreen rendition for Warner Bros. came out in 1989. Undoubtedly the fact that “Dark Knight” is about to pass the $1 billion mark at the B.O. played a part in stirring the ire of the Turkish hamlet.
Geo/Geo Ambiguity Example: Java, Georgia

Russia thrusts into South Ossetia; clashes with Georgia reported
5 hours ago

JAVA, Georgia (AFP) — Russian tanks and troops surged into Georgia’s breakaway South Ossetia province on Friday to repel a Georgian offensive to reclaim the region amid fighting said to have left hundreds dead.

"Fierce clashes" between Russian and Georgian troops in the southern suburbs of South Ossetia's capital Tskhinvali were reported by Russian news agencies as night fell on the city.

Moscow had vowed retaliation to defend Russians in Tskhinvali who had come under fire by the Georgian artillery and air assault -- the worst fighting since the 1992-94 separatist war in the region.

"Georgian forces are controlling the entire territory of South Ossetia except Java," a city north of Tskhinvali, Georgian President Mikheil Saakashvili said in a televised address.

"We are fully controlling Tskhinvali," he added, although the rebels shortly after said that they were in control, according to the Interfax news agency.
Oops, wrong Vancouver
Thu, Feb 4 2010
By Teresa Carson
VANCOUVER, Washington (Reuters) - Sallie Reavey picked up the phone at her charming Briar Rose Inn and the caller asked about rooms in mid-February. "We have a nice selection of rooms for those dates," she replied, to which the caller gasped: "You still have rooms during the Olympics?"
Reavey had to tell him: wrong Vancouver.
The Briar Rose is in Vancouver, Washington, not Vancouver, British Columbia, the Canadian city that will host the 2010 Winter Olympics starting on February 12.
"America's Vancouver," as a former town mayor liked to describe it, sits 250 miles south of the Olympic host Vancouver and has a population of some 155,000 people -- far fewer than the Canadian city.
The Hilton Vancouver Washington has also fielded Olympic enquiries and trained its reservations staff to be sensitive to the possible mistake and, naturally, turn it into a marketing opportunity.
"We absolutely want them to come here," Gerry Link, the hotel's general manager said, adding of the Vancouver mix-ups: "So far it has all been pretty good-natured."
Geotagging

Geotagging: Understanding textual references to spatial data
1. Identifying or recognizing
2. Classifying (is “Michigan” a state or a lake?)
3. Disambiguating or resolving
4. Localizing (geocoding to GPS coordinates)

Context of textual references
1. Queries - use prior queries and location
   - Ex: Query “Alexandria” when in “College Park, MD”
2. Underlying data being queried - need context
Mechanics of Geotagging

1. Goal: high recall in toponym recognition (i.e., not missing toponyms) at expense of precision
   - Rectify by subsequent use of toponym resolution which can (and will) also be used to filter erroneous location interpretations
2. Toponym recognition: 2 stages
   - Finding toponyms
   - Filtering toponyms: postprocessing to remove errors in recognition
3. Toponym resolution
   - Use local lexicons containing locations that can be specified without all of their containers (derived from articles from a particular news source) to determine spatial reader scopes for particular sources
     - E.g., "Dublin" implies “Dublin, Ohio” for readers of a news source in “Columbus, Ohio”
   - Use Wikipedia articles to find concepts related to particular locations so that the presence of these concepts in conjunction with an ambiguous reference to a location can be properly resolved
     - E.g., mention of “White House” in conjunction with “Washington” to provide evidence for resolving as “Washington, D.C.”
Local Lexicon Example
Finding Toponyms

1. Use entity tables of well-known locations (e.g., names of continents, countries, etc.), abbreviations (e.g., “CA”, “FL”, etc.), and demonyms (words used to refer to people from particular places such as "German")

2. Use entity dictionaries containing names of entities that appear frequently in news thereby precluding their interpretation as toponyms (e.g., "Apple")

3. Use a Part of Speech (POS) tagger to find proper noun phrases which could denote names even with possessives like "Prince George’s County"

4. Use Named Entity Recognition (NER) package which helps avoid geo/non-geo errors by making use of entity types such as name, place, organization, etc.

5. Compensate for NER errors
   - Boundary expansion (e.g., "Guinea" and "Equatorial Guinea")
   - Fragmented references such as names where parts can be interpreted as locations (e.g., "Paul Washington" and "Washington")
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

- **True toponyms:** Texas, Paris, Houston, Lamar County, Austin, Dish
- **Potential mistakes:** Weems, Homer, Friday (all in Texas)
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Initial text

2. Entity tables: [LOC Texas], [PER Jeff Weems], [DAY Friday], [PER Mark Homer]
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Initial text

2. Entity tables: [LOC Texas], [PER Jeff Weems], [DAY Friday], [PER Mark Homer]

3. Cue words: Rep. [PER Mark Homer], D-[LOC Paris], [LOC Lamar County]
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Initial text
2. Entity tables: [LOC Texas], [PER Jeff Weems], [DAY Friday], [PER Mark Homer]
3. Cue words: Rep. [PER Mark Homer], D-[LOC Paris], [LOC Lamar County]
4. Proper noun phrases: [NP Democratic], [NP Railroad Commissioner Jeff Weems], [NP Paris], [NP Rep. Mark Homer], [NP Railroad Commission], [NP Houston], [NP Weems], [NP Lamar County], [NP Democrats], [NP Austin], [NP Dish], [NP Texas], [NP Midcontinent Express Pipeline]
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Initial text

2. Entity tables: \([\text{LOC Texas}]\), \([\text{PER Jeff Weems}]\), \([\text{DAY Friday}]\), \([\text{PER Mark Homer}]\)

3. Cue words: \(\text{Rep. [PER Mark Homer]}, \text{D-[LOC Paris]}, \text{[LOC Lamar County]}\)

4. Proper noun phrases: \([\text{NP Democratic}]\), \([\text{NP Railroad Commissioner Jeff Weems}]\), \([\text{NP Paris}]\), \([\text{NP Rep. Mark Homer}]\), \([\text{NP Railroad Commission}]\), \([\text{NP Houston}]\), \([\text{NP Weems}]\), \([\text{NP Lamar County}]\), \([\text{NP Democrats}]\), \([\text{NP Austin}]\), \([\text{NP Dish}]\), \([\text{NP Texas}]\), \([\text{NP Midcontinent Express Pipeline}]\)

5. Named-entity recognition:

\[
\begin{array}{ll}
[\text{PER Jeff Weems}] & 0.999 \\
[\text{ORG Railroad Commission}] & 0.995 \\
[\text{ORG Midcont. Expr. Ppln.}] & 0.973 \\
[\text{LOC Houston}] & 0.917 \\
[\text{LOC Lamar County}] & 0.737 \\
[\text{ORG Democratic}] & 0.539 \\
[\text{LOC Paris}] & 0.997 \\
[\text{LOC Austin}] & 0.995 \\
[\text{PER Mark Homer}] & 0.920 \\
[\text{PER Weems}] & 0.849 \\
[\text{LOC Texas}] & 0.557 \\
\end{array}
\]
Filtering Toponyms

1. Toponym refactoring:
   - Account for different suffixes and prefixes for same entity
   - Ex: "Fort" and "Ft", "County Kildare" and "Kildare County", "Fairfax Hi" and "Fairfax High School", etc.

2. Active verbs
   - People are active while locations are passive
   - Account for metonymy where an entity like a government is referenced by its location (e.g., "Washington expects ...") and is active but there are usually other references to the location in the text so no harm in ignoring some instances

3. Use Knowledge of noun adjuncts to avoid mistaken container relationships such as "In Russia, U.S. officials ..." due to presence of comma

4. Type propagation to make unknown types consistent within a group as long as there is just one known type in the group
   - E.g., name of streets “Federalist”, “Market”, “Edgewood” while the type entity of “Paul Revere” and “First” are not identified and thus could interpret them as names of streets
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Toponym refactoring: [LOC Lamar County] → [LOC County of Lamar]
Democratic candidate for Texas Railroad Commissioner Jeff Weems **stumped** in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston **attorney** with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems **labeled** Lamar **County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems **said** he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. **Toponym refactoring**: [LOC Lamar **County**] → [LOC **County of Lamar**]
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. **Toponym refactoring:** [LOC Lamar County] → [LOC County of Lamar]
2. **Active verbs:** [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said
3. **Noun adjuncts:** [LOC Houston] attorney
Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]

2. **Active verbs**: [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said

3. **Noun adjuncts**: [LOC Houston] attorney

4. **Final location entities**: Texas, Paris, Houston, Lamar County, Austin, Dish
Toponym Resolution

1. Dateline

2. Relative geography which is usually vague
   - Ex: "Just outside Lewiston"

3. Comma group where use prominence, proximity, or sibling where share a parent in a geographic hierarchy
   - Prominence: Ex: New York, Philadelphia, Chicago
   - Proximity: Ex: Milwaukee, Chicago, Minneapolis, St. Paul
   - Sibling: Queens, Brooklyn, Manhattan

4. Location/Container – Ex: “College Park, MD”

5. Local lexicon – Ex: “Dublin” in the case of “Columbus, Ohio”

6. Global lexicon
   - Gazetteer with names of places that are known regardless of their geographic location

7. One sense
   - Consistency with previously resolved instances of same name in same source article
News-Specific Geotagging Issues

1. Name of news source
   - Identify a geographic focus (also known as a “spatial reader scope”) for a particular news source in terms of the container(s) of the articles in the source and use this to resolve geotagging ambiguities

2. Perform some preliminary clustering by focusing on the headline

3. Multiple vs: a single interpretation as a geographic location
   - Multiple: evidence that it is a geographic location
   - Single: may be an error, verify by checking
     - population
     - presence of containers
     - presence of proximate locations
Near-Duplicate Images

- Images with slight modification are similar to each other
- News Images about the same event are often near-duplicate

Use hierarchical color histograms

- Pros
  - Retains certain color layout information
  - Compressed and efficient to compute
- Cons
  - Not robust to occlusions or significant cropping, which can dramatically affect the intensity and color layout of the image
Sources of Near-Duplicate Images

- Different news sources may use the same “official” photo with some slight modifications (e.g., cropping, rotation, resizing)

- Photographer may produce many photos in short time period

- Different photographers capture the same scene from different views or different lighting conditions
Detection Result

- 1st Row: Similar time instances
- 2nd Row: Different imagecroppings
- 3rd Row: Changes in image brightness, contrast, and hue
- 4th Row: Similar grayscale and color images

Example of similar images not detected by hierarchical color histogram
Hard Cases (Misclassified as Near-Duplicate)

- Backgrounds match and face recognition might be useful

- Hard for pure vision algorithms to detect the difference
  - Context textual information should be used
TwitterStand: News from Tweets

- News gathering system using Twitter
- Twitter is a popular social networking website
  - Tweets are 140 character messages akin to SMS
  - Mostly non-news, often frivolous
- TwitterStand is a spontaneous news medium
  - Idea: users of Twitter help to gather news
    - Distributed news gathering
  - Scooping tool bypassing reporters or newspapers
    - E.g., Michael Jackson’s death, Iranian election, Haitian earthquake
TwitterStand: News from Tweets

- News gathering system using Twitter
- Twitter is a popular social networking website
  - Tweets are 140 character messages akin to SMS
  - Mostly non-news, often frivolous
- TwitterStand is a spontaneous news medium
  - Idea: users of Twitter help to gather news
    - Distributed news gathering
  - Scooping tool bypassing reporters or newspapers
    - E.g., Michael Jackson’s death, Iranian election, Haitian earthquake
- Key challenges:
  - Managing the deluge
    - Twitter is a noisy medium as most of the Tweets are not news
    - Challenge: extract news Tweets from mountain of non-news Tweets
  - Tweets are coming at a furious pace
  - Tweets capture the pulse of the moment
    - So, not a good strategy to store and process them in batches
    - TwitterStand uses online algorithms
    - Works without access to entire dataset (i.e., being offline)
    - Determine spatial focus of stories enabling news reading on map
Access to Twitter

1. Whitelisted which means TwitterStand can access Twitter 20K times per hour
2. Access to Gardenhose which yields many Tweets but not clear what percentage
3. Birddog enables TwitterStand to obtain feeds from up to 200K users
4. Seeders are 2000 handpicked users who are known to publish news
Ex: Tweets about Michael Jackson’s Death

- A 00:00: 911 call
- B 00:20: 1st tweet
- C 02:30: LA Times reports death
Ex: Tweets about Michael Jackson’s Death

Notice that Twitter beat the LA Times by more than two hours.
Live Demo: TwitterStand System

http://twitterstand.umiacs.umd.edu/

- What people are tweeting about rather than where they are tweeting from
STEWARD: A Spatio-Textual Search Engine

1. Spatio-Textual Extraction on the Web Aiding Retrieval of Documents
2. Sample spatio-textual query:
   - Keyword: “rock concert”
   - Location: near “College Park, MD”
3. Result documents are relevant to both keyword and location
   - Mention of rock concert
   - Spatial focus near “College Park, MD”
4. Issues with results from conventional search engines:
   - Is it the intended “College Park”?
   - What about spatial synonyms such as rock concerts in “Hyattsville” or “Greenbelt”?
   - Don’t usually understand the various forms of specifying geographic content
   - More than just postal addresses!
   - Results often based on other measures, e.g., link structure
5. Applied to HUD USER, PubMed, ProMED-mail, and news
STEWARD Is Not Google Local

1. Google Local geocodes postal addresses into points on the map
   - Address strings are well-formatted
   - Most results drawn from online yellow pages
2. STEWARD works on unstructured text documents
   - Document is a bag of words
3. STEWARD goals:
   - More than searching for addresses in documents, which is easier
   - Identify all geographic locations mentioned in document (i.e., Geotagging)
   - Identify geographic focus of document
   - Retrieve documents by spatio-textual proximity
STEWARD is Different from NewsStand

1. STEWARD focuses on determining the geographic focus or foci of single documents.
2. NewsStand focuses on finding clusters of articles on a single topic and associating them with the geographic locations that they are about and to a lesser extent that they mention.
3. NewsStand may choose to ignore some locations as being irrelevant to the central topic of the article.
4. The common topic of the cluster is used to improve the geographic foci determination process in NewsStand.
5. In STEWARD, the user selects the keywords that determine the documents (could be news articles) that are retrieved.
6. In NewsStand, the topics are more general than keywords and are determined by the clustering process independent of the user.
7. NewsStand uses the functionality of STEWARD to enhance the process of reading particular articles in the cluster:
   - Search the cluster for keywords
   - Browse the geographical foci of elements of the clustering
Live Demo: Using STEWARD to Monitor Disease Reporting over Time

http://steward.umiacs.umd.edu
Spatio-Textual Spreadsheets

Motivation
1. Web is full of structured tables with spatial information in the cells
2. Google’s ranking algorithm cannot index this spatial information
3. Understanding the structure of these spatio-textual tables enables a more intelligent search engine

Objectives:
1. Identify spatial attributes in spreadsheets
2. Enable web crawlers to take advantage of spatial information in spreadsheets
3. Enable web-based queries on the tuples of spreadsheets
4. Visualize spreadsheets based on their spatial attributes
5. Process spreadsheets in contrast to HTML relational tables as in Google’s WebTables
Spatial Coherence in Spreadsheets

- **Column coherence**: cells in a spatial column share the same spatial type
- **Row coherence**: containment relationships among spatial data in a row
- **Spreadsheet coherence**: locations in adjacent rows are usually geographically proximate

<table>
<thead>
<tr>
<th>State</th>
<th>Zip Code</th>
<th>County Name</th>
<th>Project or Program Type</th>
<th>Book 1 - 3</th>
<th>Loan</th>
<th>Grant</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>725429471</td>
<td>Sharp</td>
<td>City of Highland - Sewer</td>
<td></td>
<td>$128,000</td>
<td>$297,000</td>
</tr>
<tr>
<td>AR</td>
<td>726539699</td>
<td>Baxter</td>
<td>City of Salesville - Sewer</td>
<td></td>
<td>$832,000</td>
<td>$1,479,000</td>
</tr>
<tr>
<td>AZ</td>
<td>853620727</td>
<td>Yavapai</td>
<td>Yarnell Water Improvement Assoc.</td>
<td></td>
<td>$767,000</td>
<td>$533,000</td>
</tr>
<tr>
<td>CA</td>
<td>936090218</td>
<td>Fresno</td>
<td>Caruthers CSD</td>
<td></td>
<td>$1,515,000</td>
<td>$988,000</td>
</tr>
<tr>
<td>CA</td>
<td>959482117</td>
<td>Butte</td>
<td>City of Gridley</td>
<td></td>
<td>$2,750,000</td>
<td>$2,300,850</td>
</tr>
<tr>
<td>CA</td>
<td>936152125</td>
<td>Tulare</td>
<td>Cutler PUD</td>
<td></td>
<td>$1,761,000</td>
<td>$1,169,000</td>
</tr>
<tr>
<td>CA</td>
<td>961309786</td>
<td>Lassen</td>
<td>Leavitt Lake CSD</td>
<td></td>
<td>$182,000</td>
<td>$0</td>
</tr>
<tr>
<td>CA</td>
<td>952520284</td>
<td>Calaveras</td>
<td>Valley Springs Utility District</td>
<td></td>
<td>$1,300,000</td>
<td>$130,000</td>
</tr>
<tr>
<td>CA</td>
<td>961370319</td>
<td>Lassen</td>
<td>Westwood Community Services District</td>
<td></td>
<td>$500,000</td>
<td>$59,000</td>
</tr>
<tr>
<td>CT</td>
<td>62601831</td>
<td>Windham</td>
<td>Town of Putnam</td>
<td></td>
<td>$7,511,000</td>
<td>$5,989,000</td>
</tr>
<tr>
<td>CT</td>
<td>62601831</td>
<td>Windham</td>
<td>Town of Putnam Wellfield Impr.</td>
<td></td>
<td>$3,680,000</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>60760101</td>
<td>Tolland</td>
<td>Town of Stafford</td>
<td></td>
<td>$6,566,000</td>
<td>$5,333,700</td>
</tr>
<tr>
<td>FL</td>
<td>32463</td>
<td>Washington</td>
<td>Town of Wausau - water</td>
<td></td>
<td>$664,000</td>
<td>$1,691,000</td>
</tr>
<tr>
<td>ID</td>
<td>835390126</td>
<td>Idaho</td>
<td>City of Kooskia</td>
<td></td>
<td>$425,000</td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>623121303</td>
<td>Pike</td>
<td>City of Barry</td>
<td></td>
<td>$747,000</td>
<td>$0</td>
</tr>
</tbody>
</table>
Applications

- Tuple retrieval from spreadsheets
  - Find the population of India:
    ```
    SELECT population FROM SCHEMA_WITH(country, population) WHERE country = 'India';
    ```
  - Find closest restaurants to a lat/long point:
    ```
    SELECT business_name, phone FROM SCHEMA_WITH(business_name, type, address, phone)
    WHERE type = 'restaurant' ORDER BY distance(address, ``(x,y)'') LIMIT 10;
    ```

- Mining spreadsheets
  - Given an attribute (ZIP code, GDP), find its type (number, percentage)
  - Find aliases of spatial column names ("state name", "State_name", "StName", ...)
  - Use spatial attributes as join keys to merge tuples from different spreadsheets
  - Gazetteer generator: gather names and related neighborhood names of large cities in the state of Maryland
Ex: Mapping US County Rent Information

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>1436</td>
<td>Union County</td>
<td>Cranford</td>
<td>New Jersey</td>
</tr>
<tr>
<td>22</td>
<td>1918</td>
<td>Union County</td>
<td>Cranford</td>
<td>New Jersey</td>
</tr>
<tr>
<td>23</td>
<td>1775</td>
<td><strong>Union County</strong></td>
<td><strong>New Providence</strong></td>
<td>New Jersey</td>
</tr>
<tr>
<td>24</td>
<td>4296</td>
<td>Rockland County</td>
<td>Suffern</td>
<td>New York</td>
</tr>
<tr>
<td>25</td>
<td>1131</td>
<td>Westchester County</td>
<td>Elmsford</td>
<td>New York</td>
</tr>
</tbody>
</table>

- Generate a consistent location for a row entry
  1. 23 instances of “Union County” (red)
  2. One “New Jersey” (blue) consistent with a “Union County”
  3. One “New Providence” (orange) consistent with both “Union County” and “New Jersey”

- Notice use of colors to differentiate attributes
Ex: Mapping Stimulus Money Spending

| DEPT OF THE AIR FORCE | MARION          | IN  | 462414812 | $744,880.00 |
| DEPT OF THE AIR FORCE | MARION          | IN  | 462414812 | $744,880.00 |
| DEPT OF THE AIR FORCE | DAVIDSON        | TN  | 372011815 | $21,982.21  |
| DEPT OF THE ARMY      | SANTA BARBARA   | CA  | 934371499 | $249,951.00 |
| DEPT OF THE AIR FORCE | WICHITA         | TX  | 763112716 | $245,783.00 |
| DEPT OF THE AIR FORCE | WICHITA         | TX  | 763112716 | $428,090.00 |
| DEPT OF THE AIR FORCE | WICHITA         | TX  | 763112746 | $772,141.00 |
| DEPT OF THE AIR FORCE | BEXAR           | TX  | 782365253 | $1,570,941.63 |
| DEPT OF THE AIR FORCE | WICHITA         | TX  | 763112746 | $375,796.37 |
| DEPT OF THE AIR FORCE | LOWNDES         | GA  | 316991794 | $68,761.26  |
| DEPT OF THE AIR FORCE | WICHITA         | TX  | 763112746 | $519,029.00 |

- $ indicate locations where stimulus money has been spent
Ex: Mapping Census Response Rates

- Spreadsheet only contains location names and values.
- Display reveals that locations are in Maryland.
- Colors indicate ranges of Census response rates.

<table>
<thead>
<tr>
<th>Location</th>
<th>Response Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Frederick County</td>
<td>74</td>
</tr>
<tr>
<td>16 Garrett County</td>
<td>55</td>
</tr>
<tr>
<td>17 Harford County</td>
<td>75</td>
</tr>
<tr>
<td>18 Howard County</td>
<td>80</td>
</tr>
<tr>
<td>19 Kent County</td>
<td>62</td>
</tr>
<tr>
<td>20 Montgomery County</td>
<td>77</td>
</tr>
<tr>
<td>21 Prince George's County</td>
<td>68</td>
</tr>
</tbody>
</table>

Copyright 2014: Hanan Samet
Future Research Topics/Projects

- Machine learning for individual location classifiers
- Error feedback for learning better location classifiers
- Geocrowdsourcing for geotagging and building corpuses
- Window query efficiency and caching
- Near duplicate image detection
- Incorporation of foreign languages, translation, and clustering using them
- Cluster creation/death, measurements, and GPU implementation
- Keyword searching in NewsStand using pyramid model of news so use first paragraph instead of just headline as is currently the case
- TwitterStand seeder identification
- Improve precision of toponym resolution
- Devise corpuses for evaluating geotagging process
- Cloud-based implementation and quality of service for many users
- Android and Windows 8 ports
- Spatiotemporal visualization such as disease spreading
- Automatic incorporation of ontologies as done for diseases and people