The Picture of Health: Map-Based, Collaborative Spatio-Temporal Disease Tracking

Hanan Samet*

hjs@cs.umd.edu

Department of Computer Science

University of Maryland

College Park, MD 20742, USA

*Based on joint work with Rongjian Lan and Michael D. Lieberman

Introduction

- Disease tracking
  - US Centers for Disease Control and Prevention
  - World Health Organization
  - Volunteered geographic information in media such as blogs and tweets are increasingly coming into their own
  - ProMED web-based online alert system
- Goal: spatio-temporal querying and retrieval of ProMED documents
  - Extend STEWARD system for spatio-textual retrieval of documents from hidden web to handle ProMED documents
    - Yields automatic geotagging of ProMED documents
  - Extend STEWARD to allow for spatio-temporal querying through the use of a time slider
    - Enables varying the temporal components of queries by changing the time range under consideration
    - Pressing a “Play” button results in slider moving automatically across entire time spanned by all documents that satisfy the textual component of the query

Pre-Web Related Work

- Highly curated and verified information limited in scope to general public health information rather than real time disease surveillance
  - OutbreakNet from US Centers for Disease Control and Prevention
  - Eurosurveillance journal from European Centre for Disease Prevention and Control
  - Global Public Health Intelligence Network (GPHIN) from Public Health Agency of Canada
- Also client-driven
  - Specific diseases
  - Specific organizations such as hospitals
  - Groups of individuals such as travelers

Post-Web Related Work

- Wide range of nontraditional sources not specifically dedicated to health information
- Extract health information of possible interest
- Much larger and noisier than pre-web sources
- Aggregators from all sources including news, ProMED, etc. like HealthMap
- Collaborative disease tracking like GermTrax that relies primarily on disease reports from ordinary people who are sick
- None have geotagging as usually assign locations to disease reports based on minimal metadata which often has nothing to do with locations of the disease occurrence
- Infer locations of disease occurrences by monitoring IP addresses of keyword searches related to a disease such as flu and relevant medications

ProMED

- Online alert system intended to act as a clearinghouse by quickly disseminating news of infectious disease outbreaks to medical professionals and other subscribers around the world
- ProMED's editors monitor news, official government reports, and online disease summaries to learn of new cases or updates to existing cases
- Monitored diseases are limited to those of humans, animals, and plants
- Often medical professionals send local reports of diseases, or commentary related to existing disease reports, directly to ProMED's editors
  - Editors vet and verify each report
  - If find it accurate, then republish to ProMED's subscribers on one or several mailing lists, organized by topic, such as animal or plant diseases, emerging disease reports, broad location-oriented posts (e.g., Africa, Latin America, Southeast Asia), etc.
  - Prior to republication, editors add metadata to each report, including report's date and time, organisms of relevance (i.e., human, animal, plant), and location or locations affected by the disease outbreak
  - May modify report's text or add suitable commentary from contributors
- Synthesizes reports to single digests for easier republishing 1–2 times/day

Example of Typical ProMED Posting

```
Published Date: 2003-03-15 23:50:00
Subject: PRO/ALL> Severe acute respiratory syndrome - Worldwide:alert
Archive Number: 20030315.0637
[...]

World Health Organization issues emergency travel advisory
Severe Acute Respiratory Syndrome (SARS) Spreads Worldwide
----------------------

15 March 2003 | GENEVA -- During the past week, WHO has received
reports of more than 150 new suspected cases of Severe Acute
Respiratory Syndrome (SARS), an atypical pneumonia for which cause has
not yet been determined. Reports to date have been received from
Canada, China, Hong Kong Special Administrative Region of China,
Indonesia, Philippines, Singapore, Thailand, and Vietnam. Early
today, an ill passenger and companions who traveled from New York,
United States, and who landed in Frankfurt, Germany were removed from
their flight and taken to hospital isolation.
[...]
```


Structure of ProMED Posting

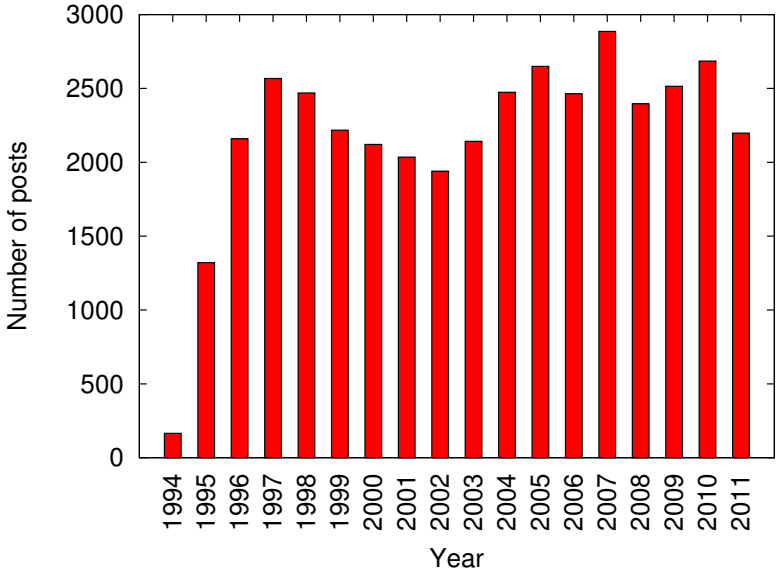
- Post released during the 2003 outbreak of severe acute respiratory syndrome (SARS) in southeast Asia and other parts of the world
- A reposting of a World Health Organization (WHO) travel advisory
  - Relating details of disease AND
  - Warning for travelers to and from locations affected by outbreak
- Post begins with ProMED metadata that is present in all postings
  - PUBLISHED DATE
  - SUBJECT
    - Tag identifying mailing lists of relevance—in this case, PRO/ALL indicating posting's relevance to all of ProMED's mailing lists
    - Disease of relevance: SEVERE ACUTE RESPIRATORY SYNDROME
  - Geographic location of relevance, which is WORLDWIDE here
    - Geographic locations are prominent due to the heavily geographic nature of reports of disease outbreaks
  - ARCHIVE NUMBER : relates date of posting, subject of post, and a unique post identifier which can be used to retrieve the original post from ProMED's website

Rest of ProMED Posting

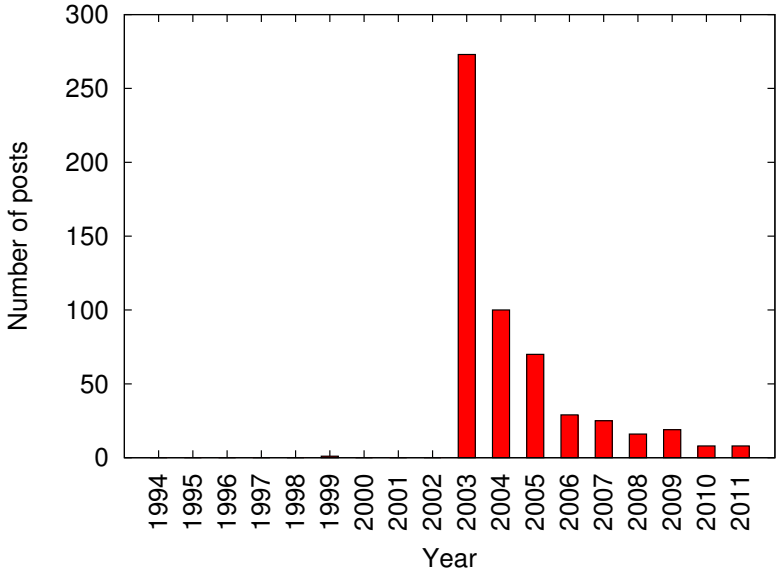
- Full text of posting
- Similar structure to news articles
- Dateline: date and location of posting
  - Date of posting
    - Not necessarily relevant to events in posting
    - For digest and summary posts, data may not be relevant since really a grouping of older posts and will not represent latest information
    - Actual text may provide clue to date (e.g., DURING THE PAST WEEK a large number of cases were reported)
  - Location of posting
    - Not necessarily where occur (e.g., Geneva is location of WHO headquarters)
  - Tags not enough for spatio-temporal queries; need content
  - Clear format and both date and location are machine-parseable

ProMED Data Retrieval

- Crawled ProMED website and downloaded archive of posts from 1994 to 2011 which numbered 39,420 in total
- $\approx$ 2,000–2,500 posts/year or 6/day
- For SARS many at outbreak in 2003 and then taper off
  - Counts are too coarse as want finer temporal ranges
  - Can be obtained by zooming-in time wise



ProMED posts per year



SARS posts per year

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 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

Live Demo: STEWARD System

http://steward.umiacs.umd.edu

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

STEWARD Control Flow

- Standardize each document's format
  - Easy for ProMED as simple text but STEWARD can handle variety including PDFs, MS Word, and HTML
- Extract relevant metadata such as time of publication, title, and relevant keywords
- Insert posting into STEWARD's PostgreSQL database
- Index each posting's text using an inverted index
- Geotag the document

Geotagging: Finding Toponyms

- Post's title often contains general locations - but usually too coarse
- Find all references to locations in text (toponyms)
- Associate toponym with spatial interpretation (i.e., lat/long value)
- Difficult due to ambiguities: is "Paris" of interest?
  - Toponym recognition: city: "Paris, France" vs: person: "Paris Hilton"
  - Toponym resolution: which "Paris"? "Paris, France" or "Paris, Texas"
- Use hybrid approach involving natural language processing (NLP):
- Named-entity recognition (NER):
  - Find typed entities within free running text including persons, organizations, locations, stock symbols, dates and times, etc.
  - Leverage NER tools for toponyms by restricting to location entities
- Part-of-speech (POS) tagging:
  - Associate each word or token in text with corresponding part of speech
  - Names of locations, and other types of entities, tend to be proper nouns, so take adjacent groups of proper nouns as toponyms
  - Of course, some of these proper nouns will not be locations, but they will be filtered out in subsequent steps of processing

Geotagging: Interpreting Toponyms

- Associate toponym with one or more spatial interpretation (i.e., lat/long value)
- Use a location *gazetteer*
  - Database of locations and associated metadata
  - STEWARD merges two freely available gazetteers:
    1. Geographic Names Information System (GNIS)
       - 2.2 million US location interpretations
    2. GeoNET Names Server (GNS)
       - 5 million non-US location interpretations
  - Implies STEWARD has excellent coverage of locations but at cost of increased toponym ambiguity and hence geotagging difficulty
    - But wide coverage is necessary for obtaining smaller locations of relevance crucial in disease detection and tracking

Geotagging: Toponym Resolution

- Involves variety of heuristic evidence applied through rules
- Ex: object/container pairs such as “Paris, Texas” indicate that interpretation of “Paris” should be implied by the interpretation of “Texas”
- Use *pair strength* algorithm where pairs of toponym interpretations within the document are ranked according to their document distance, geographic distance, and population
- Ranked pairs are then sorted, and toponyms are greedily resolved using the first pair in which they appear

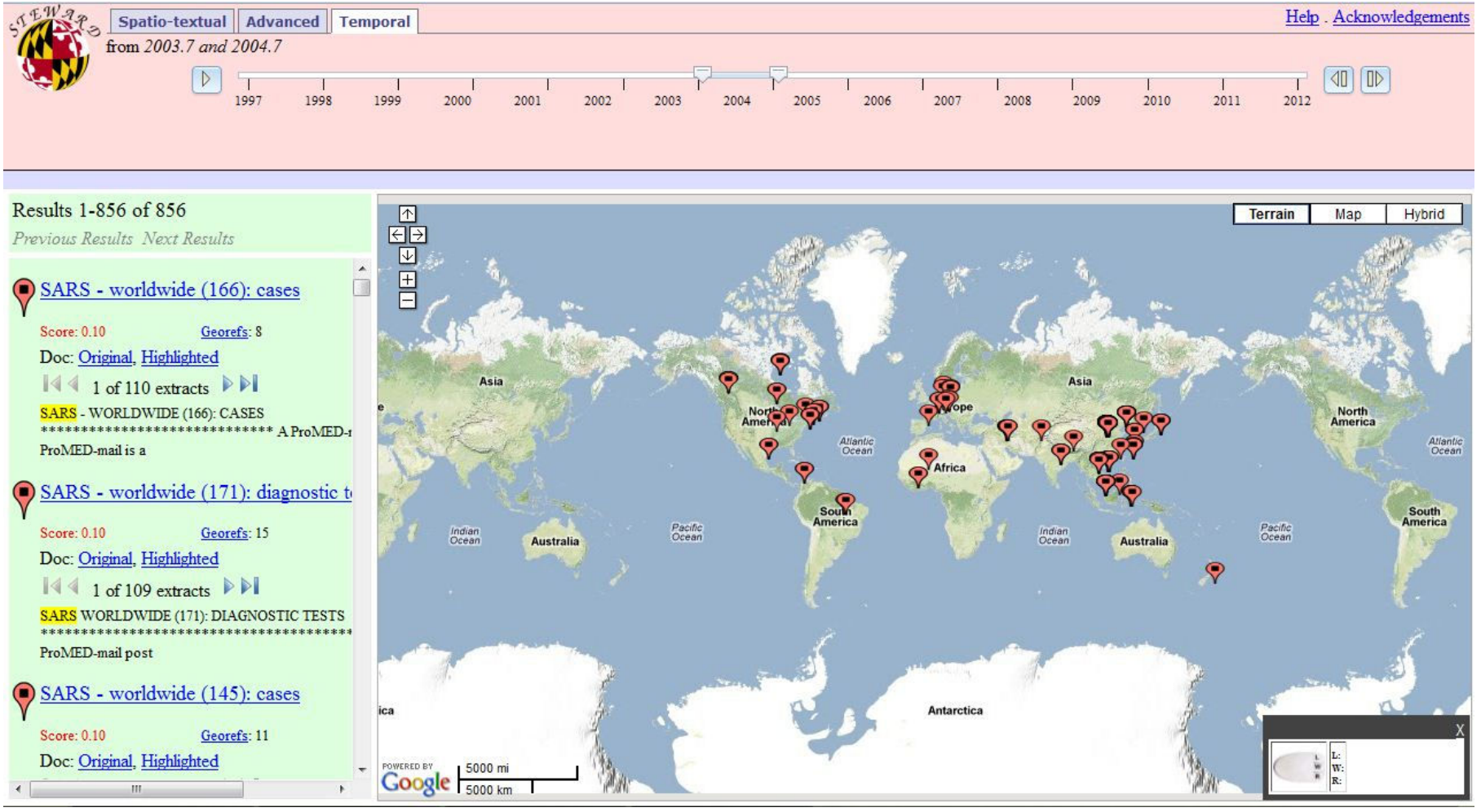
Geotagging: Document Geographic Focus

- Final stage involves finding an overall geographic focus of each document by ranking locations present in it
- Use a combination of document frequency and document position
  - Toponyms mentioned earlier are presumed to be more important to the content as a whole than those mentioned later
    - Based on “pyramid” model of writing news articles
  - Ranking locations in this way means that spatial components of queries to STEWARD return more relevant postings

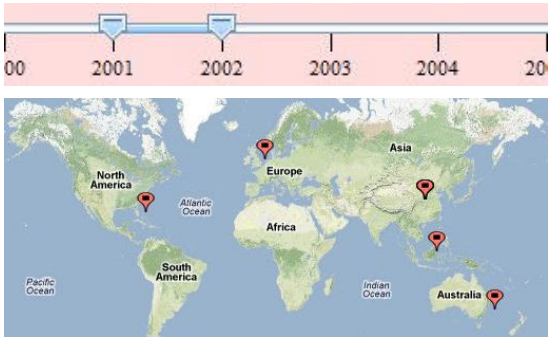
Spatio-Temporal Querying

- Execute keyword and spatial components in STEWARD's database and retrieve relevant documents R
- Can also return the first k ranked elements of R
- Use time slider to subsequently specify temporal parameters
  - Parameters determine which documents' in R to map
  - Client makes determination when rendering results
  - Thus temporal query component currently functions as a client-side post processing filter, although it would be better to integrate such queries more closely with the spatio-textual components
- Dynamic and configurable time slider demands a correspondingly fast implementation of temporal querying
- But for queries with relevance to a large part of documents, a large number of result documents R is returned from the initial spatio-textual query
- Initial implementation had interactivity issues for large query results, due to screen's limited update rate, as well as Web browser's slow script handling
- Prior to temporal querying via slider, index result documents by placing them in a list and sort by time, using an in-place Quicksort implementation
  - Temporal range querying reduces to binary search in sorted list

Using STEWARD to Monitor Disease Reports over Time



Example of Time Slider



2001

2002

2003

- Origin and spread of 2003 SARS outbreak
- Markers correspond to geotagged ProMED postings
- As move time slider across times of interest, locations affected by outbreak are easily apparent

Rendering Temporal Query Updates

- When users move slider, remove markers falling outside range and add those that are now relevant
- Could implement by clearing map of all markers after every update, and then reading only the markers that are relevant to the time range
- Unfortunately, this incurs unacceptable performance penalties and noticeably reduces interactivity, especially when using automatic slider
- Instead, leave in place those markers that are unaffected by the query update, i.e., those that remain in the range after the update

Improvements to Web Interface

- Often all documents within the time slider's range are displayed on map
  - Some time periods have too many disease outbreaks
  - Results in map being completely filled with markers that overlap significantly
  - Impedes understanding
- Rank and display documents on map using mix of
  - *Importance* measured by prevalence of disease at location, and
  - *Geographic spread* where we ensure good coverage of map
  - Both are useful from a usability perspective as
    - Don't overwhelm with too many markers on the map, and
    - Fewer markers to retrieve and better performance
- Can use timeline to augment users' exploratory capability by examining distribution of documents throughout timeline and suggesting times of interest for users to query
  - Do by clustering documents in the time dimension, and searching for clusters and outliers

Concluding Remarks

- Can augment to consider additional data sources for processing
  - Postings in language other than English and use machine translation capabilities to find relevant disease reports in these languages
  - Use PubMed data with case studies of disease incidences as well as more recent disease reports
    - Retrieved all PubMed documents from 2011, consisting of 885,316 documents, and currently processing them for inclusion
  - Use tweets from individuals with particular diseases
    - As each tweet contains a large amount of metadata, including GPS values and time stamp, this would provide a source of time-stamped geographic information
    - Need to develop methods to filter out vast majority of tweets which are not disease-related and likewise to determine the veracity of disease-related tweets and, in turn, the tweeters!
- Recognizing dates and time in addition to locations
  - E.g., “yesterday”, “last week”
- Distinguish between digests, daily reports, and incidences
- Develop and integrate spatio-temporal indexes that combine times and locations and incorporate into database’s query execution plan generator