Augmenting Spatio-Textual Search With an Infectious Disease Ontology

Michael D. Lieberman
Jagan Sankaranarayanan
Hanan Samet
Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742 USA
{codepoet,jagan,hjs}@cs.umd.edu

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

***Abstract*— A system is described that automatically categorizes and classifies infectious disease incidence reports by type and geographic location, to aid analysis by domain experts. It identifies references to infectious diseases by using a disease ontology. The system leverages the textual and spatial search capabilities of the STEWARD system to enable queries such as reports on "influenza" near "Hong Kong", possibly within a particular time period. Documents from the U.S. National Library of Medicine (http://www.pubmed.gov) and the World Health Organization (http://www.who.int) are tagged so that spatial relationships to specific disease occurrences can be presented graphically via a map interface. In addition, newspaper articles can be tagged and indexed to bolster the surveillance of ongoing epidemics. Examining past epidemics using this system may lead to improved understanding of the cause and spread of infectious diseases.**

I. Introduction

Technology can be used to understand the source and spread of disease epidemics to contain future outbreaks, thereby reducing a potentially massive toll on human life. Even though epidemiological information is available for many pathogenic microbes, disease incidence reports are scattered and difficult to summarize. In this paper, we describe the workings of an infectious disease monitoring system that automatically classifies and organizes disease incidence reports, based on geographic location and type, for analysis by domain experts. The system searches documents on the web for references to infectious disease names, as well as references to geographic locations. If a document mentions "cases of Avian Flu in Indonesia", our system is able to identify "Avian Flu" as an infectious disease and "Indonesia" as a geographic location. The system then associates that document with the appropriate disease type, as well as the set of latitude/longitude coordinates of the geographic locations found in the document, after which the document is displayed on a map interface. We refer to those web documents containing references to infectious diseases as *incidence reports*.

This work was supported in part by the U.S. National Science Foundation under Grants EIA-00-91474, CCF-05-15241, and IIS-07-13501, as well as the Office of Policy Development & Research of the Department of Housing and Development (HUD PD&R) and Microsoft Research.

Our system is distinguished from existing disease monitoring systems by making use of the fact that infectious disease outbreaks have strong geographic components. That is, the agent responsible for a disease's propagation follows a marked trajectory in space. We hypothesize that local cases of infectious diseases are generally first reported by local newspapers. By scanning and monitoring thousands of local and regional newspaper websites, we can monitor infectious disease outbreaks more effectively than existing systems and respond to incidents more quickly than ever before. Presently, we include documents from the National Library of Medicine and the World Health Organization. The system has a search interface where results are presented graphically via a map.

At this point, we distinguish our system from two other prominent existing disease monitoring systems: the Global Public Health Intelligence Network (GPHIN) (http://www.phac-aspc.gc.ca) and the International Networked System for Total Early Disease Detection (INSTEDD) (http://google.org/publichealth.html). Once completed, our system will be similar to the GPHIN and INSTEDD systems, in that it will continuously scan thousands of newspaper websites for incidence reports. Note that in GPHIN and INSTEDD, domain experts examine incidence reports in detail to determine that they are not false positives. However, our system also takes into account the geographic foci of incidence reports, which are useful for infectious disease tracking. It takes advantage of the *visual computing* aspects of maps; human eyes are adept at identifying spatial patterns that cannot be easily identified by a computer program. An expert using our system would quickly find disease outbreaks by seeing a cluster of incidence reports mapped to a particular geographic area. For example, the presence of a large number of Avian Flu reports on the map in and around Indonesia might indicate a contagious strain of Avian Flu in the region. By displaying incidence reports on a map interface, we can understand the spatial and temporal aspects of the spread of an infectious disease, and possibly even predict its future trajectory.

Our infectious disease monitoring system identifies textual

references to geographic locations by leveraging on the STEWARD system [1], a spatio-textual search engine built by us. Note that this problem is not trivial. For example, a reference to "London" in a document could refer to "London, UK", "London, Ontario, Canada", or 1500 other locations named London around the world. Moreover, the term "Washington" could refer to *persons*, *organizations*, or hundreds of geographic locations around the world. We also use STEWARD to compute the geographic focus of each incidence report — the set of important geographic locations in the document. For example, if an incidence report of Dengue fever in India appears in the Singapore Strait Times, "Singapore" might appear in the report, but would not appear as the report's geographic focus. For a more complete description of the STEWARD system, as well as related work, refer to [1].

Our system identifies textual references to infectious diseases by using an *ontology* of infectious diseases. An ontology is a hierarchical database of the important concepts and relationships in some knowledge domain, which in our case is infectious diseases. For a particular infectious disease, our ontology includes the disease's medical name, common name, scientific classification of the disease-causing pathogen (in terms of *class*, *order*, and *genus*), common symptoms, and relationships to other diseases. In this paper, we describe a simple technique to identify references to infectious diseases in documents using this ontology. Note that our technique is generalizable for use with other ontologies.

The rest of the paper is organized as follows. Section II contains the architecture of our system. We begin by describing the STEWARD system's architecture, as well as the additional modules that enable STEWARD to identify references to infectious diseases. Our system's querying capabilities are demonstrated in Section III. Finally, Section IV discusses future work and presents concluding remarks.

II. Architecture

This section provides a brief overview of the STEWARD architecture, and describes modifications to the original pipeline that enable our application's use of a disease ontology. For a more in-depth discussion of STEWARD's architecture, see [1].

A. Ontology Structure

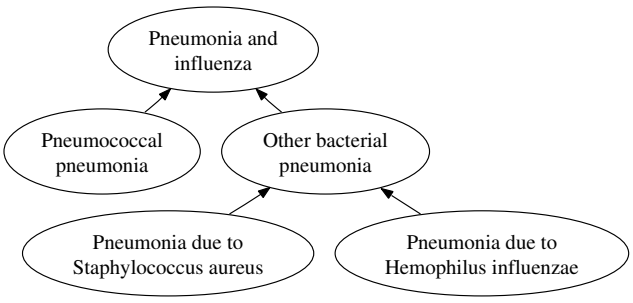


Fig. 1. A subset of our disease ontology, showing relationships between the various forms of pneumonia.

The most important enhancement to STEWARD's architecture is a domain-specific knowledge database known as an *ontology*. For our application, we used a *disease ontology*, a database of infectious diseases and associated metadata. An important challenge that we address in our system is the automatic integration of ontology information with document content. For our system, we adapted the ontology used by The Institute for Genomic Research (TIGR) as part of their Gemina project (`http://gemina.tigr.org`). This ontology is ordered in a hierarchical manner with diseases arranged in order of increasing specificity of disease descriptions. The ontology provides standardized disease names, as well as commonly-used synonyms for the disease as used in medical literature. Figure 1 shows a small subset of the ontology used in our system.

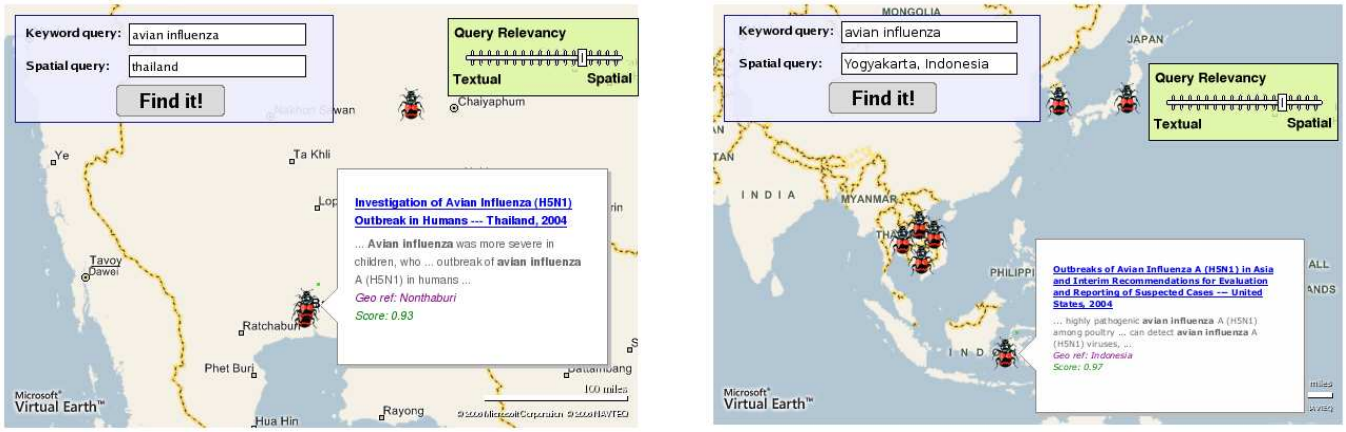
A disease ontology designed for human use can contain grammatical textual descriptions of diseases. For example, our disease ontology has descriptive names, such as "Pneumonia due to *Streptococcus pneumoniae*". However, an ontology of this type does not lend well to automated computer processing. It may be difficult to match these textual descriptions to document text, because there may not be an exact match for the ontological descriptive text. However, note that not every word is important when it comes to matching document text and ontology descriptions. That is, the words "due" and "to" in the above example are not relevant to a correct match, while "Pneumonia" and "Streptococcus pneumoniae" are related. This means that a document mentioning the disease name (*i.e.*, "Pneumonia", and possibly its standard name), but not the superflous words (*i.e.*, "due" and "to") is considered a good match.

To discount these superfluous words, we apply a preprocessing stage using the *Inverse Document Frequency* (IDF) measure [2]. IDF for a particular word w is computed as the logarithm of the number of all documents in a corpus divided by the number of documents that mention w. It emphasizes those words that do not appear frequently in the document corpus. For all entry names in the ontology, we weight each word with its IDF score. Furthermore, we compute a maximum potential score for each entry in the ontology by summing the IDF values of each word in the entry name. IDF performs well for our disease ontology, as names in the ontology mainly consist of either very specific and thus high-scoring words (*e.g.*, *Pneumonia*, *Streptococcus*, ...) or language articles with low IDF scores (*e.g.*, a, the, by, ...). STEWARD uses these IDF scores during document processing to determine the importance of partial matches of ontology entries (see Section II-E).

B. Document Retrieval and Standardization

Documents come in a variety of formats, such as text, HTML, Microsoft Word, and PDF. However, to simplify document processing in later stages of the pipeline, the initial phase of document processing involves retrieving the document and standardizing it. Later stages in the pipeline will therefore operate on a uniform document format.

While STEWARD is designed to work on unstructured or untagged documents, we can make use of information provided by metadata such as Medline tags to produce more



(a) (b)

Fig. 2. A prototype of our disease tracking and monitoring system, showing cases of "avian influenza" in the vicinity of (a) Thailand and (b) Indonesia.

accurate tagging. In particular, many medical documents about infectious diseases contain the disease name in their title. If the document uses the Medline format and has a tag containing the document title, STEWARD gives more importance to diseases in its ontology that appear in the title. STEWARD also uses the title metadata to give more weight to geographic locations found in the title.

C. Feature Vector Extraction

STEWARD continues with its geographic location extraction by discarding most of the words in the document that most likely are not textual references to geographic locations (*e.g.*, "the", "and", ...). It uses two *Natural Language Processing* (NLP) based techniques, known as *Part-Of-Speech* (POS) tagging and *Named-Entity Recognition* (NER) tagging, to aid in extracting the document's *features*, collectively called its *feature vector*. Intuitively, the feature vector is the set of most interesting words in a document — that is, the words most likely to refer to geographic locations.

D. Geotagging

The next stages in STEWARD's pipeline are responsible for *geotagging* of the document — associating the document with all references to geographic locations contained therein. After searching a *gazetteer*, or database of geographic locations, STEWARD augments each document's feature vector, adding geographic references from the gazetteer to each feature that is a potential geographic location. It then runs a *disambiguation* algorithm to choose the most likely gazetteer record to associate with each feature. Finally, STEWARD determines the subset of geographic locations that are most prevalent in the document, known as the document's *geographic focus locations* or simply its *geographic focus*. These locations are associated with the document in STEWARD's database to enable spatio-textual queries on the document collection. For a more thorough explanation of STEWARD's disambiguation and focus algorithms, refer to [1].

In our application to disease monitoring, we process reports of disease outbreaks, which tend to have strong geographic foci. Focus locations will therefore be instrumental in restricting disease searches to particular geographic areas.

E. Ontology Tagging

STEWARD next uses its ontology to find document keywords that match names of entries in the ontology, a process termed *ontology tagging*. Each word in the document is searched against the ontology, which returns potential matches of ontology entries. When a document partially matches an entry in the ontology, STEWARD computes a normalized score for the match by summing the IDF scores of the matching words and dividing by the entry's maximum potential score. Those entries with scores above STEWARD's predefined threshold are reported as relevant for the document, and are collectively termed the document's *ontology features*. Ontology features are ordered by decreasing match score. Furthermore, if multiple ontology features have the same match score, the feature with the largest IDF score (having the least common name) is reported first, as the words in that feature name are less likely to occur by chance.

F. Ontology Focus Determination

After the document has been associated with ontology features, STEWARD determines the subset of ontology features that are most prevalent in the document's text. For each feature, STEWARD extends the previously-described IDF score by multiplying it with a *Term Frequency* (TF) [2] term. The TF term for a word w is computed by dividing the number of occurrences of w in the document by the total number of words in the document. It therefore places emphasis on those words that occur frequently in the document. By multiplying the terms, we obtain the *Term Frequency-Inverse Document Frequency* (TF-IDF) [2] score for each word of the ontology feature, which will be large for words that occur frequently in the document but infrequently in the corpus. For a document with several ontology features, the features with the largest TF-IDF scores are selected as the document's *ontology focus*.

In our application, the document's ontology focus will correspond to the disease that is most prevalent in the document.

Disease outbreak reports tend to focus on a single disease, so the TF-IDF score for one disease most often stands out from other diseases reported as noise.

III. Application: Disease Tracking

To create our application for tracking infectious disease outbreaks, we retrieved documents from the ProMED-mail database (`http://www.promedmail.org`). ProMED-mail is an e-mail reporting system with several moderated mailing lists, where medical professionals around the world post reported cases of infectious diseases. This dataset is useful for an infectious disease monitoring system, as outbreaks of disease are reported quickly, sometimes within days of their occurrence. Disease reports are also available in several languages, including English, Spanish, Portuguese, and Russian.

We also applied our techniques to a subset of the data available through PubMed (`http://www.pubmed.org`), a service funded by the U.S. National Library of Medicine. PubMed provides access to Medline, a database of 16 million abstracts of documents published in medical journals. We downloaded 43,000 abstracts whose content was relevant to infectious diseases, and processed them using the modified STEWARD pipeline described in Section II. We are also working to index articles from thousands of online newspapers.

Figure 2 shows two screenshots of our disease tracking and monitoring system's user interface. Users enter spatio-textual searches using the input form shown at the top. The form has separate textual and spatial input fields, so that a user can specify a text query, spatial query, or a combined spatio-textual query. For combined spatio-textual queries, STEWARD's query engine determines the best order to process each query component by estimating the size of each result.

In addition, a query relevancy slider allows the user to choose how relevant query results should be to the keyword or spatial query components. For example, consider a spatio-textual query for outbreaks of "bovine spongiform encephalopathy", more commonly known as "mad-cow disease", in the United Kingdom. If a user was also interested in recent disease outbreaks of other diseases in the UK, she would place more emphasis on the spatial aspect of the query. However, if she was more interested in related diseases such as Creutzfeldt-Jakob disease, but not necessarily in the UK, she would emphasize the textual query.

The screenshots in Figure 2 show spatio-textual queries for outbreaks of "avian influenza" near Thailand and Yogyakarta, Indonesia. Notice that even though these locations were given as spatial query specifiers, STEWARD found other documents that mentioned nearby locations as well. Because the query relevancy slider was set toward spatial, these result documents need not have mentioned avian influenza; their proximity to the query location was enough to include them in the result. This query demonstrates that STEWARD allows a user to discover geographic relationships between disease outbreaks, indicating possible correlations.

IV. Future Work and Concluding Remarks

STEWARD can incorporate the knowledge imparted by an ontology using the methods described in Section II. However, several improvements can be made that would make STEWARD a more effective ontology tagging and focus determination tool. We will quantitatively evaluate our system's effectiveness by measuring its precision and recall for various disease queries. Also, we currently use IDF and TF-IDF scores to determine what words in the document are ontology features, but these measures do not take word context found in language into account. For example, the phrase "infection of" might be a good indicator that the next word or phrase is an infectious disease. Thus, a more appropriate method might be to train a named-entity recognizer to find references to disease names, similar to how STEWARD currently recognizes references to geographic locations. This would require a corpus of documents pre-tagged with infectious diseases, which might be difficult to obtain or create.

We do not currently process documents from ProMED-mail that are written in languages other than English. This is a potentially useful set of data, as disease reports are sometimes only available in a certain language. For example, many disease reports from areas in former Soviet states are posted only in Russian. While our methods should work for documents in any language, STEWARD would need additional part-of-speech and named-entity models, trained separately for each language. We plan to train and use these models in subsequent versions of our disease monitoring system.

Our disease ontology is organized hierarchically, which provides a useful way to group related diseases, rather than simply relying on disease names to determine relationships. Instead of filtering search results according to a single disease, a user might be interested in disease reports for a family of related diseases. STEWARD would thus benefit from an additional query module that returns disease reports with mentions of diseases that are close in the hierarchy, without necessarily sharing words in the disease name. For example, a search for "*Streptococcus pneumoniae*" could return other diseases caused by the related pathogen "*Staphylococcus aureus*".

Our system currently provides basic spatial searching functionality, but it could be extended by incorporating more queries from the SAND database system [3].

V. Acknowledgements

The authors wish to thank Adam Phillippy for his help with the ontology tagging module.

References

[1] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling, "STEWARD: Architecture of a Spatio-Textual Search Engine," in *ACMGIS 2007*, Seattle, WA, Nov. 2007, pp. 186–193.

[2] D. Jurafsky and J. H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. Upper Saddle River, NJ: Prentice Hall, Jan. 2000.

[3] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin, "Use of the SAND spatial browser for digital government applications," *Communications of the ACM*, vol. 46, no. 1, pp. 63–66, Jan. 2003.