Geospatial and Entity Driven Document Visualization for Non-proliferation Analysis

Jeffrey Baumes, Jason Shepherd, Aashish Chaudhary

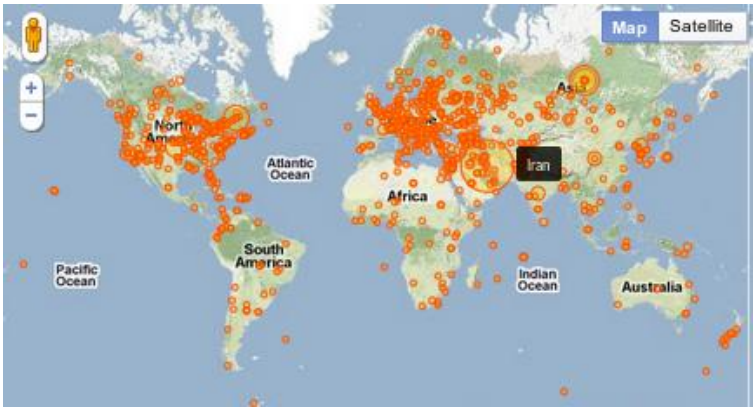

Figure 1. Overview of locations mentioned in a large corpus of non-proliferation documents.

Abstract—Non-proliferation researchers are tasked with understanding complex interactions between countries and nuclear technologies that encompass the world and span decades. This field poses a challenge for producing effective visual interfaces that enable informed decision making. In this paper we discuss the custom tools that were generated in response to this problem and the rationale behind these decisions. In particular, we present simple but effective mechanisms for presenting information related to the interplay between raw text, named entities, geospatial locations, and time in a single analytics framework.

Index Terms—non-proliferation, text analysis, geospatial visualization, entity recognition

1 Introduction

Due to the potential destructive power nuclear that weapons possess, the global community must keep a vigilant watch on the spread of this technology. Such continuous vigilance will remain a critical factor in securing peace and stability in the world. It is for this reason that an infrastructure is needed to effectively monitor and understand current and future threats.

News articles, reports, interviews, and surveys are just some of the types of data that may be useful in understanding where potential threats reside. The flood of potential data sources, too many for any group of individuals to manually understand, warrants the use of software tools that can pre-process and organize this information. There are already many databases, metrics, news articles, and other resources for proliferation data. But in order to make more informed decisions, constant cross-referencing of these data sources for corresponding dates, places, and people is necessary. Existing data sources must be effectively combined such that this information is automatically connected, and can be quickly and efficiently traversed.

This work presents the methodology and results of building a full prototype analysis and visualization system for the very specific audience of non-proliferation researchers in order to assist them in decision making. We present processes both on how we made choices on what analyses to run, as well as how we integrated these analyses into effective visualizations. To achieve this level of dynamic information management of proliferation data, we have developed a novel web-based analytical tool that processes data from several sources and automatically extracts important features useful for querying, navigating, and ultimately understanding large corpora of documents. Entity recognition routines discover places, names, organizations, dates, and other relevant content from the data. Documents and other records in this database are linked together through matching people, locations, and/or other features and presented to users as web pages with hypertext links among them, along with maps for geospatial reference.

A linked system of documents would allow individual documents to be better understood and placed into proper context with other related documents. However, the problem of data discovery from a higher level remains, for example discovering trends in the data. At this level, important questions include: How have the proliferation activities of country X changed over the last two decades? Where has technology Y has been transferred? In what places has organization Z been active? Currently, search queries and intuitive visual representations are lacking for effectively exploring existing databases in this way. Using search queries, the system gathers results from different public databases and provides not just a static list of documents, but interactive visualizations showing where and when these events took place.

- *Jeffrey Baumes is with Kitware Inc., email: jeff.baumes@kitware.com.*
- *Jason Shepherd is with Sandia National Laboratories, email: jfsheph@sandia.gov.*
- *Aashish Chaudhary is with Kitware Inc., email: aashish.chaudhary@kitware.com..*

Manuscript received 09/02/2011; accepted 09/19/2011.
For information on obtaining reprints of this article, please send *email to: tvcg@computer.org.*

In this paper, we present the process we used to develop a custom web application to enhance decision-making and highlight areas of concern in non-proliferation activities. The work combines data from different sources, performs named entity recognition to link documents, and finally enables multiple methods for traversing documents, combining geospatial visualization with entity networks. This tool enables researchers to better understand the state of proliferation in the world, and aids in the development of accurate models for predicting future proliferation activities. Figure 2 presents the general system architecture.

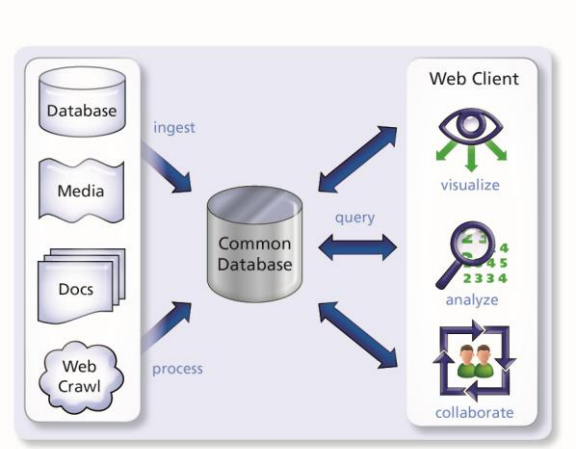

Figure 2. General system architecture for the non-proliferation document analysis application.

2 Related Work

NIS Database. One existing resource for proliferation data is being built by the James Martin Center for Non-proliferation Studies at the Monterey Institute of International Studies (cns.miis.edu). The many data sources collected there have been placed online at the Nuclear Threat Initiative website (www.nti.org) into the NIS Nuclear Trafficking Database, from which we imported documents (www.nti.org/db/nistraff/index.html). This is a collection of hundreds of abstracts on reports of nuclear proliferation activity in and out of the Newly Independent States, i.e. the former U.S.S.R.

GeoTracker: Chen et.al from AT&T Labs published their work [4] on geospatial and temporal browsing techniques that allowed users to aggregate and navigate RSS-enabled content efficiently. GeoTracker presented RSS data along geospatial and temporal dimensions, aggregated RSS feeds and mined RSS data to provide search capabilities.

Geografikos: The Geografikos package is a software system that creates a database of Wikipedia articles in which each article has associated geospatial entities extracted from the article. Geografikos allows geospatial query, information retrieval and geovisualization. Witmer, et al. [10] explained how they were able to extract geospatial named entities from a corpus of Wikipedia articles using Geografikos. Their algorithm provided context-based disambiguation which resulted in excellent geospatial entity extraction. Woodward [11] used Stanford NER for named entity recognition and used Geografikos along with the Google Maps API for basic geovisualization with limited interaction capabilities.

Web-a-Where: Amitay et al. developed a geotagging system that would locate mention of places in the web page and assign appropriate geofocus. The Web-a-Where systems lack visualization capabilities but achieved high level of accuracy for correctly tagging individual name place occurrences.

SEM: Xu et. al [12] built a prototype based on the Spatial Event Miner (SEM) algorithm. SEM consisted of two sub-algorithms: Event Location Retrieval (ELR) and Event Topic Mining (ETM). Given a query, SEM extracted the most relevant locations from search results retrieved by a query engine, gathered events topics from collected snippets and outputs, and classified locations together with associated events topics and descriptions. Simple web based visualization is then constructed using the output of SEM and the Google Maps API.

ManyEyes. ManyEyes (manyeyes.alphaworks.ibm.com) is an excellent example of how interactive, collaborative visual exploration can work. At the ManyEyes website, users may take individual datasets and plot the data using a set of visualizations. These visualizations then may be saved, viewed and discussed by others. The site has resulted in fruitful collaboration activities. ManyEyes provides limited capability for analyzing text data. Data values can be drawn as overlays on geographic regions using colors or bubbles.

3 Methodology

3.1 Data Collection

A first task for this project was to assimilate multiple data sources into a combined proliferation database. We were able to collect and process a variety of proliferation resources to show that the system can handle a variety of data formats and integrate them in a meaningful way.

In order to get a more substantial set of relevant documents, we coordinated work with Juliana Freire at NYU. She and her colleagues have performed significant research into web crawling techniques and classifiers used for identifying relevant document sets on the web. Using the NTI articles as a starting point, her group was able to gather two very relevant document collections. The first dataset was gathered using their interface with the Bing search engine. Through this, they retrieved about 2,000 documents. From manual inspection of a sample, a little more than half were very relevant pages of events, while others were simply dictionary definitions or wiki pages relating to the words nuclear, weapon, etc. The second method interfaced with Bing News, which retrieved 1,100 additional documents, almost all of which were very relevant.

In addition to the semi structured NTI documents and unstructured web crawl documents, we also utilized data from the PubMed medical journal article database in a structured XML format for our scalability tests [8]. Since there is no comparable proliferation database that matches PubMed in scale, PubMed was used for its size and organization of data to ensure our import capabilities are sufficient.

The process of extracting relevant and meaningful textual information from HTML involved careful algorithmic approaches that provide good balance between high probability of detection (via flexibility) and a low occurrence of false alarms (via high threshold). HTML documents belonging to the Nuclear Trafficking Database contain well-defined HTML tags. Accordingly, we were able to extract the relevant texts indicated by these specific tags using the open source tool BeautifulSoup [3]. Our algorithmic pipeline processes the incoming HTML data as follows:

- All the text found within the HTML element is extracted
- Most of the non-relevant HTML meta characters are replaced with string representations
- A line of text is kept or discarded based on the count of the number of words/characters. This heuristic was accurate in distinguishing article main text from peripheral elements such as menus, advertisements, and headings.
- Lines of text are processed in sequence, and every line is kept if it meets a specified minimum threshold.

Additionally, for the HTML containing news articles, empirical analysis indicated that relevant texts are better found by only considering text contained in paragraph tags (i.e. pairs). For

this particular dataset, we first extracted the text contained by these paragraph tags and then followed the procedure of text selection on a line-by-line basis.

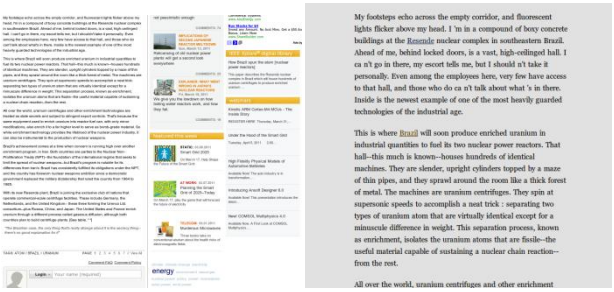

Figure 3. Comparison of a source web page (left) to a basic HTML version of the main text (right).

We see the capability of extracting text from arbitrary HTML as critical to this project for two reasons. First, in order to effectively use text analysis algorithms such as named entities and topic detection, the input must be clean and devoid of superfluous data such as ads and header text, which would skew and add noise to the results making them ineffective. Second, ideally we would like to focus on the content of the article instead of on the ads and menus that clutter the view. There are a few noteworthy tools that are taking similar measures to provide a distraction-free view for reading the content of web pages. The Safari browser has a "Reader" button that brings up a clean view of the current page for reading. The website Readability.com similarly has a method for importing arbitrary URLs into a view that is easy to read. In designing our own document text view, we took cues from these applications to make a pleasant reading experience for users. Figure 3 compares a source web page with the results of our automatic content extraction.

3.2 Visualization Design

Once the documents were collected, cleaned and stored in the database, we produced several interaction views and visualizations in our web application to navigate and quickly understand the content of the documents without needing to read every article. To accomplish this, we processed the documents with automated tools such as named entity extraction, geo-coordinate lookup, and topic models. In addition, we produced several interactive visualizations to present the results of these analyses in intuitive ways.

Within the document text itself, we automatically flagged items such as places and people. These items would appear as links in the system, allowing for simple navigation through related documents. We investigated the open-source Stanford Named Entity Recognition (SNER) toolkit as well as the Natural Language Toolkit (NLTK). Using out-of-the-box classifiers, SNER proved to be more reliable at detecting entities correctly after manually comparing results. These algorithms use a probabilistic model to flag people, places, and organizations automatically for each document. We also used custom pattern matching to detect and flag dates in the documents. All of these are stored in the database in a generic way so that in the future additional features (e.g. material names) can be detected and stored.

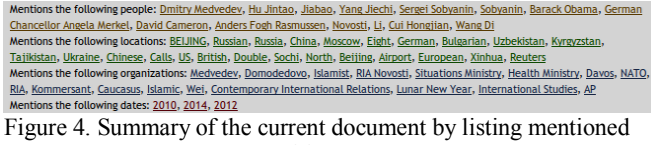

Figure 4. Summary of the current document by listing mentioned entities.

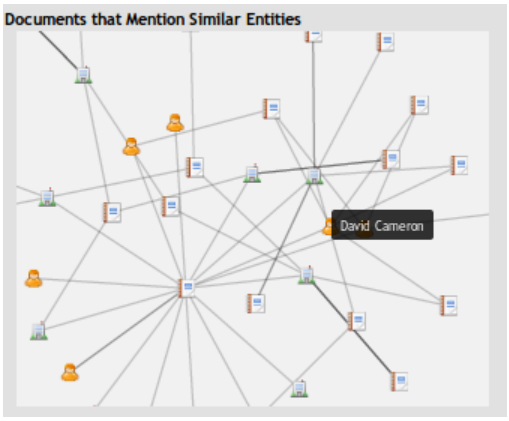

Figure 5. Interactive document-entity network.

In addition to adding hyperlinks to the text itself, we used the entity extraction information in two other views. The first view, shown in Figure 4, is a simple textual list of all entities found in the current document grouped by type. The user may quickly skim this list to determine whether the article is of interest before reading the entire document. The other entity view that we prototyped is an entity network view, which is not simply a summary view, but is also a navigation view enabling direct access to related documents (See Figure 5). This view contains all entities mentioned in the document, in addition to other documents that reference those entities. The relationships between documents and entities forms a node-link diagram where documents that share more entities in common with the current document can easily be seen as having more connections. Hovering the mouse over a document shows its title, and clicking on a document in this view navigates to the page with that document's details. These and other summary views are valuable to analysts as they reduce time necessary to analyze each search query.

Once the data has been processed and ingested into a database, we developed an efficient keyword search. Since our data was stored in a Postgres database, we investigated the full-text search capabilities of that storage engine. The full-text search features of Postgres were found to be more than adequate with automatic ranking and highlighted text snippets for content. Postgres also scaled to significantly sized document sets well, as described in Scalability Testing below. Figure 6 shows a sample search results page. In our search we included the ability to restrict searches to specific sets of documents, which we believe will be an important feature in a multi-user application with many different data sources, only some of which will be relevant to any particular user.

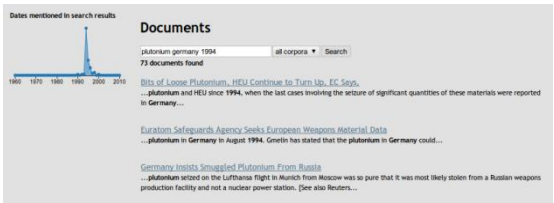

Figure 6. Sample query with a timeline (left) indicating what dates were mentioned in documents matching the query.

To offer a more usable interface for showing the characteristics of the documents retrieved from a search query, we researched various visualization techniques that will supplement the traditional flat list of data. The first was an interactive chart showing the documents categorized by date. We grouped the query results by the year they were written and by years mentioned in the text and presented the resulting histogram in a chart. Hovering highlights the number of articles for a particular year, and clicking in the chart will take the user to the list of documents associated with that year (see the left of

Figure 6). This chart was implemented using the Protovis javascript framework [7], which allows you to easily create rich visualizations including interaction and animation support.

For location references, such as cities or countries, we used the public GeoNames web service [5] to lookup geospatial coordinates for plotting on a map. The results of the raw data processing and named entity recognition were loaded into the database as a cache to reduce the number of look-ups.

Based on the geospatial tags detected in the documents, we developed intuitive views to show how the search results are distributed across the globe. For this visualization, we used the open Google Maps API. We started by placing standard icons, or placemarks, on the map using the built-in API functions. While effective, we saw that the icons were static and uniform, not conveying the locations that were truly the most important in the document. So in a later revision, instead of just using this API to put generic placemarks onto the map, we overlaid a richer Protovis visualization on top of Google Maps. This provided a much more intuitive view and icons can be scaled based on the number of times that location appears in the current document or set of documents. The resulting view for an overview of the entire document corpus is shown in Figure 1.

Figure 7. Interactive geospatial map with linked documents.

For exploring connections related to a single document or query, the view was also made more dynamic and interactive. To enable a user to visit other related documents, we added the capability to show documents mentioning that location in a ring when clicked with the mouse. The documents show their title when hovered over and clicking on the icon brings you immediately to that document's page (see Figure 7).

4 Results and Conclusions

As a result of this work, we have produced a live, publically accessible website [6]. To use the site, simply search for a term such as "Ukraine" or "India", and then click on documents that appear in the search results. Since this is a proof of concept, we have experimental views on the site and the web page load times and animations still need optimization. Also, we tested on the Google Chrome browser and Firefox; however, due to limitations in Internet Explorer (Protovis requires an SVG-based context that is not supported by IE) we do not yet support it.

Feedback on the site has been positive, specifically with regard to automatic detection and plotting of geospatial locations to give additional context to the article. Additionally, the visualizations and entity summaries provide ways to save valuable time by providing quick summarizations allowing irrelevant articles to be quickly skipped or disregarded. It was elucidating to discover that the simplest of visualizations, such as listing mentioned entities, can be considered as important or even more important than, elaborate visualizations.

Because we are developing a modular and open-source architecture, these analysis and visualization components may be easily integrated into other existing databases and tools.

Acknowledgments

We would like to thank Juliana Freire from New York University for assisting us in collecting document corpora from web crawls.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] **Amitay, Einat, Nadav Har'El, Ron Sivan, Aya Soffer**. Web-a-Where: Geotagging Web Content. SIGIR'04, July 25-29, 2004. Sheffield, UK.

[2] **Baumes, Jeff, Timothy Shead, Jason F Shepherd, Brian Wylie,** "Multi-model Analysis of Document Caches," *Report,* June 2011. (available at https://cfwebprod.sandia.gov/cfdocs/CCIM/docs/finalP2paper.pdf)

[3] Beautiful Soup. http://www.crummy.com/software/BeautifulSoups

[4] **Chen, Yih-Farn, Fabbrizio, Giuseppe Di, David Gibbon, Rittwik Jana, Serban Jora**. GeoTracker: Geospatial and Temporal RSS Navigation. WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.

[5] GeoNames. http://www.geonames.org.

[6] Nonproliferation Search Application Prototype. http://paraviewweb.kitware.com:83/Midas/midas/nonproliferation/documents

[7] Protovis. http://www.protovis.org

[8] PubMed. http://www.ncbi.nlm.nih.gov/pubmed

[9] **Wilson, Andy T, Michael W Trahan, Jason F Shepherd, Thomas J Otahal, Steven N Kempka, Mark C Foehse, Nathan D Fabian, Warren L Davis, IV, Gloria Casale,** "Text Analysis Tools and Techniques of the PubMed Data Using the Titan Scalable Informatics Toolkit," *Report,* May 2011. (available at https://cfwebprod.sandia.gov/cfdocs/CCIM/docs/pubmed_paper.pdf)

[10] **Witner, Jeremy and Kalita, Jugal**. Extracting Geospatial Entities from Wikipedia. 2009 IEEE International Conference on Semantic Computing.

[11] **Woodward, Daryl**. Extraction and Visualization of Temporal Information and Related Named Entities from Wikipedia. Report. (Available at http://www.cs.uccs.edu/~kalita/work/reu/REUFinalPapers2010/Woodward.pdf)

[12] **Xu, Kaifeng, Rui Li, Shenghua Bao, Dingyi Han, and Yong Yu**. SEM: Mining Spatial Events from the Web. PAKDD. pp. 393-404. 2008.