Health GeoJunction: Geovisualization of news and scientific publications to support situation awareness

Michael Stryker[1], Ian Turton[1], Alan M. MacEachren[1]

[1]GeoVISTA Center, Department of Geography

The Pennsylvania State University

302 Walker Building

University Park, PA 16802
{mzs114, ijt1, maceachren}@psu.edu

ABSTRACT

The Health GeoJunction application is a web portal that extracts information from scientific literature, public health reports, and news feeds to support geographically contextualized searches and the discovery of cross-connections between documents through a map-based interface of coordinated views. Text extracted from documents can be queried and represented as visual artifacts within a map, timeline, or an extended tag cloud. These linked views enable the user to progressively filter a collection of documents and provide an intuitive means for expressing queries in terms of the conceptual dimensions of location, time, and theme. In this paper we focus specifically on aspects of the Health GeoJunction client development, including support for the space-time-attribute perspective, information visualization and cartographic visual representations, and access to computational reasoning and information extraction tools on the server. To guide development, we have created scenarios for situation awareness and knowledge construction within a visual analytic environment for an analyst assessing the evolution of Avian Influenza as a public health threat from the exploration of RSS (Real Simple Syndication) feeds from the World Health Organization (WHO), World Animal Health Organization (OIE) reports on outbreak incidents, and the PubMed database of biomedical scientific abstracts.

1 BACKGROUND

The influenza pandemic of 1918-19 estimated to have resulted in 50 million deaths provides some indication of the impact an equally virulent influenza strain might have on the current human population that is now more mobile, interconnected and in more dense living arrangements. Infectious diseases with high mortality rates such as SARS and cases of animal-to-human contraction of Avian Influenza have drawn attention from the public. The scientific community that has noted a greater frequency of new infectious diseases and the role that wild animal populations serve as reservoirs for new zoonotic diseases (Jones, Patel et al. 2008). In response, a number of organizations contribute to a surveillance network monitoring the evolution of the Avian Influenza threat including the WHO, OIE, and Food and Agriculture Organization (FAO) of the United Nations. The prevalence of online mapping sites (Boulos 2003) for these data and the recent launch of the GISAID EpiFlu database (http://platform.gisaid.org) indicate that tracking this information and collaboration among researchers is an open challenge. Achieving situation awareness of the avian influenza threat provides a case study for the Health GeoJunction web portal and the visual analytic techniques employed.

2 VISUAL ANALYTICS SCIENCE AND RELATED WORK

Key objectives within visual analytic science are to leverage visual perception and cognition to manage large volumes of information through effective visual representations and visual interfaces to information repositories, aid the reasoning process and discourse with information through appropriate modes of interaction, and facilate the linkage of computational methods with application of human judgment(Thomas and Cook 2005). The development efforts for Health GeoJunction are focused on the primary technologies for extracting entities from text, disambiguating and geolocating placenames, enabling the expression of space-time-attribute queries through a visual interface, and investigating new multi-view geographic and information visual displays that correspond to the underlying conceptual perspective (Peuquet 1994; Andrienko, Andrienko et al. 2003). Situation awareness—the process of perceiving environmental cues, integrating that with an existing image schema of prior knowledge or creating new schemata, and then using this to estimate future conditions—grounds the Health GeoJunction feature design and implementation in an analytic process and in particular the process for tracking query histories with the state of the interface (Endsley 1995; Livnat, Argutter et al. 2005).

3 HEALTH GEOJUNCTION ARCHITECTURE OVERVIEW

Health GeoJunction is a client-server application developed for the Flash player web browser plug-in using Adobe Flex™ featuring a port of the Open Layers map viewer to Flex and support for Open Geospatial Consortium standard WMS and WFS. The FactXtractor web service parses documents to extract keywords and place names which are then disambiguated and georeferenced. Key word tags, place names, and other text attributes are then stored in a spatial database for additional analysis. These data are accessible through Geoserver for spatial queries and a custom java servlet for non-spatial queries.

4.1 Geographically contextualized searches

The map provides multiple filter options at country level geographic resolution (Figure 1). Divided proportional symbols indicate the number of publications by source, OIE report or PubMed abstract. When selected, a filter by country is applied including papers about places within that country. Ctrl-click extends the filter to include the selected country plus neighbors (defined as those sharing a physical boundary. PubMed abstracts provide the additional distinction of filtering by place of writing ('from'). For non-adjacent place selection or spatial relationships better understood within a hierarchy, we provide the MeSH place tree (not shown) and the geonames place hierarchy for specific spatial filters not constrained to the country administrative level.

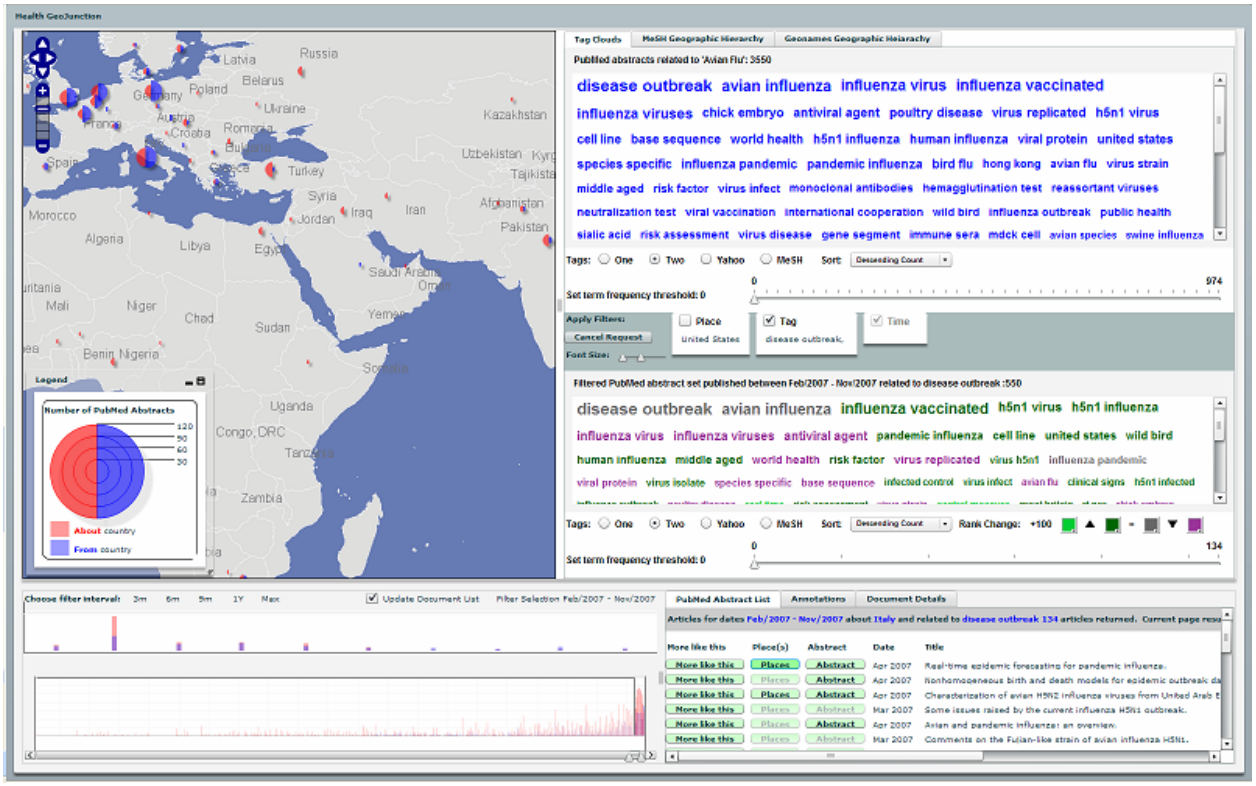

Figure 1. In this instance the map shows PubMed articles by country distinguishing 'from' or 'about'. Tag clouds provide a summary of the most frequent keywords in all the relevant documents (upper) and the filtered subset (lower). A timeseries and list of individual articles returned from the latest query fill the portion of the display.

The document footprint provides arcs on the map linking places and distinguishing between 'from' and 'about' place types. With this technique we are interested in determining how effectively one or more documents can be contextualized in terms of the geographic scope of interest, for how many documents the technique remains useful, and if querying for related documents via these properties is more easily expressed from this display than in the main map or geographic hierarchy tree control.

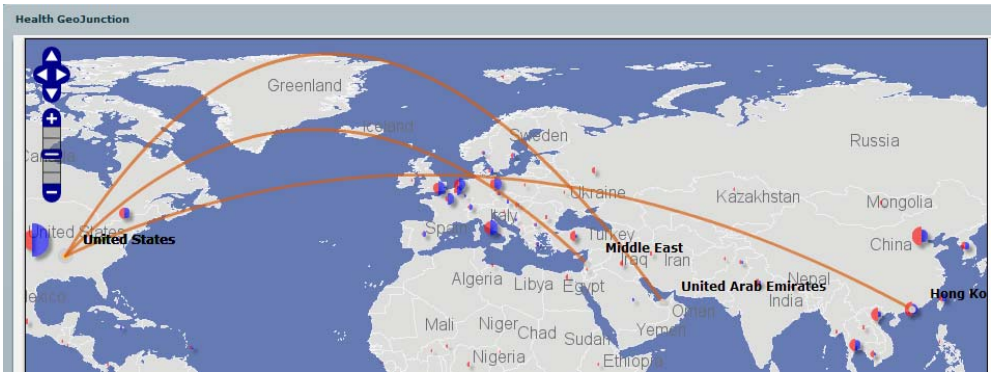

Figure 2. The arcs display a document footprint linking the publication site with the locations mentioned in the body of an abstract for the selected articles. This approach provides a drill-down approach for relating a smaller subset of documents to place or querying related documents with similar spatial properties.

4.2 Extending tag clouds to compare document collections

Tag clouds have become a common information visualization method for representing the most important topics mentioned in text documents. We have extended the tag cloud method to support comparison of a full document collection to a filtered subset. As with all tag clouds, font size is assigned in proportion to term frequency, with more frequent terms displayed in a larger font. Documents have been processed with four different tagging schemes (one, two, Yahoo, MeSH). One word tags are common keywords extracted by frequency from the document, two is the same technique applied to pairs of words. The Yahoo tags were generated by the Yahoo tagging web service which takes documents and returns a list of key terms. The MeSH terms are added to each paper in the PubMed system by human indexers. If the same scheme is displayed for both tags, the lower cloud uses color to encode changes to the rank of given term; the default is a diverging color scheme with no change in rank order of frequency in gray, higher frequency in green and lower in purple (bright green represents terms that are not in the top 100 for the entire document set but are for the filtered set). A slider highlights terms above a user defined frequency. Other available user interactions include sorting and use of a second dual slider for dynamic font size scaling to handle both small and large differences between the most frequent and least frequent terms for a given query.

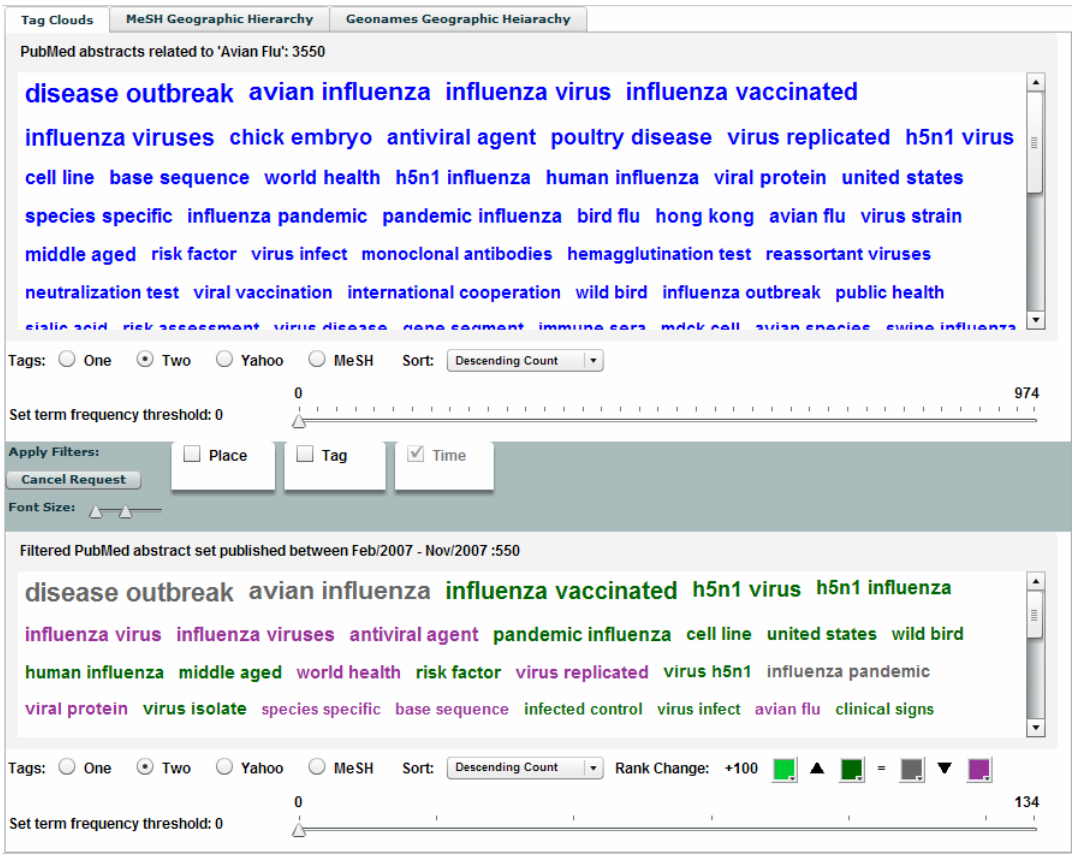

Figure 3. The paired tag cloud arrangement shows the most frequent terms in a larger font. The top cloud summarizes all Avian Flu related abstracts and the bottom cloud shows a subset for a shorter time interval with promoted terms shown in green and demoted terms shown in purple.

4.3 THE 'MORE-LIKE-THIS' FEATURE

As a supplement to progressive space-time-attribute filtering of document collections, a query-by-example search has been incorporated. Once an analyst has identified an article of interest through manual exploration, she might want the system to suggest other related articles. The server makes use of the Lucene indexing system to provide term vector based searching which provides a "more like this" capability to the system. This allows a user who has found an interesting paper to discover other papers which have similar content.

4.4 A USAGE CASE SUPPLEMENTED BY THE RELATED DOCUMENTS TOOL

In a typical usage case, a public health analyst for the Center for Disease Control (CDC) is interested in prioritizing regions vulnerable to an avian influenza outbreak and the level of expertise in the scientific community for these locations. The analyst begins by reviewing a map of RSS feeds filtering by time and navigating to identify several regions of interest. She then applies temporal and spatial filters to OIE reports of documented cases of human and animal exposure to the H5N1 virus and skims several reports in the details pane. These reports indicate the severity of an incident, whether follow-up activity was pursued in a given location, or if the outcome of a reported incident remains unclear. As she works, she captures a snapshot of the interface to archive the times and places of interest. Next she turns to the PubMed documents selecting another country of interest, Turkey, and scans the filtered tag cloud for the full time period to see if any terms are more prominent. She then applies a temporal filter to the tag cloud for each period of interest and again incorporating a time lag following incidents. Mousing over the title *The situation of highly pathogenic avain influenza (HPAI) outbreaks in Turkey* in the results list, she views the abstract and then selects the 'places' button to see the geographic scope of the article. Finally, she selects 'more like this' to explore whether other authors have a regional focus and whether the keywords suggest a unique topical focus within these publications.

5 FUTURE DEVELOPMENT

Future development work includes plans to extend the annotation capabilities (not shown) to insert events in the timeseries and map to mark the user's knowledge of key events and times. The annotation functionality allows the development of collaborative situation awareness documenting analyst insights as they are made and communicating them to others within the visual analytic environment. Query parameter summary labels that are currently organized by dimension (location-time-attribute), will become interactive artifacts providing facet-like query modification for removing individual parameters. Also underway and key to support for situation assessment is a bookmarking mechanism for capturing a document result list, a text expression of the associated query, and an image of the state of the visualization.

ACKNOWLEDGEMENTS

This work is supported by the National Visualization and Analytics Center, a U.S. Department of Homeland Security program operated by the Pacific Northwest National Laboratory (PNNL). PNNL is a U.S. Department of Energy Office of Science laboratory.

REFERENCES

Andrienko, N., G. Andrienko, et al. (2003). "Exploratory spatio-temporal visualization: an analytic review." Journal of Visual Languages and Computing **14**: 503 - 541.

Boulos, M. N. K. (2003). "The use of interactive graphical maps for browsing medical/health internet information resources." International journal of Health Geographics **2**(1).

Endsley, M. R. (1995). "Toward a Theory of Situation Awareness in Dynamic Systems." Human Factors **37**(1): 32-64.

Jones, K. E., N. G. Patel, et al. (2008). "Global trends in emerging infectious diseases." Nature **451**(21): 990-994.

Livnat, Y., J. Argutter, et al. (2005). "Visual Correlation for Situational Awareness." INFOVIS '05: Proceedings of the 2005 IEEE Symposium on Information Visualization.

Peuquet, D. (1994). "It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems." Annals of the Association of Americann Geographers **84**(3): 441-461.

Thomas, J. J. and K. A. Cook, Eds. (2005). Illuminating the Path: The R&D Agenda for Visual Analytics, IEEE Computer Society.