

Thrust 1: Information retrieval, extraction, and contextualization

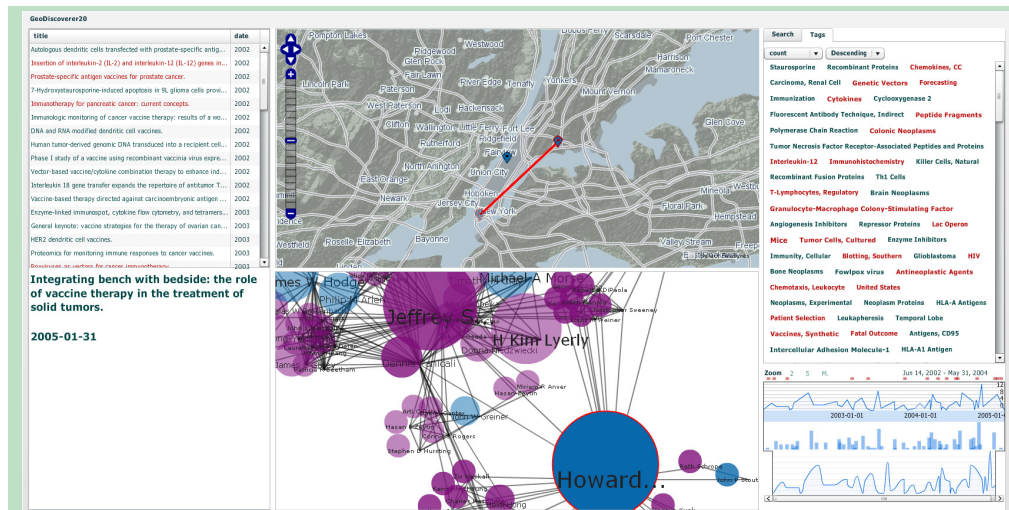
Benefit: In 2006, the digital information generated totaled approximately 3 million times the information in all books ever written. A fundamental challenge for homeland security, given this nearly incomprehensible volume of data, is to quickly and reliably find information relevant to, and essential for, threat and situation assessment and subsequent decision making. The NEVAC information retrieval, extraction, and contextualization thrust addresses this challenge directly through fundamental research linked to development, implementation, testing, and deployment of visual analytics tools.

A key challenge for visual analytics in support of both threat assessment and crisis management is assembling and organizing relevant information and grounding the information analysis and decision-making process in appropriate geographic, temporal, and thematic contexts. NEVAC is addressing this challenge directly through a research thrust targeted at information extraction and contextualization focused on leveraging implicit references to place and time in unstructured text documents, using these place-time references to build connections to formal geographic and other data sources, and linking information fragments derived through map-based analytic interfaces. The core goal is to support the activities of understanding terrorism, engaging in threat and situation assessment, and monitoring ongoing events (whether related to terrorism or natural disasters).

Thus far, NEVAC has made progress on several, complementary information extraction and contextualization strategies. These include: (a) locating geographically and topically relevant documents from distributed sources, (b) extracting entities and relations from both semi-structured and unstructured text in situation reports, scientific publications and news stories, (c) identifying geographic references in text and contextualizing document sets through geocoding and mapping, (d) extracting social network information from text sources, and (e) identifying and supporting assembly of evidence for and against “storylines” derived from multiple text reports and related sources.

This research is producing conceptual strategies and principles for targeting these challenges along with a set of software applications that focus on particular sub-tasks. The prototypes range from early development to *GeoDiscoverer*, an application being developed under separate contract from Raytheon Corporation, see page 12-13 in http://www.raytheon.com/technology_today/archive/2007_issue4.pdf.

Collaborator(s): NVAC, Pacific Northwest National Laboratory.



GeoDiscoverer is a web-based tool that allows analysts to discover patterns and networks in complex data that have concept, space and time attributes. By selecting a particular author in the concept graph, an analyst can immediately see the author’s interests, co-authors, places of publication, and how these elements change over time. The analyst may then drill-down to the abstract of a publication of interest.

Funded by: The Department of Homeland Security; Raytheon Corporation; National Geospatial Intelligence Agency

Early Development

Lab Prototype

Commercial Product

Jan 2009 – tools within this thrust cover a range from early development to commercialization; the latter is represented by GeoDiscoverer

Entity Recognition and Relationship Extraction

To take advantage of relevant but unstructured information found in text documents, NEVAC has developed FactXtractor, a tool that extracts named entities and their relationships from text. It performs relationship extraction at two levels. First, it extracts terms that describe relationships among named entities directly from the sentences in which the named entities appear. Second, FactXtractor uses machine learning to enhance its ability to extract positive examples of relationships of interest. For example, if we want to extract the relationship “located_in”, FactXtractor would identify that “situated at”, “headquartered at”, etc., are synonyms of that relation and extract all entities between which the relationship holds.

NEVAC researchers have built multiple tools incorporating FactXtractor and it is also used in the Scalable Reasoning System (SRS) developed in the NVAC at PNNL. Our tool, FemaRepViz, parses and FEMA situation update reports, segments the reports, identifies the report topics and location of incidents mentioned, and maps the result. A related tool, GeoJunction, uses FactXtractor to support a visual analytic web portal that identifies concepts in scientific and public health documents and places they are about, maps information extracted together with health event data and supports identification of cross-connections among extracted information fragments.

FactXtractor utilizes the open-source named entity extraction tool GATE, and improves the precision of the named entity extraction using additional rules; with evaluations indicating acceptable accuracy. A web-service interface allows external tools to submit a set of documents to FactXtractor whose named entities (and relationships among them) need to be extracted.

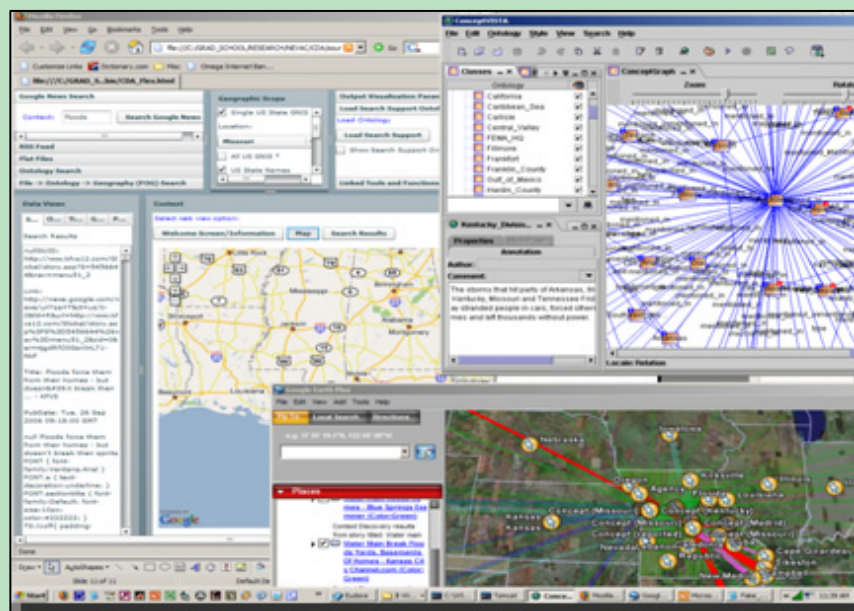
Knowledge-enabled information foraging and geo-contextualization

Our work on entity recognition and relation extraction becomes relevant to real problems when connected to strategies for retrieving potentially relevant documents and putting results of extraction into context.

Geographic Information Retrieval (GIR) is a challenge being addressed by a wide array of research teams internationally. NEVAC researchers are leveraging recent advances targeting three core GIR problems: (a) balancing geographic and topical relevance rankings for what gets retrieved, (b) determining functional geographic ‘footprints’ for retrieved documents containing references to place, and (c) enhancing success of GIR through integration with both knowledge representation and management tools and visual interfaces to support query and interpretation of results.

Once relevant documents are retrieved and entities and relations are extracted, they are made useful by grounding the results in relevant geographic, temporal, and topical context. NEVAC researchers are developing a conceptual strategy to achieve this task and prototype applications with which to assess and refine the approach. The conceptual strategy specifically delineates how select aspects of local model semantics (LMS), in conjunction with geographic and historic categories, can be used to theoretically inform the development of geo-historical context that serves as the platform to support situation assessment.

More specifically, our approach and tools allow analysts to retrieve, extract, combine, and visualize situation information from multi-scale, open source web documents. Our intent with this work is to support situation assessment in the crisis management domain through tools that link and geographically contextualize information contained in text documents retrieved from multiple sources.



Knowledge enabled info-foraging & geo-contextualization: Integration of multiple NEVAC tools support search for information relevant to an evolving crisis situation; in this example the focus is on a major flood event in Missouri. Specifically, this figure illustrates a knowledge-enabled search of news media reports about floods in Missouri guided by concepts extracted from a FEMA situation report. The steps in the process include:

- apply entity-relation extractor (ERE) to FEMA National Situation Report → generate: people, places, organizations; display extracted info in ConceptVISTA
- use ERE results to ID places of interest
- supplement Google News query with ERE derived Owl (web ontology language) file
- map results & relations

For more information, contact:

Alan MacEachren, (814) 865-7491, maceachren@psu.edu

www.geovista.psu.edu/NEVAC