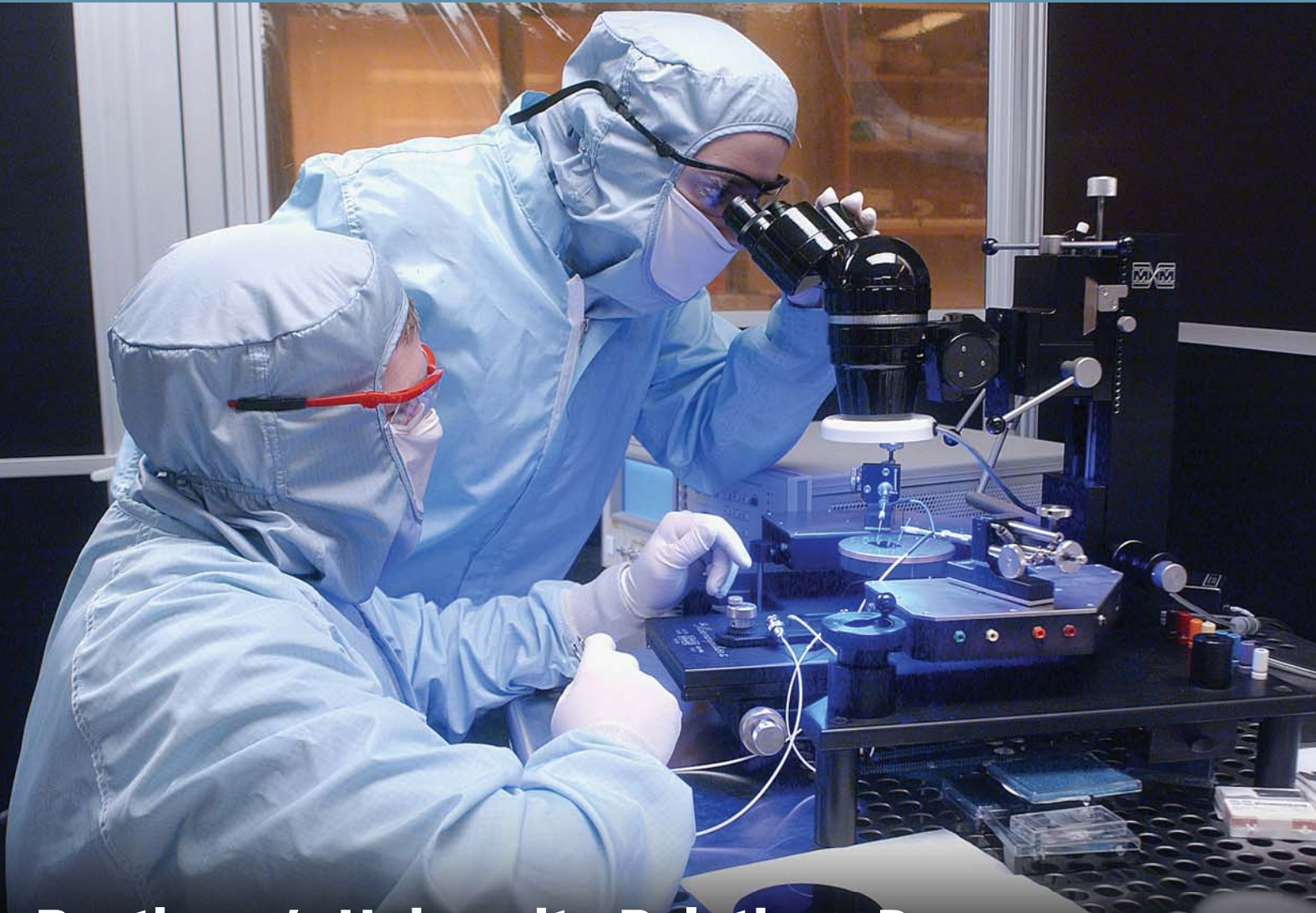


Technology Today

HIGHLIGHTING RAYTHEON'S TECHNOLOGY

2007 Issue 4



Raytheon's University Relations Program
Collaborative Research to Advance State-of-the-Art
Technologies for Our Customers

Raytheon

Customer Success Is Our Mission

GeoDiscoverer: A Search Engine to Integrate Social Networks With Geospatial Information

Part of the focus of the ongoing Raytheon Intelligence and Information Systems research program in Knowledge Management and Knowledge Discovery is developing new ways to enhance exploitation of unstructured text data. One of the research program's goals is to team with academia to provide cutting-edge research supporting various aspects of text data enrichment.

One such activity involving Pennsylvania State University combined the disciplines of information sciences and geographic visualization to automatically discover and display social network information that is embedded in text documents. The principal investigators from Penn State are the College of Information Science and Technology's Dr. Lee Giles, an ACM fellow who directs the CiteSeer search engine project (<http://cite-seer.ist.psu.edu/>), and the College of Earth and Mineral Sciences' Dr. Alan MacEachren, the director of the North-East Visual Analytics Center (<http://www.geovista.psu.edu/NEVAC/index.html>), which is part of the NVAC consortium led by the Pacific Northwest National Laboratory.

Combining and integrating social network discovery techniques with geographic information retrieval, indexing and visualization techniques will enable the discovery of new location-based social networks. Such methods can reveal more of the semantic structure of social networks enhanced by the topology of communication. By crawling the Web of communicated messages (or documents) or using Web logs of people, geospatial information and their relationships, significant and novel social connections and their geographic patterns can be discovered.

A core goal for the GeoDiscoverer application is to develop and implement search methods for collecting and recognizing information about people, geospatial contents and purposes from a message pool, using the CiteSeer search engine as a platform of testing. It enables geo-referencing

and mapping of social networks and their geo-semantics using visual-analytics tools that support knowledge analysts' efforts to discover new evidence of social interactions and their geographic characteristics.

The CiteSeer search engine has a large collection of academic documents obtained from the Web using an automatic crawler. A tool has been developed for segmenting, recognizing and disambiguating the text in these documents into meaningful author profiles. Geospatial information is then extracted from these author profiles and ingested into a newly-developed, Web map services-based visualization tool. CiteSeer can be considered a data gather that acquires data based on a need-to-know basis. The CiteSeer data is a good analog for a range of other real-world data; data included are flawed and imperfect, like that obtained from search engines such as Google and Yahoo. Dates of acquisition for publications, for example, should be interpreted as found dates, not dates of publication.

The tool developed for author profiling consists of several components, including (1) header segmentation; (2) refined field recognition; and (3) disambiguation. For header segmentation, an easy and fast tool based on keyword match and heuristic rules has been implemented. Heuristics are used to determine tokenizing and extraction points in the text, where such terms as "introduction," "abstract," "keywords," etc. denote important text structure. For refined field recognition, the headers are segmented into fields to fill in the predefined ontology of documents and authors. A typical author ontology includes individuals' names, affiliations, addresses, etc. Several support vector machine (SVM) classifiers are trained from the part-of-speech tags provided by a natural language processing (NLP) software component, word text features, position features, surrounding text features, etc. The classifiers determine whether a word is within the boundary of a field and which field that is, yielding the

mapping between the text pieces and the predefined ontology fields.

The fields obtained so far can be very ambiguous, especially for the author names. We have developed various methods for author disambiguation. On the disambiguated text fields, we run a regular expression recognizer to extract e-mails and zip codes, which are later used for visualization.

To support analysis of the retrieved and extracted information, we are developing and implementing client-server, Web services-based tools that support the application of geovisual analytics strategies to analysis and presentation. The methods incorporate a particular focus on handling the complex interactions among place, time, person, organization and attribute components of information as these relate to social networks of interest.

The parts of GeoDiscover directed to analysis of information extracted from CiteSeer by the methods discussed above build upon a range of open source and (for the geospatial information) Open Geospatial Consortium-compliant tools and methods. In particular, the multilayer base maps displayed in the interface are generated with a Web-map service (WMS) request and the filtered publications that are drawn on the client are accessed with a Web-feature service (WFS) request.

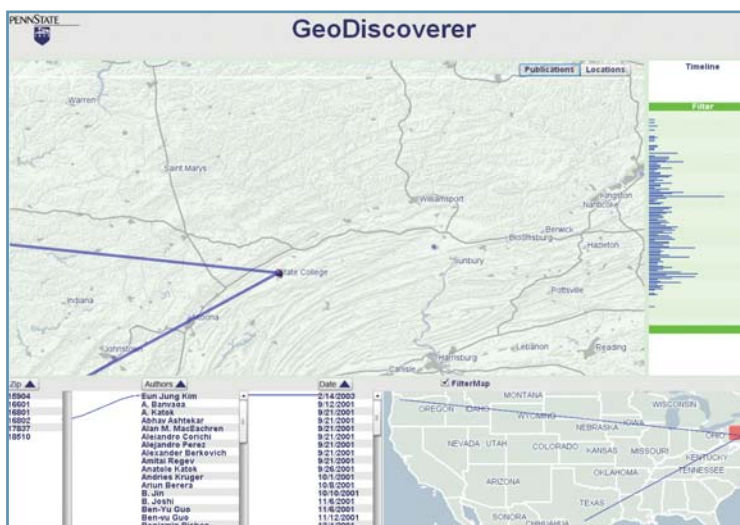
WFS filters are also used extensively to query the geographic data based upon user selections. For example, identifying locations of all co-authors for individuals at any location selected on the map is supported through WFS filter requests. Geographic data of various kinds can be integrated into the Web client view using Geoserver and Mapserver WMS and WFS requests. PostgreSQL/PostGIS is used to store local geographic data and MySQL is used as the backend database for the CiteSeer author/publication data.

The focus of the user interface and display components of the GeoDiscover application is on presenting and exploring cross-connections for the author networks in a geographical-temporal context. The functionality being implemented in the GeoDiscover application includes:

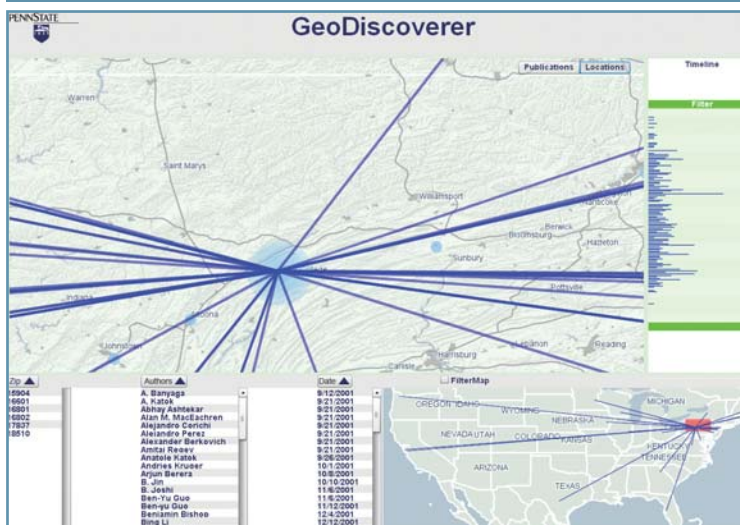
- Geographic coordinate retrieval for locations extracted from documents. The initial implementation is focused on zip codes for places in the United States, but is being extended to support text-geography matching using a gazetteer approach implemented to be flexible on the source of geographic names information.
- Display of overview information in linked map and table forms. The tables include the following information extracted as outlined above: location (zip codes in the example shown), author names and publication dates.
- Compound filtering on data subsets and multiple forms of result display. Filtering can be on place, person or time.

The following analysis scenario illustrates a subset of developed tool functionality:

An analyst is interested in determining when a particular individual has published one or more documents. Her goal is to determine when others had access to the information, thus when it had the potential to make an impact. To pose this question, she simply scrolls to the author name, clicks on it, and views a display of publication dates associated with this author (reference top Figure). If she is interested in a particular publication, she can click on it to determine (by highlighting in the list) who the co-authors are and (by viewing the map) where the co-authors are in relation to the main author.



The user drills down to individual publications by clicking on the Author list. Connections are displayed between lists for that author's location and publication dates. By filtering the map to show individual publications and then clicking the desired publication, the user can see the geographic connections between the lead author of that publication and all collaborators at different locations (in this case there is one collaborator). The inset map zooms to an extent that includes all relevant locations.



The user selects the Locations layer to display a map with graduated symbols showing total publication counts at each location. By clicking on the desired location, geographic connections between lead authors at the location and all remote collaborators with those authors are displayed on both the detail map and the overview map.

A second analyst notices a series of publications of interest being generated from one institution/location. He is curious about who the authors are collaborating with. To find out, he uses the geographic view into the data and selects a place by clicking its location on the map (reference bottom Figure). The map display shows the number of publications (represented by circle size) and the locations of all collaborators from other locations (shown in both the main

map and context map with the lines emanating from the selected location).

For future work, we intend to perform other intelligent analysis of the text document data with special focus on the author social networks. First, we will investigate the temporal dimensions in addition to the geographical locations in the distribution of documents and authors. We will use and propose topic analysis techniques on CiteSeer documents to show how topics and actors move over time. We will also examine how the underlying social networks correlate with the observed movements of topics covered in papers. These correlations will be modeled using statistical learning techniques (e.g., Bayesian networks). Social networks of entities at different levels will be studied including authors, conferences and institutes.

Finally, we propose to evaluate the importance of authors and their documents based on their heterogeneous relationships. CiteSeer has used citation counts for ranking documents and accumulated citation counts for ranking authors. This ranking method can be undermined due to the lack of domain categorization and insufficient considerations of social networks. We intend to improve the ranking in CiteSeer by combining the various topological relationships among

those entities. As a result, we can discover the major players with respect to certain topics in the social network embedded in the document space. This will enable us to see the flow of topics and authors in geo-space describing what topics are created, which authors are prominent and how a discipline changes over time. ●

BJ Simpson, bj.simpson@raytheon.com

Contributors: Dr. Lee Giles, Dr. Alan MacEachren (Penn State University)

Copyright © 2007 Raytheon Company. All rights reserved.
Approved for public release. Printed in the USA.
Customer Success Is Our Mission is a registered trademark of Raytheon Company.
Raytheon Six Sigma, MathMovesU and NoDoubt are trademarks of Raytheon Company.
MATHCOUNTS is a registered trademark of the MATHCOUNTS Foundation.
Capability Maturity Model, CMM and CMMI are registered in the U.S. Patent and
Trademark Office by Carnegie Mellon University.

Raytheon
Customer Success Is Our Mission