Preventing Terrorism with Geographic Text Searches

Randy Ridley and John-Henry Gross, MetaCarta Inc.

By analyzing text messages in geographic context, law enforcement personnel can detect otherwise unseen relationships and better focus domestic intelligence analyses, surveillance, and emergency preparedness activities

Almost all agree that, in the future, terrorists will again attack the United States. Numerous times the federal government has elevated the threat warning to yellow and even orange. Key to preventing such an attack is providing law enforcement, intelligence, and homeland security personnel with timely and relevant information about threats. That concept is partly behind the creation of the Department of Homeland Security (DHS). The unprecedented reorganization under DHS is designed to better coordinate interagency information sharing and facilitate better coordination among federal, state, and local law enforcement groups. But as the intelligence and law enforcement agencies tasked with homeland security collect, share, and analyze information, they can become overwhelmed by the massive amounts of data available. Even as the agencies undergo change, the basic means of collecting and analyzing data remain largely unaltered. Future domestic security, therefore, depends on the successful implementation of innovative information technology (IT) systems within DHS and within agencies that have homeland security missions.

The glacial pace of manually reading and analyzing volumes of unstructured text-based information remains a huge obstacle to basic analytical tasks and real-time reaction capabilities. Significant information contained in such documents as e-mail, incident reports, visa applications, and immigration data can contain vital clues to a pending attack. Given that there has been little change in the IT tool infrastructure in many agencies to date, perhaps all the clues to the 9/11 attack have yet to be discovered. It is reasonable to ask what can be done to cost-effectively integrate dynamic, unstructured data from disparate sources while improving the timeliness and accuracy of decision-making. Whereas Geographic Information Systems (GIS) and unstructured text-based information systems have been powerful but separate tools in the mission to prevent terrorism, the fusion of these two tools, dramatically helps analysts and agents as a powerful lens for synthesizing and analyzing documents and adding greater context to traditional geospatial information . This lens can bring new value and a new viewpoint to homeland security efforts. A geographic text search (GTS) — an automated, IT-based method to fuse text and geographic data — bridges the gap between geographic information systems (GIS) and text search tools, creating an innovative method to reveal new information.

Greater Context and Performance

Using a GTS approach, homeland security analysts, agents and organizations can identify previously unseen relationships — such as patterns of activity close to nuclear power plants, national monuments, or critical infrastructure nodes, or within cities, counties, and regions of interest. Adding one or more layers based on specific information in massive amounts of unstructured content brings even greater context to analysis from a GIS solution. Homeland security efforts focus, in part, on specific places. But sifting through massive amounts of documents to identify one or two relevant geographic references is a time-consuming manual endeavor. For example, it takes an analyst approximately three eight-hour days to identify all geographic references in a single 250-page PDF text file. By contrast, an advanced real- time GTS system could identify all geographic references in as many as 30,000 250-page

documents in the same timeframe. Productivity gains, however, are only part of GTS's benefits.

How It Works

To better pinpoint GTS value, it is useful to examine how a user interacts with a GTS system as well as look at specific applications within the homeland security mission that could be best served by the technology. Let's look at a GTS system that supports a potential national security mission. To use the GTS, the user first opens the ESRI ArcMap desktop GIS product, the central ESRI Desktop application for all map-based tasks including cartography, map analysis, editing and viewing. ArcMap is a comprehensive map authoring application for ArcGIS Desktop. ArcMap has additionally been enabled to access the GTS server that points to a collection of several million text documents. The collection can include e-mail messages, public records, newswires, and proprietary communications. Next, the user adjusts the map display within ArcMap to specify the area that is of interest and enters one or more keywords into a search window that is provided by an ArcMap plug-in from the GTS, to designate a location and subject based filter. If the user does not know the specific location, he or she can also enter decimal degrees for the bounding area of the map to narrow the geographic focus. The GTS system then presents results in the form of a new layer with symbols that designate individual and groups of documents called "stacks" displayed on a map of the region in question (Figure 1). By placing the mouse directly over the stack, the analyst can read the first part of the document. He or she can then read the entire document simply by clicking on a hyperlink. To fine-tune the analysis, the analyst can perform other searches that create their own layers and see the relationship between various text searches and the underlying GIS data.

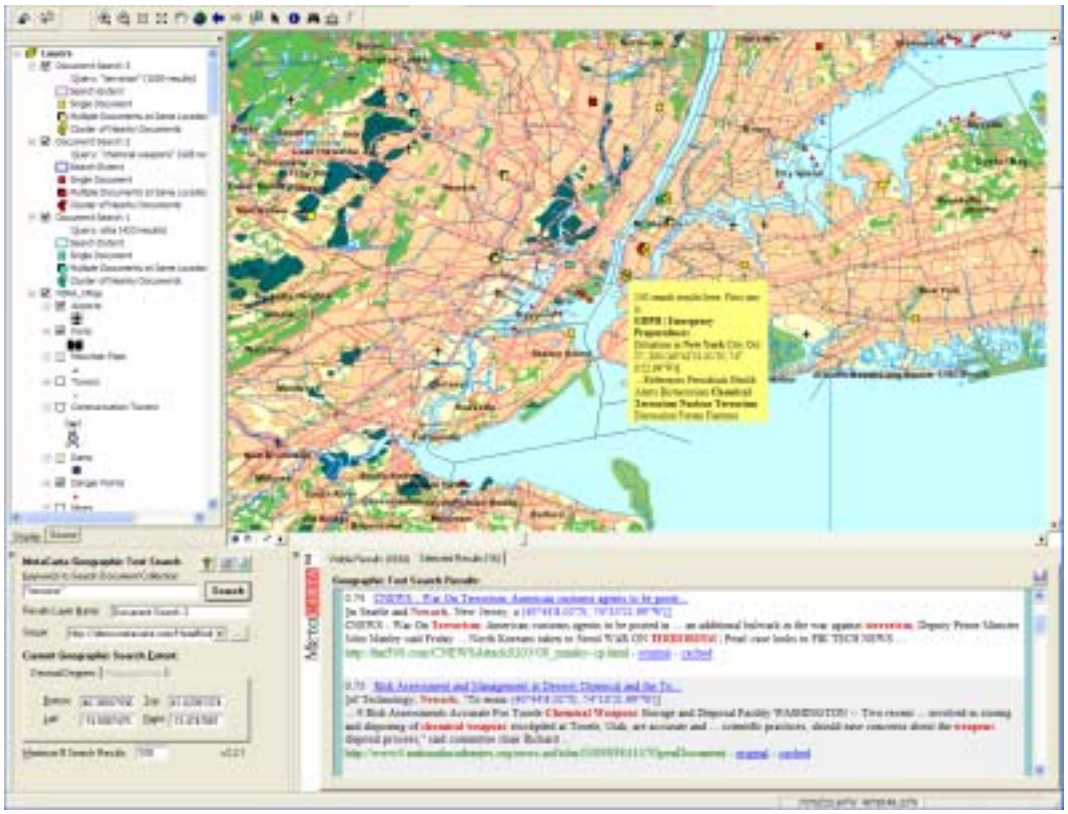

Figure 1 ArcMap interface to GTS

What is the GTS doing to present documents that contain keywords and reference a location of interest? The GTS system has previously ingested the documents, "reads" text within documents, and uses Natural Language Processing (NLP) to compare possible location matches to a Geographic Knowledge Module, that contains gazetteer data and a probabilistic textual model for the location names in that gazetteer — to extract geographic references from a document. The probabilistic textual model is created by analyzing nearly a billion documents for the statistics and patterns of human language. The NLP engine is a critical capability contained in the GTS and identifies both implied (8 miles southwest of New York City) and explicit (1600 Pennsylvania Avenue, Washington, D.C.) geographic references, bridging the gap between document management systems and GIS. The use of NLP greatly accelerates the GTS system's ability to read and identify geographic references. By combining GIS and document management tools, the GTS presents a comprehensive, spatially organized picture of all information available to homeland security analysts. Before the GTS can read the documents, it must first quickly, cleanly, and accurately ingest documents in different formats. To this end, the system relies on standards-based ingestion facilities that support both document "push and pull" methods, using Simple Object Access Protocol (SOAP) and Open Data Base Connectivity interfaces, to upload documents stored in external databases. Because the GTS relies on a single optimized index that contains text and location data, it operates quicker than standard index systems, making it possible to search across millions of documents per second. Once the GTS system ingests and tags documents for geographic references, the results are immediately available to authorized users on the system. (Figure 2) The GTS also supports ongoing ingestion of documents at a scale that supports the existing and ongoing needs of the homeland security application. This amounts to millions of documents per day, with limited system administrator interaction.

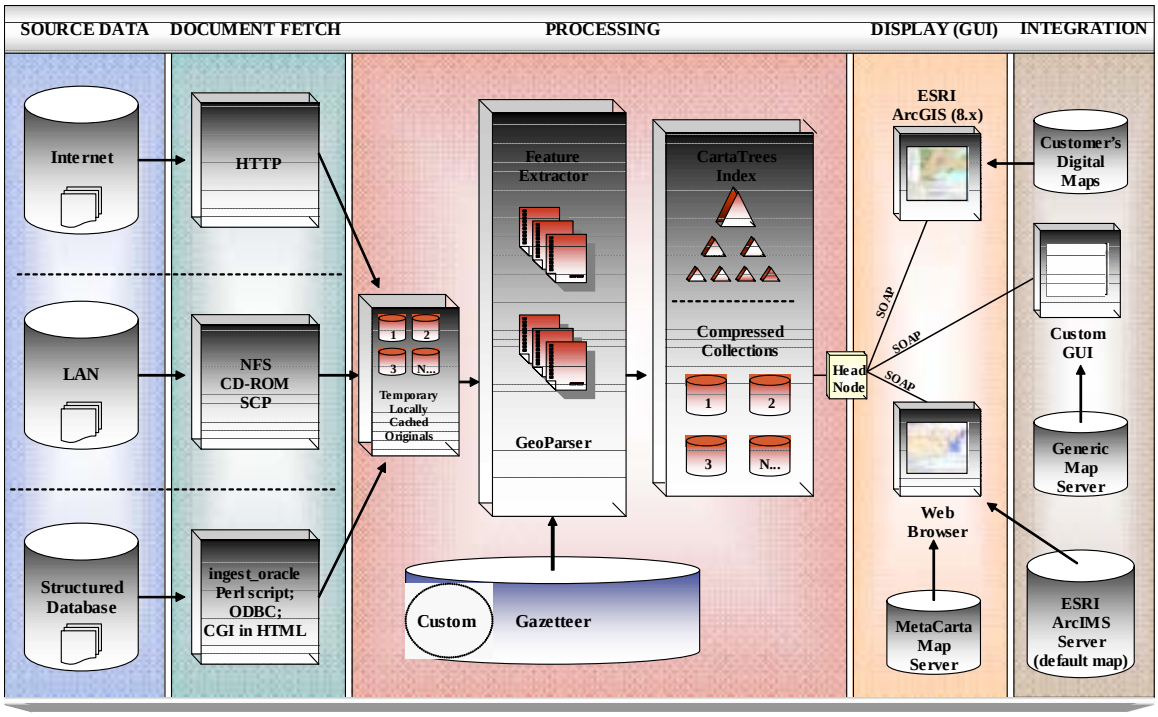

Figure 2 Document processing

Enterprise Integration
The GTS system also integrates into existing enterprise applications (Figure 3). In the government enterprise environments, in particular, agency CIOs and program managers are faced with the challenges of integrating disparate data sources, including legacy sources. Thus, the advantages of any new tool must not be outweighed by the complexity of integration. To maximize investments in existing commercial or proprietary document and knowledge management systems, GTS user interface capabilities can be added directly into ESRI's ArcGIS product for integration across the enterprise. In this way, the GTS can "see" these documents without requiring changes in how these applications operate.

Figure 3 Architecture

For enterprise integration, the GTS additionally operates within the LDAP/X.509 and Active Directory environments to authenticate and authorize users. Through LDAP or Active Directory, system administrators can identify which users can access the GTS system to conduct searches, and to determine which document collections should be made available to each user. This functionality is paramount for homeland security application, documents, and systems security. Next, let's look at how such functionality can be applied.

Why Use GTS for Homeland Security?
Homeland security missions are many, and agency roles are diverse. Organizations perform tracking and surveillance, analyze threat indications and warnings (I&W), provide law enforcement support, prepare for and respond to emergencies, and conduct investigations. All of these responsibilities can benefit from GTS.

Investigations - To answer a "who attacked us?" question after an event, a GTS can be used to search for historical information about the names of attackers (when known). In this manner, analysts can track back geographically to find the security holes at points of entry or to identify terrorist-supporting entities that may still exist and present a continuing threat. As the people who conducted the terrorist attack are highlighted, investigators can also use GTS technology to discover safe houses and other support nodes (both physical and financial). Geographic patterns might also resolve the extent of training and holding areas that terrorist groups might use for the next attack.

Tracking and Surveillance - Federal agencies expend tremendous effort looking for terrorists and their associates. Intelligence agencies need a way to know who is a terrorist in waiting. With estimates ranging from hundreds to as many as 5,000 people in the United States connected to al-Qaeda, intelligence analyses must be comprehensive and at least partially automated. Tracking suspected terrorists requires coordination, planning, and execution of prevention and response tactics. Agents must be able to find quickly documents that have geographic relevancy to locations throughout the United States. For example, reports detailing

the theft of explosives in related geographic locations contain critical patterns that a GTS can illuminate. Authorities and analysts need to understand as much information about the thefts as possible — in the shortest amount of time. By using GTS, an analyst can leverage geography to locate other relevant reports that might link common terrorist actions to a potential plan of attack.

Threat Indications and Warnings - I&W watch teams stationed throughout the United States rely on many sources to alert the public and government agencies about potential threats. Currently, watch teams must read through every single message and document to identify information regarding possible terrorist activities for specific geographic areas. This is very time consuming and inefficient. By first using a GTS to narrow the focus to a specific U.S. region or city, watch teams can more effectively identify critical pieces of information that might otherwise be overlooked or not found in time. The Homeland Security Advisory System is one example of an I&W approach. Watch teams use geography to help support the justification of alert levels. When analysts discover intelligence about potential attacks, they communicate that information to the decision makers responsible for providing early-warning notifications. With a GTS solution, advisories could be very focused on specific geographic areas — cities, states, regions, and known high-value targets.

Emergency Preparedness and Response -Preparing our nation for the aftermath of terrorist attacks involves a vast network of federal, state, and local emergency management organizations. Intra-agency cooperation is vital to domestic disaster response and recovery. Federal agencies, such as the Federal Emergency Management Agency, as well as state and local emergency managers, maintain a large volume of contingency and emergency response plans. Each plan has a specific focus and scenario. During a crisis, a GTS could be used to help present a complete tactical picture by ingesting real-time text news feeds. Emergency response operations managers could then see the attack and its consequences unfold in real time based on available news reports. Also, they can place plans, estimates, and other relevant documents in the GTS for identifying such locations as fallout shelters (yes, they still exist), schools, hospitals, and potential triage sites.

Law Enforcement Support - Investigators need a way to search for documents relevant to a case or to connect related cases. Geography can be an important filter for looking for related information or cases. By focusing on geography and bringing all relevant text information to bear, investigators can instantly discover the criminal history of a postal address and answer such questions as “were there any previous crimes committed at this site?” “What is the crime level in this neighborhood?” and “What types of crimes have been committed here before?” Using temporal features, investigators can even find out what crimes were committed when or during what period of time.

Border Security – The oceans no longer provide border security for the United States. Federal Customs and Border agents spend large amounts of effort looking for terrorists, illegal aliens and other criminals. The intelligence agents within those agencies need a way to see patterns of people and events to reduce threats and increase safety. Tracking suspected illegal aliens, drug smugglers, suspected terrorists, shipping containers, aircraft and other things that cross our borders, requires organization, preparation, and execution of deterrence and reaction tactics. Custom and border agents must be able to locate efficiently arrest reports, visas, bills of lading and emails that all contain geographic relevancy to locations at thousands of points of entry throughout the United States. The ability of system administrators to add the unique names of locations that agents use, such as Hanging Noose to the geographic knowledge base module of the GTS is essential to success. For example, these unique toponyms (place names) appear in reports detailing multiple arrests of groups of individuals of ten or more in specific geographic locations. This is a critical pattern that a GTS can display and could be used by customs and border patrol.

A Comprehensive View

Clearly, the benefits of GTS for homeland security are many. A GTS system accelerates analysis and helps analysts and agents see trends without missing vital information. With a more comprehensive view of all data available, a GTS system implementation can improve the quality and timeliness of information provided to decision makers. Agents and analysts involved have already successfully used the GTS system, and, as the system continues to develop and evolve, it may become a new mission critical tool for preventing terrorism and enhancing homeland security.

For more information on GTS visit the MetaCarta website at www.metacarta.com or contact:

Randy Ridley, VP & GM Public Sector
randyr@metacarta.com
(703) 760-7800
John-Henry Gross, Product Manager Public Sector
jhgross@metacarta.com
7043 760-7800

Steve Zeoli, Partner Manager Public Sector
stevez@metacarta.com
703-760-7800

Glossary

DHS: Department of Homeland Security
GIS: Geographic Information Systems
GTS: Geographic Text Search
I&W: Indications and Warnings
IT: Information Technology
LDAP: Lightweight Directory Access Protocol
NLP: Natural Language Processing