12th ICCRTS
“Adapting C2 to the 21st Century”

Real Time News Analysis for Improved Social Relationship Discovery

Topic: Cognitive and Social Issues

Joan Forester
Janet O’May

POC: Janet O’May
U.S. Army Research Laboratory
ATTN: AMSRD-ARL-CI-CT
Building 321
Aberdeen Proving Ground, MD 21005-5067

410-278-4998
410-278-4988 (fax)

janet.omay@us.army.mil

Real Time News Analysis for Improved Social Relationship Discovery

ABSTRACT: Intelligence analysts and military intelligence officers have the arduous responsibility of processing large amounts of data to determine trends and relationships. Intelligence Analysts must be able to gather traditional information (signal, human, and measurement and signature intelligence) and nontraditional data (financial and social context) to form actionable intelligence. An Intelligence Estimate is part of a Commander's mission plan and includes social, political, and the demographic information. Current data from news sources may enhance traditional intelligence data. The U.S. Army Research Laboratory (ARL) currently has two projects in this arena. The first is the Real Time News Analysis (RTNA) project. RTNA is being developed to harvest real-time streaming data from World Wide Web-based news sources and pre-process it by filtering, classifying, tagging, and fusing. This data will then be fed to ARL's Social Network Analysis (SNA) for Actionable Intelligence (SNAAI) project. This project is developing a testbed of currently available SNA software and working on enhanced algorithms and visualization techniques to improve relationship discovery. ARL is not trying to develop SNA software but to utilize software currently available to better suit the needs of the Intelligence Community. This paper discusses the reasoning, interrelationship, and possible future expansions of the two projects.

Real Time News Analysis for Improved Social Relationship Discovery

1. Introduction

“Command and control refers to procedures used in effectively organising [sic] and directing armed forces to accomplish a mission.”[1] When a mission is given, the actions necessary to carry out the mission are devised by the Commander and staff. Included in the staff is the S2 or Intelligence Officer. The S2 is responsible for the preparation of an intelligence estimate, which “is a logical examination of the intelligence factors affecting the mission.”[2] Included in an intelligence estimate is cultural information regarding the areas of operation, such as factors of economics, psychology, politics, religion, and the population. As the Army moves away from traditional methods of warfare and moves to operations other than war, the cultural features of an area of interest become more critical. In addition to traditional sources of data such as human, sensor and imagery information, real-time data could benefit the intelligence estimate process. The World Wide Web contains large amounts of non-traditional data and if made available for intelligence analysis would augment traditional sources.

The large amount of data accessible from the World Wide Web will overwhelm most Intelligence Officers. Generally, the operational tempo of current military missions does not allow enough time to process the amount of data typically available on the World Wide Web. The information must be filtered and consolidated. Also once an Intelligence Officer has the data, techniques and methodologies must exist to change the large amount of data into meaningful knowledge.

2. Background

Intelligence analysts and military intelligence officers have the arduous responsibility of processing large amounts of data to determine trends and relationships. Intelligence Analysts must be able to gather traditional information (signal, human, and measurement and signature intelligence) and nontraditional data (financial and social context) to form actionable intelligence. An Intelligence Estimate is part of a Commander’s mission plan and includes information on sociology, politics and the population.

Current data from news sources may enhance traditional intelligence data. The U.S. Army Research Laboratory (ARL) currently has two projects in the Intelligence arena. The first is the Real Time News Analysis (RTNA) project. RTNA is being developed to harvest real-time streaming data from world web-based news sources

[1] Command and Control definition, http://www.army-technology.com/glossary/command-and-control.html, accessed 19 December 2006

[2] “Preparation of the Intelligence Estimate,” U.S. Army Intelligence Center, Subcourse IT0565, Edition B, The Army Institute for Professional Development (Army Correspondence Course Program), p. 1-2.

and pre-process it by filtering, classifying, tagging, and fusing. The data, now transformed to knowledge, is ready for interpretation using ARL's Social Network Analysis (SNA) for Actionable Intelligence (SNAAI) project. SNAAI is looking at methods to take this knowledge and apply it to varying SNA software packages based upon the type of data and the required information. Algorithms must be developed to provide heuristics for determining the correct SNA software package to provide the most comprehensive analysis. ARL is not trying to develop SNA software but to utilize software currently available to better suit the needs of the Intelligence Community. Additional work will be completed on improving the visualization of the data from the SNA packages.

The goal of these projects is to increase the amount of current data available to an S2 or Intelligence Analyst and then provide the best overview of the data for further analysis. This paper discusses the reasoning behind the two projects, their interrelationship, and possible expansions.

3. Real Time News Analysis (RTNA)

a. Overview

The RTNA project is being developed by ARL as a networked service for multiple projects, one of which is the SNAAI project. RTNA is a complete system that locates news stories, extracts text data, and creates a repository. Three target audiences exist for the RTNA data – the Soldier in the field, the Intelligence Analyst or S2, and the laboratory researcher.

RTNA addresses three types of data. The first type is actionable data, or knowledge. This is data that is timely and meaningful and has gone through extensive processing. The target audience for this data is a Soldier in the field. For example, a Soldier would request a piece of information through a Google application programming interface (API) and RTNA would search news sources for an answer. This is not just another search engine. RTNA will provide a user-directed feature that will allow a Soldier to better describe his data needs. The graphical user interface (GUI) will provide the capability for a user to enter parameters to tailor the search. Such parameters include: geographic areas of interest; characteristics (such as social, economic, or political); dates to restrict the search to a specific time period; multiple key words; and selection of only user-defined news sources.

The second type of data is reference data, or information to include data that has been partially processed. This is a large set of pre-processed pertinent news stories. It would be filtered, tagged, and readily be loaded into various software packages, including SNA software. The tagging process incorporates a generic methodology, such as extensible markup language (XML), to allow the greatest ease of loading into disparate software packages. A future direction will increase usability by adding different methods of tagging or organizing the data to fulfill specialized software requirements. The reference data forms a large repository and is readily available for researchers and Intelligence Analysts.

The third type of data is all pertinent news stories. This is a large data store of the unprocessed news stories. This is a future direction and the data for this repository will be collected in real-time using Web crawler software. “A Web crawler (also known as a Web spider or Web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner.”[3] The web crawler will interrogate available news sources to obtain data that benefits the analyst or researcher. The web crawler would be trained to access U.S. news sites along with other English language sources.

b. Process

The RTNA process is comprised of several steps. The first step is to locate news stories, this is accomplished through 1) a Google API based upon a user-defined query, 2) alerts that forward articles as they are posted to news sites, or 3) the use of a web crawler that searches out and retrieves the stories.

The second step is the actual news extraction. This is first accomplished by obtaining only the pertinent information from or “scraping” the page. A news web site will contain more information than just the actual story. There are advertisements, links to other stories, and other extraneous information (e.g., site banners or video clips). This extra information is scraped away to just leave the text of the story. The news extraction process also includes identifying duplicate or near-duplicate stories. Only one instance is needed, however RTNA will include heuristics to assure that the selected news story has all the information contained in the duplicates. After the “best” story is selected, it will be saved in text and XML formats. A future direction will provide the capability for an analyst to see all the duplicate stories for further analysis.

The third and final phase is multi-level knowledge extraction from the saved stories and can involve multiple processes. One process is text mining. “Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, sentiment analysis, document summarization, and entity relation modeling (*i.e.*, learning relations between named entities).”[4] Other processes in this stage include: parsing, tagging, filtering, and possibly fusing the data. Some of these processes will be dependent upon how the data will be used and what software

[3] Definition from Wikipedia, http://en.wikipedia.org/wiki/Web_crawler, accessed 21 December 2006.
[4] Definition from Wikipedia, http://en.wikipedia.org/wiki/Text_mining, accessed 21 December 2006.

packages will be receiving the data. It will be possible to store the same news story at different levels of processing. For example, a story can be saved with just tagging completed or can be saved along with other similar stories in a fused format.

Individual steps in the RTNA project are currently being done by other software packages and developers; however this combines all of these steps within a single data-creation process. The process is being designed as a fully automated service with no human-in-the-loop necessary. Gathered text data will be saved in a generic format to be accessible by disparate applications.

4. Review of Current Social Network Analysis (SNA) Software

a. Background

The Army has changed the way that it fights. It no longer is involved in strictly tank-on-tank conflict, but is involved more typically in asynchronous warfare and humanitarian efforts. Military mission completion is more dependent than ever before on the Commander's and staff's understanding of the culture and social climate within their area of operation. The cultural climate of a region is just as important as the terrain. The S2 must be able to access the most up-to-date information in order to provide correct information to the Commander in the operations planning stages.

RTNA will provide data, however data only has value if it is useful to the intelligence requirements. RTNA is being used to provide data to multiple projects at ARL. One project is SNAAI. Actionable intelligence is defined as "having the necessary information immediately available in order to deal with the situation at hand."[5] As the Army is placed in situations where lives depend on knowing as much about the population in areas of operation as possible, information on social contacts must be quickly and reliably obtained.

SNA examines the relationships between people and how they communicate and interact. An offshoot of SNA is dynamic network analysis (DNA) which "varies from traditional social network analysis in that it can handle large dynamic multi-mode, multi-link networks with varying levels of uncertainty."[6] Knowing which groups of the population are connected and providing a basic understanding of who can be "trusted" will be invaluable to not only the Soldier on the ground but also to the Commander and staff in planning operations. ARL is working to incorporate DNA into improved mission planning.

[5] PCMAG.com encyclopedia, http://www.pcmag.com/encyclopedia_term/0,2542,t=actionable+intelligence&i=37443,00.asp, accessed 22 December 2006.

[6] Kathleen M. Carley, Dynamic Network Analysis, Institute for Software Research International, http://www.si.umich.edu/stiet/researchseminar/Winter%202003/DNA.pdf, accessed 22 December 2006.

b. Current Status

ARL's SNAAI project is a new start for fiscal year 2007. Currently three tasks are associated with the project. The first task is using concept maps to improve tactical intelligence. "Concept maps offer a method to represent information visually… Concept maps harness the power of our vision to understand complex information 'at-a-glance.'"[7] ARL is partnering with the Institute for Human and Machine Cognition (IHMC) at the University of West Florida. Using IHMC's software, Cmap, ARL is exploring the feasibility of creating concept maps that will enhance the military decision making process within scenarios.

A second task is exploring the use of machine translated documents to feed SNA software packages. The hypothesis is that if a Soldier in the field obtains foreign language documents and sufficient translation capabilities exist, the text from the documents can be sent to SNA software. Use of the SNA software will provide actionable intelligence given valuable translated text. Current work involves experimenting with translated documents and processing the text in two software packages, ORA from Carnegie Mellon University (CMU) and UCINET from Analytic Technologies, Inc.

The third task is applying DNA to provide actionable intelligence. ARL is working with Dr. Kathleen Carley and Mr. Robert Behrman from CMU to work on improved methods of extracting information based upon context. The task includes using existing SNA software from CMU and others to process data. The initial subtasks include setting up a SNA testbed, exploring methodologies for improved algorithms to uncover social relationships, building APIs to access disparate software, and investigating visualization techniques to improve software output. ARL will not be developing new SNA software, but will be enhancing existing software to adapt for military intelligence applications.

5. Army Research Laboratory's RTNA/SNA Testbed

ARL is currently developing a software testbed to evaluate SNA packages. The process started in the summer of 2006, when a preliminary examination of existing packages began. Ms. Ashley Foots, a student under the George Washington University apprenticeship program, performed a cursory review of multiple free and low cost SNA packages. Based upon her recommendation, two packages have been installed for experimentation. The testbed will be expanded to incorporate more SNA packages. Also work will begin on how to select the right SNA software to fit the data provided. Heuristics will be developed to fit data with software and to provide the most effective visualization based upon the software output. This future work will involve close contact with Intelligence Analysts and S2s to ensure that the formulations developed in the laboratory do enhance the intelligence analysis process through SNA software selection and output display.

[7] Introduction to Concept Maps, http://classes.aces.uiuc.edu/ACES100/Mind/CMap.html, accessed 22 December 2006.

6. Conclusion

Army Intelligence Analysts must have a basic understanding of the cultural situation of their area of interest. The days of bringing in large formations of armored equipment to win a conflict are over. Careful attention must be paid to understand the cultural and social features of an area. Analysts must be able to quickly obtain and analyze large amounts of data to provide the Commander necessary information to operate safely in an area. The RTNA and SNAAI projects are working jointly to obtain large amounts of current data from World Wide Web news sources, process the data, and then provide the appropriately formatted data to SNA software and visualization routines to provide a quick and concise overview. If the Analyst has a clearer picture of the current social climate, he/she can then provide better information to a Commander to improve mission effectiveness.