Annual Project Summary

Proposal Number and Title
05274 - Inferring and Predicting the Social Dynamics of Groups via Psycho-Textual and Communications Flow Analysis

Authors
Jack Schryver; Edmon Begoli, Yu Jiao, Tracy Warren

Project Description
The goal of this LDRD project is to develop and demonstrate a fundamentally new capability in social network analysis using electronic communications data that can be analyzed with text data mining techniques. We extend traditional social networking analysis tools by including deep-dive text analysis of message contents to obtain inferences of group state and group processes. We consider all lifecycle phases including group formation, recruitment, coalition, threat, conflict and schism. Psychosocial phenomena like group identification, in-group bias and out-group antipathy are important indicators that can be discovered in text. However, they are rarely coded explicitly; instead, they are embedded in connotative/affective meanings. Semantic analysis of text has generally focused on denotative rather than connotative meaning, creating a huge potential for data mining and knowledge discovery. Most related research is in the field of sentiment analysis - detection of positive/negative orientations toward a pre-determined subject. Some research has been performed at the document level to extract emotions from text. Our work departs from previous work in three main directions. First, we perform a more fine-grained analysis by narrowing the focus from document to sentence level and from pure sentiment to 22 affective states. Second, we integrate common sense affective knowledge. Third, we link affect with entities identified in documents for deeper understanding of affective meaning. This kind of analysis capability will be an indispensable aid for improving our nation’s ability to protect itself from terrorism and has the potential to reshape the way information on individual interactions is stored and analyzed allowing us to peer deeper in the relationships formed within organizations and across the globe.

Mission Relevance
This research extends ORNL's Knowledge Discovery mission objective. In addition, this research is directly aligned with the mission of the IARPA, the Intelligence Community, (IC) and the DHS. These agencies need innovative and reliable tools to help them analyze the deep contents of massive quantities of electronic messaging data (e-mail, text chat, sms, and transcribed talk). The proposed research will help establish a capability that is essential for long-term human factors analysis and intelligence analysis (DHS), social-network analysis (IC), and computational and psycholinguistics (IARPA). With internationally recognized experience and expertise in knowledge discovery and data mining, high performance computing, human factors, social networking technology and geospatial sciences, ORNL is strategically positioned to develop this capability.
Results and Accomplishments

The main first year research accomplishments are: (1) the development of algorithms required for affect extraction from text, and (2) the implementation of algorithms and research ideas in a software prototype that accepts real text documents as input. We selected a vector model to represent the intensity of affect in text. The model was adapted from a cognitive evaluation or appraisal model of affect. We developed a spreading activation algorithm that assigns vector values for affect to specific words in the English lexicon. The algorithm requires seed lists for each of the 22 specific affect categories in the vector model. The seed lists contain up to 200 words saturated in affective meaning for each affect category. Seed lists were constructed manually using dictionaries, thesauri, and WordNet. The PageRank algorithm was adapted to regulate the flow of affective meaning from the seed list words to many other English words using WordNet as a lexical resource. We also developed a sentence level algorithm that propagates affective meaning from the affective lexicon to words and entities used in the actual text being analyzed. A prototype Java software application was developed on the subversion server and the affect algorithms described above were implemented in the Java application. An automated search interface for document retrieval is currently being implemented in the Java prototype. When a real text document is loaded into the application, the software will analyze it with the extended affective lexicon. The affective words are identified and highlighted in the scrolling text window, and their intensities are plotted on a timeline representation of the document in another window. Stop words are pruned from the graph to reduce visual clutter. A graph of the document appears in a third pane. Finally, a list of entities discovered in the document is displayed. The application integrates existing code that performs entity extraction, clause partition, word stemming, part-of-speech tagging, and other processes needed for text analysis. A recent project accomplishment is the integration of a more robust named entity extractor called Lbj NER Tagger. We are also implementing a capability to perform an independent document-level sentiment analysis from an existing application in order to compare the results to the roll-up of our own methods for affect extraction to document-level sentiment extraction.

Accomplishments in program development are still in their initial phases because we emphasized scientific and technical achievements during the first year of the project. We have held several discussions with Jennifer O’Connor and other staff at DHS S&T Human Factors Division about their division’s research needs. ORNL has worked with the DHS Human Factors Division previously to develop computer-based tools to aid intelligence analysts in the task of evaluating group violent intent. A new BAA for DHS S&T is likely to be released early in CY10 and group violent intent will continue to be emphasized as a research topic of great interest to the agency. Upon evaluating the BAA for 2010, we will submit a proposal in response to the BAA.

We are collaborating with Pennsylvania State University (PSU) to further develop affect extraction algorithms and write conference papers describing the prototype computer system for affective text analysis under development as part of this LDRD project. We plan to team with PSU for submission of unsolicited proposals to DHS, IARPA, and DoD.

We have initiated contacts with several potential sponsoring agencies and collaborators. These individuals include: (1) Kathleen Carley – Carnegie Mellon University; (2) Heather McCallum-Bayliss – IPA and Incisive Analysis Program Manager at IARPA; (3) Rebecca Goolsby – Office
of Naval Research, and (4) Alenka Brown-VanHoozer – IPA at DoD-CIO/ASD-NII (Architecture & Interoperability Directorate). Carley is a potential collaborator, whereas the other contacts control funding sources. As these discussions go forward, our objective is to identify relevant future BAAs that are scheduled for release in the coming year, or prepare unsolicited proposals.

Publications
Under preparation.