Semantic Wiki for Visualization of Social Media Analysis

Daniel Reininger, David Ihrie, and Bob Bullard

Semandex Networks Inc., 5 Independence Way, Suite 309,
Princeton, NJ 08540 (609) 681-5382
{djr, dihrie, bob}@semandex.net

Abstract. A semantic wiki provides visualization of social media analysis applicable to military Information Operations and law enforcement counter-terrorism efforts. Using inputs from disparate data sets, semantic software exports data to link analysis, geospatial displays, and temporal representation. Challenges encountered in software development include the balance between automated and human assisted entity extraction, interoperability with existing visualization systems and ontology management.

1 Introduction

Social media analysis is an important part of military and law enforcement operations [1] [2]. The analysis requires the ability to model and extract significance from the social media interactions of persons and organizations of interest. This analysis must be done in real time and in the virtual, collaborative workspaces of the law enforcement and intelligence communities.

This paper outlines issues identified during the development and demonstration of a software tool to provide shared visualization for social media analysis in selected government environments. We developed and tested a software application pursuant to a federally sponsored program titled Information Networking for Operational Reporting and Monitoring (INFORM). The project was designed to facilitate collaborative analysis and workflows for elements of the U.S. Marine Corps, the U.S. Special Operations Command, and the U.S. Department of State.

Existing information sharing applications available to the user community included the Combined Information Network Data Exchange (CIDNE) [3], Intellipedia [4], and the Net-Centric Diplomacy portal [5]. Each of these programs provided an avenue for information sharing and multi-agency collaboration, primarily by making documents—whether finished reports or community-updated web pages—available to a broad community. However, each of these systems exhibited a common disadvantage that the INFORM program was designed to help mitigate: tactical users needed to model information of local interest that could not be easily captured in CIDNE, NCD, or Intellipedia in a way that facilitated efficient and dynamic query, retrieval and display. A solution had to provide three advantages over the existing systems. First, the solution had to provide the user with a means to rapidly tailor the

information model to handle novel concepts encountered at the lowest tactical echelons. Second, the solution had to allow for the dynamic assembly of documents so that views of information were automatically and continually updated throughout the knowledge base; new social links had to be instantly recognized and published as soon as these links were discovered by the system. Third, the solution had to provide a means of efficient manual and automated query and display, including the ability to export data extracts to specific visualization applications (external to this software solution) designated by the user community.

The goal of the INFORM program was to create a web-based application with these capabilities that supported Information Operations. The technical approach was to develop a semantic wiki for data capture, analysis, and display. The desired end state was the ability to link entities contained in reports, open source articles and other sources encountered by users, creating a semantic graph that helped with social media analysis rather than simply serve as a document management system. A semantic approach met the end state requirements and offered additional advantages. First, data could be combined from disparate sources. Some data were highly structured and amenable to computer processing, while other data were unstructured, with syntactic incompatibilities that inhibited automated data ingestion to the system. We used a semantic schema and domain-specific ontology to parse information, generate concept instances, and represent relationships identified in the data. Second, a web-based wiki provided distributed access and rapid dissemination of information for multi-user collaboration. It provided a platform that generated and transmitted alerts based on changes in collective knowledge, such as the discovery of additional relevant information. Individual users set their own alert parameters and received individual notifications, by email or web-based chat, that included embedded links for one-click viewing of updated information.

2 Discussion

The use case in which the software was applied involved social media analysis supporting psychological operations. Specifically, we used the new semantic application to perform analysis of a selected target audience in accordance with existing doctrinal procedures [6]. This involved the review of data from open source media, combined with data from additional sources, to support an assessment of social groups, subgroups, and individuals within a population.

Visualization of the results of social media analysis is essential to effective target characterization for influence operations. The analysis in this project supported the initial study of a subject audience and evaluated measures of effectiveness to determine behavioral change, as evidenced in differences in social media behavior. Variations were observed both in social media content generation and activity patterns. The approach taken to determine changes in social media behavior was driven by the data available, which was a function of available sensors and information access.

2.1 Input Interfaces

The semantic software developed supported interfaces with databases, emails, RSS feeds, web pages, and spreadsheet files customized to support existing concepts of operation. An issue we faced was achieving the optimal balance between automated and human-assisted data ingestion. Uploading spreadsheets is a simple means of automated input; however, extracting pertinent information and context from unstructured text is also an important component of social media analysis, since statistical display of themes extracted from social media (e.g., blogs) are an indicator of social sentiment. Research comparing human-assisted entity extraction from text with automated methods, in efforts to enable automated network node/edge determination, indicate that the methods are complementary [7]. A mix of human involvement and automated processes provides the ideal balance of speed, ease, and validity. We used automated entity recognition in text coupled with human-validated associations to input data into a semantic graph. Additionally, for statistics not requiring additional human validation (such as summations of statistical analysis of sentiment), we incorporated automated data ingestion and automated visualization.

A customized data loader feature was developed for the software to facilitate automated upload of information. The primary challenge encountered during automated input from source databases was gaining an understanding of the structure of source data without a descriptive model of that data. This understanding was integral to writing the appropriate SQL statements to retrieve data in the desired format. This obstacle was overcome by manual inspection of columns in the source database, looking for promising column names, and examining the contents to verify that the mapping was appropriate. This mapping and tailored ingestion [detailed in reference 8] was critical to harmonize geospatial and temporal data from disparate datasets. This process would have been greatly aided by a mechanism to find similar names, similar content, and matching enumerations in order to help understand how the source data mapped to the target, as well as a mechanism to build and test the necessary SQL statements that ultimately retrieved data from the source. Once complete, the data loader mapped data into the semantic database while ensuring that incoming data met certain standards. The data loader template built and loaded concept instances and properties, and then built relationships between the concepts to form the semantic graph.

2.2 Output Interfaces

Visualization was delivered in three distinct categories: link analysis, geospatial representation and temporal representation. Rather than duplicate efforts to develop capabilities that would require user training, we focused on utilities for export of relevant data to existing third party applications already commonly employed by the user community.

Figure 1 shows the system's architecture and input/output interfaces. Input interfaces included databases, emails, RSS feeds, web pages, spreadsheet files. External applications included, but were not limited to, link analysis, and geospatial visualization of select data. Figure 2 shows the output displays of results into Link

Analysis tools and geospatial visualization. Results summarized in the person's page can be visualized as a link chart and in geospatial representations.

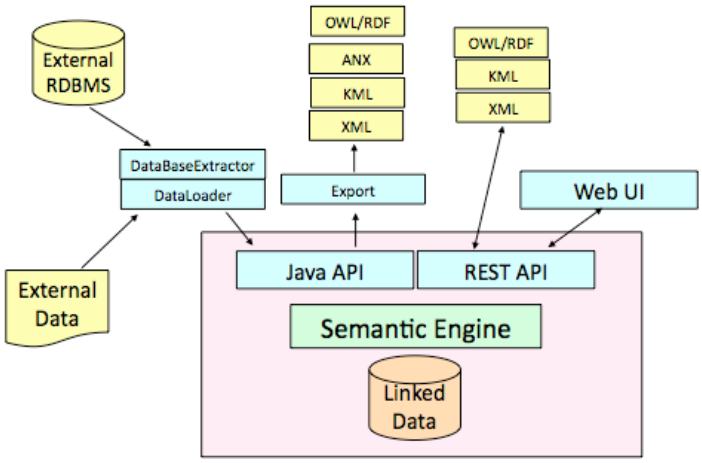

Fig. 1. Semantic wiki software architecture.

Fig. 2. Export to external applications.

Figure 3 shows a geographic display, using Google Maps, of population sentiment derived using statistical aggregation of data related to individuals, exported as KML. Geographic clustering and display of sentiment statistics derived from social surveys is an accepted methodology within the military for obtaining "ground truth" [9].

Temporal views complemented geographic displays. Software views based on adjusting time frames can indicate periods of high and low centrality, productivity,

and information dissemination; however, contextual cues that compliment temporal views are critical to gaining a true understanding of social interactions [10]. We developed a tailored display to fit the unique requirements of temporal visualization for social communications between individuals, which we could not obtain using existing external applications.

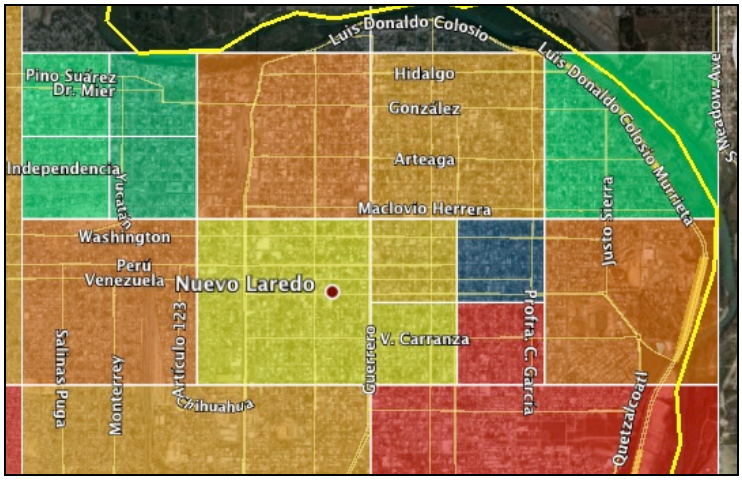

Fig. 3. Geographic display of population group sentiment using Google Maps.

Tailored representations included a heat map that showed activity by time and day of week to identify changes in individual social communications behavior. Filters provided adjustable date ranges and the ability to select the type or types of interaction (phone call, text message, etc.) displayed by the software.

Such visualization of the results of social media monitoring and analysis offers direct application to addressing the challenges and opportunities that result from the widespread use of social media, and its necessary inclusion in an environment of Information Operations. The utility of these visualizations applies equally to law enforcement, particularly in a counter-terrorism role.

During the course of this project, we encountered several salient issues that merit further research to expand the capabilities for social media analysis and visualization. We next present some possible approaches to ontology management, but leave the recommendations as open-ended avenues for the development of the field.

2.3 Ontology Management

This project developed a common schema for representation of information of interest to multiple potential user communities, including psychological operations, civil affairs, and intelligence information related to people, regions, countries, events, threats, and similar topics. This common schema provided the foundation for semantic information modeling that resulted in the ability of users to contribute to and draw on a common information picture expressed in a semantic graph.

There has been discussion of implementing a high-level, domain-independent ontology to provide a framework from which disparate systems in the government and military arenas could derive domain-specific ontologies [11]. Lacking such a foundational ontology of universal application, we developed a semantic schema and a domain-specific ontology for this project. Our solution did not need to provide formal inferencing; accordingly, our application did not require a formal ontology. We did need enough structure to capture associations in the data to present, for example, the optimal path to get a message to influential individuals within a community of interest.

Certainly, the integration of social media analysis software with existing enterprise information systems requires either ontological commonality or ontological bridging to enable effective interface across domains. Such a bridging ontology [12] might be useful to more easily ingest additional data sets into the semantic graph, or export data from the semantic graph to other databases without the need for customizing data loader templates or data export functions.

Even within a single domain and system, users required a method of tailoring the ontology by extending it "on the fly" to accommodate new concepts. For example, while examining media inputs, a user identified the need to add a "Tweet" as a subtype of instant message. Fortunately, the user interface provided a means by which the user could, without the need for software programming, spontaneously add relevant concepts and integrate this data into the analysis picture without corrupting the structure of the database or impacting functionality of the software. While the ontology remained "informal" in that users could extend it, the software-enforced adherence to that ontology "formalized" its acceptance by the user community.

The requirement for an adaptable ontology poses conflicting challenges. First, a user faced with a new classification of information must be able to define the new entity in the ontology. Second, this process must be managed collaboratively. If every user is continually modifying the structure of the ontology, it will rapidly cease to function; data calls for visualization will fail. Instead, designated stewards of the ontology in a user organization must make necessary modifications without recourse to technical support. Developers must take the requirements for agile schema into account in the planning and design phases of the software design life cycle to ensure that users can keep the application relevant in the ever-evolving conditions of social media analysis.

Two lessons learned, and successfully applied, were that recognition engines and recommendation engines can assist with ontology management. A recognition engine was incorporated that functions on free text input from files or websites, and preprocesses the text before presenting it to the user's view. It highlights entities in the text that are already known to the semantic wiki, such as the name of a specific person or place. The engine then uses a semi-automated process to help the user delineate relationships between existing entities and new entities created by the user. This human-to-machine interface prevents the user from unnecessarily creating new entities, and prevents the software from making errors of association that are a common byproduct of fully automated text recognition and database ingestion.

A recommendation engine provides assistance if the operator is still not satisfied with existing subtypes (as in the above example of the "Tweet"). First, the user interface allows the addition of a new subtype to the ontology. The software then

repopulates the modification to all user displays, allowing collaborative awareness and use of the new subtype. Now a recommendation engine can provide the user with awareness of alternatives, including newly created subtypes; when the software processes incoming text, this recommendation engine presents the user with a list of available entity types with which to tag new entities. "Tweet" is now recommended as an option that the user can select. This utility optimizes visibility of the ontology and limits the unnecessary duplication of subtypes.

2.4 Results

A source corpus of data composed of 204 files in four formats (.jpg, .html, .txt, .doc) produced a semantic graph of 4310 concept instances in a semantic database totaling 102 MB. This represented the interrelations of 585 events and 196 persons.

The domain-specific ontology expanded from eight basic concepts (Person, Organization, Place, Event, File, Characteristic[1] and Watchlist[2]) to 262 types of pages defined by use case analysis and by direct user additions. For example, "Communication Event" is a type of "Event" and "Tweet" is a type of "Communication Event". Also a "Facility" is a type of "Place" and a "Broadcasting Station" is a type of "Facility". However, actual data modeling for social media analysis during the practical application phase of this program utilized only 63 types of pages, or 24% of the total available. This suggests that users will, even when presented with a myriad of choices in modeling data, often use commonly recognized entity types. It also evidences the effectiveness of recognition and recommendation engines in limiting the inclination of users to modify an adequately developed ontology.

3 Conclusion

This project has resulted in the development of software that provides social media analysis and visualization for specific customers in the government community. We have developed and tailored a commercially available semantic software solution to integrate disparate social media data sources using automated and machine-assisted techniques that promote data validity and collaborative accessibility. While the results apply directly to military Information Operations and law enforcement counter-terrorism efforts, we believe that the issues faced are widely applicable to researchers, software developers, and program managers in other domains related to the semantic exploitation of social media.

An issue of interest to the reader community is the delicate task of finding the balance between automated and machine-assisted (human-validated) data ingestion. This is a balance that all data analysis applications must attain to preserve data

[1] Characteristics model distinctive features of any entity (e.g., person, thing, event, place). For example, "tall", "long hair" and "caucacian" can be person's characteristics.

[2] A Watchlist page has two links: (Has Member) *Page* and (Is Watchlist Of) *User*. When a member page is updated, the user will be notified of the update.

validity, without sacrificing scalability. Input methods such as spreadsheet and unstructured text ingestion promote speed and utility, while entity recognition and user-supervised text exploitation validate input to a collective database.

We have leveraged existing third party applications preferred by the user community to visualize information, including relations between abstract concepts in social media, using link analysis and geographic display. Additionally, selected user requirements have been met by the development of tailored temporal displays.

Ongoing challenges include ontology management, where the requirement is to provide the user with a mechanism to continually refine a domain-based ontology. While adaptable software, with recognition and recommendation engines, allows for a user-extensible ontology, the broader issue of cross-domain mapping using a bridging ontology offers opportunity for further study.

References

1. U.S. Department of Defense (USDoD) Joint Publication 2-01.3: Joint Intelligence Preparation of the Operational Environment, pg. xii. Department of Defense, Washington, DC (2009)
2. International Association of Law Enforcement Intelligence Analysts (IALEIA). 2004 Law enforcement analytic standards, published in association with the U.S. Department of Justice (2004)
3. Intelligent Software Solutions, Inc.: CIDNE, http://www.issinc.com/solutions/cidne.html
4. Wikipedia: Intellipedia, http://en.wikipedia.org/wiki/Intellipedia
5. Pack, D.: Profiling and testing procedures for a net-centric data provider. SPAWAR System Center Charleston, North Charleston, SC (2005)
6. U.S. Department of the Army (USDoA). FM 3-05.302: Tactical psychological operations tactics, techniques, and procedures, pp. 6-4 to 6-11. Headquarters, Department of the Army, Washington, DC (2005)
7. Graham, J., Carley, K., Cukor, D.: Intelligence database creation & analysis: network-based text analysis versus human cognition. In: Proceedings of the 41st Hawaii International Conference on System Sciences (2008)
8. Semandex Networks, Inc: Rapid semantic integration of data using the Tango DataLoader framework, http://www.semandex.net/servlet/DownloadServlet?id=397
9. U.S. DoA 2005: FM 3-05.302, pp. B1-B12
10. Gloor, P., Laubacher, R., Zhao, Y., Dynes, S.: Temporal visualization and analysis of social networks. In: Proceedings of the North American Association for Computational, Social and Organizational Science Conference (2004)
11. Semy, S., Pulvermacher, M., and Obrst, L.: Toward the use of an upper ontology for U.S. Government and U.S. military domains: an evaluation. MITRE Corporation, Bedford, MA (2004)
12. Gilson, O., Silva, N., Grant, P.W., Chen, M.: From web data to visualization via ontology mapping. Computer Graphics Forum, 27, no. 3 (2008)