Semantic Analytics in Intelligence: Applying Semantic Association Discovery to determine Relevance of Heterogeneous Documents

Boanerges Aleman-Meza, Amit Sheth, Devanand Paliniswami, Matthew Eavenson, I. Budak Arpinar
{boanerg, amit}@cs.uga.edu, {devp,durandal}@uga.edu, budak@cs.uga.edu

Large Scale Distributed Information Systems (LSDIS) Lab,
Computer Science Department, University of Georgia,
415 Graduate Studies Research Center
Athens, GA 30602-7404
http://lsdis.cs.uga.edu

Keywords: Semantic Web technology, semantic analytics, semantic association, semantic metadata, knowledge discovery, semantic applications for homeland security, content analytics, ontology, RDF, insider threat

INTRODUCTION

Creating applications that allow users to gain insightful and actionable information or mine for interesting patterns from vast amounts of heterogeneous information is one of the most exciting new areas of information systems research. This information to be analyzed may come from numerous sources spanning proprietary, trusted, and open-source information, including intranets, the deep Web and the open Web. The fast emerging markets of business intelligence as well as national and homeland security are finding themselves in increasing need of a class of applications called risk and compliance (Sheth 2005). One representative example of this class of applications is the *Insider Threat* application, which involves validation of legitimate access of documents. While physical security measures may help reduce malevolent access to documents by employees within an organization, the development of new information-based security systems provides additional capabilities for defense against insider threat attacks. The intent of this application is to monitor that analysts who are assigned various investigation tasks access the information on a "*need to know basis*" and that the system should identified access to irrelevant information in an attempt to reduce the chances that confidential information is leaked or released inappropriately.

Research in techniques for search of documents was a critical component of the first generation of the Web, and has gone from academia to mainstream. A second generation "Semantic Web" will be built by adding semantic annotations to Web content that software can understand and from which humans can benefit. Large-scale semantic annotation of data (domain-independent and domain-specific) is now possible because of numerous advances in the areas of entity identification, automatic classification, taxonomy and ontology development, and metadata extraction (Dill et al. 2003; Hammond et al. 2002; Shah et al. 2002). Relationships are at the heart of semantics (Sheth et al. 2003; Woods 1975). The next frontier, which fundamentally changes the way we acquire and use knowledge, is to automatically identify complex relationships between entities in this semantically annotated data. Instead of a search engine that merely returns documents containing terms of interest, we propose an approach that supports semantic analytics of heterogeneous content to return actionable information that gives useful insight into the connection between documents and real-world entities, thus providing better-than-ever support for important decisions and actions. This approach is demonstrated using a prototype application that supports the task of ensuring that an intelligence analyst accesses documents on a "need to know" basis, that is, documents that are relevant to the analyst's investigation assignment. This is one of many semantic applications as part of advanced information technology necessary to support homeland security.

From the research perspective, one of the challenges was to devise a framework for the formal definition and representation of meaningful and interesting relationships, which we call "semantic associations". Semantic associations are at the core of our research in content analytics and knowledge discovery using an ontology-driven process. Other challenges arise from the large scale of metadata sets and the need for complex data structures containing entities and relationships that are used to perform query processing against those sets. Lastly, we need to utilize a notion of context to capture an analyst's investigation assignment using an ontology. These challenges call for a fresh look at indexing, query processing, ranking, as well as tractable and scalable graph algorithms that exploit heuristics. This book chapter describes a prototype supporting the identification of insider threats for documents-access based on the underlying concept of exploiting semantic associations among real-world entities. Our work addresses the aforementioned challenges, building on our previous research in the following areas:

– semantic metadata extraction and annotation (Hammond et al. 2002),
– practical domain-specific ontology creation (Aleman-Meza et al. 2004), Glyco Ontology: http://lsdis.cs.uga.edu/Projects/Glycomics/; ProPreO: http://lsdis.cs.uga.edu/projects/glycomics/propreo/
– semantic association definition and computation (Aleman-Meza et al. 2003; Anyanwu and Sheth 2003; Milnor et al. 2005; Perry et al. 2005), and
– main-memory query processing (Janik and Kochut 2005)

This book chapter provides a description of a prototype for the legitimate access problem of Insider Threat. Extended descriptions in both technical and theoretical aspects are provided in more detail than our previous work (Aleman-Meza et al. 2005a). In particular we highlight the following:

– An ontological approach to capture an investigation assignment of an analyst into a *context of investigation*.
– Semantic discovery techniques to identify the relevance of documents based on the explicit relationships existing between a document and the *context of investigation*.
– An 'inspection' visualization interface that supports exploration of the *need to know* of otherwise legitimate access documents.

We also discuss how a commercial Semantic Web technology product, Semagix Freedom based on SCORE technology (Sheth et al. 2002) developed in our lab, is used for metadata extraction technology in designing and populating an ontology from trusted sources. The ontology contains relevant metadata extracted from different information resources including government watch-lists, sanction-lists, gazetteers, organizations, etc.

BACKGROUND

Ontologies

Ontologies are at the heart of most approaches and technologies (Sheth et al. 2003) that seek to realize the the Semantic Web vision[1] (Berners-Lee et al. 2001). The Resource Description Framework (RDF) data model (Lassila and Swick 1999) is a proposed framework to capture the meaning of an entity (or resource) by specifying how it relates to other entities (or classes of resources). In the RDF model, concepts of entities are linked together with relations (properties). The classes and/or relationships can be defined with an RDF Schema vocabulary (Brickley and Guha 2000). The properties are denoted by arcs and labeled with the relation name. Thus, the metadata can be represented as a graph together with a graph for the vocabulary of the classes and relationships (Karvounarakis et al. 2002).

[1] http://www.w3.org/2001/sw/

A key feature needed in semantic technologies is the capability to create and maintain ontologies. Semi-automatic creation of metadata based on specific domain has been researched in the S-CREAM framework (Handschuh et al. 2003), and other tools have been developed (Vargas-Vera et al. 2002). An ontology populated with domain knowledge provides an important asset for applications in semantic analytics. While a schema of the ontology needs to be designed by an expert, our work shows that given trusted and high quality knowledge sources, coupled with a set of disambiguation techniques, can largely automate the process of populating domain ontologies, often with millions of instances.

Semantic annotation is referred to as both the metadata added to a document and the process of generating such metadata (Popov et al. 2003). The Semantic Enhancement Engine (Hammond et al. 2002) of Semagix Freedom also provides this capability. In industry, SemTag has demonstrated large scale annotation of over 1 billion of documents (Dill et al. 2003). Similarly, annotations for specific domains have been also developed such as BioAnnotator for biomedical domain (Subramaniam et al. 2003).

Semantic Associations

The conceptual basis of the system is based on what we have termed semantic associations (Anyanwu and Sheth 2003). A semantic association represents a direct or indirect relationship between two entities. "Semantics" here specifically involves those relations that are meaningful to the application and can be inferred either based on the data itself or with the help of additional knowledge. Semantic associations are meaningful and relevant complex relationships between entities, events and concepts. They lend meaning to information, making it understandable and actionable, and provide new and possibly unexpected insights. Different entities can be related in multiple ways. For example, a Professor can be related to a University, students, courses, and publications; but s/he can also be related to other entities by different relations like *hobbies*, *religion*, *politics*, etc. Relationships that span several entities may be very important in domains such as national security, because they may enable analysts to see the connections between seemingly disparate people, places and events.

Semantic associations are based on intuitive notions such as connectivity and semantic similarity. Different semantic associations in an RDF graph have been formally defined in our previous work (Anyanwu and Sheth 2003). Here we present a a simplified definition of semantic associations:

Definition 1 (*Semantic Association*): Two entities e_1 and e_n are *semantically associated* if there exists a sequence $e_1, p_1, e_2, p_2, e_3, \ldots, e_{n-1}, p_{n-1}, e_n$ in an RDF graph where e_i, $1 \leq i \leq n$, are entities and p_j, $1 \leq j < n$, are properties.

Semantic associations have proven to be a foundational layer in real world applications, most usefully in the area of homeland security such as Passenger Threat Assesment (Sheth et al. 2005a). Additionally, semantic associations have been used in retrieval of biomedical patents (Mukherjea and Bamba: 2004), knowledge discovery and composition in peer-to-peer networks (Aleman-Meza et al. 2005c; Perry et al. 2005), and geospatial semantic analytics (Arpinar et al. 2004). Ranking of semantic associations has also been addressed (Aleman-Meza et al. 2005b; Aleman-Meza et al. 2003; Anyanwu et al. 2005) as well as efficient algorithms focusing on performance, scalability and efficiency (Janik and Kochut 2005; Milnor et al. 2005). Measures of credibility of semantic associations from multiple sources have been proposed (Ding et al. 2005).

Discovery of indirect relationship gains importance for detecting, for example, potential terrorist cells, which remain distant and avoid direct contact with one another in order to defer possible detection (Krebs 2002) or money laundering (2003) involves deliberate innocuous looking transactions. Some examples of applicability of semantic associations in national security domain include the following:

1. Is a person known to be associated with an organization on watch lists?
2. Does a document contain people names that work for an organization that is known to sponsor an organization on a watch-list?
3. Is there a connection between a document on the Sri Lankan group "LITE" and terrorist organizations located in Middle East?

Document Access Problem of Insider Threat

In the context of the intelligence community, one of many security aspects involves that of detection of "malevolent actions by an already trusted person with access to sensitive information and information systems" (Anderson and Brackney 2004). For document access, the goal is to ensure that an analyst only accesses documents on a *need-to-know* basis. Typically, data of an analyst's activities are often analyzed after the fact, done reactively rather than proactively. This may be due to a "culture of trust", but more often it has to do with the prohibitive costs of creating/defining methods to detect malevolent actions, as well as of their implementation and maintenance. There are various techniques that can be applied to determine if a collection of documents is relevant to a given domain. Some of these techniques exploit statistical, natural language processing, machine learning, document clustering, and documents classification techniques; typically referred to as implicit semantics (Sheth et al. 2005b) because cannot or do not name specific relationships among concepts or entities.

One of related approaches uses a list of positive and negative examples to generate a set of weight vectors that determine the permission of each document for an analyst (Rectenwald et al. 2004). When an analyst selects a document, the authorization agent determines whether it is viewable to the analyst based on the generated weight vectors. These techniques typically do not provide the reasons on why a document is or not relevant to the investigation objective of the analyst. Similarly, these techniques have none or limited support for exploiting named relationships between concepts (e.g., an organization is *located-in* a country). Our strategy includes the use of an ontology to capture the semantics of the domain to process named relationships both for identification of relevance of documents as well as to provide a visualization on why and how a document is related to the analyst investigation objective (i.e., for auditing purposes).

The document access problem of Insider Threat has also been referred to as 'misuse'. Misuse detection systems have been built based on building a user profile based on legitimate queries to access information (Cathey et al. 2003). Subsequent queries are then compared against the user profile to detect misuse.

Traditional data mining (Chen et al. 1996; Fayyad et al. 1996) has mainly focused on discovering patterns and relationships out of their repetition in the data. However, data mining has been applied for misuse detection in order to eliminate manual adjustment of weights on the different levels of misuse (Ma et al. 2005).

DESCRIPTION OF THE APPROACH

Overview of our System for Insider Threat Document

Our prototypical system demonstrates a workflow involving a supervisor and an analyst performing the following tasks:

– The supervisor specifies an assignment for an analyst
– The supervisor specifies (and modifies) a context of investigation for the assignment

– The analyst performs tasks related to the assignment. As part of this, the analyst accesses various heterogeneous documents using a system that can keep track of the documents that were viewed
– The supervisor can verify if the documents accessed by the analyst are within the context of investigation specified for the assignment. The system analyzes the relevance of the documents and ranks them accordingly.

Ontology Specification and Development

As part of the ongoing Semantic Discovery project at the LSDIS lab, we have created and are maintaining a test-bed (SWETO) for evaluating ontological management and semantic technologies (Aleman-Meza et al. 2004). SWETO contains an ontology schema covering various domains, and it is populated using factual data or knowledge from multiple knowledge sources. To serve the purposes of the document access problem of insider threat, we refined a part of SWETO schema to sufficiently capture the domain of National Security and Terrorism to meet our prototyping and evaluation goals. A schematic of this part of the ontology is provided in Figure 1.

This ontology provides a conceptualization of organizations, countries, people, watch-lists, terrorists, events, terrorist acts, etc. that are all inter-related by named relationships to reflect real-world knowledge about the domain (i.e. "terrorist" *belongs-to* "terrorist organization"). The sources used to populate the ontology were selected for their information richness, semi-structured format and aptitude to quickly populate the ontology with a large number of entities and (more importantly) relationships in the domain of terrorism. For example, publicly available data maintained by intelligence agencies and Int'l. organizations, such as watch lists containing publicly declared "bad" persons and organizations. For ontology design and population, we used Semagix's Freedom, a commercial software which itself is based on a technology developed at and licensed from the LSDIS lab (Sheth et al. 2002). The same technology was used to build the Glycomics ontology (http://lsdis.cs.uga.edu/Projects/Glycomics/). Other large scale ontologies have been developed elsewhere with other methods yet not in the domain of national security. For example, TAP (Guha and McCool 2003) is an ontology about places, musicians, sports, movies, etc.

The ontology schema and the populated instances data were exported from Freedom and modeled in RDF. The part of SWETO used by the methods described in this paper consists of about 40 classes in the schema part of the ontology; the instances part consists of about 32,000 entities and about 35,000 explicit relationships.

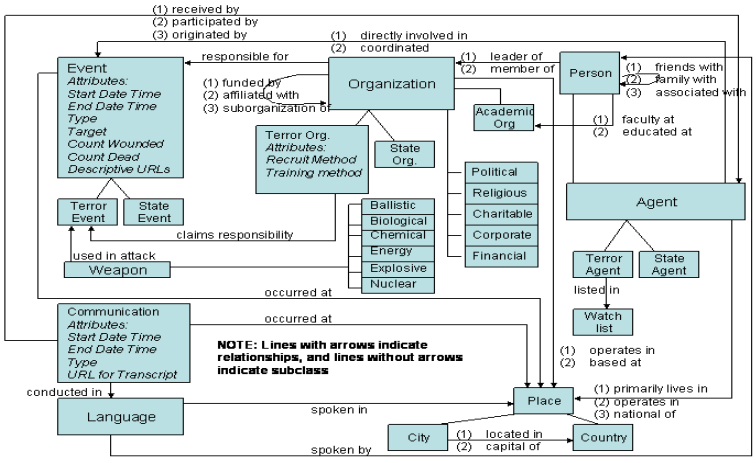

Fig. 1. National Security and Terrorism Part of SWETO Ontology

Ontological approach to the document access problem of insider threat

Figure 2 provides a schematic view of our approach. The first step utilizes a large ontology populated from trusted sources to semantically annotate a collection of documents. A *context of investigation* can be defined by a supervisor to capture (in ontological terms) the scope of an investigation assignment given to an intelligence analyst.

The main processing involves computing a measure of the relevance of each document (using the annotations), with respect to the context. Given that a collection of documents needs to be processed, additional technical challenges include the need to compute a potentially large number of semantic associations per document and use them to measure their relevance with respect to the context.

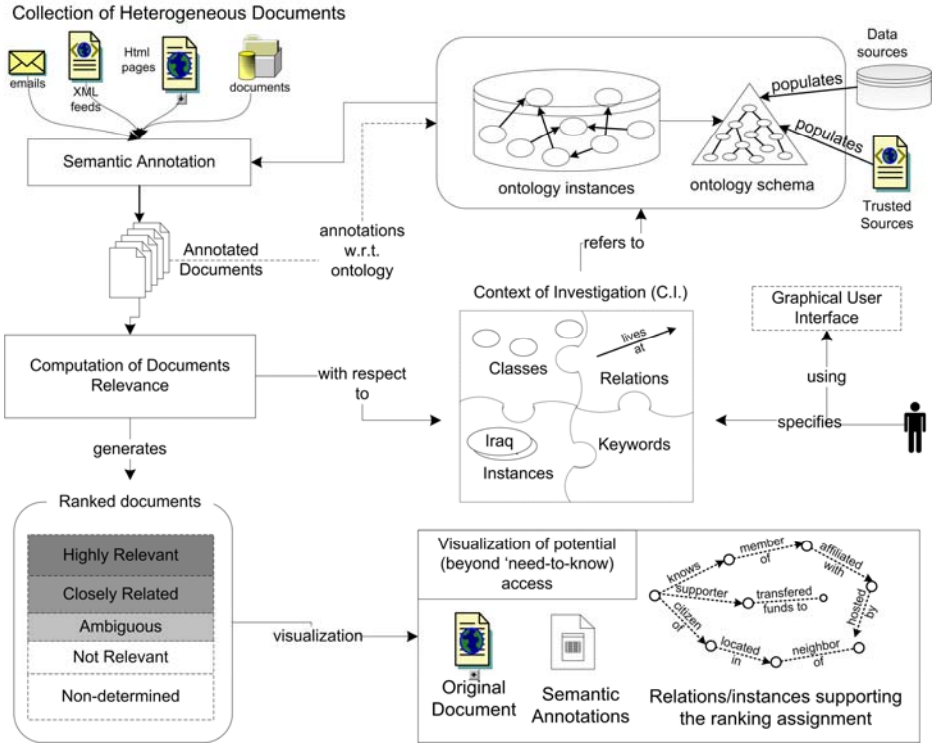

Fig. 2. Ontological Approach to the Legitimate Access Problem

Finally, the documents are ranked according to the computed relevance measure. Each document can be viewed/inspected to gain insight on the intention of access by the analyst beyond the "need to know". It has been noted that "finding relationships among suspects is vital in law enforcements applications" . A graphical user interface (accessible in the form of a Java Applet) displays the semantic associations that interconnect entities within a document to those that form part of the context of investigation. Only the relationships regarded relevant in the given context are displayed. Such semantic associations form the basis of identifying connections between two or more seemingly unrelated entities.

Context of investigation

Based on our previous work (Aleman-Meza et al. 2003) we define context as follows:
Definition 2 (*Context*): A context is a non-empty set of entities, relationships, and/or classes from an ontology.

The intuition is that a context captures a set of types of entities, relationships, and entities (at an ontology level) that are to be considered relevant (for example, in a query). In the case of legitimate document access problem of Insider Threat, the context is used to capture the investigation assignment given to an intelligence analyst. We refer to this as *context of investigation*, which is a combination of the following:

- A set of entity classes and relationships
- A set of entity instance names, and/or a negation of a set of entity instance names
- A set of keyword values that might appear at any attribute of the populated instance data, and/or a negation of a set of keyword values.

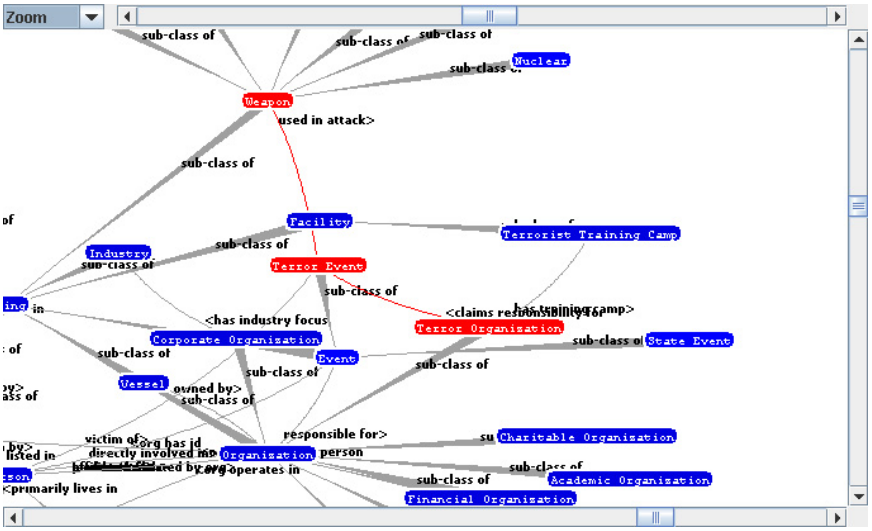

Fig. 3. Graphical interface for defining a Context of Investigation

Related research has mentioned an 'ontological foundation' for context yet did not provide a means of expressing context with respect to an ontology (Coutaz et al. 2005). One of the key components is providing a means for graph-based creation of a context of investigation.

We expanded upon our previous prototype for capturing the context of a user's interest with respect to an ontology (Halaschek et al. 2004) and implemented a graphical user interface. This was done by extending a version of the TouchGraph (http://www.touchgraph.com) applet to display graphs. Figure 3 displays an example of a context of investigation where the classes 'Airport', 'Event', and 'Person' have been added to the context. In addition, the context can be further defined in order to specify a more rigid set of semantic constraints. For example, it can be specified that a relation 'affiliated with' is part of the context only when it is connected with an entity that belongs to a specific class, say, 'Terror Organization'.

Figure 4 illustrates this example by highlighting with a thick line the combination of (a sample) entity and a relationship fit the context. (The gray nodes represent classes of the ontology; the 'rdf:type' relation indicates the class type of an entity).

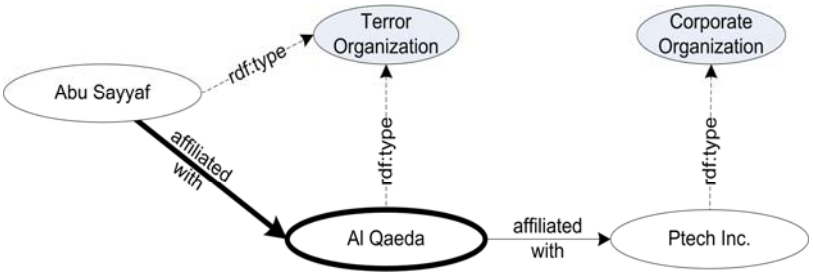

Fig. 4. Context Constraint of a Specific Relation-Entity Combination

Semantic Annotation

The documents viewed by an analyst are processed to generate a set of *semantically annotated documents*. Semantic annotation is referred to as both the metadata added to a document and the process of generating such metadata. We utilized Semagix's Freedom software to semantically enhance the documents that an analyst accessed as part of the assignment. The Freedom software annotates a document by passing it through the Semantic Enhancement Server module (Hammond et al. 2002). Entity names or synonyms within the document that are contained in the ontology are recognized.

The output of the semantic annotation is an XML document listing the identified entities; an 'enhanced' document is also produced by highlighting recognized entities. A fragment of a semantically annotated document is provided in Figure 5 (both in XML and with highlighting of recognized entities in the original document).

Fig. 5. Fragment of a Semantically Annotated Document

Relevance Measure for Documents

The measure the relevance of annotated documents with respect to the context of investigation is intended to help a supervisor determine whether the work of the analyst on a particular assignment poses an Insider Threat. At a high level, a relevance-engine module takes as input the set of semantically annotated documents (accessed by the intelligence analyst as part of his/her investigation assignment), the context of investigation for the assignment, and the ontology population and schema represented in RDF. The engine discovers semantic associations among entity annotations in the annotated document and the entity classes, entity instances, and/or keywords specified in the context of investigation. The discovery algorithm traverses the RDF graph searching for semantic associations up to a sequence of (predefined) length *k* (set to 9 by default). In order to perform this semantics analytics task we build upon previous work on discovery of semantic associations. We extend the definition of ρ-operators for semantic associations (Anyanwu and Sheth 2003) to introduce a ρ_κ operator for expressing queries for semantic associations using context.

Definition 3 (*ρ_κ-Query*): A ρ_κ-Query, expressed as $\rho_\kappa(x, c)$, where *x* is an entity and *c* is a context, results in the set of all semantic associations that exist between *x* and *c*. A semantic association between *x* and *c* exists if there is a sequence $e_1, p_1, e_2, p_2, e_3, \ldots, e_{n-1}, p_{n-1}, e_n$ in an RDF graph where $e_1 = x$ and e_i, $1 \le i \le n$, are entities and p_j, $1 \le j < n$, are properties, and either $e_n \in c$ or $type(e_n) \in c$ or $p_i \in c$, where type(e) is the class type (or concept) of entity e. Figure 6 illustrates an example of a ρ_κ-Query.

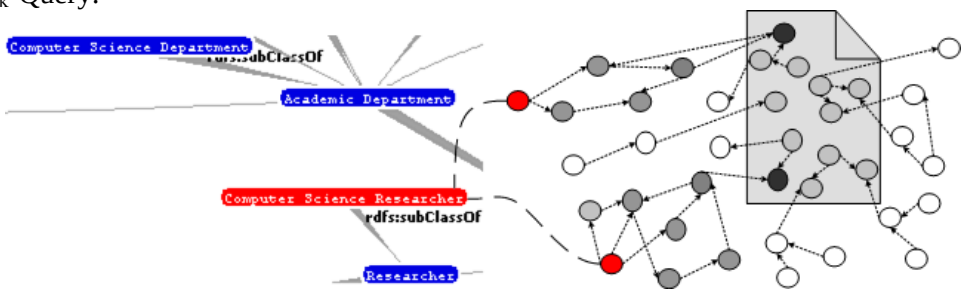

Fig. 6. ρ_κ-Query from entities within a document and a context.

Once the semantic associations among entities within a document and the context of investigation have been discovered, the relevance of a document *d* with respect to a context of investigation *CI* is computed as follows:

$$Relevance(d) = C_{CI} + R_{CI} + E_{CI} + K_{CI} \quad (1)$$

where, C_{CI} is the component of matching classes with respect to the context of investigation, *CI*. Similarly, R_{CI}, E_{CI}, and K_{CI} are the components for matching relations, entities, and keywords, respectively, with respect to the context of investigation. The discovered semantic associations interconnecting a document to a context can be seen as a neighborhood of *k* hops – similar to the intuition of a 'semantic neighborhood' (Rodriguez and Egenhofer 2003).

Each of the components in Equation 1 is computed based on the proximity of a match of the types of the entities of the document and its neighborhood with respect to the context of investigation. The computation of C_{CI} is as follows:

$$C_{CI} = \frac{\sum_{e_j \in d} \left[\sum_{i=1}^{|ng(e_j)|} \frac{1}{dist(e_j, v_i)+1} \right]}{|d|} \quad (2)$$

where, $ng(e)$ is the set of nodes and relations in the neighborhood of entity e; and the function $dist(e, v)$ computes a the distance between e and v. For the particular case of the component for keywords, K_{CI}, the formula also considers all attributes of each entity v_i with those keywords specified in the context of investigation. As part of the future work we plan to incorporate into the formula for K_{CI} a simplified version of that introduced by a hybrid search approach (Rocha et al. 2004).

Experimental Results

Increasingly, publicly available ontologies are being developed in various domains such as scientific publications (http://www.semanticweb.org/library/), the Open Directory Project (http://www.dmoz.org/), SWETO (Aleman-Meza et al. 2004) and TAP (Guha and McCool 2003). However, we were limited to the national security domain for which we had to develop our own ontology as described earlier. As document collection, we utilized a small but representative collection of 1,000 documents carefully chosen to test different scenarios of context of investigation. We observed that high scores were computed for documents containing entities directly or strongly fitting the context of investigation. The score values can get quite low when weak or rather long associations connect entities in the document to the context. A subset of 100 documents was carefully chosen to detect whether four cases of relevance: (i) directly related documents; (ii) strongly related documents; (iii) loosely related documents; and (iv) non-related documents. Cases at the extremes were easily verified to work correctly (i, iv). However, the strongly and loosely related cases required inspection and analysis by human to identify why a document has a medium to high ranking or medium to low ranking. Figure 7 illustrates the color pattern used to group documents according to their score value (instead of rank position).

We present the following examples on why documents get a high or low scores with respect to a context of investigation. A document on the terror organization Hizballah for which the algorithm was able to establish the following three semantic associations: Hamas –*operates in*→ Middle East, Al Qaeda –*operates in*→ Middle East, and Hizballah –*operates in*→ Middle East received a high score of 0.915.

A document on the terror organization Jemaah Islamiyah for which the algorithm was able to establish the single longer semantic association: Abu Sayyaf Group –*affiliated with*→ Al Qaeda –*operates in*→ Middle East received a lower score of 0.735.

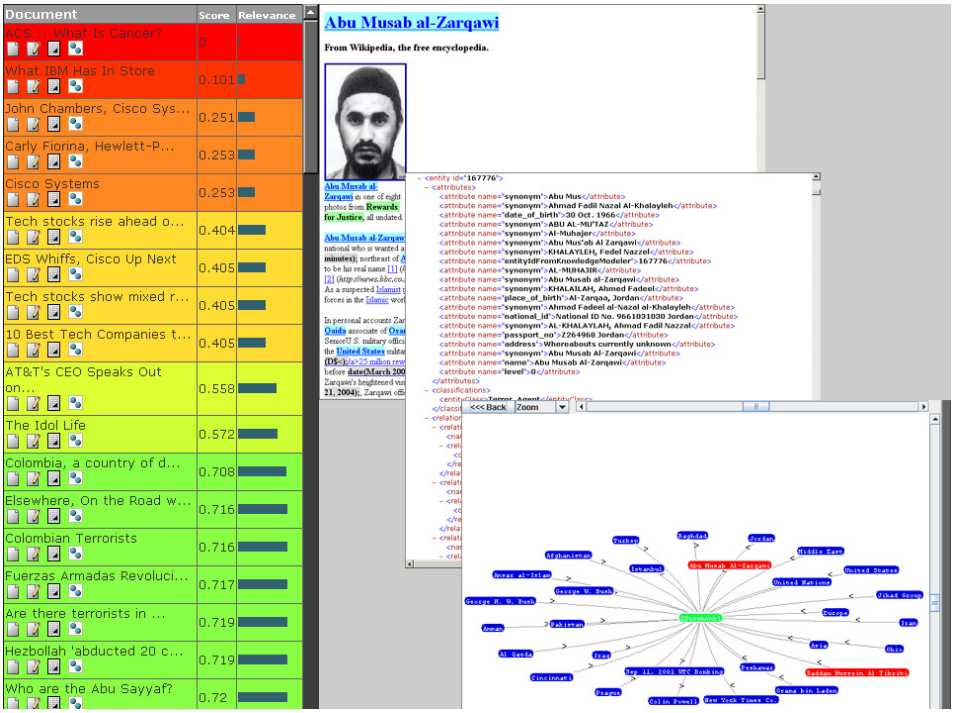

Fig. 7. Ranked documents and a few 'inspection' views.

A document on the terror organization Liberation Tigers of Tamil Eelam (LTTE) for which the algorithm was able to establish a few long semantic associations such as: Rajiv Gandhi –*victim of*→ May 21, 1991 LTTE Suicide Bomb –*located in*→ India ←*national of*– Dawood Ibrahim –*affiliated with*→ Al Qaeda –*operates in*→ Middle East received a very low score of 0.273. Additionally, an online demo of the application is available (http://lsdis.cs.uga.edu/Projects/SemDis/NeedToKnow/).

Semagix Freedom

Semagix Freedom is built around the concept of ontology-driven metadata extraction, allowing modelling of fact-based, domain-specific relationships between entities. It provides tools that enable automation in every step in the content chain - specifically ontology design, content aggregation, knowledge aggregation and creation, metadata extraction, content tagging and querying of content and knowledge. Figure 8 below shows the domain-model driven architecture of Semagix Freedom.

Semagix Freedom operates on top of a domain specific ontology that has classes, entities, attributes, relationships, a domain vocabulary and factual knowledge, all connected via a semantic network. The domain specific information architecture is dynamically updated to reflect changes in the environment, and it is easy to configure and maintain. The Freedom ontology maintains knowledge, which is any factual, real–world information about a domain in the form of entities, attributes and relationships (e.g., Figure 1).

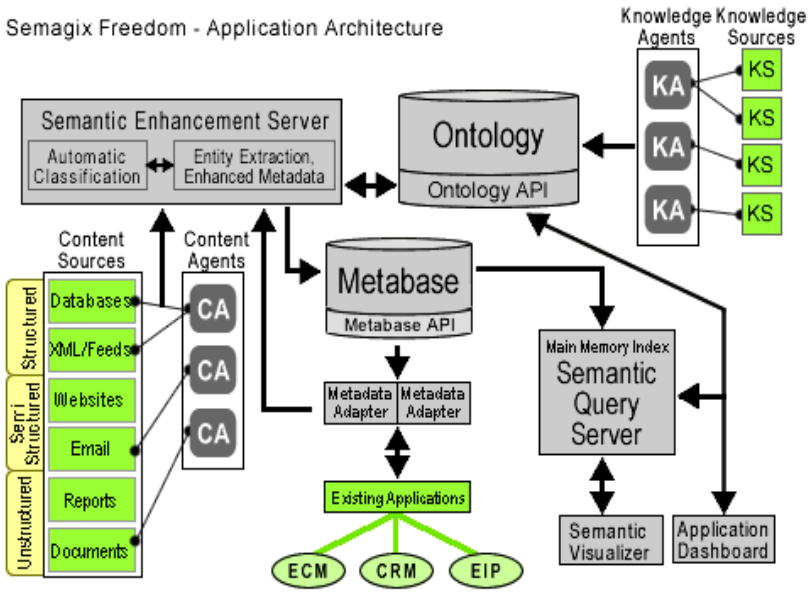

Fig. 8. Semagix Freedom Architecture

The ontology forms the basis of semantic processing, including automated categorization, conceptualization, cataloging and enhancement of content. Freedom provides a modeling tool to design the ontology schema (the assertional component of the system) based on the application requirements. Specifically, it allows flexible designing of the domain model by offering features like definition of customized entity types, relationships between entity types, entity attributes, cardinality constraints, class membership, etc. The ontology is automatically maintained by Knowledge Agents. These are software agents created without programming that traverse trusted knowledge sources and exploit structure to extract useful entities and relationships for populating the ontology automatically. Once created, they can be scheduled to perform knowledge extraction automatically at any desired interval, thus keeping the ontology up-to-date. Freedom also aggregates structured, semi-structured and unstructured content from any source and format, by extracting syntactic and contextually relevant semantic metadata. Much like Knowledge Agents, Content Agents are software agents created without programming using extraction infrastructure tools that extract useful syntactic and semantic metadata information from content and tag it automatically with pre-defined metatags. Incoming content is further “enhanced” by passing it through the Semantic Enhancement Server module (Hammond et al. 2002).

The Metabase stores both semantic and syntactic metadata related to content in either custom formats or one or more defined multiple metadata formats such as RDF, PRISM, Dublin Core, and SCORM. The Metabase stores content into a relational database as well as a main-memory checkpoint. At any point in time, a snapshot of the Metabase (index) resides in main memory (RAM), so that retrieval of entities is accelerated using the patented Semantic Query Server. The Semantic Query Server is a main memory–based front–end query server that enables the end–user to retrieve relevant content. A variety of semantic applications that exploit this technology can be built including Anti Money Laundering identification and risk assessment (2003), Financial Analyst Workbench, Homeland Security, and Citizen Portal applications. The Semantic Enhancement and Query Servers operate on the Metabase and ontology; they yield high quality query results because they provide the basis for in-context querying, whereas common search engines lack context and ambiguity resolution, and therefore relevance and accuracy.

CONCLUSIONS

This paper discussed a challenging problem of detecting illegitimate access of documents beyond *need to know*, one of the problems of Insider Threat. The approach involved processing of documents to produce semantic annotations and then use the semantic annotations to measure the relevance of a document with respect to the investigation assignment of an intelligence analyst. This measure computes an aggregated score of a set of semantic associations. A notion of *context* is defined to capture such assignment. A graphical representation of the ontology is used within a graphical user interface to specify the context. A new semantic association query is introduced to query for semantic associations among an entity and a context. A prototype is described by discussing the technical challenges involving this type of text and content analytics. The prototype provides an interface for inspection of the explicit relations that make a document relevant to an intelligence analyst's investigation assignment. Thus, it provides insight to a supervisor of the analyst on the need-to-know reason for access to a document.

This effort illustrates a unique attempt of driving research from a realistic application, core research issues in semantic association discovery, and use of commercial Semantic Web technology in building a populated ontology over publicly available data. This research demonstrates an example of collaboration involving academic research, industry technology, and government priorities, to address unique and technically demanding challenges.

Acknowledgements: We thank Semagix, Inc. for providing access to Freedom, which is based on the SCORE technology and related research performed at the LSDIS Lab. The work on semantic associations is funded in part by National Science Foundation (NSF) Award IIS-0325464 ("SemDis: Discovering Complex Relationships in Semantic Web"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. The insider-threat prototype was developed as part of the Advanced Research Development Activity (ARDA) Insider Threat Initiative, contracted through the Department of the Interior, Ft. Huachuca, contract # NBCHC030083.

REFERENCES

(2003). "Anti Money Laundering." Semagix, Inc.

(2004). "To Catch a Thief." Visual Analytics Inc.

Aleman-Meza, B., Burns, P., Eavenson, M., Palaniswami, D., and Sheth, A. P. "An Ontological Approach to the Document Access Problem of Insider Threat." *IEEE International Conference on Intelligence and Security Informatics (ISI-2005)*, Atlanta, Georgia, USA.

Aleman-Meza, B., Halaschek-Wiener, C., Arpinar, I. B., Ramakrishnan, C., and Sheth, A. P. (2005b). "Ranking Complex Relationships on the Semantic Web." *IEEE Internet Computing*, 9(3), 37-44.

Aleman-Meza, B., Halaschek, C., and Arpinar, I. B. (2005c). "Collective Knowledge Composition in a Peer-to-Peer Network." Encyclopedia of Database Technologies and Applications, L. C. Rivero, J. H. Doorn, and V. E. Ferraggine, eds., Idea-Group Inc.

Aleman-Meza, B., Halaschek, C., Arpinar, I. B., and Sheth, A. "Context-Aware Semantic Association Ranking." *First International Workshop on Semantic Web and Databases*, Berlin, Germany, 33-50.

Aleman-Meza, B., Halaschek, C., Sheth, A., Arpinar, I. B., and Sannapareddy, G. "SWETO: Large-Scale Semantic Web Test-bed." *16th International Conference on Software Engineering and Knowledge Engineering (SEKE2004): Workshop on Ontology in Action*, Banff, Canada, 490-493.

Anderson, R., and Brackney, R. (2004). *Understanding the Insider Threat*, RAND Corporation, Rockville, MD, USA.

Anyanwu, K., and Sheth, A. P. "r-Queries: Enabling Querying for Semantic Associations on the Semantic Web." *12th Int'l World Wide Web Conference*, Budapest, Hungary, 690-699.

Anyanwu, K., Sheth, A. P., and Maduko, A. "SemRank: Ranking Complex Relationship Search Results on the Semantic Web." *14th International World Wide Web Conference*, Chiba Japan, 117-127.

Arpinar, I. B., Sheth, A. P., Ramakrishnan, C., Usery, E. L., Azami, M., and Kwan, M.-P. (2004). "Geospatial Ontology Development and Semantic Analytics." Handbook of Geographic Information Science, J. P. Wilson and A. S. Fotheringham, eds., Blackwell Publishing.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." *Scientific American*, 284(5), 34-+.

Brickley, D., and Guha, R. V. (2000). "RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation."

Cathey, R., Ma, L., Goharian, N., and Grossman, D. "2003 ACM CIKM International Conference on Information and Knowledge Management." New Orleans, Louisiana, USA.

Chen, M. S., Han, J. W., and Yu, P. S. (1996). "Data mining: An overview from a database perspective." *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883.

Coutaz, J., Crowley, J. L., Dobson, S., and Garlan, D. (2005). "Context is key." *Communications of the ACM*, 48(3), 49-53.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R. V., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., and Zien, J. Y. "SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation." *Twelfth International World Wide Web Conference*, Budapest, Hungary, 178-186.

Ding, L., Kolari, P., Finin, T., Joshi, A., Peng, Y., and Yesha, Y. "On Homeland Security and the Semantic Web: A Provenance and Trust Aware Inference Framework." *AAAI Spring Symposium on AI Technologies for Homeland Security*, Stanford University, CA, USA.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press.

Guha, R. V., and McCool, R. (2003). "TAP: A Semantic Web Test-bed." *Journal of Web Semantics*, 1(1), 81-87.

Halaschek, C., Aleman-Meza, B., Arpinar, I. B., and Sheth, A. P. "Discovering and Ranking Semantic Associations over a Large RDF Metabase." *30th International Conference on Very Large Data Bases*, Toronto, Canada.

Hammond, B., Sheth, A., and Kochut, K. (2002). "Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content." Real World Semantic Web Applications, V. Kashyap and L. Shklar, eds., Ios Press, 29-49.

Handschuh, S., Staab, S., and Studer, R. (2003). "Leveraging metadata creation for the semantic web with CREAM." *Ki 2003: Advances in Artificial Intelligence*, 2821, 19-33.

Janik, M., and Kochut, K. "BRAHMS: A WorkBench RDF Store And High Performance Memory System for Semantic Association Discovery." *4th International Semantic Web Conference*, Galway, Ireland.

Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., and Scholl, M. "RQL: A Declarative Query Language for RDF." *The Eleventh International World Wide Web Conference*, Honolulu, Hawaii, USA, 592-603.

Krebs, V. (2002). "Mapping Networks of Terrorist Cells." *Connections*, 24(3), 43-52.

Lassila, O., and Swick, R. R. (1999). "Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation." W3C.

Ma, L., Goharian, N., and Meyers, C. "Detecting Misuse of Information Retrieval Systems Using Data Mining Techniques." *IEEE International Conference on Intelligence and Security Informatics (ISI-2005)*, Atlanta, GA, USA, 604-605.

Milnor, W. H., Ramakrishnan, C., Perry, M., Sheth, A. P., Miller, J. A., and Kochut, K. J. (2005). "Discovering Informative Subgraphs in RDF Graphs." LSDIS Lab, Computer Science, University of Georgia.

Mukherjea, S., and Bamba:, B. "BioPatentMiner: An Information Retrieval System for BioMedical Patents." *Thirtieth International Conference on Very Large Data Bases*, Toronto, Canada, 1066-1077.

Perry, M., Janik, M., Ramakrishnan, C., Ibanez, C., Arpinar, I. B., and Sheth, A. P. "Peer-to-Peer Discovery of Semantic Associations." *2nd International Workshop on Peer-to-Peer Knowledge Management (P2PKM)*, La Jolla, California, USA.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. "KIM - Semantic Annotation Platform." *2nd International Semantic Web Conference (ISWC2003)*, Sanibel Island, Florida, USA, 484-499.

Rectenwald, M., Lee, K., Seo, Y., Giampapa, J. A., and Sycara, K. (2004). "Proof of Concept System for Automatically Determining Need-to-Know Access Privileges: Installation Notes and User Guide." *CMU-RI-TR-04-56*, Robotics Institute, Carnegie Mellon University.

Rocha, C., Schwabe, D., and Aragao, M. P. "A Hybrid Approach for Searching in the Semantic Web." *13th International World Wide Web*, New York, New York, USA, 374-383.

Rodriguez, M. A., and Egenhofer, M. J. (2003). "Determining Semantic Similarity Among Entity Classes from Different Ontologies." *IEEE Transactions on Knowledge and Data Engineering*, 15(2), 442-456.

Shah, U., Finin, T., Joshi, A., Cost, R. S., and Mayfield, J. "Information Retrieval on the Semantic Web." *10th International Conference on Information and Knowledge Management*, McLean, Virginia, USA, 461-468.

Sheth, A. P. "Enterprise Applications of Semantic Web: The Sweet Spot of Risk and Compliance." *IFIP International Conference on Industrial Applications of Semantic Web*, Jyväskylä, Finland.

Sheth, A. P., Aleman-Meza, B., Arpinar, I. B., Halaschek, C., Ramakrishnan, C., Bertram, C., Warke, Y., Avant, D., Arpinar, F. S., Anyanwu, K., and Kochut, K. (2005a). "Semantic Association Identification and Knowledge Discovery for National Security Applications." *Journal of Database Management*, 16(1), 33-53.

Sheth, A. P., Arpinar, I. B., and Kashyap, V. (2003). "Relationships at the Heart of Semantic Web: Modeling, Discovering and Exploiting Complex Semantic Relationships." Enhancing the Power of the Internet Studies in Fuzziness and Soft Computing, M. Nikravesh, B. Azvin, R. Yager, and L. A. Zadeh, eds., Springer-Verlag.

Sheth, A. P., Bertram, C., Avant, D., Hammond, B., Kochut, K., and Warke, Y. (2002). "Managing semantic content for the Web." *IEEE Internet Computing*, 6(4), 80-87.

Sheth, A. P., Ramakrishnan, C., and Thomas, C. (2005b). "Semantics for the Semantic Web: the Implicit, the Formal and the Powerful." *International Journal on Semantic Web and Information Systems*, 1(1), 1-18.

Subramaniam, L. V., Mukherjea, S., Kankar, P., Srivastava, B., Batra, V. S., Kamesam, P. V., and Kothari, R. "Information extraction from biomedical literature: methodology, evaluation and an application." *2003 ACM CIKM International Conference on Information and Knowledge Management*, New Orleans, Louisiana, USA, 410-417.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., and Ciravegna, F. "MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup." *13th International Conference on Knowledge Engineering and Management (EKAW 2002)*, Sigüenza, Spain.

Woods, W. (1975). "What's in a link: Foundations for Semantic Networks." Representation and Understanding, D. Bobrow and A. Collins, eds., Academic Press, New York, 35-82.