(To appear in Proceedings of the) IEEE Intl. Conference on Intelligence and Security Informatics (ISI-2005), May 19–20, 2005; (link to: Insider Threat paper)

# An Ontological Approach to the Document Access Problem of Insider Threat

Boanerges Aleman-Meza[1], Phillip Burns[2], Matthew Eavenson[1],
Devanand Palaniswami[1], Amit Sheth[1]

[1]LSDIS Lab, Department of Computer Science, University of Georgia, Athens, GA 30602
boanerg@cs.uga.edu, durandal@uga.edu, devp@uga.edu, amit@cs.uga.edu
[2]Computer Technology Associates, 7150 Campus Drive, Ste 100, Colorado Springs, CO 80920
phillip.burns@cta.com

**Abstract.** Verification of legitimate access of documents, which is one aspect of the umbrella of problems in the Insider Threat category, is a challenging problem. This paper describes the research and prototyping of a system that takes an ontological approach, and is primarily targeted for use by the *intelligence community*. Our approach utilizes the notion of *semantic associations* and their discovery among a collection of heterogeneous documents. We highlight our contributions in (graphically) capturing the scope of the investigation assignment of an intelligence analyst by referring to classes and relationships of an ontology; in computing a measure of the relevance of documents accessed by an analyst with respect to his/her assignment; and by describing the components of our system that have provided early yet promising results, and which will be further evaluated more extensively based on domain experts and sponsor inputs.

## 1. Introduction

Insider Threat refers to the potential malevolent actions by employees within an organization, a specific type of which relates to legitimate access of documents. In the context of the intelligence community, one of the goals is to ensure that an analyst accesses documents that are relevant to his/her assigned investigation objective, i.e., accesses the data on a "need to know" basis.

In this paper we discuss our work as part of an Advanced Research and Development Activity (ARDA) funded project, in which we have developed an ontological approach to address the *legitimate document access* problem of Insider Threat. There is a range of techniques that support determining if a collection of documents is relevant to a particular domain. Such techniques can be applied to determine if documents accessed by an intelligence analyst are relevant to his/her job assignment. Examples include statistical, NLP, and machine learning techniques such as those leading to document clustering and/or automatic document classification that exploit implicit

semantics[1]. A concern with these approaches is that they generally do not support an ability to clearly understand the reasons behind why an accessed document is relevant (or not relevant) to the investigation objective of the intelligence analyst. Most of these techniques have also focused on mapping documents to a predefined taxonomy, which is found to be a rather limited method of representing knowledge when named relationships between concepts (e.g., a person *works-for* an organization) represent an important part of the domain knowledge. In this context, we pursue a strategy that uses ontology to capture domain semantics and semantic metadata to capture semantics of heterogeneous domains.

In our approach, we utilize *semantic associations*, which aim to capture meaningful and possibly complex relationships between entities (in a large dataset of metadata based on a graph model) [3]. Initially we sought to leverage our previous experience where we have applied such associations to a class of national security and homeland security applications (e.g., Passenger Threat Assessment [7]). The need to represent the scope of the investigative assignment given to an analyst required us to take a fresh look at our previous work in capturing a user's interest with respect to an ontology (or subset thereof) [1]. Additional technical challenges include the need to compute a large number of semantic associations per document. Scalability becomes an issue given the potentially large collection of documents to be analyzed. For our ontological approach, a starting point was the building of a populated ontology. In doing so, we have built upon our significant experience in the development of large populated ontologies (e.g., [2], Glycomics Ontology[2]).

This paper presents the following novel conceptual and technical contributions:

- A practical yet flexible notion of capturing the scope of the investigation assignment of an analyst in terms of semantic constraints over an ontology. We call it the *context of investigation*, and we specify it using a graphical user interface to be used by the supervisor or investigator associated with an analyst's assignment.
- A computational measure that exploits *semantic associations* in a novel way to determine the relevance of a document with respect to a context of investigation.
- A prototype tested with a small-to-medium but representative document set.

Since we have not completed a comprehensive evaluation and have not fully evaluated scalability challenges, we present this work as a short paper. A comprehensive literature overview is also not presented for brevity.

## 2. Our Ontological Approach to the Legitimate Access Problem

Figure 1 provides a schematic of our approach. We use a large ontology populated from trusted sources to semantically annotate a collection of documents (viewed by an intelligence analyst). The system provides a means to define a *context of investigation* that aims to capture, in ontological terms, the scope of an investigation assign-

[1] Implicit semantics (as used here) capture possible relationships between concepts, but cannot or do not name specific relationships between the concepts. Explicit semantics use named relationships between concepts, and in the context of recent Semantic Web approaches, often use ontologies represented using a formal language; for further discussion, see [8].

[2] http://lsdis.cs.uga.edu/Projects/Glycomics/

ment given to an intelligence analyst. Hence, the goal is to measure the relevance of each document (using the annotations), with respect to the context of investigation. The documents are then grouped based on that measure (using a user-customizable threshold). Additionally, each document can be inspected by a supervisor to gain insight on the purpose of access by the analyst (beyond the "need to know"). The system supports this task by graphically displaying the *semantic associations* that interconnect entities in a document to those that form part of the context of investigation.

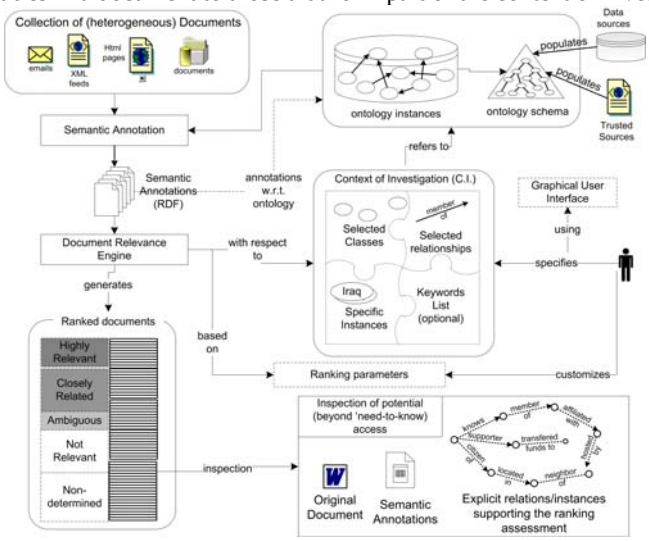


**Fig. 1.** Schematic of Ontological Approach to the 'Legitimate Document Access' Problem

### 2.1 Ontology Specification and Development

The ongoing Semantic Discovery project at the LSDIS lab has created (and maintains) a test-bed (SWETO) for evaluating semantic technologies [2]. We used and refined a subset of SWETO focusing on the domain of National Security and Terrorism. It was populated with real-world publicly available data maintained by international organizations. For ontology design and population, we used Semagix's Freedom[3], a commercial software based on earlier research developed at and licensed from the LSDIS lab [6]. The ontology consists of about 40 classes, populated with about 32,000 entities and about 35,000 explicit relationships among them.

### 2.2 Context of Investigation

The intuition behind a context of investigation lies in capturing, at an ontology level, the types of entities and relationships that are to be considered important. The context

[3] http://www.semagix.com

can contain 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'. The *context of investigation* is a combination of (i) entity classes; (ii) entity instances; (iii) named relationships between entity classes. Our prototype supports a graph-based user interface for defining a *context of investigation* (using TouchGraph[4]).

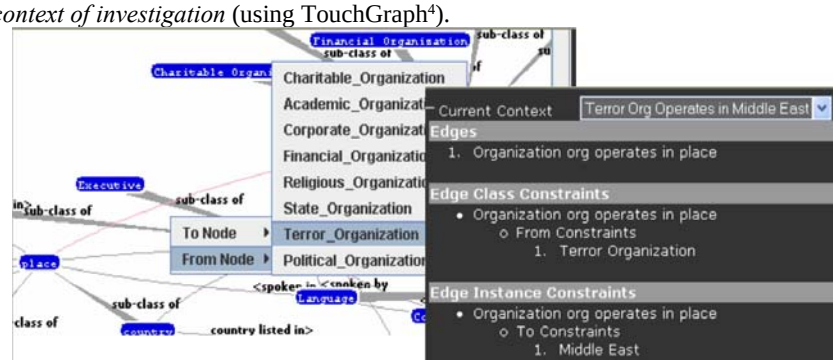


**Fig. 2.** Specifying Context of Investigation

Figure 2 displays an example of a context of investigation where *Middle Eastern Terrorism* is specified using the relationship "*Organization*" *operates_in* "*Place*" constrained into: "*Terror Organization*" *operates_in* "*Middle East*"

### 2.3 Semantic Annotation

The documents viewed by the analyst are processed to produce *semantically annotated documents*. Semantic annotation is metadata (data that describes data) added to a document, and also the process of generating such metadata[5]. Semagix's Freedom software was used to semantically enhance the documents that an analyst accessed as part of the assignment. The Freedom software searches the document and picks out entity names or synonyms within the document that are contained in the ontology.

### 2.4 Relevance Measure for Documents

The *Documents Relevance Engine* measures the relevance of annotated documents with respect to (w.r.t.) the context of investigation. The engine takes as input the set of semantically annotated documents, the context of investigation for the assignment, the ontology schema represented in RDF[6], and the ontology instances represented in RDF. The goal is to provide a ranked list of the documents based on their relevance to the assignment (represented using context of investigation described in Section 2.2). The documents relevance engine measures the relevance of the entity annotations in

[4] http://www.touchgraph.com
[5] For example, the KIM Platform http://www.ontotext.com
[6] Resource Description Framework, http://www.w3.org/RDF/

an annotated document w.r.t. the context of investigation. The relevance score of each document to the context of investigation is computed using semantic associations. A formalization of Semantic Associations over metadata represented in RDF was presented in [3]. Here we provide an adapted definition.

*Definition 1* ( $\rho$ *−Semantic Association*): Two entities $e_1$ and $e_n$ are semantically associated if there exists a sequence $e_1, P_1, e_2, P_2, e_3,..., e_{n-1}, P_{n-1}, e_n$ in an RDF graph where $e_i, 1 \leq i \leq n$, are entities and $P_j, 1 \leq j < n$, are relationships.

The relevance measure of a document *d* considers four components 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 *CI*. Similarly, R, E, and K are the components for matching relations, entities, and keywords, respectively. In our system, we pre-compute semantic associations for each document up to a (fixed) association length *n*. That is, a neighborhood of *n* hops from the entities on the document (similar to the intuition of a 'semantic neighborhood' described in [5]).

$C_{CI}$ is computed based on whether there is a match of the types of the entities of the document and its neighborhood with respect to the context of investigation,

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

where, $ng(e)$ is the set of nodes and relationships in the neighborhood of entity *e*; and the function $dist(e, v)$ computes the distance between *e* and *v*. Computing components $R_{CI}$, and $E_{CI}$ proceeds in similar fashion. In the component for keywords, $K_{CI}$, the formula differs by considering all attributes of each entity $v_i$ with those keywords specified in the context of investigation. We plan to incorporate into the formula for $K_{CI}$ a simplified version of the ideas presented in [4].

## 3. Initial Results and Conclusions

Our initial experiments were conducted on a collection of 1000 documents. A few example results in the context of Middle-Eastern Terrorism discussed in Section 2.2 are provided here. A high score of 0.91 was calculated for a document on 'Ansar al-Islam' where the semantic association *Ansar al-Islam –operates in→ Middle East* relates it to the context. A score of 0.735 for a document on Abu Sayyaf was the result of the (longer) semantic association *Abu Sayyaf Group –affiliated with→ Al Qaeda –operates in→ Middle East*. A low score of 0.425 was calculated for a document on the Sri Lankan group 'LTTE' due to the long semantic association *Sriperumbudur –located in→ India ←national of– Dawood Ibrahim –affiliated with→ Al Qaeda –operates in→ Middle East*.

We acknowledge that further evaluations are needed, but early results are promising and provide useful insights for our future work. An online demo is available[7].

Our approach has several advantages, including: (a) capability to keep the ontology updated (this becomes particularly important in dealing with changing and/or new information, e.g., new data being posted in watch-lists); (b) a means to support inspection of the explicit relationships on why a document is relevant to the context of investigation. Thus the supervisor of the intelligence analyst is able to gain insight on the need-to-know reason for access to the document. Our next steps in this project include conducting extensive evaluations, and addressing quality and scalability issues.

**Acknowledgements.** This work is conducted as part of the Advanced Research Development Activity (ARDA) Insider Threat Initiative, contracted through the Department of the Interior, Ft. Huachuca, contract # NBCHC030083. The larger projects at the LSDIS lab which provide the basis for research in Semantic Association Discovery are funded by the National Science Foundation through Awards 0219649 (“Semantic Association Identification and Knowledge Discovery for National Security Applications”), and IIS-0325464 ("SemDis: Discovering Complex Relationships in Semantic Web"). We also acknowledge our collaboration with Semagix, Inc, which enabled our use of Semagix Freedom.

## References

1. B. Aleman-Meza, C. Halaschek, I.B. Arpinar, A. Sheth, Context-Aware Semantic Association Ranking. Proceedings of Semantic Web and Databases Workshop, Berlin, September 7-8, 2003, pp. 33-50
2. B. Aleman-Meza, C. Halaschek, A. Sheth, I.B. Arpinar, and G. Sannapareddy. SWETO: Large-Scale Semantic Web Test-bed. Proceedings of the 16th International Conference on Software Engineering and Knowledge Engineering (SEKE2004): Workshop on Ontology in Action, Banff, Canada, June 21-24, 2004, pp. 490-493
3. K. Anyanwu and A. Sheth ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web. The Twelfth International World Wide Web Conference, Budapest, Hungary, 2003, pp. 690-699
4. C. Rocha, D. Schwabe, M.P. Aragao. A Hybrid Approach for Searching in the Semantic Web, In Proceedings of the 13th International World Wide Web, Conference, New York, May 2004, pp. 374-383.
5. M.A. Rodriguez, M.J. Egenhofer, Determining Semantic Similarity Among Entity Classes from Different Ontologies, IEEE Transactions on Knowledge and Data Engineering 2003 15(2):442-456
6. A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke. Managing Semantic Content for the Web. IEEE Internet Computing, 2002. 6(4):80-87
7. A. Sheth, B. Aleman-Meza, I.B. Arpinar, C. Halaschek, C. Ramakrishnan, C. Bertram, Y. Warke, D. Avant, F.S. Arpinar, K. Anyanwu, and K. Kochut. Semantic Association Identification and Knowledge Discovery for National Security Applications. Journal of Database Management, Jan-Mar 2005, 16 (1):33-53
8. A. Sheth, C. Ramakrishnan, and C. Thomas, Semantics for the Semantic Web: the Implicit, the Formal and the Powerful, International Journal on Semantic Web and Information Systems, 2005, 1(1):1-18

[7] http://lsdis.cs.uga.edu/Projects/SemDis/NeedToKnow/