

An Ontological Approach to the Document Access Problem of Insider Threat

Boanerges Aleman-Meza¹, Phillip Burns², Matthew Eavenson¹,
Devanand Palaniswami¹, Amit Sheth¹

¹LSDIS 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

²Computer 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 large meta-database of heterogeneous documents, as well as ranking semantic associations based on relevance. 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 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 “malevolent actions by an already trusted person with access to sensitive information and information systems” [3]. A specific type of Insider Threat relates to document access and involves effective monitoring of the actions of an intelligence analyst. 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. One of the goals of addressing Insider Threat is to ensure that an intelligence analyst accesses documents that are relevant to his/her assigned investigation objective, i.e., accesses the data on a “needs to know” basis. Furthermore, when extraneous access is made, the systems or methods in place should help in identifying and, if possible, auditing such access.

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 specifically 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 the Insider Threat problem 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 semantics¹. In [9], a list of positive and negative examples is used by a classifier agent to generate a set of weight vectors that determine the permission of each document for an analyst. When an analyst selects a document, the authorization agent determines whether it is viewable to the analyst based on the generated weight vectors. 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 an alternative strategy that uses ontology to capture domain semantics and semantic metadata to capture semantics of heterogeneous domains. This approach captures semantics more explicitly and comprehensively.

Focusing on the *legitimate access problem* of Insider Threat, we utilize the notion of *semantic associations* which aim to capture meaningful and possibly complex relationships between entities (in a large dataset of metadata based on a graph model) [4]. Initially we sought to leverage our previous experience in algorithms for discovery of semantic associations where we have applied such associations to a class of national security and homeland security applications (e.g., Passenger Threat Assessment [13]). The need to represent the scope of the investigative assignment given to an analyst and its relevance to a collection of documents required us to take a fresh look at our previous work in capturing the context of a user's interest with respect to an ontology (or subset thereof) aimed at ranking a set of semantic associations interconnecting two entities [1], [7]. However, addressing the *legitimate document access* problem of Insider Threat brings additional technical challenges, given 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. Additionally, new strategies are required to measure the relevance of the collection of documents with respect to the scope or context of the investigation assignment of an intelligence analyst.

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 Ontology²). This capability is enabled in significant part by a semantic technology infrastructure commercialized from our earlier research which supports creation of large populated ontologies, and automatic semantic metadata extraction/annotation of heterogeneous documents [12].

This paper presents the following novel conceptual and technical contributions:

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 [14].

² <http://lsdis.cs.uga.edu/Projects/Glycomics>

- 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, monitor or investigator associated with an analyst’s assignment. The context of investigation consists of classes and named relations from the ontologies, a combination of them, specific entities from the ontology, and/or a set of (optional) keywords.
- 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.
- An initial description of the first stage of the components of our system, with the design decisions made to face scalability constraints. Key ideas were implemented and tested with a small-to-medium but representative set of documents; we describe the initial results of this first stage that look promising.

Since we have not completed more comprehensive evaluation and have not fully evaluated the scalability challenge (albeit recent advances in semantic association discovery algorithms [5, 6], coupled with main memory Resource Description Framework (RDF³) graph processing implemented in the LSDIS lab make us quite hopeful), we are presenting this work as a short paper rather than a full paper. A comprehensive literature overview is also not presented for brevity.

Remainder of this paper is organized as follows. Section 2 describes our ontological approach to the *legitimate access problem*. Section 3 details the ontology specification and development. Section 4 gives an initial description of the first stage of our system, and Section 5 provides concluding remarks and insights on future work.

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. This produces a collection of *semantic annotations* for the 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 assignment 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 (customizable based on several parameters, e.g., see [5, 6]), and a threshold (selectable by the user) is applied to identify unrelated documents. 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 (or their lack of relevance to the scope of the investigation assignment).

³ <http://www.w3.org/RDF/>

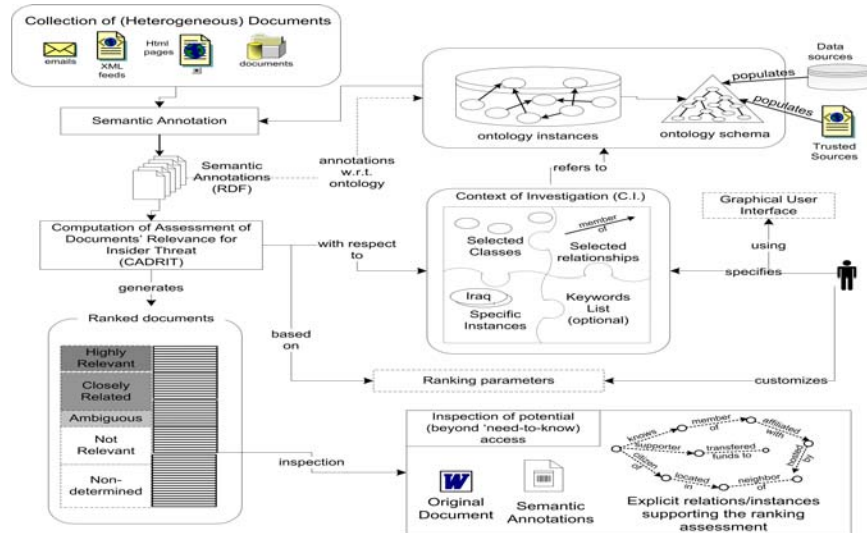


Fig. 1. Schematic of Ontological Approach to the Legitimate Access Problem

3. Ontology Specification and Development

As part of the ongoing Semantic Discovery project at the LSDIS lab, we had created and are maintaining a test-bed (SWETO) for evaluating ontological management and semantic technologies [2]. 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 this project, 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 2.

In particular, we populated the terrorism part of the ontology schema with real-world publicly available data maintained by international organizations. For ontology design and population, we used Semagix’s Freedom⁴, a commercial software which itself is based on a technology developed at and licensed from the LSDIS lab [12].

The ontology schema and the populated instances data were then exported from Freedom and modeled according to the (W3C technical recommendation) RDF. Entities in RDF are described using properties which have values, and by explicitly relating them to other entities by means of named relationships. 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.

⁴ <http://www.semagix.com>

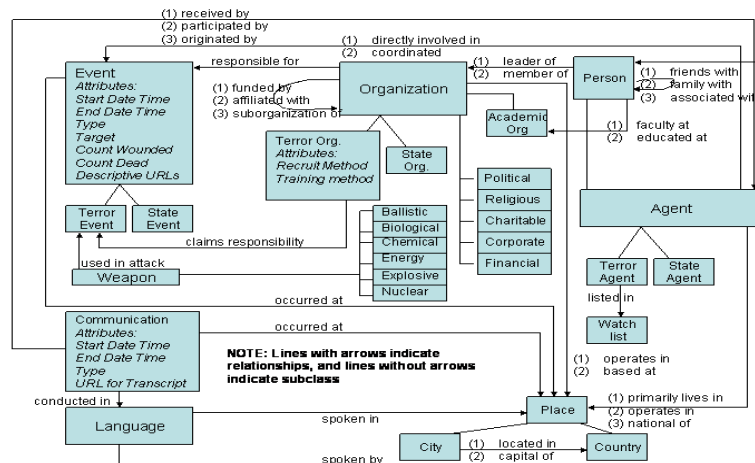


Fig. 2. National Security and Terrorism Part of SWETO Ontology

4. System for Insider Threat Document Protection

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 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 back-end of this step involves the *Documents Relevance Engine* that we developed. The engine analyzes *semantically annotated* documents that corresponds to the documents accessed by the analyst (the *semantic annotation* process is described later).

4.1 Context of Investigation

For the problem on hand, we define the *context of investigation* as a combination of the following:

- A set of entity classes and relationships, and/or a negation of 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.

In all of the above, the sets imply that they are relevant to the assignment whereas a negation of a set implies that the contents of those sets are to be considered outside the scope of the investigation assignment of an intelligence analyst. One of the key components of the system is to support a graph-based creation of a *context of investigation* by means of a user interface. We expanded upon our previous ideas in capturing the context of a user's interest with respect to an ontology [1] and implemented a graphical user interface (by extending a version of the TouchGraph⁵ applet to display graphs).

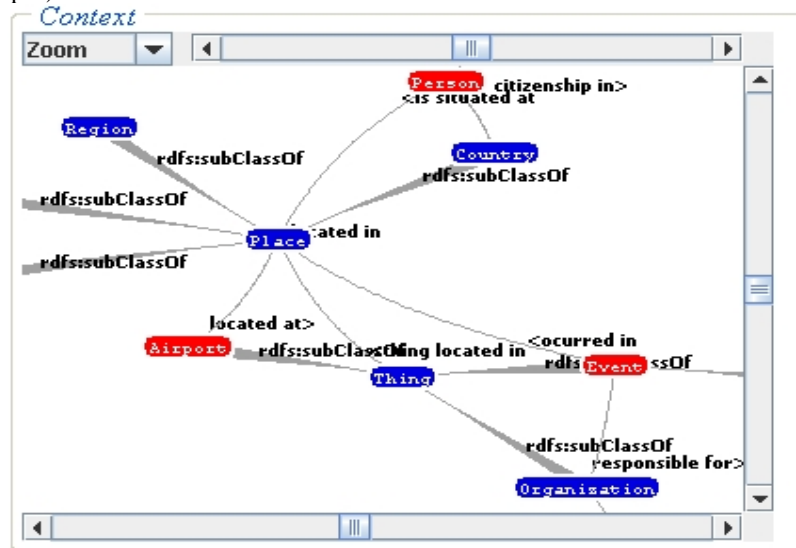


Fig. 3. Specifying Context of Investigation

Figure 3 displays an example of a context of investigation where the classes ‘Airport’, ‘Event’, and ‘Person’ have been added to the context.

The intuition behind specifying a *context of investigation* that refers to classes and relationships of the ontology lies in the goal of capturing, at a high level, the types of entities, (or relationships), that are to be considered important. However, 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-relation that fits the context. (The gray nodes represent classes of the ontology; the ‘rdf:type’ relation indicates the class type of an entity).

⁵ <http://www.touchgraph.com>

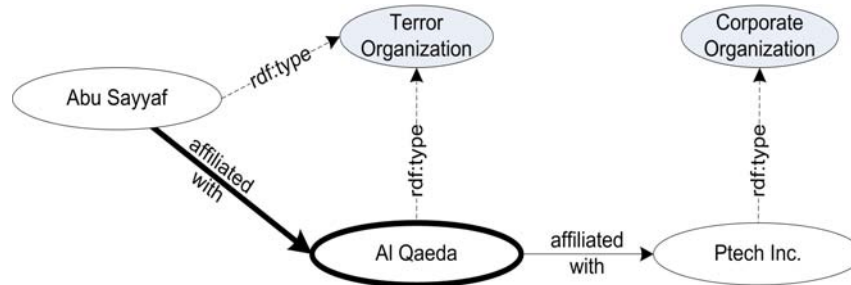


Fig. 4. Context Constraint of a Specific Relation-Entity Combination

4.2 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 metadata⁶. 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. Among other things, Freedom uses user-defined rules to extract metadata out of unstructured and (semi)-structured sources, and to then pick out entities (relating them to a class in the ontology), their attributes, as well as named relationships to other entities [8]. A fragment of a semantically annotated document is provided in Figure 5.

<pre> - <attribute name="synonym" type="string"> <CDATA[USS Cole]> </attribute> - <attribute name="synonym" type="string"> <CDATA[USS Cole attack]> </attribute> - <attribute name="synonym" type="string"> <CDATA[attack on the USS Cole]> </attribute> - <attribute name="type" type="string"> <CDATA[Suicide Bomb]> </attribute> - <attribute name="target" type="string"> <CDATA[Military Personnel]> </attribute> - <attribute name="description" type="string"> <CDATA[A U.S. warship, the U.S.S. Cole, was targeted by a suicide squad, ship with an explosives-laden boat. Thirteen American sailors were killed]> </attribute> - <attribute name="start_date_time" type="date"> <CDATA[Oct 12 2000 00:00:00]> </attribute> - <attribute name="count_wounded" type="long"> <CDATA[33]> </attribute> </pre>	<pre> - <attribute name="synonym" type="string"> <CDATA[Bin Laden]> </attribute> - <attribute name="synonym" type="string"> <CDATA[Bin Laden]> </attribute> - <attribute name="synonym" type="string"> <CDATA[Abu Abdullah and Al-Motah]> </attribute> - <attribute name="synonym" type="string"> <CDATA[Osama Bin Laden]> </attribute> - <attribute name="synonym" type="string"> <CDATA[Osama Bin Laden]> </attribute> - <attribute name="place_of_birth" type="string"> <CDATA[Yemen, Saudi Arabia]> </attribute> - <attribute name="date_of_birth" type="string"> <CDATA[10 Jul 57]> </attribute> - <attribute name="name" type="string"> <CDATA[Osama Bin Laden]> </attribute> - <relationships> <relationship id="1039"> <class> <instance id="1"> <CDATA[person]> </instance> </class> </relationship> </relationships> </attribute> - <relationships> <relationship id="2001"> <instance id="16113" position="2"> <CDATA[Osama]> </pre>
--	---

Osama [Bin Laden](#) is both one of the [CIA's](#) most wanted men and a hero to his associates were already being sought by the [US](#) on charges of international [year\(1998\)](#); bombing of American embassies in Africa and last year's [attack on the USS Cole in Yemen](#). In May this year a [US](#) jury convicted four men believed to be linked with [Bin Laden](#) of plotting the embassy bombings in [Kenya](#) and [Tanzania](#). [Bin Laden](#), an immensely wealthy and private man, has been granted a safe haven by [Afghanistan's](#) ruling [Taliban](#) movement. During his time in hiding, he has called for a holy war against the [US](#), and for the killing of [Americans and Jews](#). He is reported to be able to rally around him up to 3,000 fighters. He is also suspected of helping to set up Islamic training centres to prepare soldiers to fight in [Chechnya](#) and other parts of the [former Soviet Union](#).

Fig. 5. Fragment of a Semantically Annotated Document

⁶ For example, the KIM Platform <http://www.ontotext.com/>

4.3 Relevance Measure for Documents

The purpose of the *Documents Relevance Engine* is to measure the relevance of annotated documents with respect to the context of investigation. This would help a supervisor determine whether the work of the analyst on a particular assignment poses an Insider Threat. At a high level, the engine 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, the ontology schema represented in RDF, and the ontology instances represented in RDF. By using a relevance measure function, the engine verifies whether the entity annotations in the annotated document can be fit into the entity classes, entity instances, and/or keywords specified in the context of investigation. After analysis, it groups the collection of documents within one of the following categories:

- *Highly relevant*, when the entities annotated in the document directly fit the context of investigation.
- *Closely related*, when the entities in the document fit the context of investigation through strong *semantic associations* with other entities.
- *Ambiguous*, when the entities in the document fit the context of investigation through weak *semantic associations* with other entities.
- *Not relevant*, if the entities in the documents fall outside the context of investigation and/or fall into the ‘negative’ sets (see Section 4.1) of the context of investigation (if any had been specified).
- *Undeterminable*, if no entities that correspond to concepts in the domain ontology were found in the document.

The computation of the relevance measure for documents is in part based on the notion of semantic associations. A formalization of Semantic Associations over metadata represented in RDF was presented in [4]. Here we provide an adapted definition.

Definition 1 (ρ-Semantic Association): Two entities e_1 and e_n are semantically associated if there exists a sequence $e_1, P_1, e_2, P_2, e_3, \dots, e_{n-1}, P_{n-1}, e_n$ in an RDF graph where $e_i, 1 \leq i \leq n$, are entities and $P_j, 1 \leq j < n$, are relationships (the sequence of relationships is not constrained to direction).

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, E, and K are the components for matching relations, entities, and keywords, respectively, with respect to CI . For our initial version of the formula to compute $Relevance(d)$ we traverse semantic associations up to a sequence of (predefined) length n . Thus, a neighborhood of n hops from the entities on the document is obtained (this is similar to the intuition of a ‘semantic neighborhood’ described in [11]).

C_{CI} is computed as follows, (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 relations 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. Alternatively, in the case of the component for keywords, K_{CI} , the formula differs in the sense that it considers 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 [10].

5. Conclusions

The early results from our ontological approach to address this problem, including the use of a graphical representation of the ontology to specify the scope of investigation, and the use of a computational measure that builds on semantic associations to determine relevance, are promising, and provide useful insights for our future work. Our approach has several advantages, including (a) the technology we used to populate an ontology from heterogeneous sources includes capabilities for updates (this becomes particularly important in dealing with changing and/or new information, e.g., new data being posted in watch-lists); and (b) a means to support inspection of the explicit relations that make a document (highly/closely/ambiguously/not) 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 research include conducting further extensive evaluation with domain expert and sponsor inputs, leading to better understanding of quality and scalability issues.

Acknowledgements

This work is conducted as part of the Advanced Research Development Activity (ARDA) (<http://www.ic-arda.org>) 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 (NSF) through Awards 0219649 ("Semantic Association Identification and Knowledge Discovery for National Security Applications"), and 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 funding organizations. 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. R. Anderson and R. Brackney. Understanding the Insider Threat. Proceedings of a March 2004 Workshop. Prepared for the Advanced Research and Development Activity (ARDA). <http://www.rand.org/publications/CF/CF196/>
4. 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
5. K. Anyanwu, A. Maduko, A. Sheth, SemRank: Ranking Complex Relationship Search Results on the Semantic Web, In Proceedings of the 14th International World Wide Web Conference, Japan 2005 (accepted, to appear)
6. K. Anyanwu, A. Maduko, A. Sheth, J. Miller. Top-k Path Query Evaluation in Semantic Web Databases. (submitted for publication), 2005
7. C. Halaschek, B. Aleman-Meza, I.B. Arpinar, A. Sheth Discovering and Ranking Semantic Associations over a Large RDF Metabase Demonstration Paper, VLDB 2004, 30th International Conference on Very Large Data Bases, Toronto, Canada, 30 August - 3 September, 2004
8. B. Hammond, A. Sheth, and K. Kochut, Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content, in Real World Semantic Web Applications, V. Kashyap and L. Shklar, Eds., IOS Press, December 2002, pp. 29-49
9. M. Rectenwald, K. Lee, Y. Seo, J.A. Giampapa, and K. Sycara. Proof of Concept System for Automatically Determining Need-to-Know Access Privileges: Installation Notes and User Guide. Technical Report CMU-RI-TR-04-56, Robotics Institute, Carnegie Mellon University, October, 2004. http://www.ri.cmu.edu/pub_files/pub4/rectenwald_michael_2004_3/rectenwald_michael_2004_3.pdf
10. C. Rocha, D. Schwabe, M.P. Arago. 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.
11. 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
12. 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
13. 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

14. 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