An Ontological Approach to the Document Access Problem of Insider Threat

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

[1] LSDIS Lab, Department of Computer Science, University of Georgia, Athens, GA 30602
{boanerg, amit}@cs.uga.edu
{durandal, devp}@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

P. Kantor et al. (Eds.): ISI 2005, LNCS 3495, pp. 486 – 491, 2005.

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

[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/