2004

# Semantic Analysis for Monitoring Insider Threats

Svetlana Symonenko
*Syracuse University*

Elizabeth D. Liddy
*Syracuse University*

Ozgur Yilmazel
*Syracuse University*

Robert Del Zoppo
*Syracuse Research Corporation, Information Technologies Center*

Eric Brown
*Syracuse Research Corporation, Information Technologies Center*

# Semantic Analysis for Monitoring Insider Threats

Svetlana Symonenko[1], Elizabeth D. Liddy[1], Ozgur Yilmazel[1],
Robert Del Zoppo[2], Eric Brown[2], Matt Downey[2]

[1]Center for Natural Language Processing, School of Information Studies,
Syracuse University, Syracuse NY 13444, +1 315.443.5484
{ssymonen,liddy,oyilmaz}@mailbox.syr.edu

[2]Information Technologies Center, Syracuse Research Corporation
6225 Running Ridge Road, North Syracuse NY 13212 +1 315.452.8000
{delzoppo,brown,downey}@syrres.com

## Abstract

Malicious insiders' difficult-to-detect activities pose serious threats to the intelligence community (IC) when these activities go undetected. A novel approach that integrates the results of social network analysis, role-based access monitoring, and semantic analysis of insiders' communications as evidence for evaluation by a risk assessor is being tested on an IC simulation. A semantic analysis, by our proven Natural Language Processing (NLP) system, of the insider's text-based communications produces conceptual representations that are clustered and compared on the expected vs. observed scope. The determined risk level produces an input to a risk analysis algorithm that is merged with outputs from the system's social network analysis and role-based monitoring modules.

**Introduction**

Natural Language Processing (NLP) has proven successful in a range of applications of significance to the intelligence community (IC). Most of these applications support the IC's need for improved representation of, and access to, large amounts of textual information for tasks such as information retrieval, question-answering, cross-language information retrieval, cross-document summarization, and information extraction. In the research we are herein reporting, we adapt our proven NLP capabilities to provide fine-grained content representation and analysis of text-based communications in a novel application – detecting insider threats via semantic analysis of text-based artifacts either produced by or accessed by IC analysts.

Our full insider threat solution integrates evidence from social network analysis and role-based access monitoring of system usage with our semantic analysis of insiders' cyber communications as inputs to a risk analysis algorithm that assesses these inputs and produces an indication of the potential risk of an insider threat within the organization. This research is being conducted as part of ARDA's Information Assurance for the Intelligence Community Program, and therefore, it is being modeled on and tested in a simulated IC malicious insider threat scenario developed by Subject Matter Experts (SMEs) on our project with years of experience with the community. A malicious insider is someone who, while a valid user of IC systems, decides to perform unauthorized malicious acts, including sharing of information with groups unfriendly to the US. The goal of this ARDA program is to develop solutions for efficiently detecting and monitoring such unwanted behaviors. While the IC is the main focus of our development efforts, the banking and securities industries have the same need to recognize potential insider threats and will be able to utilize this model and methodology for distinguishing between normal and abnormal cyber behavior of their employees.

To accomplish our goal, we are developing and testing an Insider Threat Model that integrates Context, Role, and Semantics, here defined as: Context – the tasks the analyst is assigned; Role – the analyst's assigned job functions within that context, and; Semantics – the content of the information produced or accessed by the analyst. Given these inputs, the model will detect levels of insider threat risk by comparing expected cyber behaviors against observed cyber behaviors. This paper reports on the Semantic Analysis module.

**Operational Scenario**

Intelligence analysts operate within a mission-based context, focused mainly on specific topics of interest (TOIs) and geo-political areas of interest (AOIs) that they are assigned. The role the analyst plays dictates the TOI/AOI, organizational relationships, communication patterns, intelligence products and information systems needed, and the intelligence work products created, thereby the need for monitoring Context, Role, and Semantics. The demonstration scenario we will be testing within is based on an organizational network of analysts working in various groups. Our scenario is based on a fictitious government agency with fictitious information targets. However, our SMEs will ensure that the scenario will be representative of the information assurance problem of malicious insider threats in the U.S. Intelligence Community.

**Related work**

To the best of our knowledge, there is no account of the integrated social context, role, and semantics approach that we are taking. While some projects have addressed these dimensions individually, most research appears to be focused on *cyber threat* and *cyber security*. When semantics has been utilized, it is applied to describe the role-based access policy of an organization (RAND, 1999; Upadhyaya et al., 2001). In related work, a research project by Raskin et al. (2002) aims to use a natural language-based ontology to scan texts for indicators of possible intellectual property leakage.

The 2003 NSF / NIJ Symposium on Intelligence and Security Informatics marked an increased interest in the research community in applying linguistic analysis to the problems of cyber security. Stolfo et al. (2003) mined subject lines of email messages for patterns typical for particular user groups (e.g. software developers vs. the legal department). Patman & Thompson (2003) reported on the implementation of a personal name disambiguation module that utilizes knowledge of cultural contexts. Burgoon et al. (2003) looked for linguistic indicators of deception in interview transcripts. Zhou et al. (2003) conducted a longitudinal study of linguistic cues of deception in email messages. Zheng et al. (2003) compared machine-learning algorithms on the task of recognizing the authorship of email messages, and evaluated the efficiency of using different semantic features such as style markers, structural features, and content-specific features. Sreenath et al. (2003) employed latent semantic analysis to reconstruct users' original queries from their online browsing paths and applied this technique to detecting malicious (e.g. terrorist) trends.

The novelty of our approach is in both the problem and the scope. The patterns that we are seeking to detect may not look malicious. In fact, they may look perfectly legitimate but, when considering the user's role and context of their assignment (i.e. topics and geo-political focus), they indicate that the insider's activities are out of range of "expected behavior". Our task is to assess the semantic distance between the content of the documents that the insider is currently accessing and creating and the expected content, given the analyst's assigned TOI and AOI. For this purpose, concept-based semantic analysis will be applied to the wide range of textual documents that analysts use and produce while working on a task, e.g. documents provided by other organizations or from internal collections, email communication, or database or Internet query logs.

## Approach

The insider threat scenario described above presents the following *problem* amenable to the semantic analysis module of our system. Given the set of textual data available electronically and ranging in genre from news articles to analyst reports, official documents, email messages, query logs, and so on, the system should be able to identify the TOI / AOI mentioned in the documents and compare them against the expected TOI and AOI. In other words, the task is to detect an outlier, i.e. a TOI / AOI, which is significantly different from the expected ones.

Our approach is based on a number of assumptions developed in the course of our talks with members of the IC. First, we assume that analysts are assigned relatively long-term tasks and dedicate most of their work time to it.[1] Next, we assume, there may be more than one analyst who is assigned the same main topic and that each would then work on particular subtopics. Finally, we assume that the analysts work with documents and engage in email communication on topics related to their assigned task. We can also expect that the analysts working on subtopics of the same main topic would access different, but topically related, documents. Given the above assumptions, we can expect that clustering documents that the analysts work with would yield a larger cluster(s) containing on-topic documents, and a few smaller clusters of off-topic documents. Further, we can train a clustering model on the dataset containing mainly on-topic documents. The topical description of a cluster will be generated from the *n* most frequent concepts in the clustered documents. Then, we can assess whether the documents accessed or created by the analyst fall within the scope of on-topic cluster(s) or whether they are significantly far from such topical cluster(s).

We will experimentally compare and select from the range of available hierarchical clustering methods[2] the most appropriate one for our task of developing a model of expected TOI/AOI for the documents that the analyst accesses/generates. Then, each new document will be assessed in terms of its semantic distance from the existing cluster(s). As a result, the document will be merged with on-topic cluster(s), or existing off-topic cluster(s), or will start a new off-topic cluster. It is important to note that not every off-topic cluster should raise an alert flag. First, clustering algorithms can generate sporadic clusters. Also, realistically, analysts cannot be expected to work on their assigned topic 100% of their time. Finally, the emergent topic can be a legitimate development in the analyst's work. Therefore, the system will check the semantic distance between the off-topic cluster and the on-topic cluster(s), and also the size of the off-topic cluster. When both parameters exceed thresholds, the semantic analysis module emits an indicator to the insider threat monitoring application. A human (e.g. an information assurance engineer) can then review the indicators for their relevancy. Documents assessed as being on-topic will be added to the model; thus, adjusting the semantics of the expected TOI/AOI and the on-topic cluster parameters.

We intend to boost the efficacy of the clustering methods by use of multiple ontologies, which will enable mapping of individual terms and locations to appropriate categories and, thus, will reduce the high dimensionality of data[3] and, more importantly, contribute to the conceptual coverage of the resulting clusters.

---

[1] This assumption does not cover analysts working on time-critical requests that need to be turned in within a couple of hours. Such analysts are *expected* to change topics quickly. A different TOI / AOI model would be needed for them.

[2] See Ward, 1963; Zhao & Karypis, 2002 for details on methods.

[3] Known to negatively affect computational effectiveness of clustering algorithms (Hotho et al., 2003)

**Resources**

*Data*

One of the challenges of this project is to develop a test collection of questions / topics and related documents for training and testing to adequately represent the spectrum of textual data accessed / generated during the analyst's work processes. Such data collection is bound to be diverse in both, format (such as *txt, html, doc, tabular*) and genre (e.g. formal documents, analytic reports, online news stories, email messages). Being aware of the constraints on data procurement from operational settings, we gathered resources that would best fit the context of the IC. The resulting collection, discussed in greater detail below, is an example of collaboration and sharing among different research teams involved in ARDA and DARPA funded projects.

The analysts' tasks were modeled on scenarios developed by Center for Non-Proliferation Studies (CNS)[4] experts for use in ARDA's AQUAINT (Advanced Question and Answering for Intelligence) Program. We also make use of the scenario-based questions generated at the 2003 ARDA-NRRC workshop on Scenario-Based Question-Answering (Liddy, 2003). A scenario consists of a question (i.e. particular task that the analyst is charged with) and a set of sub-questions, thus, modeling the analyst's decomposition of the main question into a set of contextually related sub-topics and posing them iteratively against the appropriate information resources (Figure 1).

| Main Question/Topic |
| --- |
| *Despite having complete access, to this day UN inspections have been unable to find any biological weapons, or remnants thereof, in Iraq. Why has it proven so difficult to discover hard information about Iraq's biological weapons program and what are the implications of these difficulties for the international biological arms control regime?* |
| **Question Decomposition / Subtopics *(selected from 15)*** |
| 1. What does it take to determine/find signatures of a biological weapons program?<br>2. What are UN capabilities and procedures for inspection?<br>3. Are they looking for the right thing?<br>4. Where are they likely to be?<br>5. Signature of the inspections: how predictable were they? Did they lend themselves to deception?<br>6. What is the Iraqi denial and deception capability? How much effort is involved in hiding it? What evidence is available? |
| **Sources to Answer the Question(s)[5]** |
| • *Arms control agreements*<br>• *UN databases, guidelines, and procedures*<br>• *UNSCOM report*<br>• *CNS data for weapons info*<br>• *Office of Technology Assessment reports*<br>• *Foreign press reports*<br>• *General search*<br>• Talk to inspectors<br>• Geospatial sources |

**Figure 1. Sample AQUAINT scenario**

[4] http://cns.miis.edu/

[5] Italicizing indicates data amenable to semantic analysis

From our conversations with intelligence analysts, we have learned that these scenarios fairly accurately represent actual analysts' tasks.

Another benefit of the AQUAINT scenarios is that they were developed under the premise that much of the needed information can be found in the CNS collection, in particular, in: datasets on nuclear weapons and missile proliferation; country profiles for North Korea and China; NIS Nuclear Profiles; a Nuclear Trafficking Database; the news archive on CBW / WMD. The resources are of various genres: news (including translations); analytic reports by various agencies, and; treaties. Our data set also includes a collection of online news topically related to the CNS data, compiled by the AQUAINT team at SUNY-Albany[6].

*Ontology*

In the semantic analysis approach, rather than using the literal words in texts, we develop algorithms to augment the document terms selected for clustering with appropriate concepts. Given that the focus will be on TOI and AOI, we needed an ontology for the nonproliferation domain, as well as a gazetteer.

Through collaboration with ISI / SAIC / Ontolingua, we obtained access to an ontology of CNS concepts[7], which also includes topics from non-CNS knowledge bases on terrorism. We will adjust this ontology to incorporate our currently employed taxonomy. Figure 2 illustrates the current semantic mapping of the terms *sarin* and *mustard gas* to a type *cweap* (chemical weapon) and its augmentation with CNS topics (*WMD, weapons*).

| |
|---|
| cbw092502<br>*.. the regime has accumulated substantial stockpiles of deadly liquid agents such as* ***mustard gas****, and ominous nerve agents, such as* **sarin** *and* ***VX****, the report said.* |
| *entity = mustard_gas\|NN*<br>*type = cweap*<br>Cat = WMD<br>Top Cat = weapon<br><br>*entity = sarin\|NN*<br>*type = cweap*<br>Cat = WMD<br>Top Cat = weapon |

**Figure 2. Example of term-mapping**

For the conceptual organization of AOI, we will utilize the SPAWAR Gazetteer, also developed under the AQUAINT Program. It combines resources of four publicly available gazetteers (NGA[8]; USGS; CIA World Factbook; DARPA's TIPSTER Program), and is dynamically updated. The gazetteer uses a single comprehensive categorization scheme based on the Alexandria Digital Library thesaurus[9]. When tested on text annotation tasks, it was shown to cover 90% of geographic references in texts.

**Preliminary Example**

To exemplify our methods, consider the following example that we developed in order to familiarize ourselves with the data collection we were assembling. We selected a small set (five) of documents from the North Korea collection compiled by CNS. All documents

---

[6] http://www.hitiqa.albany.edu/index.html
[7] http://ontolingua.stanford.edu.
[8] National Geographic Intelligence Agency; former name is National Imagery and Mapping Agency
[9] www.alexandria.ucsb.edu/~lhill/FeatureTypes

were of a similar genre, namely, chronology of proliferation events. Two documents came from the *Missile* subset, and three documents came from the *Chemical* subset. We ran the documents through CNLP's text processor and analyzed the extracted entities and named entities[10]. The analysis led to a few important observations. First, selecting only entities to represent the conceptual scope of the document reduces it by about 3/4$^{th}$, and further limiting to the named entities cut it to about 1/10$^{th}$ of its original size (Table 1):

| **Tokens** | **Doc92** | **Doc95** | **Doc47_96** | **Doc97_00** | **Doc01_02** |
|---|---|---|---|---|---|
| Words | 5356 | 4102 | 2736 | 1787 | 690 |
| Entities + Named Entities | 1399 | 1136 | 748 | 462 | 181 |
| Named Entities only | 420 | 405 | 252 | 161 | 81 |

**Table 1. Count of document terms**

Second, using a gazetteer to resolve individual location names to their upper level geographic concept appears beneficial for identifying important AOIs. For instance, out of 39 *Russia*-related place names in Doc92, 23 were literally *Russia[n]*. The rest (one third) constituted city names (*Moscow* – 11, *Miass* – 4) and a region name (*Ural*). Another example: of 13 mentions of *South Korea*, 8 (two thirds) referred to *Seoul.* Assuming that locations are almost exclusively proper names, we estimated AOI frequencies against the named entities only. Table 2 shows prevalent AOIs (in %) for the two *Missile* documents.

| **AOI** | **Doc92** | **Doc95** |
|---|---|---|
| North Korea | 29.05 | 19.01 |
| South Korea | 3.1 | 4.44 |
| United States | 4.29 | 4.94 |
| Syria | 6.19 | 0 |
| Iran | 8.57 | 4.94 |
| Russia | 9.29 | .25 |

**Table 2. AOI frequency for *Missile* documents**

Next, we wanted to compare the topicality of *Missile* vs. *Chemical* documents. Table 3 shows TOI frequency across all five documents. Obviously, Doc92 and Doc95 focus on the *Missile* topic, whereas the other three documents mainly discuss *Chemical/Biological Weapons*. Again, the concept-based approach seems promising. For example, out of 174 *Missile*-related terms in Doc92, 131 were literal *missile[s]*. The document also contained 40 mentions of a topically important term, *Scud* (a ballistic missile); including 23 cases where the term was used just as a proper name. Applying the TOI ontology would group these and other[11] terms under the *Missile* concept, thus, increasing its frequency by 24.7%[12].

| **TOI** | **Doc92** | **Doc95** | **Doc47_96** | **Doc97_00** | **Doc01_02** |
|---|---|---|---|---|---|
| Missile | 12.44 | 14.35 | 3.21 | 2.6 | 1.66 |
| Chem/Bio | .07 | .7 | 4.95 | 5.19 | 6.63 |

**Table 3. TOI frequency for *Missile* and *Chemical* documents**

[10] Extracted entities include nouns (*missile*) and noun phrases (*biological warhead*), as well as named entities, which are proper names (*China, Scud*).
[11] Such as: *launcher, gun, nuclear, Nodong* (a proper name for the nuclear missile)
[12] For Doc95, the TOI frequency would be boosted by 31.5%

## Conclusion

This project further extends the idea of combining NLP and machine learning (clustering) techniques to an application in the field of information security. This merging presents a few challenges, as well as potential areas of contribution, to the problem of knowledge acquisition. First, the majority of the prior research focused on a particular genre (news stories, or email messages, or query logs). Our data collection combines various genres, differing in style, syntax, and semantics[13]. We will, therefore, be enhancing our existing NLP tools to deal with genre specifics at the term extraction, term mapping, and term/concept-weighting stages[14]. Next, we will further investigate benefits and issues related to an ontology-driven approach to identifying important topical structures in large and stylistically diverse datasets.

While this is a nascent project, we believe that the application area, the approach, and the model described herein should be of interest to researchers in the area of insider threats as well as other NLP teams dealing with analogous situations.

**Acknowledgements**
This work is supported by the Advanced Research and Development Activity (ARDA).

## References

Allan, J., Lavrenko, V., Malin, D., & Swan, R. 2000. *Detections, Bounds, and Timelines: UMass and TDT-3*. http://citeseer.nj.nec.com/455856.html

Anderson, R. 1999. *Research and Development Initiatives Focused on Preventing, Detecting, and Responding to Insider Misuse of Critical Defense Information Systems: Results of a Three-Day Workshop.* http://www.rand.org/publications/CF/CF151/CF151.pdf

Berkhin, P. 2002. *Survey Of Clustering Data Mining Techniques*. http://citeseer.nj.nec.com/berkhin02survey.html

Burgoon, J., Blair, J., Qin, T., Nunamaker, J., Jr. *Detecting Deception Through Linguistic Analysis.* Presentation at 2003 NSF/NIJ Symposium on Intelligence and Security Informatics.

Hotho, A., Staab, S. & Stumme, G. 2003. *Text clustering based on background knowledge.* University of Karlsruhe, Institute AIFB. http://citeseer.nj.nec.com/hotho03text.html

*Intelligence and Security Informatics: First NSF/NIJ Symposium*, Tucson, AZ, USA, June 2-3, 2003. Proceedings. (Eds.). Chen, H., Miranda, R., Zeng, D.D., Demchak, C., Schroeder, J., & Madhusudan, T. Volume 2665 / 2003. Springer-Verlag Heidelberg.

Liddy, E.D. 2003. *Scenario Based Question-Answer Systems*. AQUAINT 2003 PI Meeting. http://cnlp.org/presentations/present.asp?show=conference

Patman, F. & Thompson, P. 2003. A New Frontier in Text Mining. In *Intelligence and Security Informatics*, p. 27-38.

Raskin, V., Hempelmann, C., Triezenberg, K., & Nirenburg, S. 2001. *Ontology in Information Security: a Useful Theoretical Foundation and Methodological Tool*. Proceedings of the 2001 Workshop on New Security Paradigms, p.53-59.

Sreenath, D.V., Grosky, W.I., & Fotouhi, F. 2003. Emergent Semantics from Users' Browsing Paths. In *Intelligence and Security Informatics*, p. 355-357.

[13] Compare, for example, the style of email communication (informal, abundant in morphologic and syntactic shortcuts) and official briefing reports.
[14] For example, in query logs, every word is assumed to be on topic, which is not true for a news story where most content-indicative terms are located in the lead sentence/paragraph.

Steinbach, M., Karypis, G., & Kumar, V. 2000. *A comparison of document clustering techniques*. http://citeseer.nj.nec.com/steinbach00comparison.html

Stolfo, S., Hershkop, S., Wang, K., Nimeskern, O., & Hu, C. 2003. *Behavior Profiling of Email*. Presentation at 2003 NSF/NIJ Symposium on Intelligence and Security Informatics.

Upadhyaya, S., Chinchani, R., Kwiat. K. 2001. *An Analytical Framework for Reasoning About Intrusions*. 20th IEEE Symposium on Reliable Distributed Systems, p. 99-108.

Ward, J. H., Jr. 1963. *Hierarchical grouping to optimize an objective function*. Journal of the American Statistical Association, vol. 58, p.236-244.

Zhao, Y. & Karypis, G. 2002. E*valuation of Hierarchical Clustering Algorithms for Document Datasets.* http://citeseer.nj.nec.com/zhao02evaluation.html

Zheng, R., Yi, O., Zan, H., & Hsinchun, C. 2003. Authorship Analysis in Cybercrime Investigation*.* In *Intelligence and Security Informatics*, p. 59-73.

Zhou, L., Burgoon, J.K., & Twitchell, D.P. 2003. A Longitudinal Analysis of Language Behavior of Deception in E-mail. In *Intelligence and Security Informatics*, p. 102-110.