Technical Memorandum # 03-009, LSDIS Lab, Computer Science, the University of Georgia, August 15, 2003. Prepared for Special Issue of JOURNAL OF DATABASE MANAGEMENT on Database Technology for Enhancing National Security, Ed. Lina Zhou. (Invited paper, subject to revision).

Semantic Association Identification and Knowledge Discovery for National Security Applications[1]

Amit Sheth[1,2], Boanerges Aleman-Meza[1], I. Budak Arpinar[1], Clemens Bertram[2], Yashodhan Warke[2], Cartic Ramakrishnan[1], Chris Halaschek[1], Kemafor Anyanwu[1], David Avant[2], F. Sena Arpinar[2], Krys Kochut[1]

[1] Large Scale Distributed Information Systems (LSDIS) Lab, Computer Science Department, University of Georgia, Athens, GA 30602-7404

[2] Semagix, Inc. Athens, GA 30601

Abstract. Enterprises have access to vast amount of internal, deep Web and open Web information. Transforming this heterogeneous and distributed information into actionable and insightful information is the key to the emerging new class of business intelligence and national security applications. This paper attempts to bring together novel academic research and commercialized Semantic Web technology to provide these new capabilities. In particular, we discuss academic research on semantic association identification, use of commercial Semantic Web technology for semantic metadata extraction, and a prototypical demonstration of this research and technology through an aviation security application of significance to national security.

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

1 Introduction

Creating applications that allow users to gain insightful and actionable information from vast amounts of heterogeneous information is one of the most exciting new areas of information systems research. This information 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

[1] This work is funded by NSF-ITR-IDM Award # 0219649 titled "Semantic Association Identification and Knowledge Discovery for National Security Applications."

Technical Memorandum # 03-009, LSDIS Lab, Computer Science, the University of Georgia, August 15, 2003. Prepared for Special Issue of JOURNAL OF DATABASE MANAGEMENT on Database Technology for Enhancing National Security, Ed. Lina Zhou. (Invited paper, subject to revision).

are finding themselves in increasing need of such applications. One of the clear manifestations of such a need occurs in aviation safety, which became a critically important issue for national security after the tragic events of September 11. While the current efforts for enhanced physical security measures may help reduce the risk of a similar future event, it is generally accepted that the development of new information-based security systems is a necessary additional capability for defense against such attacks.

Research in search techniques was a critical component of the first generation of the Web, and has gone from academe 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. 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 [21]. Instead of a search engine that merely returns documents containing terms of interest, PISTA can return actionable information to a user or application that gives useful insight into the connection between documents and real-world entities, thus providing better-than-ever support for important decisions and actions.

From a scientific perspective, we face several challenges. One of them is to devise a framework for the formal definition and representation of meaningful and interesting relationships, which we call "semantic associations". 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 the context to select relevant subsets of metadata to process. These challenges call for a fresh look at indexing, query processing, ranking, as well as tractable and scalable graph algorithms that exploit heuristics. Our work addresses these challenges, building on our preliminary results in semantic metadata extraction, practical domain-specific ontology creation, semantic association definition, and main-memory query processing.

In this paper, we examine a prototypical aviation security application[2] called "Passenger Identification, Screening, and Threat Analysis application" (PISTA) and discuss applications of a semantic technology infrastructure that

[2] PISTA is loosely based on Semagix's efforts in applying its Freedom products to homeland security applications [14]

supports information integration and our research in semantic associations that is at the core of our research in content analytics[3] and knowledge discovery. Specifically, we discuss a semantic technology infrastructure for discovering and preventing threats for such applications as aviation safety, and for gaining insights and actionable information. In the process, we define our primary research technique of semantic association identification that we use in building complex ontology-driven information systems. We also discuss how a commercial Semantic Web technology product is used for metadata extraction technology in creating a test bed for PISTA. Next two paragraphs explain the two key parts of this paper.

PISTA extracts relevant metadata from different information resources including government watch-lists, flight databases, and historical passenger data. Using the extracted metadata, PISTA's semantic-based knowledge discovery techniques can identify suspicious patterns and categorize passengers into high-risk groups, low-risk groups, no-risk groups and positive groups (i.e., passengers increasing the safety). The level of physical inspection and optional interrogation of a passenger can be determined at various planned checkpoints accordingly.

PISTA's theoretical fundamentals are semantic associations. 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. The term, "knowledge discovery" is used in this paper to refer to the process of identifying what types of semantic associations are meaningful for the application. Of particular interest to the PISTA type of application are those semantic associations that identify passengers that pose a security risk, and discovering various types of semantic associations, such as a passenger's relationship to a terrorist organization.

With the use of a commercial semantic web technology, Semagix Freedom, we developed a prototype aviation security application. The prototype demonstrates the use of semantic associations in the calculation of possible risk of passengers in a given flight.

This paper is organized as follows. Section 2 presents a formal description of semantic associations of various types. Section 3 describes the creation of PISTA's ontology and how it was populated with a large number of instances. It also shows the PISTA architecture and implementation together with pre-

[3] Text analytics, plus support for semi-structured and unstructured data

liminary results. Semagix Freedom and a national security application based-on semantic Freedom architecture are presented in Section 4. Section 5 summarizes the related work, and Section 6 concludes the paper.

2 Semantic Associations

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. When we consider data on the Web, different entities can be related in multiple ways that cannot be pre-defined. 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. In the semantic Web vision [5], the Resource Description Framework (RDF) data model [18] provides a framework to capture the meaning of an entity (or resource) by specifying how it relates to other entities (or classes of resources). Each of these relationships between entities is what we call a "semantic association". Some examples of semantic association queries in flight security domain include the following:

1. Is the passenger known to be associated with an organization on the watch list?
2. Does the passenger work for an organization that is known to sponsor an organization on a watch-list?
3. Is there a connection between the passenger and one or more passengers on the same flight or different flights?

Most useful semantic associations involve some intermediate entities and associations. 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. In [4], we have presented a formalization of semantic associations over metadata represented in RDF. Concepts are linked together by properties denoted by arcs and labeled with the property name. Different types of semantic associations in an RDF graph are formally defined in the following:

Technical Memorandum # 03-009, LSDIS Lab, Computer Science, the University of Georgia, August 15, 2003. Prepared for Special Issue of JOURNAL OF DATABASE MANAGEMENT on Database Technology for Enhancing National Security, Ed. Lina Zhou. (Invited paper, subject to revision).

Definition 1 (Semantic Connectivity): Two entities e_1 and e_n are *semantically connected* 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.

A sequence of entities and properties represent a *semantic path*.

Definition 2 (Semantic Similarity): Two entities e_1 and f_1 are semantically similar if there exist two semantic paths $e_1, P_1, e_2, P_2, e_3, \ldots e_{n-1}, P_{n-1}, e_n$ and $f_1, Q_1, f_2, Q_2, f_3, \ldots, f_{n-1}, Q_{n-1}, f_n$ semantically connecting e_1 with e_n and f_1 with f_n, respectively, and that for every pair of properties P_i and Q_i, $1 \leq i < n$, either of the following conditions holds: $P_i = Q_i$ or $P_i \subseteq Q_i$ or $Q_i \subseteq P_i$ ($\subseteq$ means rdf:subPropertyOf). We say that the two paths originating at e_1 and f_1, respectively, are semantically similar[4].

Definition 3 (Semantic Association): Two entities e_x and e_y are *semantically associated* if e_x and e_y are *semantically connected* or *semantically similar*.

We use the following operators for expressing queries about *semantic associations*.

Definition 4 (r-Query) A *r-Query*, expressed as $r(x, y)$, where x and y are entities, results in the set of all semantic paths that exist between x and y.

Definition 5 (s-Query) A *s-Query*, expressed as $s(x, y)$, where x and y are entities, results in the set of all pairs of semantically similar paths originating at x and y.

3 PISTA Architecture and Preliminary Results

In PISTA, we have designed an ontology for the national security domain (see **Fig. 1**). This ontology provides a conceptualization of organizations, countries, people, terrorists, 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").

[4] In the future, this restrictive form of semantic similarity definition will be relaxed.

Technical Memorandum # 03-009, LSDIS Lab, Computer Science, the University of Georgia, August 15, 2003. Prepared for Special Issue of JOURNAL OF DATABASE MANAGEMENT on Database Technology for Enhancing National Security, Ed. Lina Zhou. (Invited paper, subject to revision).

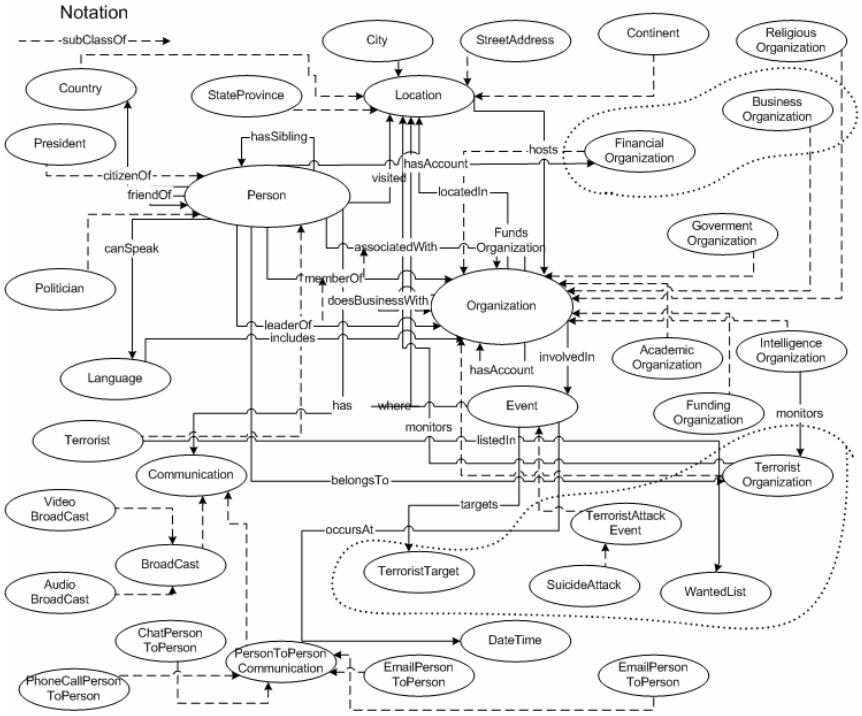


Fig. 1. Sample Ontology

The sources, from which metadata were extracted, were selected for their information richness and aptitude to quickly populate the ontology with a large number of entities and relationships related to terrorism. The metadata is represented in RDF, and semantic association queries are performed based on this RDF. For information extraction we have used Semagix's SCORE-based Freedom suite, which includes a set of tools for extraction of entities from (semi)-structured sources [20]. Freedom can programmatically extract entities found on Web pages and can establish relationships between them. This extraction is based on the ontology thereby placing an extracted entity into its appropriate place in a hierarchy of classes. Currently there are over 6,000 entities and more than 11,000 explicit relations among them. A test bed with an order of magnitude larger dataset is planned over the next year.

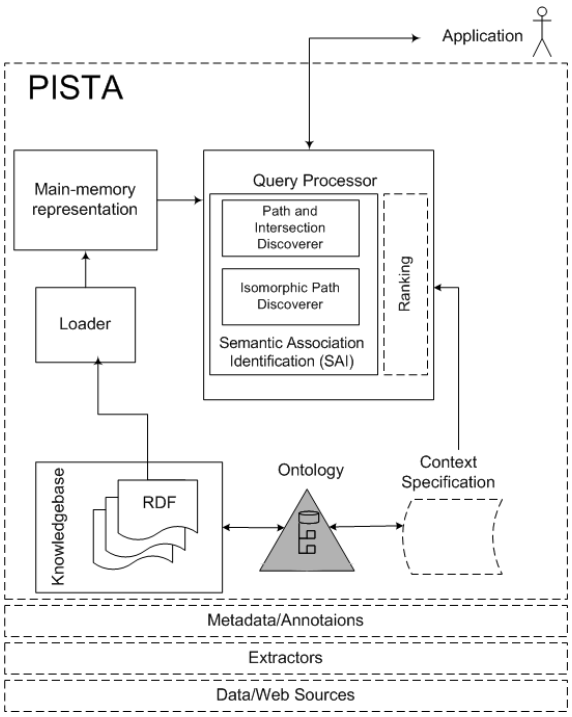


Fig. 2. PISTA Architecture

Figure 2 shows the components of PISTA. Data sources are extracted by using Semagix's SCORE Knowledge Agents [22]. Entities and relationships from trusted sources compose the knowledge base. We are able to create a large test bed due to the advantage of automatic extraction of entities and relationships for populating the ontology. By using a provided API, we convert the knowledge base and ontology from SCORE to RDF and RDF Schema (RDFS) [6], respectively. The Query Processor module interacts with a main-memory representation of the RDF data and RDFS as directed graphs based on the JENA [14] model. The Ranking module processes the results of the query processor. Context-aware ranking is guided by context preferences specified by the user.

Heuristic based search

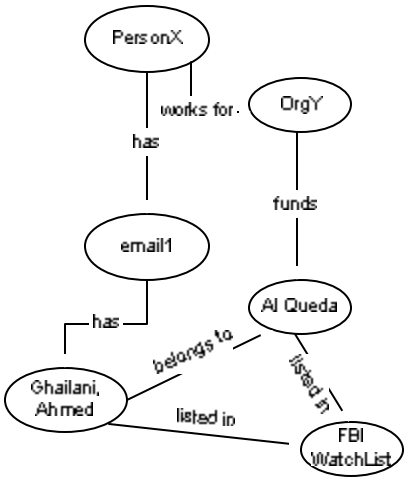
We have implemented simple search algorithms for three operators, namely ?-path, ?-intersect and σ-iso which are explained below. *?-path* and *?-intersect*

are used discover semantic connections (Definition 1) and σ-*iso* is used for finding semantic similarities (Definition 2).

?-path

The naive algorithm to find all paths between two nodes in a directed graph is a recursive implementation of a depth-first search. Our first implementation of the ?-path operator is based on [21].

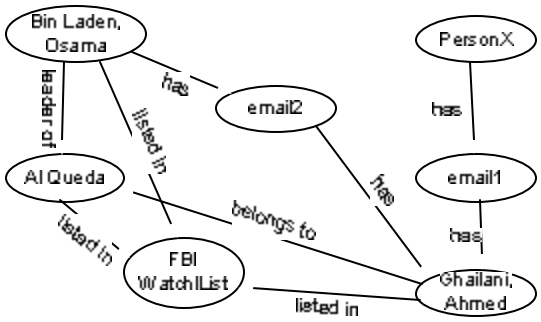
The test suite has at least one RDF schema and RDF data based on this schema. The basic idea in reducing the complexity of the above algorithm is to use the information from the schema level to prune the search at the data level. The nodes at the schema level are far fewer than those at the data level. Hence a search running at the schema level will take less time than a search at the data level. When looking for all paths between entities e_1 and e_2 in the graph representing the RDF data, G_{data}, we check if the classes c_1 to which e_1 belongs and c_2 to which e_2 belongs have a path between them in the schema graph G_{schema}. If there is such a path then we find all such paths first. These schema path expressions will then be used to prune the list of successor nodes for every state in the search through G_{data}. Based on this concept, we have implemented the ?-path operator. Figure 3 shown below shows the visualization of such a set of paths that our algorithm finds with respect to the instance data.



PersonX → email1 → Ghailani, Ahmed → Al Queda → FBI WatchList
PersonX → email1 → Ghailani, Ahmed → FBI WatchList
PersonX → OrgY → Al Queda → FBI WatchList

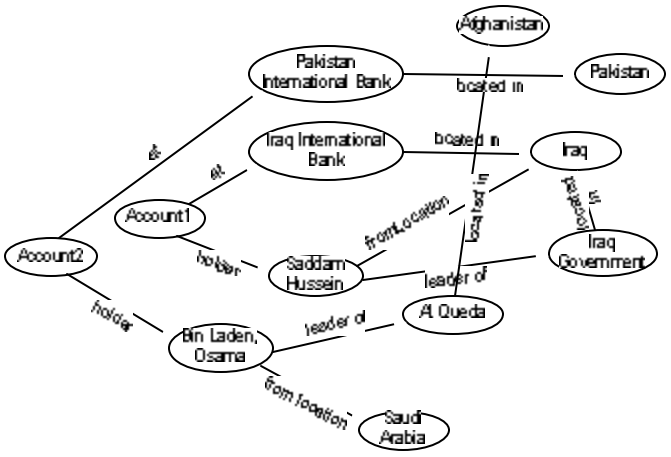
r -Intersect

Our initial implementation of the ρ-Intersect operator is based on the ρ-path operator. It searches for nodes where two ρ-paths intersect (see Figure 4).



specialization of "memberOf", that is, a leader of an organization is as well a member of the organization. The resources connecting properties may also be "similar". If two resources are of the same class they are considered similar. However, we also consider two resources similar if they belong to different classes as long as the classes to which they belong share a common parent. From a *hierarchy* perspective this means that two resources that are "siblings" are similar in our implementation. The third and last similarity is a situation where a resource belongs to a subclass of the other resource. That is, one of the resources is ancestor of the other in the hierarchy

An example of a σ-Iso path relating two persons is two persons that received training, where one took firearms training courses and the other took flight courses. This could be considered a possible threat. σ-Iso, however, discovers paths that may span several relations and resources. Thus, σ-Iso finds that two persons are related to a terrorist organization through a series of associations that span *similar* relations and resources.



Technical Memorandum # 03-009, LSDIS Lab, Computer Science, the University of Georgia, August 15, 2003. Prepared for Special Issue of JOURNAL OF DATABASE MANAGEMENT on Database Technology for Enhancing National Security, Ed. Lina Zhou. (Invited paper, subject to revision).

Context-aware Ranking

A typical semantic query can result in many paths that semantically link the entities of interest. It is likely that many of these paths would be regarded as irrelevant with respect to the user's domain of interest. Thus, the semantic associations need to be filtered according to their perceived relevance. A customizable criterion needs to be imposed upon the paths representing semantic associations to focus only on relevant associations. Additionally, the user should be presented with a ranked list of resulting paths to enable a more efficient analysis. The issues of filtering and ranking raise some interesting and challenging scientific problems.

To determine the relevance of semantic associations it is necessary to capture the context within which they are going to be interpreted (or the domains of the user interest). For example, consider a sub-graph of an RDF graph representing two biology scientists who belong to the same university and were both involved in the same biological weapon development program. If the user is interested in the terrorism domain, the semantic associations involving university-related information can be regarded as less relevant. This can be accomplished by enabling the user to browse the ontology and mark a region (sub-graph) of classes and/or properties of interest. If the discovery process finds some associations passing through these regions then they are considered relevant, while other associations are ranked lower or are discarded. More formally, *ontological regions* can represent context.

Ranking of semantic associations effectively requires more than using the "ontological context" for relevance determination. The ranking process needs to take into consideration a number of criteria which can distinguish among associations which are perceived as more and less meaningful, more and less distant, more and less trusted, etc. Furthermore, different weights can be given to different parameters according to users' preferences (e.g., trust could be given more weight than others, which means a more trustworthy semantic association can be ranked higher than another one which falls perfectly within the ontological context). This is a new and different problem than ranking documents using traditional search engines where documents are usually ranked according to the number of (sometimes subject-specific) references to them (e.g. Google [7], Teoma [23]).

The formalisms of the ranking formulas are presented in [2], as well as details about the context specification. The context specification is the means to specify (with degrees of flexibility) which properties and entities are to be

considered relevant for ranking in a given user's context. A ranking mechanism that considers the user-assigned weights to particular regions of ontology allows for context-aware ranking.

4. 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 6 below shows the domain-model driven architecture of Semagix Freedom.

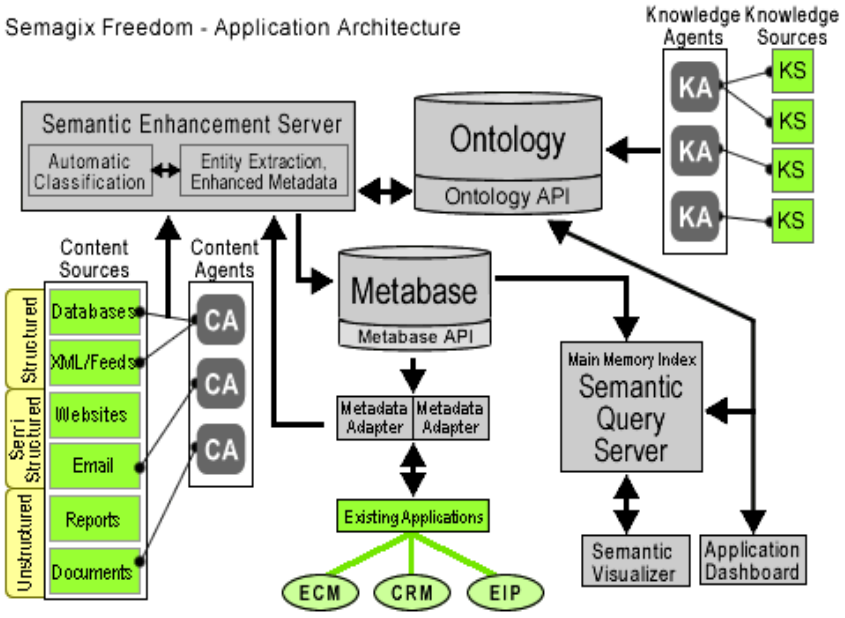


Fig. 6. Semagix Freedom Architecture

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). The ontology forms the basis of semantic processing, including automated categorization, conceptualization, cataloging and enhancement of content. Freedom enhances a content item by organizing it into a structured format and associating it with instances within the ontology. 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. Custom meta-tags, driven by business requirements, can be defined at a schema level. 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 [10].

The Semantic Enhancement Server tools classify aggregated content into the appropriate topic/category (if not already pre–classified) and subsequently perform entity identification, extraction and content enhancement with semantic metadata from the ontology. Semantic associations in the ontology are leveraged to derive tag values if such metadata is not explicitly mentioned in the content. The Semantic Enhancement Server can further identify relevant

document features such as currencies, dates, etc., perform entity disambiguation, and produce XML tagged output.

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. Page: 14 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 assets 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 contextually relevant content. A variety of semantic applications that exploit the SCORE technology can be built including Anti Money Laundering identification and risk assessment, 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. Freedom facilitates in-context querying through semantic metadata associated with individual content items and associations between semantic metadata. Freedom supports querying of the ontology as well as the Metabase.

Homeland Security Application based on Semantic Associations

Semantic associations have proven to be the foundational layer in real world applications, most usefully in the area of homeland security. We present here a Semagix application implemented for a government organization that is related to Passenger Security and Threat Assessment, and is fully based on the underlying concept of exploiting semantic associations between real-world entities.

The primary aim of the government organization is to provide a robust solution to aviation security by addressing the following types of requirements:

- Analysis of government watch lists containing publicly declared “bad” persons and organizations
- Security applications for the sequence of kiosks at the airport departure location

- Aggregation and intelligent analysis/inference of valuable information from multiple sources to provide valuable and actionable insight into identifying high-risk passengers
- Scalable and near real-time system that can co-relate multiple pieces of information to detect the overall risk factor for the flight before departure

The main idea behind the strategy of the application is to automatically attach a threat score to every passenger that boards any flight from any national airport, so that flights and airports could be assigned corresponding threat levels. This threat is based extensively on semantic associations of passenger entities with other entities in the ontology like terrorist organizations, watch lists, travel agents, etc. The following semantic associations are considered in the generation of a passenger's threat score:

- appearance of the passenger on any government-released watch-list of bad persons or bad organizations
- relationship of the passenger to anyone on any government-released watch-list of bad persons or bad organizations
- deviation from normal methods of ticketing, flight scheduling, use of a travel agent in reservation of tickets
- origin of the passenger and his flight
- appearance of the passenger's name with that of a known bad person in any public content, etc.

Based on the threat score for each passenger, the passenger will either be either allowed to proceed from one checkpoint to another in a normal manner, or would be flagged for further interrogation along concrete directions as indicated by the semantic associations in the application.

Figure 7 below shows a 'passenger profile' screen that provided a 360-degree view of information related to a passenger.

Fig. 7. Actionable information related to passenger profile

Section 1 in Figure 7 above presents a listing of the semantic associations of the passenger to numerous other entities in the ontology. More precisely, it provides a passenger (entity)-centric view of the ontology, thus unearthing a number of semantic associations, both direct and indirect (hidden), as shown in Figure 8 below. 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.

Fig. 8. Semantic Associations and Semantically Relevant Content

Section 2 in the Figure 8 above presents a listing of all the content that is contextually relevant to the passenger, but not necessarily mentioning the name of the passenger. Once again, this approach exploited semantic associations in the ontology in order to decide relevance of content. All content stored in Metabase is enhanced with the use of Semantic Enhancement Engine resulting in semantic relationships to the entities in the ontology. A piece of content was perceived as relevant to a passenger even if it was about an entity that was associated closely with the passenger name in the ontology.

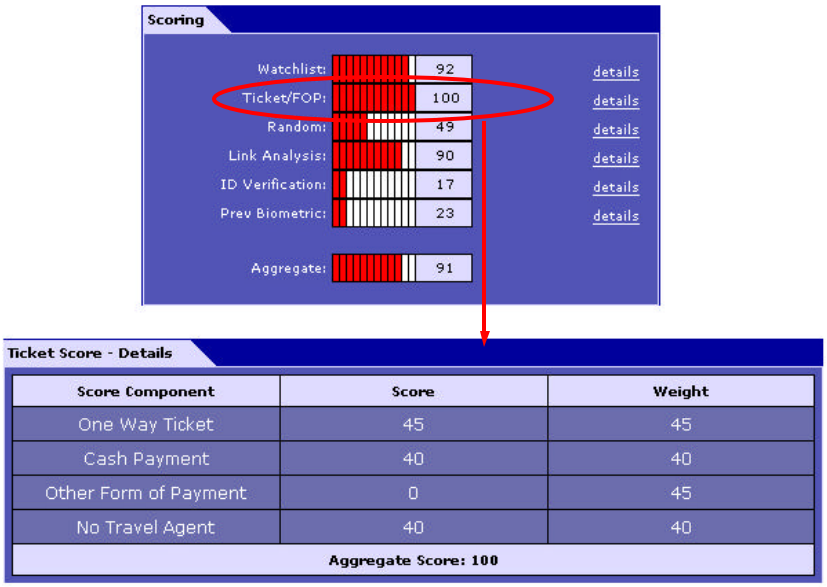


Fig. 9. Scoring

Section 3 in the Figure 9 above presents the comprehensive scoring mechanism for arriving at the overall threat score of each passenger. The score was comprised of a number of components like Link Analysis, Watch-list Analysis, Ticket/Form of Payment Analysis, etc. Each of these components was based on deductions from specific semantic association ?-paths between the passenger entity and a number of interesting entities such as terrorists, watch-lists, terrorist organizations, etc. For example, watch list analysis for a passenger indicated that "the passenger worked for an organization, which appeared on a publicly declared watch-list"; and that proved to be reason enough to assign a high-threat value to the watch list analysis component of the passenger (even though the passenger himself may not be directly associated with a watch-list). The link analysis score for a passenger is calculated by examining the relevant content (from Metabase) for that passenger. For example, if the passenger's name is mentioned in a document which is about a terrorist organization or if the passenger is closely related to another person mentioned in such a document, the link analysis results in a higher score. Finally, aggregate score for a passenger is the weighted sum of all previous scores watch list analysis score having the highest weight of all.

The ?-Intersect and σ-Iso operators further enhance the functionality provided in this application, by identifying possible links between two passengers, who may have both met a known terrorist, let's say, around the same time; or who may have similar association patterns in their links to two different terrorist organizations. The application provided an ability to visually detect the seating proximity of such high-threat score passengers, and if necessary dynamically decide to recommend the assignment of an air marshal to the flight.

5. Related Work

In [1] a tool that identifies communities of practice is described. This is done by analyzing ontologies of different domains. The application, Ontocopi, discovers and clusters related instances by following paths not explicit between them. Their work differs from ours in the dataset size. We aim at large scale algorithms that take advantage of the large metadata extracted from data sources. A crucial differing aspect is that Ontocopi's algorithms have a *link threshold* that limits the depth and length of paths that can be discovered. Our approach does not consider a limit in the length of the semantic associations. Long paths may be more significant in the domain where there may be deliberate attempts to hide relationships; for example, potential terrorist cells remain distant and avoid direct contact with one another in order to defer possible detection [13] or money laundering [3] involves deliberate innocuous looking transactions.

In their approach to context, which they call *selection mode*, the automatic selection mode considers instances with many connecting relations as important whereas a manual mode allows selection of relations and/or instances to be considered relevant when discovering paths. Though they also studied a semi-automatic approach, we believe that direction does not benefit discovery of semantic associations in national security applications. Their semi-automatic approach considers entities with many links (e.g. *country*) as not important because of many entities have relations to it (e.g. *locatedIn, basedIn, visitedPlace*). This would give more preference to finding paths that include a popular entity as compared to a possible semantic relation of interest involving an entity with only two relations connecting a potential path of interest (e.g. *supportsOrganization, transfersFundsToOrganization*).

In [24] the problem of finding relevant information is approached by following the intuition of social networks. Agents search data, based on referral

graphs which get updated according to answers received as well as the discovered connections to other agents that they are referred to. Their approach to search the network efficiently differs with our approach mainly because we try to get multiple paths connecting entities of interest whereas their approach aims at locating relevant information.

Our work differs from traditional data mining [8][9] because instead of focusing on discovering patterns and relationships out of their repetition in the data, we approach discovery as goal driven and we do not intend to develop models of the data, we provide search techniques that find out whether the associations exist in the data.

Finally, aviation security applications such as PISTA have been discussed at a non-technical level in [16], [17] as well as in a white paper by Semagix, Inc [15]. This paper discusses the technical aspects of developing such an application in much more detail.

6. Conclusions

This paper discussed a challenging problem of finding new insights and actionable information from large amounts of heterogeneous content. We particularly discuss the technical challenges in developing a prototypical aviation security application, but similar requirements and challenges exist in business intelligence as well as national and homeland security applications involving large scale text and content analytics. This paper makes 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 scalable test bed over open source data. We also hope it demonstrates an example of collaboration involving academic research, industry technology, and government priorities, to address unique and technically demanding challenges.

For future work, we plan on using the reification approach of RDF to include provenance information. We found that the similarity measure for s-Iso might be too restrictive and relaxed similarity versions of it are being now considered.

Though we did not use DAML+OIL [12] to represent extracted data, we have been following the development of OWL [11]. If it is beneficial then it will replace RDF in our application. We plan to incorporate a context-driven

approach to the search to increase performance in the discovery of semantic associations.

Acknowledgements: We thank Semagix, Inc. for providing its Freedom product, which is based on the SCORE technology and related research performed at the LSDIS Lab [20].

References

[1] Harith Alani, Srinandan Dasmahapatra, Kieron O'Hara, and Nigel Shadbolt. Identifying Communities of Practice through Ontology Network Analysis. IEEE Intelligent Systems, March/April 2003, pp. 18-25.

[2] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar, and Amit Sheth. Context-Aware Semantic Association Ranking, Int. Conf. on Semantic Web and Databases, 2003, Berlin Germany, to appear.

[3] Anti Money Laundering, Application White Paper, Semagix, Inc. http://www.semagix.com/pdf/anti_money_laundering.pdf

[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)

[5] T Berners-Lee, J.Hendler, and O.Lassila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities", Scientific American, May 2001

[6] D. Brickley and R.V. Guha. Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation. 2000.

[7] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proc. 7th International World Wide Web Conference (1998)

[8] M. Chen, J.Han and P. Yu. Data Mining: An Overview from the Database Perspective. IEEE Trans. On Knowledge and Data Engineering. Vol. 8. No. 6. December 1996.

[9] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R Uthurusany. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press 1996.

[10] 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, pp. 29-49, December 2002.

[11] F. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. Stein. OWL Web Ontology Language 1.0 Reference, W3C Working Draft, 31 March 2003.

Technical Memorandum # 03-009, LSDIS Lab, Computer Science, the University of Georgia, August 15, 2003. Prepared for Special Issue of JOURNAL OF DATABASE MANAGEMENT on Database Technology for Enhancing National Security, Ed. Lina Zhou. (Invited paper, subject to revision).

[12] F. Harmelen, P. F. Patel-Schneider, I. Horrocks, eds. Reference Description of the DAML+OIL (March 2001) ontology markup language.
[13] V. Krebs, "Mapping Networks of Terrorist Cells" Connections 24(3): 31-34, 2001.
[14] Brian McBride. Jena: A Semantic Web Toolkit. IEEE Internet Computing, November/December 2002, pp. 55-59.
[15] National Security and Intelligence, Semagix White Paper, 2003. http://www.semagix.com/pdf/national_security.pdf
[16] R. O'Harrow Jr., Intricate Screening Of Fliers In Works, WashingtonPost.com, http://www.washingtonpost.com/wp-dyn/articles/A5185-2002Jan31.html, 2002.
[17]B. Online, The Price of Protecting the Airways, http://www.businessweek.com/technology/content/dec2001/tc2001124_0865.htm, Dec. 4, 2001.

[18] Ora Lassila & Ralph R. Swick: "Resource Description Framework (RDF) Model and Syntax Specification", W3C Recommendation, World Wide Web Consortium, Cambridge (MA), February 1999
[19] http://www.python.org/doc/essays/graphs.html
[20] Semagix. http://www.semagix.com.
[21] A. Sheth, I. B. Arpinar, and V. Kashyap, "Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships," Enhanceing the Power of the Internet Studies in Fuzziness and Soft Computing, M. Nikravesh, B. Azvin, R. Yager and L. Zadeh, Springer-Verlag, 2003 (in print).
[22] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, Y. Warke, Semantic Content Management for Enterprises and the Web, IEEE Internet Computing, July/August 2002, pp. 80-87.
[23] Teoma: http://sp.teoma.com/docs/teoma/about/searchwithauthority.html
[24] Bin Yu, Munindar P. Singh. Searching Social Networks. Proceedings of Second International Joint Conference on Autonomous Agents and Multi-Agent Systems, 2003, to appear.