Semantic Association Identification and Knowledge Discovery for National Security Applications

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

{amit, boanerg, budak, ch, cartic, anyanwu, kochut }@cs.uga.edu,
{clemens.bertram, yash.warke, david.avant, sena.arpinar}@semagix.com

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

[2] Semagix, Inc.
297 Prince Ave, Suite 11
Athens, GA 30601

Abstract. Public and private organizations 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. Although role of semantics in search and integration has been often talked about, in this paper we discussed semantic approaches to support analytics on vast amount of heterogeneous data. In particular, we bring together novel academic research and commercialized Semantic Web technology. The academic research related to semantic association identification, is built upon commercial Semantic Web technology for semantic metadata extraction. A prototypical demonstration of this research and technology is presented in the context of an aviation security application of significance to national security.

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

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 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 academia 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 (Dill et al., 2003; Shah, Finin, Joshi, Cost, & Mayfield, 2002; Hammond, Sheth & Kochut, 2002). Relationships are at the heart of semantics (Woods, 1975; Sheth, Arpinar, & Kashyap, 2003).

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. Instead of a search engine that merely returns documents containing terms of interest, we propose an approach that supports semantic analytics of heterogeneous content to return actionable information that gives useful insight into the connection between documents and real-world entities, thus providing better-than-ever support for important decisions and actions. This approach is demonstrated using a prototypical aviation security application[1] called "Passenger Identification, Screening, and Threat Analysis application" (PISTA) that involves discovering and preventing threats for aviation safety. This is one of many semantic applications as part of advanced information technology necessary to support homeland security.

From the research perspective, one of the challenges was to devise a framework for the formal definition and representation of meaningful and interesting relationships, which we call "semantic associations". Semantic associations are at the core of our research in content analytics[2] and knowledge discovery using an ontology-driven process. 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 a notion of 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 previous research in semantic metadata extraction, practical domain-specific ontology creation, semantic association definition, and main-memory query processing. We also discuss how a commercial Semantic Web technology product is used for metadata extraction technology in creating a test bed for PISTA. The 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 an application like that of PISTA 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 direct or indirect relationship to a terrorist organization. With the use of a commercial Semantic Web technology, Semagix Freedom based on SCORE technology (Sheth et al., 2002), 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 preliminary 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.

---

[1] PISTA is loosely based on Semagix's efforts in applying its Freedom products to homeland security applications ("National Security and Intelligence", 2003)

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

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 (Berners-Lee, Hendler, & Lassila, 2001), the Resource Description Framework (RDF) data model (Lassila & Swick, 1999) is introduced as 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 (interconnected) relationships between entities are examples of "semantic association". Some examples 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? Is such connection in the context of aviation safety?

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 the RDF model, concepts of entities are linked together with relations (properties). The classes and/or relationships can be defined with an RDF Schema vocabulary (Brickley & Guha, 2000). The properties are denoted by arcs and labeled with the relation name. Thus, the metadata can be represented as a graph together with a graph for the vocabulary of the classes and relationships (Karvounarakis, Alexaki, Christophides, Plexousakis, & Scholl, 2002). Different semantic associations in an RDF graph have been formally defined (Anyanwu & Sheth, 2003). We build upon such semantic associations theory for the implementing the r and s operators:

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 represents 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[3]. An example of '`rdf:subPropertyOf`' is given later in Section 3.

Definition 3 (Semantic Association): Two entities e_x and e_y are *semantically associated* if e_x and e_y are either *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] In the future, this restrictive form of semantic similarity definition will be relaxed.

3 PISTA Architecture and Preliminary Results

In PISTA, we have designed an ontology which covers some of the aviation security aspects of the national security domain (see Figure 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"). The ontology is populated with a set of tools for knowledge extraction that instantiate different parts of the ontology from trusted knowledge sources with (semi-)structured data. The sources extracted to populate the ontology, were selected for their information richness and aptitude to quickly populate the ontology with a large number of entities and (more importantly) relationships related to terrorism.

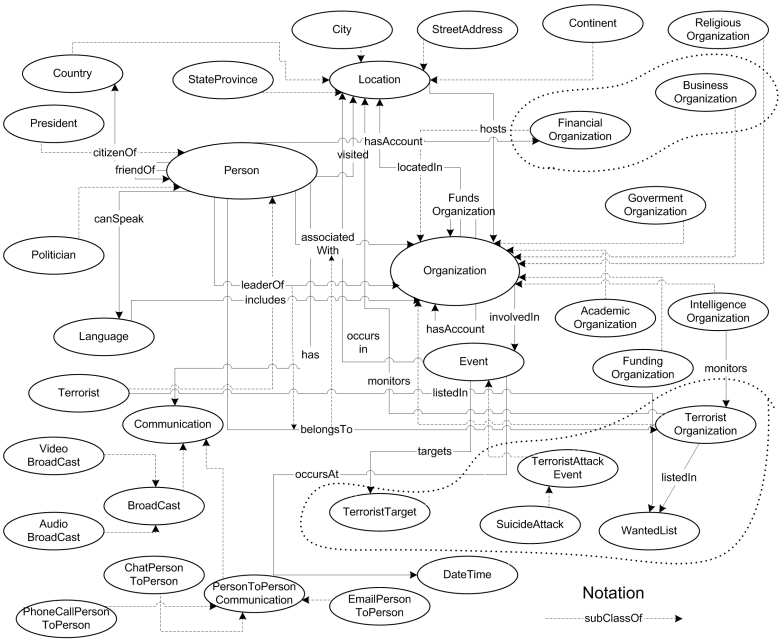

Fig. 1. (Subset of) PISTA Ontology

The populated ontology can further be extended with semantic metadata by extracting information from unstructured, semi-structured or structured content sources. This metadata extraction is based on the ontology thereby placing an extracted entity into its appropriate place in a hierarchy of classes. The extended populated ontology is represented in RDF, and semantic association computations performed based on it.

Semi-automatic creation of metadata based on specific domain has been researched in the S-CREAM framework (Handschuh, Staab, & Studer, 2003), and other tools have been developed (Vargas-Vera et al., 2002). In PISTA, by utilizing Semagix Freedom, we had completed the initial testbed with over 100,000 entities. We are able to create a large testbed due to the advantage of automatic extraction of entities and relationships for populating the ontology. A testbed with an order of magnitude larger dataset has also been developed since the time we implemented the PISTA system reported in this paper (http://lsdis.cs.uga.edu/proj/SemDis/). In this context, this paper presents the results obtained before the larger testbed was completed.

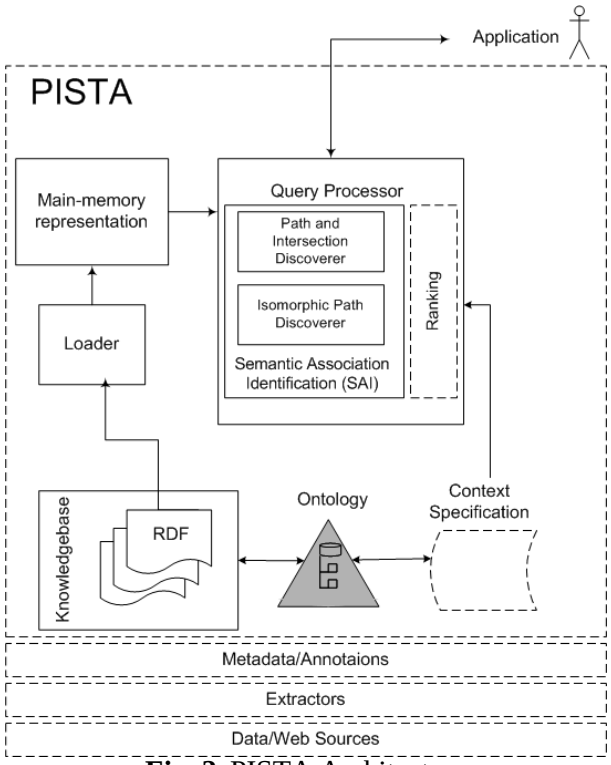

Fig. 2. PISTA Architecture

Figure 2 shows the components of PISTA. Data sources are extracted by using Semagix Freedom's Knowledge Agents. Entities and relationships from trusted sources compose the knowledge base. By using a provided API of the Semagix Freedom toolkit, we convert the knowledge base to RDF and the ontology definition to RDF Schema. The Query Processor module interacts with a main-memory RDF representation of the populated ontology as directed graphs based on the JENA model (McBride, 2002). 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 to 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 (*Python Patterns - Implementing Graphs*, 2003). The foundations for our first implementation of the ?-path operator are based on our previous research on semantic associations (Anyanwu & Sheth, 2003). 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}. Figure 3 shows the visualization of such a set of paths that our algorithm finds with respect to the instance data.

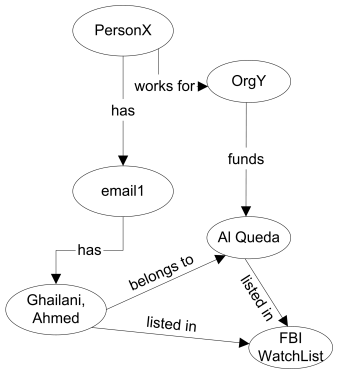

Fig. 3. ?-path results between Person-X and FBI-WatchList

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

ρ-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).

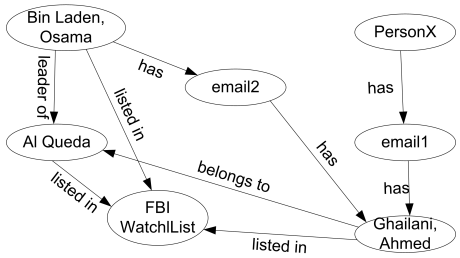

Fig. 4. ?-Intersect originating at *Bin Laden, Osama* and *PersonX*

Bin Laden, Osama → Al Queda
Person X → email1 → Ghailani, Ahmed → Al Queda
Bin Laden, Osama → FBI Watch List
Person X → email1 → Ghailani, Ahmed → FBI Watch List
Bin Laden, Osama → email2 → Ghalani, Ahmed
Person X → email1 → Ghailani, Ahmed

σ-Iso

The goal of σ-Iso is to take two entities as input and discover all paths that are "isomorphic" in both entities (see Figure 5). A path discovered by σ-Iso is a path that is similar in both entities. By similar we mean that a labeled edge (representing a property) exists in both nodes. For example, the property "participatesIn" appears in two "Person" entities. However, there is more flexibility in how two properties are similar. In the RDF model, it is possible to define a *hierarchy* of properties. An example is a property "belongsTo" with a sub-property (`rdf:subPropertyOf`) "leaderOf". The property "leaderOf" would then be a specialization of "belongsTo", that is, a leader of an organization is as well a member of the organization. The entities connecting properties may also be "similar". If two entities belong to the same class they are considered similar. However, we also consider two entities 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 entities that are "siblings" are considered to be similar in our implementation. The third and last similarity is a situation where an entity belongs to a subclass of a class to which the other entity belongs. Figure 5 is an example of semantic similarity results between two entities.

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 entities. Thus, σ-Iso finds that two persons are related to a terrorist organization through a series of associations that span *similar* relations and entities.

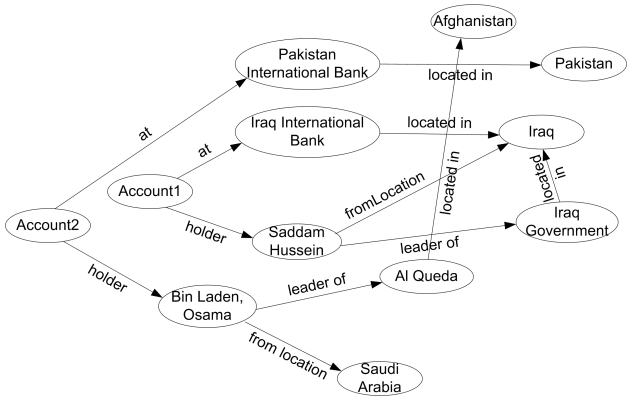

Fig. 5. σ-Iso between *Account1* and *Account2*

Account1 → IraqInternationalBank → Iraq Account2 → PakistanInternationalBank → Pakistan
Account1 → SaddamHussein → Iraq Account2 → OsamaBinLaden → SaudiArabia
Account1 → SaddamHussein → IraqGovernment → Iraq Account2 → OsamaBinLaden → AlQeada → Afghanistan

Path Expression approach

The path expression approach is based on a simplification of the Single Source Path Expression problem. This problem has been represented as a system of linear equations (Tarjan, 1981b) and an algorithm that uses Gaussian-Jordan Elimination to solve it has been proposed (Tarjan, 1981a).

The solution to these equations yields regular expressions that represent all possible paths between any two nodes in the graph. The way that the ? operator has been envisioned requires that the operator work on an undirected graph whereas a previously proposed algorithm assumes a directed graph (Tarjan, 1981a). Hence it needs to be adapted to work for undirected graphs.

There is a way of circumventing this problem. Given the start and the end nodes one could run the ? operator in both directions and take the union of the resulting path expressions. There is however a substantial computational cost attached to this. However the main computational cost is going to be associated with the Gaussian-Jordan Elimination step. Time complexity of this is known to be $T(n^3)$ where n is the number of edges in the graph. This can be prohibitively large considering the size of the graph that we have envisioned in practical usage scenarios. Hence the use of context described next.

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 compared to the participation in a biological weapon development program. By defining regions (or sub-graphs) of the RDF Schema (RDFS) we can capture the areas of interest of the user. Particularly important for us is the ability to define that the path of interest (semantic association) should include properties and/or classes of interest for the user. We provide a preliminary context definition below.

Context Definition

To begin, we define a *region* of interest as a subset of classes (entities) and properties of a schema. The detail to which a region of interest can be specified may vary for different applications. We have considered the following cases: class level and property level. Within the Class level, we may restrict or allow subclasses, as well as super-classes, to be considered relevant. For example, an "*Organization*" class may be considered relevant together with subclasses "*PoliticalOrganization*", "*FinancialOrganization*" and "*TerroristOrganization*", but a class "*Account*" that is parent of the class "*CorporateAccount*" may not be of importance. At a Property level, we can specify restrictions similar to those of the Class level. An interesting and powerful context restriction that can be specified in properties is indication of which classes the property can be applied to ("domain" in RDFS) as well as which classes a property points to ("range" in RDFS). An example is a property "*involvedIn*" with a domain "*Organization*" and range "*Event*" (that is, *Organization* → *involvedIn* → *Event*). Our context specification allows restriction of the type of classes for domain and/or range. For example, it is possible to indicate that the property "*involvedIn*" is relevant when the entity that it is applied to is of class "*TerroristOrganization*".

A user can define several ontological regions with different weights to specify the association types s/he is interested in. Hence, we define a *context* as a set of user defined *regions* of interest. The representation of context in RDF itself is an interesting approach for which we plan to use RVL, a recently proposed RDF view language (Magkanaraki, Tannen, Christophides, & Plexousakis, 2003).

When paths are to be ranked, the entity types contained in them can be inspected in a linear fashion. As each entity class type is traversed, a lookup can be performed to see if the type falls in the earlier defined context. If so, the corresponding context weights are used to assign an overall context weight to a path. Thus, if the discovery process finds some associations passing through highly weighted contextual regions then they are considered relevant, while other associations are ranked lower or are discarded.

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. 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. Teoma[4] and Google (Brin & Page, 1998).

We have defined preliminary ranking criteria for semantic associations (Aleman-Meza, Halaschek, Arpinar, & Sheth, 2003). The "ranking" module within PISTA Architecture (Figure 2) has implemented initial ranking ideas. The ranking module and our implementations of the semantic association operators use several indices for efficient access for querying the ontology schema and instances. RDF/XML serialization of ontology and instances was merely a means to represent the data. The indices support keyword to entity access, entity to class access, and querying for the hierarchy and relationships of the PISTA ontology. We have build upon other tools for accessing data and ontologies that are provided by the Semagix Freedom toolkit, which is introduced in the next section.

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.

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 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.

[4] Teoma: http://sp.teoma.com/docs/teoma/about/searchwithauthority.html

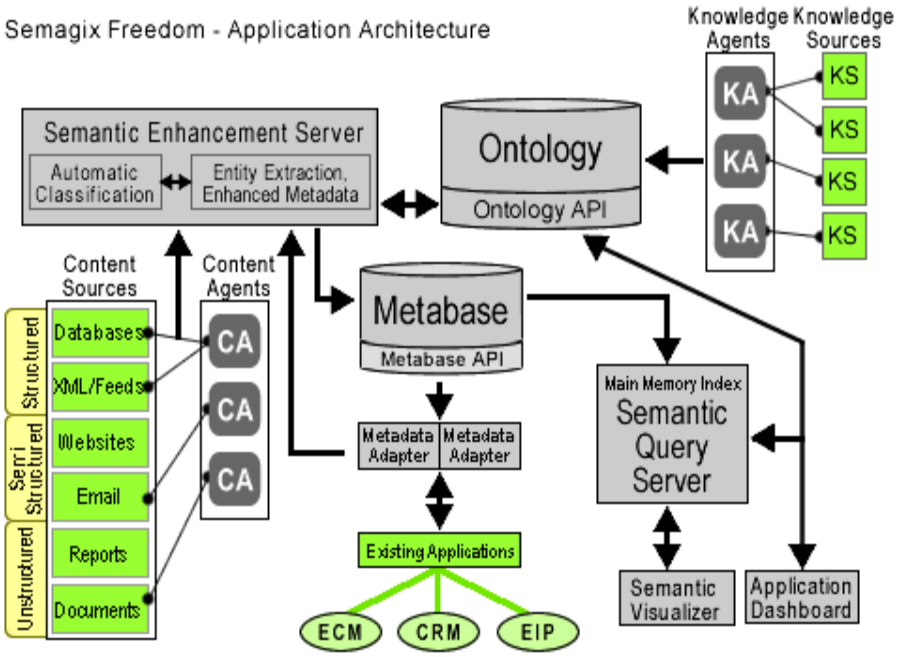

Fig. 6. Semagix Freedom Architecture

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. 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 (Hammond, Sheth, & Kochut, 2002).

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. 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 entities 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 relevant content. A variety of semantic applications that exploit this 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.

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.

One of the key homeland security objectives 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.

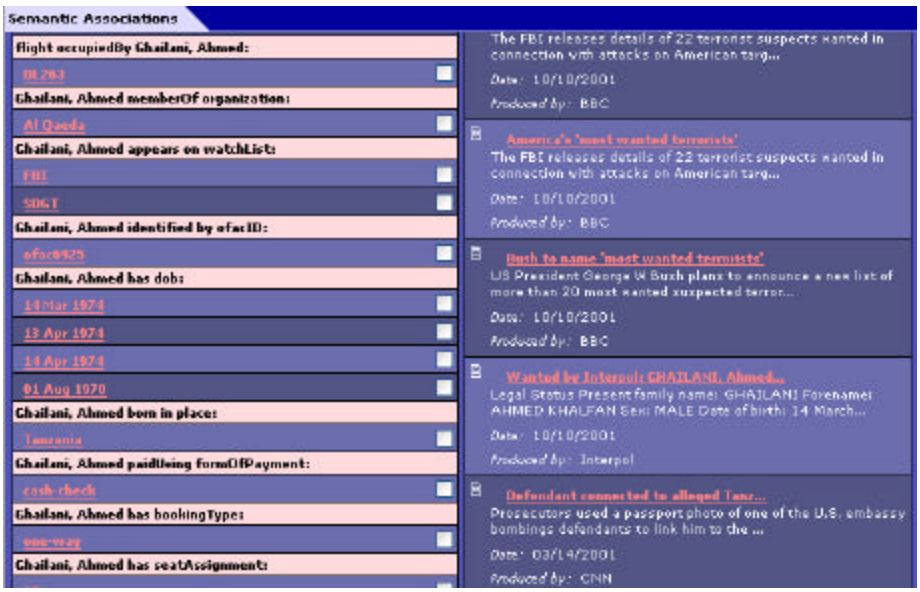

Fig. 8. Semantic Associations and Semantically Relevant Content

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.

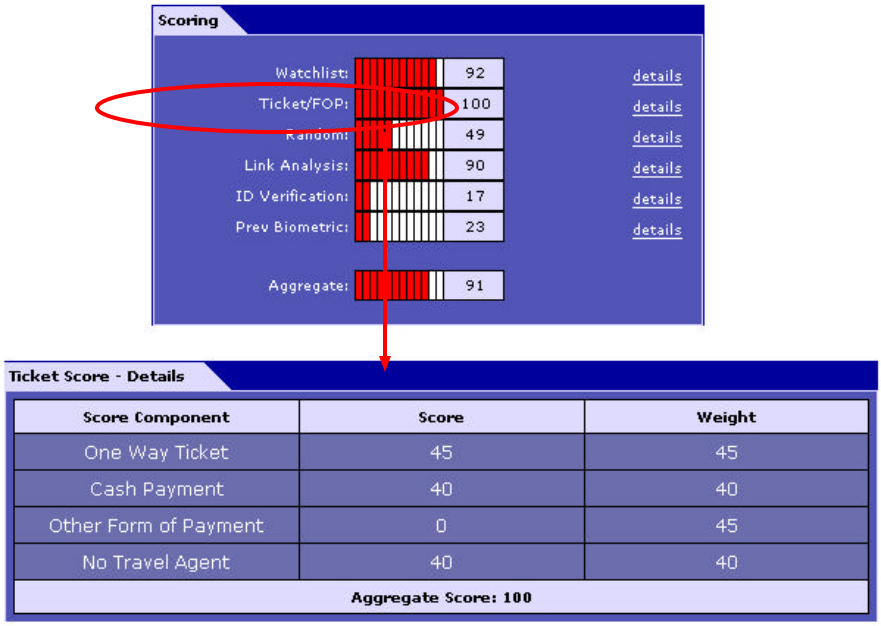

Fig. 9. Scoring

Figure 9 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

Ontocopi is an application that identifies communities of practice (Alani, Dasmahapatra, O'Hara, & Shadbolt, 2003). This is done by analyzing ontologies of different domains. 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 (Krebs, 2001) or money laundering (*Anti Money Laundering*, 2003) 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*).

The problem of finding relevant information has been approached by following the intuition of social networks (Yu & Singh, 2003). 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 efficient search in the network 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 (Chen, Han, & Yu, 1996; Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996) 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 (O'Harrow, 2002; Salkever & Cady, 2001) as well as in a white paper by Semagix, Inc (*National Security and Intelligence*, 2003). 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. This

research 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. We have been following the development of OWL (Bechhofer et al., 2003), and have started to support OWL as knowledge representation language for our test bed data (http://lsdis.cs.uga.edu/proj/SemDis/testbed/).

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. This work is funded in part by National Science Foundation (NSF) 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 NSF.

References

Alani, H., Dasmahapatra, S., O'Hara, K., & Shadbolt, N. (2003). Identifying communities of practice through ontology network analysis. *IEEE Intelligent Systems, 18*(2), 18-25.

Aleman-Meza, B., Halaschek, C., Arpinar, I. B., & Sheth, A. (2003). *Context-Aware Semantic Association Ranking*. Paper presented at the First International Workshop on Semantic Web and Databases, Berlin, Germany.

Anti Money Laundering. (Application White Paper)(2003). Semagix, Inc.

Anyanwu, K., & Sheth, A. (2003). *r-Queries: Enabling Querying for Semantic Associations on the Semantic Web*. Paper presented at the Twelfth International World Wide Web Conference, Budapest, Hungary.

Bechhofer, S., Harmelen, F. v., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., et al. (2003). *OWL Web Ontology Language Reference. W3C Proposed Recommendation*, from http://www.w3.org/TR/owl-ref/

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American, 284*(5), 34-+.

Brickley, D., & Guha, R. V. (2000). *RDF Vocabulary Description Language 1.0: RDF Schema. W3C Proposed Recommendation.*, from http://www.w3.org/TR/2003/PR-rdf-schema-20031215/

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and Isdn Systems, 30*(1-7), 107-117.

Chen, M. S., Han, J. W., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866-883.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., et al. (2003). A Case for Automated Large Scale Semantic Annotation. *Journal of Web Semantics, 1*(1).

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*: AAAI/MIT Press.

Hammond, B., Sheth, A., & Kochut, K. (2002). Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In V. Kashyap & L. Shklar (Eds.), *Real World Semantic Web Applications* (pp. 29-49): Ios Pr Inc.

Handschuh, S., Staab, S., & Studer, R. (2003). Leveraging metadata creation for the semantic web with CREAM. *Ki 2003: Advances in Artificial Intelligence, 2821*, 19-33.

Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., & Scholl, M. (2002). *RQL: A Declarative Query Language for RDF*. Paper presented at the 11th Intl. World Wide Web Conference, Honolulu, Hawaii.

Krebs, V. (2001). Mapping Networks of Terrorist Cells. *Connections, 24*(3), 43-52.

Lassila, O., & Swick, R. R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation*, from http://www.w3.org/TR/REC-rdf-syntax/

Magkanaraki, A., Tannen, V., Christophides, V., & Plexousakis, D. (2003). *Viewing the Semantic Web Through RVL Lenses.* Paper presented at the Second International Semantic Web Conference, Sanibel Island, Florida.

McBride, B. (2002). Jena: A semantic Web toolkit. *IEEE Internet Computing, 6*(6), 55-59.

National Security and Intelligence. (White Paper)(2003). Semagix, Inc.

O'Harrow, R. (2002, February 1, 2002). Intricate Screening Of Fliers In Works. *The Washington Post*.

Python Patterns - Implementing Graphs. (2003). Retrieved January 20, 3003, from http://www.python.org/doc/essays/graphs.html

Salkever, A., & Cady, J. (2001, December 4, 2001). The Price of Protecting the Airways. *BusinessWeek*.

Shah, U., Finin, T., Joshi, A., Cost, R. S., & Mayfield, J. (2002). *Information Retrieval on the Semantic Web.* Paper presented at the 10th International Conference on Information and Knowledge Management, McLean, Virginia.

Sheth, A., Arpinar, I. B., & Kashyap, V. (2003). Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships. In M. Nikravesh, B. Azvin, R. Yager & L. A. Zadeh (Eds.), *Enhancing the Power of the Internet Studies in Fuzziness and Soft Computing*: Springer-Verlag.

Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., & Warke, Y. (2002). Managing semantic content for the Web. *IEEE Internet Computing, 6*(4), 80-87.

Tarjan, R. E. (1981a). Fast Algorithms for Solving Path Problems. *Journal of the ACM, 28*(3), 594-614.

Tarjan, R. E. (1981b). A Unified Approach to Path Problems. *Journal of the ACM, 28*(3), 577-593.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., & Ciravegna, F. (2002). *MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup.* Paper presented at the 13th International Conference on Knowledge Engineering and Management (EKAW 2002), Sigüenza, Spain.

Woods, W. (1975). What's in a link: Foundations for Semantic Networks. In D. Bobrow & A. Collins (Eds.), *Representation and Understanding* (pp. 35-82). New York: Academic Press.

Yu, B., & Singh, M. P. (2003). *Searching social networks.* Paper presented at the Second international joint conference on Autonomous agents and multiagent systems, Melbourne, Australia.