ogy), Molecular Biology Ontology Working Group, Stanford University's Knowledge Systems Lab (Enterprise Ontology), are also coming together, to effectively conceptualize the domain knowledge, and enable standards for exchanging, managing and integrating data more efficiently. Research in the Semantic Web has also spawned several commercially viable products through companies such as Semagix [17,14] and Ontoprise [15] to name a few.

Given these developments, the stage is now set for the next generation of technologies, which will facilitate getting actionable knowledge and information from massive data sources thereby assisting in information analysis. Many users try to analyze information by either browsing the information space, or using a search engine. Search engine based systems only locate documents based on keywords or key phrases. These approaches are not very representative of what the user actually wants. Therefore, most of the retrieved documents are either irrelevant, or contain the information buried deep within other data. The onus is then on the user, who must decide, which of the retrieved documents are relevant, and then use their mental model, of the information they are looking for, in order to obtain the relevant information.

The main goal of this work is to ease the process of analyzing across different sources of data and enable users to uncover previously unknown and potentially interesting relations (or associations) [2,19]. In the quest for finding associations, it is also possible to find too many of them between the entities. Therefore, it is also important to locate interesting and meaningful relations and to rank them before presenting to the user.

1.1 Semantic Associations

The associations 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, the Resource Description Framework (RDF) data model [11] 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" and users can formulate queries to find the semantic association(s). For example, semantic association queries in flight security domain may 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 of 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 this may enable analysts to see the connections between disparate people, places and events.

Semantic associations are based on intuitive notions such as connectivity and semantic similarity. In [2], 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:

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, … 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.

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, … e_{n-1}, P_{n-1}, e_n and f_1, Q_1, f_2, Q_2, f_3, …, 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.

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 (ρ-Query) A *ρ-Query*, expressed as *ρ(x, y),* where *x* and *y* are entities, results in the set of all semantic paths that connect *x* and *y*.

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

We are currently working on a ranking technique for similarity associations, which is not discussed in this paper. Furthermore, it is conceptually different than ranking semantic connections because it involves ranking the set of all pairs of semantically similar paths originating at entities *x* and *y.* Thus semantic associations and semantic association queries are used to refer to only semantic connectivity and ρ-Queries respectively in the rest of the paper.

1.2 Ranking Semantic Relations

A typical semantic query can result in many semantic paths semantically linking the entities of interest. Because of the expected high number of paths, it is likely that many of them 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. Also, 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 and used (or the domains of the user interest). For example, consider a sub-graph of an RDF graph representing two soccer players who belong to the same team and who also started a new restaurant together. If the user is just interested in the *sports* domain the semantic associations involving restaurant related information can be regarded as irrelevant (or ranked lower). This can be accomplished by enabling a user to browse the ontology and mark a region (sub-graph) of nodes 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 discarded. More formally, *ontological regions* can represent context. In this paper we present a flexible method for specifying context through an ontology-based context specification language.

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. In this paper, the ranking score assigned to a particular semantic association is defined as a function of these parameters. Furthermore different weights can be given to different parameters according to users' preferences (e.g., trust could be given more weight than others). 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.

Thus our contributions in this paper are two-folds:

- Capturing users' interests semantically through an ontology-based context specification language,
- Using a ranking function incorporating user-defined semantics (e.g., context) and universal semantics (e.g., associations conveying more information).

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 introduces context specification language and discusses ranking technique. Section 4 concludes the paper.

2 Related Work

Knowledge representation approaches tried to capture relationships based on logics, or sets theory, etc. Our approach is to consider relations in the semantic Web, those that are expressed semantically using the RDF model. Then from a set of semantic associations we try to distinguish the relevant ones quantitatively. Research in the area of ranking semantic relations includes [12], where the notion of "semantic ranking" is presented to rank queries returned within semantic Web portals. Their technique reinterprets query results as "query knowledge-bases", whose similarity to the original knowledge-base provides the basis for ranking. The actual similarity between a query result and the original knowledge-base is derived from the number of similar super classes of the result and the original knowledge-base. In our approach, the relevancy of results usually depends on a context defined by users.

Our earlier work [2] introduces using "context", path length, and property relevance as a basis for ranking. Basically, [2] defines a notion of context which includes a set of ontologies and a set of relationship name pairs with a value. The value indicates the precedence level, a degree of importance for a particular context. This approach considers context based on value assignments for different ontologies. In this work instead, we provide context specification at a level (of classes and properties) that allows precise definitions of areas of interest for the user.

While the issues of ranking semantic relations are fundamentally different from those addressed in contemporary information retrieval ranking approaches, it is worth discussing some of these techniques. [5] presents the *page rank* algorithm used by Google. Page rank weights are assigned on the basis of page references, thus more popular pages have a higher rank. [21] presents Teoma's technique of *subject specific popularity*, in which a page's rank is based on the number of same-subject pages that reference it, not just its general popularity. Earlier, Northern Light had introduced the concept of folders and the documents resulting from keyword search results were segregated by these folders representing relevant categories. While relevant, these ranking algorithms lack the consideration of formal semantics (as captured through ontology representation) and explicitly specified context when assigning ranks, both of which are needed when ranking semantic associations. Although the current semantic association ranking scheme differs from ranking Web pages through not involving social contributors such as a *voting* mechanism, it is an interesting research direction to involve similar techniques for assessing importance and value of semantic associations.

Attempts to model context include [9], which proposed a context representation mechanism to solve conflicts of semantic and schematic similarities between database objects. [6] introduced an ontology that captures users' context and situation by considering goals, tasks, actions and system's context in order to observe and model human activities. The approach is mainly focused to use context to reduce user's intervention in the system.

3 Ranking Semantic Associations

In this work, we provide semantic associations which are ranked for a given semantic association query. Our approach for ranking semantic associations is primarily based on capturing the interests of a user. Therefore, a context specification is the first step towards measuring how relevant a semantic association is.

3.1 Context Specification

A context specification captures the users' interest in order to provide her with the relevant knowledge within numerous indirect relationships between the entities. We consider data in an RDF model with an associated RDF Schema [4] that describes the relationships between entities. Since the types of the entities are described in the RDF Schema, we can use the associated class and relationship types to restrict our attention

to the entities and relations of interest. Thus, by defining regions (or sub-graphs) of the RDF Schema (RDFS) we are capturing 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. A *region* of interest is 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: (*i*) Class level: paths that include instances of that class are relevant, and (*ii*) Property level: paths including the specified properties are relevant.

Within the Class level, we may also restrict or allow subclasses to be considered relevant as well as the classes higher in the class hierarchy. 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 subclass of "*Organization*").

We specify in a flexible yet detailed manner which Classes and Properties are relevant using XML. The following is an example of specifying Classes with restrictions:

```
<region id="R1" weight=".65">

  <classLevel name="TerroristAct" includeSubclasses="all"/>

  <classLevel name="TerroristOrg" includeSubclasses="no"/>

  <propertyLevel name="involvedIn" domainRestrictions="TerroristOrg"
rangeRestrictions="TerroristAct, Kidnapping, SuicideAttack" />

</region>
```

A region has a weight defining its relative importance. The particular XML example shown above captures the area of interest that is used as *region A* in **Fig. 5** in Section 3.2.2. Note that a user can define several ontological regions with different weights to specify the association types s/he is interested in.

3.2 Weight Assignments

Semantic associations represented as paths connecting two entities can span across multiple domains (or regions) and involve any number of entities and properties. Our ranking approach defines a path rank as a function of various intermediate weights.

As a path is traversed it will have many different intermediate weights which ultimately contribute to its overall rank. We classify these weights into two categories, *Universal* and *User-Defined*.

3.2.1 Universal Weights

Certain weights will influence a path rank regardless of the query or context of interest. We call them *Universal Weights*. The following subsections identify and define *Universal Weights* that contribute to the overall path rank.

Subsumption Weight. When considering entities in ontology, those that are lower in the hierarchy can be considered to be more specialized instances of those further up in the hierarchy [16]. Thus, lower entities have more specific meaning. **Fig. 1** depicts a class, "*Organization*", as well as various subclasses of it. In the figure, "*Organization*" is the highest class in the hierarchy, and thus is the most general. It is clear that a "*Political Organization*" is a more defined "*Organization*".

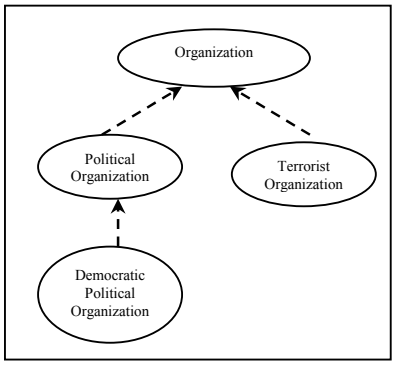

Fig. 1. Class Hierarchy Example

Similarly, a "*Democratic Political Organization*" conveys more meaning than both an "*Organization*" and a "*Political Organization*". Hence, it is very apparent that as the hierarchy is traversed from the top down, subclasses become more specialized than their super-classes. The concept of class specialization in a path is captured by a Universal Weight that we call a *Subsumption Weight*. The intuition is assigning more weights to more "specific" semantic associations because they convey more meaning then "general" associations.

We will now provide some brief definitions used to define the overall *Subsumption Weight* of a path. First, we define a component, *c*, within a path *P* to be any entity or property contained in *P*. Thus, $c = \{entity\}|\{property\}$.

Next we define a *component weight* of the i^{th} component c_i, in a path *P* such that

$$Component\ Weight_{\ i} = \frac{H_{c_i}}{|H|} \ . \tag{1}$$

where H_{c_i} is the position of the i^{th} component in its hierarchy *H* (the class at the top has value 1) and $|H|$ is the total height of the classes/properties hierarchy. Hence, $Component\ Weight_i \rightarrow (0,1]$. For example, given **Fig. 1** above, the component weight of the classes *Democratic Political Organization,* c_3, and *Political Organization,* c_2, would be

$$c_3 = \frac{H_{c_3}}{|H|} = \frac{3}{3} = 1 \text{ and } c_2 = \frac{H_{c_2}}{|H|} = \frac{2}{3} = 0.6 . \quad (2)$$

We can know define the overall *Subsumption Weight* of a path *P* such that

$$S_P = \frac{1}{|c|} \times \prod_{i=2}^{|c|+1} c_i \ . \quad (3)$$

where $|c|$ is the number of components in *P* (excluding the start and end entities because they will never change in a result set) and c_i is the *component weight* of the i^{th} component in the path. Thus the *Subsumption Weight* of a path *P*, S_P, is the product of all the component weights within *P*, normalized by the number of components in the path (to avoid bias in path length). To illustrate this, we use the ontology that has been developed for the national security domain in our lab (see **Fig. 2**).

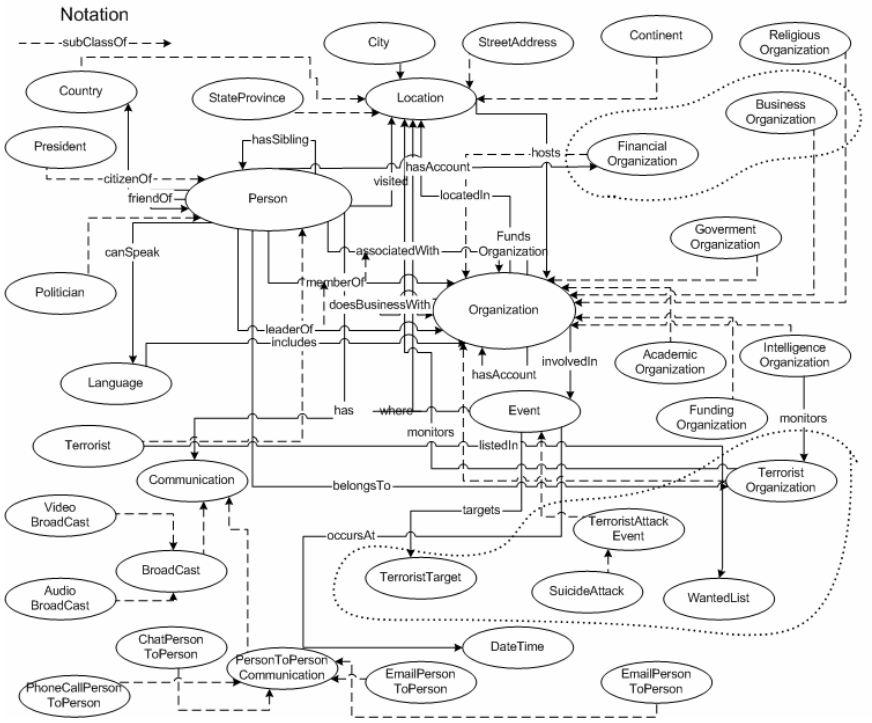

Fig. 2. Sample Ontology

Consider the following paths between entities e_1 and e_5 depicted in **Fig. 3**. First, one can see that all three paths are somewhat similar. The middle path seems to be a bit more specific that the top path, in that the person is member of a "*Terrorist Organization*," not just any "*Organization*," that is "*involvedIn*" a "*Suicide Attack*". When

inspecting the bottom path we see that this person is actually a "*leaderOf*" some "*Terrorist Organization*" that was "*involvedIn*" the same "*Suicide Attack*". Thus we assume that the third path conveys more meaning than the first two. When ranking these three paths with respect to their total meaning conveyed, one would expect to see that last path ranked higher than the others (in absence of additional user defined context/weights).

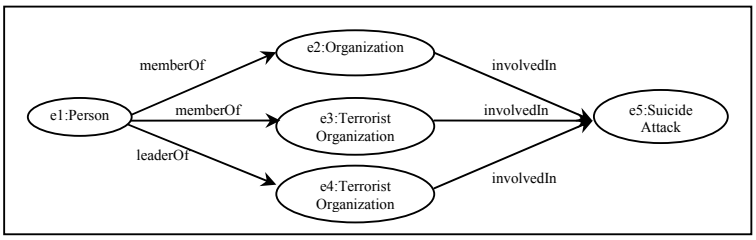

Fig. 3. Subsumption Weight Example

Now we will determine the *Subsumption Weight*, S_1, of the first path in **Fig. 3**, $e_1 \rightarrow e_2 \rightarrow e_5$. The corresponding *Subsumption Weight* for this path would be given by

$$S_1 = \frac{1}{3} \times \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{1}\right) = .083 \,. \quad (4)$$

Similarly, the middle path $e_1 \rightarrow e_3 \rightarrow e_5$ has a *Subsumption Weight* of .167 and a higher value of .334 for the path $e_1 \rightarrow e_4 \rightarrow e_5$.

Hence as desired previously, with respect to only the meaning conveyed in the path, the *Subsumption Weight* will assign higher weights to paths with a more defined meaning. Thus, quality and completeness of the ontology become important to avoid biased ranking ([16] addresses issues on explicitness and formalization of ontologies). Note that we are considering specificity of relations besides entities. This is why the third semantic association is ranked higher than the second one. Furthermore, statistical properties of ontology (e.g., connectivity of certain nodes, etc.) can contribute to *Universal Weight* yet discussion of those metrics is out of scope of this paper.

3.2.2 User-Defined Weights

In contrast to *Universal Path Weights*, some path weights will be query (or context) specific. These will be referred to as *User-Defined Weights*. The following subsections identify and define *User-Defined Weights* that contribute to the overall path rank.

Path Length Weight. In some queries, a user may be interested in the most direct paths (i.e., the shortest path). This may infer a stronger relationship between two entities. Yet in other cases a user may wish to find possibly hidden, indirect, or discrete paths (i.e., longer paths). The latter may be more significant in domains 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 [10] or money laundering [1] involves deliberate innocuous looking transactions. Hence, the user should determine which *Path Length* influence, if any, should be used (largely domain dependent).

We will now define the *Path Length Weight*, *L*, of a path *P,* where $L_P \rightarrow [0, 1]$
If a user wants to favor shorter paths, (**5a**) is used, where $|c|$ is the number of components in the path *P* (excluding the first and last nodes). In contrast, if a user wants to favor longer paths (**5b**) is used.

$$L_P = \frac{1}{|c|} \text{ (a); } L_P = 1 - \frac{1}{|c|} \text{ (b).} \quad \textbf{(5)}$$

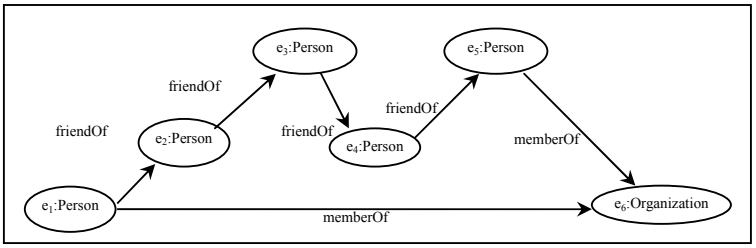

Fig. 4. Path Length Examples

To demonstrate the *Path Length Weight*, consider **Fig. 4**. This figure depicts two possible paths between a person and an organization. Given this example, suppose a user is interested in more direct path between entities. In this case, the longer of the two paths (call it P_1) should be ranked lower than the shorter one (P_2), so (**5a**) should be used.

Using (**5a**), the *Path Length Weight* of the two paths would be

$$L_{P_1} = \frac{1}{9} \text{, where as } L_{P_2} = \frac{1}{1} \text{ .} \quad \textbf{(6)}$$

Thus the shorter of the two paths has a higher rank value as initially expected. If a user were alternatively interested in longer paths, (**5b**) would be used instead. In this case

$$L_{P_1} = 1 - \frac{1}{9} = .889 \text{, where as } L_{P_2} = 1 - \frac{1}{1} = 0 \text{ .} \quad \textbf{(7)}$$

Thus, P_1 has a higher *Path Length Weight* than P_2, again as desired.

Context Weight. As discussed in Section 3.1, it is possible to capture a user's interest through a *Context Specification*. Thus, using the *context* specified, it is possible to rank a path according to its relevance with a user's domain of interest.

With the Context Specification proposed in Section 3.1, a user can assign a weight to particular regions of ontology. When considering how to use these weights many issues arise. For example, paths can pass through numerous regions of interest. Large and/or small portions of paths can pass through these regions as well. Another consideration is whether all of the nodes in a path actually lie within a specified region. While we could omit paths that contain some nodes outside of all regions, we have decided to rank them lower because they are still considered relevant since they pass through some region. Suppose a user specifies the following region *A* containing the class "*TerroristAct*" and its subclasses and region *B* containing the class "*FinancialOrganization*" and its subclasses. The resulting regions, *A* and *B,* are within the terrorist and financial domains respectively. **Fig. 5** illustrates various paths which pass through these regions.

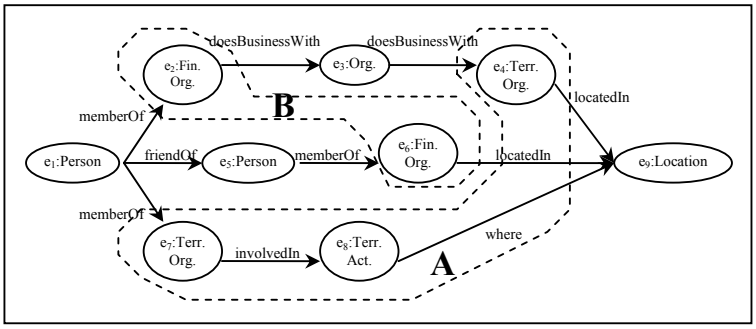

Fig. 5. Context Related Paths

The topmost path (call it P_1) passes through regions *B* and *A*, the middle path (P_2) passes through *region B*, and the third path (P_3) at the bottom passes through *region A*. Next, let the (user-defined) weight associated with a *region x* be represented as r_x. Also assume that $r_A = .75$ and $r_B = .50$.

The weight assignment illustrates the user is more interested in terrorism domain but also wants to consider financial associations, albeit with lesser priority. If we take into consideration the components of a path, excluding its start and end entities, the expected ranking of these three paths would be P_3, P_1, P_2. Path P_3 would have the highest rank because all of its components (entities and properties) are included in some context, which happens to be the context with the highest weight. P_1 would be ranked next because it has a component in *B*, but (unlike P_2) also has a component in *A*. Given this background we will define the *Context Weight* of a path. First, let the i^{th} *region* be represented by R_i. Thus, we define the *Context Weight* of a given path *P*, C_P, such that

$$C_P = \frac{1}{|c|}\left(\left(\sum_{i=1}^{\#regionsPisIn}(r_i \times (\sum c \in R_i))\right) \times \left(1 - \frac{\#c \notin R}{|c|}\right)\right) \quad (8)$$

where r_i is the user defined weight of the *region* R_i, c is a component in the path P (excluding the start and end entities), and $|c|$ is number of components in the path (again excluding the start and end entities). That is, for each context that P passes through, sum the total number of components in P that are in the *region* R_i and multiply it by the weight attributed to that *region*, r_i. In order to reward paths in which all components are included in some region, the total number of components not in any region is divided by the total number of components, which is then subtracted from 1. This is then multiplied by the previous summation. Lastly, this total is normalized by the total number of components in the path. Note that a property component is considered to be in some *region* if it is entirely included in that *region* or one of the entities it is involved with (at either end) is in that *region*. If the two entities in which some property is involved are contained in two separate *regions*, the higher of the two *region* weights will be the *region* weight for that property. Also note that due to the subclass relationship of entities, properties which do not directly appear in a *region* may actually be included in some situations. To illustrate this, we will assign a *Context Weight* to the three paths presented **Fig. 5**.

P_1 passes through both regions A and B, which have a weight of .75 and .50 respectively. In both of these regions, three components are involved. Thus the initial summation is $(0.75 \times 3) + (0.5 \times 3) = 3.75$. There is one component (*Organization*) in P_1 which is not included in a region, so we have

$$3.75 \times \left(1 - \frac{1}{7}\right) = 3.21 \quad . \quad (9)$$

This is normalized by the number of components in P_1, hence we have

$$C_{P_1} = \frac{1}{7} \times 3.21 = .458 \quad . \quad (10)$$

Next consider P_2. This path only passes through region B, which has a weight .50. In this region, three components are involved. Thus the initial summation is $(0.50 \times 3) = 1.5$. There are two components ("*friendOf*" and "*Person*") in P_2 which are not included in a region, so we have

$$1.5 \times \left(1 - \frac{2}{5}\right) = .9 \quad . \quad (11)$$

This is normalized by the number of components in P_2, so

$$C_{P_2} = \frac{1}{5} \times .9 = .18 \quad . \quad (12)$$

Lastly, consider P_3. This path only passes through region A, which has a weight .75. In this region, five components are involved. Thus the initial summation is (0.75×5)

= 3.75. There are no components in P_3 which are not included in some *region*, so we have

$$3.75 \times (1 - \frac{0}{2}) = 3.75 \ . \quad (13)$$

This is normalized by the number of components in P_3, so

$$C_{P_3} = \frac{1}{5} \times 3.75 = .75 \ . \quad (14)$$

Hence, as expected initially the ranking is P_3 (0.75), P_1 (0.458), and P_1 (0.18).

Trust Weight. Various relationships (properties) in a path originate from different sources. Some of these sources may be trusted while others may not (e.g., Reuters could be regarded as a more trusted source on international news than some of the other news organizations). Thus, trust values need to be assigned to relationships depending on the source. The process of automatically assigning trust to a specific relationship is out of the scope of this paper; instead we assume that users or other processes previously specified the trust value of relationships. Let the trust weight of the i^{th} property p_i of a path be t_{p_i}, where $t_{p_i} \rightarrow [0,1]$. We now define the Trust Weight of an overall path P as

$$T_P = \prod_{i=1}^{\#p \in c_P} t_{p_i} \ . \quad (15)$$

where c_P are all the *property* components within the path P. Thus, T_P is the product of all property weights in the P.

3.3 Ranking Criterion

Section 3.2, defines various path weight influences. We will now define the overall path rank, using these weights. As mentioned earlier, *Universal Weights* will always affect the overall path weight, while the *User-Defined Weights* will only be used when specified by the user. Let the Overall *Path Weight* of a path P denoting a semantic association be a linear function such that

$$W_P = k_1 \times S_P + k_2 \times L_P + k_3 \times C_P + k_4 \times T_P \ . \quad (16)$$

where k_i add up to 1.0 and are intended to allow fine-tuning of the different ranking criteria (e.g., *trust* can be given more weight than *path length*).

3.4 Preliminary Results

As a test-bed for querying semantic associations we have implemented a prototype named PISTA (see **Fig. 6**). In PISTA (Passenger Identification, Screening, and Threat Analysis) we have designed an ontology for national security domain (see **Fig. 2**). This ontology has names of organizations, countries, people, terrorists, terrorist acts etc. that are all inter-related to each other with named relationships to reflect real-world knowledge about the domain (e.g., "terrorist" "belongs to" "terrorist organization").

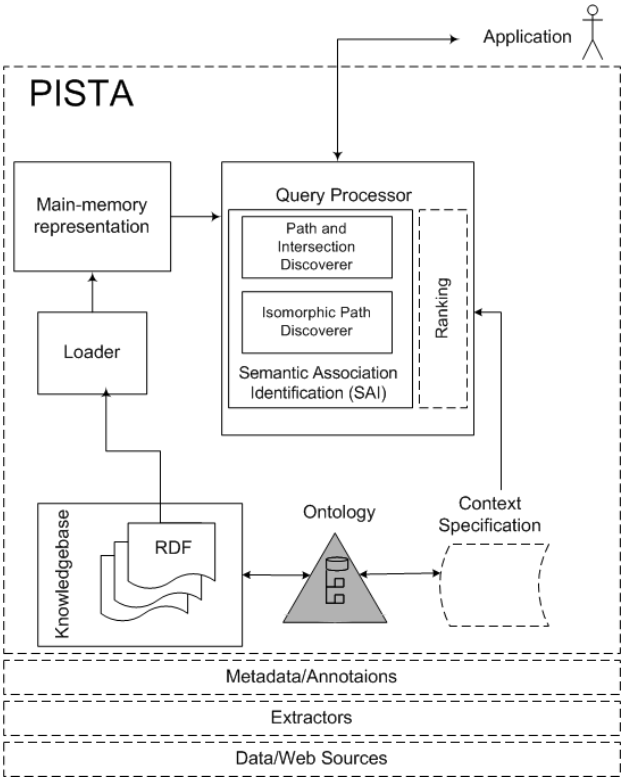

Fig. 6. PISTA Architecture

The sources from which metadata were extracted were selected to populate the ontology with entities related to terrorism. The metadata is represented in RDF, on which semantic association queries were performed. For information extraction we have used Semagix's suite which includes a set of tools for extraction of entities from (semi)-structured information sources [17]. This toolkit allows extraction of entities from Web pages and establishes relationships between them. This extraction is based on our national security ontology thereby placing an extracted entity in its appropriate

place in the hierarchy of classes. Currently, there are over 6,000 entities and more than 11,000 explicit relations among them.

For querying semantic associations, we have implemented search algorithms, which use the schema information in conjunction with the RDF data that find semantic associations (Definition 3) that represent the relationships between any two entities. We represent both the RDF Schema and the RDF data as main memory directed graphs based on the Jena model [13]. Then, search for semantic similarity recursively finds similar paths between two entities by relaying on the schema to find similar entities/relationships (i.e., which belong to same parent class) (see Definition 2). We also use a graph traversal algorithm (based on breadth-first search), which does not consider the direction of the edges when searching for semantic connectivity associations (see Definition 1).

For example, consider following semantic association query *ρ("Nasir Ali", "AlQeada")*. In PISTA this query results in 2234 associations. A small subset of these associations is shown in the table below (not in a particular order).

Nasir Ali → friendWith → T. Smith → memberOf → AlQeada
Nasir Ali → friendWith → Cabbar Ali → visited → Afganistan → hosts → AlQeada
Nasir Ali → friendWith → T. Smith → hasAccount → J. Funds → fundsOrganization → AlQeada
Nasir Ali → friendWith → OsamaBinLaden → leaderOf → AlQeada
Nasir Ali → hasAccount → J. Funds → fundsOrganization → AlQeada
Nasir Ali → associatedWith → A. G. College → hasAccount → J. Funds → fundsOrganization → AlQeada
Nasir Ali → memberOf → TRO ← memberOf ← OsamaBinLaden → leaderOf → AlQeada
Nasir Ali → associatedWith → TRO → doesBusinessWith → AlQeada

For illustration, we have a context defined by a region that captures 'terrorism' interest with weight of 0.6 (lower region in **Fig. 2**) and another region capturing 'financial' interest with weight of 0.4 (upper region in **Fig. 2**). The following table shows how the relationships are ranked when we apply our ranking formula. The ranking criteria (constants k_i in equation (16)) for this example assign values of 0.6 to *context weight*, 0.2 to *subsumption weight*, 0.1 to *path length weight* (longer paths favored), and 0.1 to *trust weight* (we assumed same trust for all entities/properties in this example).

Ranked Results	**Rank**
Nasir Ali → memberOf → TRO ← memberOf ← OsamaBinLaden → leaderOf → AlQeada	0.5560
Nasir Ali → associatedWith → TRO → doesBusinessWith → AlQeada	0.5488
Nasir Ali → has Account → J. Funds → fundsOrganization → AlQeada	0.5123
Nasir Ali → friendWith → T. Smith → has Account → J. Funds → fundsOrganization → AlQeada	0.3208
Nasir Ali → associatedWith → A. G. College → has Account → J. Funds → fundsOrganization → AlQeada	0. 2941
Nasir Ali → friendWith → OsamaBinLaden → leaderOf → AlQeada	0.2733
Nasir Ali → friendWith → T. Smith → memberOf → AlQeada	0..2511
Nasir Ali → friendWith → Cabbar Ali → visited → Afganistan → hosts → AlQeada	0.2344

The top ranked semantic association comes up first because its entities all belong to the "terrorism" region (with higher relevance than "financial") and it is one of the

longer associations. The second ranked semantic association includes entities only within the "terrorism" region as well, but it is a shorter path (longer paths are preferred in this example). The third association consists only of entities within the "financial" region, which we would expect to be ranked lower that the first two because we have weighted the "terrorism" region higher. The remaining paths contain some nodes not within any region, thus they are ranked below the previous three associations as expected. The fourth and fifth semantic associations are ranked as such because they are both longer than the rest and contain more entities within the two regions of interest. Note that the fourth association is ranked above the fifth because the "friendWith" relationship is more specific than the "associatedWith" relationship. When inspecting the last three associations, it is seen that they contain the least number of entities within a context. Thus, we would expect them to be ranked lower than the rest (due to the context being weighted so heavily). When we look at the sixth and seventh ranked associations, we see that the sixth is more specific in that entity "OsamaBinLaden" is the "leaderOf" "AlQeada", where the entity "T. Smith" is only a "memberOf" the same *Terrorist Organization*. The path ranked lowest contains the least number of entities in some region of interest, as expected.

4 Conclusions and Future Work

Semantic associations primarily capture information relating two entities. We are interested in the path that relates two entities by a sequence of interconnected links. Discovering of such relations (explained in [2]) gives results containing multiple paths connecting two entities. These paths have different meaning depending on the type of relation or the type of entities in each of components (either resource or property) of the path. The number of semantic associations between entities will grow much faster than the rate of the growth of a graph representing a knowledgebase and corresponding ontology. Also, understanding the relevance of each of the semantic association as a result of a query is arguably harder than determining a document's relevance and ranking in a result provided by a typical search engine. Hence determining a good ranking strategy is crucial.

In this paper, we defined a ranking formula that considers *Subsumption Weight* (how much meaning a semantic association conveys depending on the places of its components in the ontology), *Path Length Weight* (that allows preference of either immediate or distant relationships), *Context Weight* (how relevant is the path to the user interest – defined using our context specification framework), and *Trust Weight* (determining how reliable a relationship is according to its provenance).

Currently we are working on ranking similarity associations (Definition 2). In fact this involves discovering all semantic connections between two entities (Definition 1), and then measuring if and how these associations can be broken into semantically symmetric associations (e.g., two terrorist attacks may be similar because they might be symmetrically connected to same methods). A formal query language for semantic associations is currently under development.

In order to assess the effectiveness of the ranking scheme outlined in this paper, standard ranking metrics such as precision and recall can be employed. However, we

think metrics for context-aware ranking should be different than the traditional metrics only using precision and recall. Because we rank the results considering a context specified by the user, and the evaluation criterion would be very subjective according to user's interests. Therefore, we believe a user-oriented assessment criterion is needed.

The future work also includes improving the semantic association discovery algorithms using the ranking scheme we have described in this paper for better scalability in very large data sets. For example, some partial paths can be pruned on the fly if their (partial) rank value drops under a predefined threshold.

Acknowledgements: We thank Semagix, Inc. for providing its Freedom product, which is based on the SCORE technology and related research out at the LSDIS Lab [20]. PISTA application has benefit from Semagix Inc.'s application in this area [14]. Brainstorming and discussions with members of our research project team, specifically Krys Kochut, John Miller, Kemafor Anyanwu, and Cartic Ramakrishnan, have enriched this work.

References

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

[2] K. Anyanwu and A. Sheth, "r-Queries: Enabling Querying for Semantic Associations on the Semantic Web", The Twelfth International World Wide Web Con-ference, Budapest, Hungary (2003)

[3] 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

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

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

[6] J. L. Crowley, J. Coutaz, G. Rey and P. Reignier, "Perceptual Components for Context Aware Computing", UBICOMP 2002, International Conference on Ubiquitous Computing, Goteborg, Sweden, September 2002

[7] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien, SemTag and Seeker: Bootstrapping the semantic Web via automated semantic annotation, The Twelfth International World Wide Web Conference Budapest, Hungary (2003)

[8] B. Hammond, A. Sheth, and K. Kochut, "Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content", in Real World Semantic Web Applications, V. Kashyap and L. Shklar, Eds., IOS Press, pp. 29-49, December 2002

[9] V. Kashyap, A. Sheth. Semantic and schematic similarities between database objects: a context-based approach. VLDB Journal (1996) 5: 276–304.

[10] V. Krebs, “Mapping Networks of Terrorist Cells”. Connections, 24(3): 43-52. (2002).
[11] O. Lassila and R. R. Swick: "Resource Description Framework (RDF) Model and Syntax Specification", W3C Recommendation, World Wide Web Consortium, Cambridge (MA), February 1999
[12] A. Maedche, S. Staab, N. Stojanovic, R. Studer, and Y. Sure. SE-mantic PortAL – The SEAL approach. to appear: In Creating the Semantic Web. D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (eds.) MIT Press, MA, Cambridge (2001)
[13] B. McBride “Jena: Implementing the RDF Model and Syntax Specification”, in: Steffen Staab et al (eds.): “Proceedings of the Second International Workshop on the Semantic Web - SemWeb'2001”, May 2001.
[14] National Security and Intelligence, A Semagix White Paper, 2003. http://www.semagix.com/pdf/national_security.pdf
[15] Ontoprise® GmbH, http://www.ontoprise.com
[16] M. Rodriguez, and M. Egenhofer, “Determining Semantic Similarity among Entity Classes from Different Ontologies”, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 2, March/April 2003.
[17] Semagix. http://www.semagix.com.
[18] U. Shah, T. Finin, A. Joshi, R. S. Cost, and J. Mayfield, “Information Retrieval on the Semantic Web”, 10th International Conference on Information and Knowledge Management, November 2002.
[19] 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).
[20] 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.
[21] Teoma: http://sp.teoma.com/docs/teoma/about/searchwithauthority.html