

# Searching and Ranking Documents based on Semantic Relationships

Boanerges Aleman-Meza  
*LSDIS Lab, Computer Science*  
*University of Georgia, Athens, GA, USA*  
*boanerg@cs.uga.edu*

## Abstract

*Just as the link structure of the web is a critical component in today's web search, complex relationships (i.e., the different ways the dots are connected) will be an important component in tomorrow's web search technologies. In this paper, I summarize my research on answering the question of: How we can exploit semantic relationships of named-entities to improve relevance in search and ranking of documents? The intuition of my approach is to first analyze the relationships of named-entities with respect to a query. Second, relevance weights, which are assigned by human experts, can then be used to guarantee results within a relevance threshold. These relevance measures can be applied both for searching and ranking of documents.*

## 1. Introduction (Research Problem)

Today's Web search technologies rely on link analysis techniques that exploit the structure of the web to determine important documents. The web itself contains large amounts of information (i.e., knowledge and facts), such as the National Library of Medicine's MeSH (Medical Subject Heading) vocabulary, which is used for annotation of scientific literature. The information contained in such agreed-upon vocabularies (or ontologies) is quite valuable for determining relevance based on the 'known' facts, relations, or other data.

In some domains, there are available ontologies that were built with significant human effort. However, it has been demonstrated that large ontologies can be built with tools for extraction and annotation of metadata [18, 31]. Industry efforts have demonstrated capabilities for building large populated ontologies [28] as well as for metadata extraction from well over a billion web pages [11] (IBM's WebFountain).

There are two specific aspects of ontologies which are particularly important for improving relevance in search and ranking of documents. The first aspect is that of named-entities, such as names of countries, cities, people, research articles, artists, and politicians. Existing techniques have been developed for entity-based search of documents (i.e., [14]). The second aspect is relationships, which provide the context (or meaning) of

an entity. The value of such relationships relies on the fact that they are *named* relationships. That is, they refer to a 'type' defined in an ontology. Relationships will play an important role in the continuing evolution of the Web [27]. Additionally, it has been argued that people will use web search not only for documents, but also for information about semantic relationships [25]. In this context, the research question that I am addressing is: How we can exploit semantic relationships of named-entities to improve relevance in search and ranking of documents?

The idea of utilizing semantics for improving search results is typically referred to as semantic search [15, 23, 30]. A variety of aspects on improving search and ranking of documents have been considered, such as concept-based search of documents [8, 13]. Search techniques typically rely on some form of document processing. Thus, existing document processing techniques can be helpful for semantic search; for example, document summarization [24] and analysis of the semantics of terms [21] (e.g., hyponym).

The research problem of improving relevance in search and ranking of documents requires techniques that consider the semantics of relationships. My work builds upon the creation of large populated ontologies, entity discovery and (semantic) annotation of documents as well as discovery of 'semantic associations' on large populated ontologies. Semantic associations are the different relationships that interconnect two entities [6]. Each semantic association can be viewed as a path consisting of one or more relationships. Applications that utilized the concept of semantic associations include search of biological terms in patent databases [22], provenance and trust of data sources [12] and national security [26]. Semantic associations are applicable in the research problem that I am addressing due to the need of analyzing the relationships that interconnect two entities.

My approach for relevance in search and ranking of documents can be summarized as follows. First, the relevance of named-entities, contained within documents, is computed with respect to a query. Second, ranking of documents involves a different analysis that also exploits the semantics of relationships. The research challenges are: demonstrating the effectiveness of the proposed approach, establishing relevance threshold guarantees and addressing scalability issues.

## 2. Proposed Solution and Preliminary Results

Addressing this research problem involves a multi-step process: (i) A populated ontology should be obtained (or created). (ii) A document collection should be pre-processed to obtain semantic annotations. The result of semantically annotating documents is a set of explicit assertions indicating named-entities from the ontology that appear within a document. (iii) Indexing of documents and the named-entities within them. (iv) Retrieval of documents related to a query. (v) Ranking of the documents considering their relevance to the query.

These steps require addressing both dataset and algorithm aspects. Some of the components of this multi-step process have been addressed in my previous publications. Figure 1 provides a schematic view of the system architecture.

Firstly, the creation of an ontology intended for search of documents calls for focusing in a specific domain where populated ontologies are available or can be easily built. My preliminary work for building a test-bed ontology (called SWETO [4]) has served its initial purpose yet further improvements on this dataset are needed. In particular, I have found that the richness and diversity of relationships within an ontology is a crucial aspect.

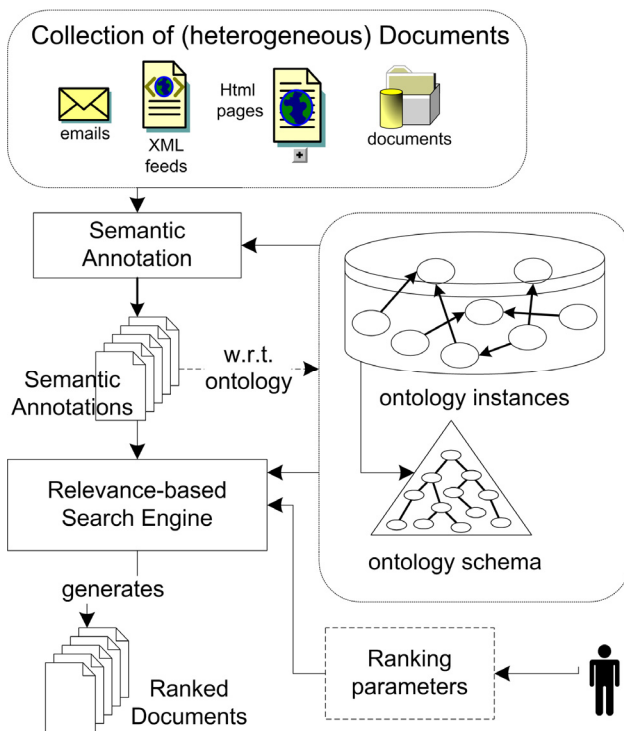


Figure 1. System Architecture.

The Second, Third, and Fourth items of the multi-step process before mentioned have been looked into in an

application where I addressed how semantics can help in the Document-Access problem of Insider Threat [1]. The utilization of a collection of documents and its corresponding semantic annotations is an aspect where we relied upon existing technologies, in particular, the Semagix Freedom toolkit. Freedom is based on technology developed at and licensed from the LSDIS Lab [28]. The Semantic Enhancement Engine [17] of Freedom was used for automatic semantic annotation of a small (i.e., 1K) collection of documents. The indexing of these documents included keeping track of the named-entities spotted by the semantic annotation process. For this application, the retrieval of relevant documents involves a very specific measure of relevance. An extended description [5] of the Insider Threat application includes a more detailed description of ranking of documents.

The fifth item in the multi-step process involves the ranking of documents by considering their relevance to a query. This requires analysis of relationships, namely, semantic associations. One of my earliest efforts in this respect was that of ranking semantic associations, where a user-defined context was used to determine relevance [3]. The feedback we obtained from a prototype demo [16] of this ranking technique led us to look further into measuring relevance using semantic associations. Thus, extended work on ranking of semantic associations included evaluations by human subjects as well as a revised ranking formula [2] (and also an updated demo: <http://lstdis.cs.uga.edu/projects/semdis/rankingAH/>). Related work by our colleagues investigated the ranking of semantic associations in the criteria of discovery vs. conventional mode, which was done by considering rare vs. common appearances of relationships in a populated ontology [7].

All these efforts provided insight on how to measure relevance by analyzing relationships among entities in documents. The proposed solution for search and ranking of documents relies upon the before mentioned multi-step process. However, the core component is concerned with how to determine relevance using semantic associations. My previous work on ranking semantic associations involved finding the list of these associations and then ranking them. The proposed solution, on the other hand, is based on measuring relevance during the discovery of semantic associations. The intuition is to look at how the relevance of a sequence of relationships grows or diminishes when considering the next relationship in the sequence. In some cases, it makes sense to continue 'discovering' an increasingly longer semantic association. In other cases, it makes sense to discard the current semantic association.

The challenging part is how to incorporate human judgment into an algorithm that could guarantee certain

relevance (i.e., above some threshold). For example, an entity representing city ‘Pasadena’ is related to a state ‘California,’ but also to a country, ‘USA.’ Also consider another city, ‘Singapore,’ which is related to a country ‘Singapore.’ These entities are themselves related by relationships such as ‘located in,’ and ‘makes business with.’ From a geo-spatial perspective, relationships such as ‘located-in’ would be more relevant than in a business perspective. For this reason, my approach makes use of subjective knowledge by a human expert. That is, the associations are considered ‘relevant’ depending upon relevance weights determined by a human expert. One of the key elements in my approach is that such weights are assigned just once and regular users do not have to be concerned about it. Another key aspect is that relevance-threshold guarantees can be part of the algorithm that computes relevance.

My preliminary results involve finding relevant semantic associations containing up to five relationships (i.e., a path of max. 5 edges). The traditional algorithms for discovery of semantic associations that we developed for other applications [26] found hundreds of semantic associations. However, with the preliminary techniques for determining relevance on the fly, the number of results is drastically reduced, in most cases, to less than ten semantic associations.

### 3. Outstanding Future Work

What remains is to demonstrate the effectiveness of the approach. Several aspects can be improved. However, the algorithm for determining relevance is the current focus. This involves further testing of algorithms that process the ontology as a graph using one of the available systems for processing data of ontologies (i.e., BRAMS [19]). Graph-based algorithms help analysts of information to understand relationships between participants of an event, or activity [9].

The ontology being used is a subset of SWETO that contains entities such as computer science articles, authors and publication venues. Some improvements on this dataset have build upon an ontology created from data of DBLP (<http://www.semanticweb.org/library/>). DBLP is a bibliographic dataset containing over 400,000 publications. Thus, the preliminary evaluations of my approach have been done in the Computer Science publications domain. Other improvements on the dataset call for the addition of more data sources such as the ACM Digital Library and arXiv archives.

Evaluation of the approach requires a combination of experimental results and comparisons with respect to related work. However, some evaluations include subjective measures, such as those involving human subjects. The goal is to show how the methods proposed here genuinely add value compared to search features of

existing systems such as Google Scholar and Citeseer. In addition, the comparisons could also include desktop search applications such as those offered by Yahoo!, Google, and Microsoft. Other aspects of evaluation include comparison this approach with those of query expansion.

The main queries to be tested involve a search engine-like query. The benefits on this technique can be utilized to re-rank results from a search engine, or to filter out non-relevant results depending upon a threshold level. Additional query constrains can potentially provide more precise search results. Examples of this are: including not/and operators, referencing classes or relationships in an ontology, explicitly indicating other entities important to the query yet not required to appear within the result set, and the relevance to a pre-defined user context. The idea of a user context is to capture more accurately the focus of the search. This idea has been mentioned in literature [10, 20] yet it has not gained much attention by major search engines.

Indexing of entities and semantic annotations of documents is another aspect where there is room for improvement in order to show that this approach can be scalable. It can be argued that the applicability of this approach is limited given that an ontology in general, has a specific and narrow domain. Thus, scalability issues could be dependant upon the domain of the ontology. Additionally, the size of the ontology could have an impact on the performance of the algorithm that computes relevance. However, the current design considers the before mentioned relevance threshold as a means to deal with decreasing performance. That is, this threshold could be updated dynamically and even modified to include conditions such as maximum number of nodes found and maximum length for semantic associations.

### 4. Discussion and Contributions

This approach depends upon a populated ontology. Although techniques for creating ontologies have improved, the creation of ontologies could be said to be a ‘weak’ link. This approach could be negatively affected by an ontology that is not complete in its domain (i.e., low-quality ontology). Measures of ontology quality [29] can serve as a guide to select a good ontology. It is also important to note that this approach also could be negatively affected by a semantic annotation process that produces poor results. However, this issue is part of the problem of entity disambiguation (also called name reconciliation), which continues to be an important research problem.

In my experience, the value of ontologies resides on a few factors: it should contain a large number of instances; these instances should be interconnected because their value lies on the context given by the relationships they

have with other entities; and, the ontology should be easy to maintain and keep up to date. In my experience building an ontology from DBLP dataset, these before mentioned factors are materialized because DBLP contains a very large number of entities, these entities are interconnected through co-authorship and through publication venues, and the semi-structure nature of this dataset allows for relatively easy data extraction or data conversion from the XML file(s) that they make available when their data is updated.

The main contributions of this work are on applying semantic analytics techniques for identifying (i.e., search and ranking) relevant documents. Different scenarios will be considered in evaluations, intended to highlight potential impact areas. That is, it is necessary to identify the applicability of this approach in related techniques such as semantic similarity, targeted advertisement and recommender systems.

The contributions can be summarized as follows. (i) A flexible approach for ranking of semantic associations based on various ranking criteria identifies the most interesting semantic associations between two entities within a populated ontology. The main criterion was that of utilizing a user-defined context to measure relevant associations. (ii) An expanded notion of such context (using classes and relationships in an ontology) identifies related documents in a prototype application that addresses the Document-Access problem of insider Threat. These two aspects set the stage for exploiting semantic relationships in the problem of searching and ranking documents. A collection of documents can be viewed through the lenses of a large populated ontology containing named-entities whereby a semantic annotation process identifies these named-entities within the documents. Thus, the third and pending contribution is as follows. (iii) A relevance measure that exploits semantic relationships for search and ranking of documents.

## 5. References

- [1] Aleman-Meza, B., Burns, P., Eavenson, M., Palaniswami, D. and Sheth, A.P., An Ontological Approach to the Document Access Problem of Insider Threat. In *IEEE International Conference on Intelligence and Security Informatics (ISI-2005)*, (Atlanta, Georgia, USA, 2005).
- [2] Aleman-Meza, B., Halaschek-Wiener, C., Arpinar, I.B., Ramakrishnan, C. and Sheth, A.P. Ranking Complex Relationships on the Semantic Web. *IEEE Internet Computing*, 9 (3). 37-44.
- [3] Aleman-Meza, B., Halaschek, C., Arpinar, I.B. and Sheth, A., Context-Aware Semantic Association Ranking. In *First International Workshop on Semantic Web and Databases*, (Berlin, Germany, 2003), 33-50.
- [4] Aleman-Meza, B., Halaschek, C., Sheth, A., Arpinar, I.B. and Sannapareddy, G., SWETO: Large-Scale Semantic Web Test-bed. In *16th International Conference on Software Engineering and Knowledge Engineering (SEKE2004): Workshop on Ontology in Action*, (Banff, Canada, 2004), 490-493.
- [5] Aleman-Meza, B., Sheth, A.P., Palaniswami, D., Eavenson, M. and Arpinar, I.B. Semantic Analytics in Intelligence: Applying Semantic Association Discovery to Determine Relevance of Heterogeneous Documents. In Siau, K.L. ed. *Advanced Topics in Database Research*, Idea Group Publishing, 2006 (In Print).
- [6] Anyanwu, K. and Sheth, A.P., r-Queries: Enabling Querying for Semantic Associations on the Semantic Web. In *Twelfth International World Wide Web Conference*, (Budapest, Hungary, 2003), 690-699.
- [7] Anyanwu, K., Sheth, A.P. and Maduko, A., SemRank: Ranking Complex Relationship Search Results on the Semantic Web. In *14th International World Wide Web Conference*, (Chiba Japan, 2005), 117-127.
- [8] Chen, H., Lynch, K.J., Basu, K. and Ng, T.D. Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval. *IEEE Intelligent Systems*, 8 (2). 25-35.
- [9] Coffman, T., Greenblatt, S. and Marcus, S. Graph-based Technologies for Intelligence Analysis. *Communications of the ACM*, 47 (3). 45-47.
- [10] Coutaz, J., Crowley, J.L., Dobson, S. and Garlan, D. Context is key. *Communications of the ACM*, 48 (3). 49-53.
- [11] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R.V., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A. and Zien, J.Y., SemTag and Seeker: Bootstrapping the Semantic Web Via Automated Semantic Annotation. In *Twelfth International World Wide Web Conference*, (Budapest, Hungary, 2003), 178-186.
- [12] Ding, L., Kolari, P., Finin, T., Joshi, A., Peng, Y. and Yesha, Y., On Homeland Security and the Semantic Web: A Provenance and Trust Aware Inference Framework. In *AAAI Spring Symposium on AI Technologies for Homeland Security*, (Stanford University, California, USA, 2005).
- [13] Graupmann, J., Schenkel, R. and Weikum, G., The SphereSearch Engine for Unified Ranked Retrieval of Heterogeneous XML and Web Documents. In *31st International Conference on Very Large Data Bases*, (Trondheim, Norway, 2005), 529-540.
- [14] Guha, R., McCool, R. and Fikes, R., Contexts for the Semantic Web. In *International Semantic Web Conference*, (Hiroshima, Japan, 2004), 32-46.
- [15] Guha, R., McCool, R. and Miller, E., Semantic Search. In *Twelfth International World Wide Web Conference*, (Budapest, Hungary, 2003).
- [16] Halaschek, C., Aleman-Meza, B., Arpinar, I.B. and Sheth, A.P., Discovering and Ranking Semantic Associations over a Large RDF Metabase. In *30th International Conference on Very Large Data Bases*, (Toronto, Canada, 2004).
- [17] Hammond, B., Sheth, A. and Kochut, K. Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In Kashyap, V. and Shklar, L. eds. *Real World Semantic Web Applications*, Ios Press Inc, 2002, 29-49.
- [18] Handschuh, S., Staab, S. and Studer, R. Leveraging Metadata Creation for the Semantic Web with CREAM. *Ki 2003: Advances in Artificial Intelligence*, 2821. 19-33.
- [19] Janik, M. and Kochut, K., BRAHMS: A WorkBench RDF Store And High Performance Memory System for Semantic

- Association Discovery. In *Fourth International Semantic Web Conference*, (Galway, Ireland, 2005).
- [20] Lawrence, S. Context in Web Search. *IEEE Data Engineering Bulletin*, 23 (3). 25-32.
- [21] Mihalcea, R.F. and Mihalcea, S.I., Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web. In *13th IEEE International Conference on Tools with Artificial Intelligence*, (Dallas, Texas, USA, 2001), 280-287.
- [22] Mukherjea, S. and Bamba, B., BioPatentMiner: An Information Retrieval System for BioMedical Patents. In *Thirtieth International Conference on Very Large Data Bases*, (Toronto, Canada, 2004), 1066-1077.
- [23] Rocha, C., Schwabe, D. and Aragao, M.P., A Hybrid Approach for Searching in the Semantic Web. In *13th International World Wide Web*, (New York, New York, USA, 2004), 374-383.
- [24] Sengupta, A., Dalkilic, M. and Costello, J., Semantic Thumbnails: A Novel Method for Summarizing Document Collections. In *22nd Annual International Conference on Design of Communication: The Engineering of Quality Documentation*, (Memphis, Tennessee, USA, 2004), 45-51.
- [25] Shah, U., Finin, T., Joshi, A., Cost, R.S. and Mayfield, J., Information Retrieval on the Semantic Web. In *10th International Conference on Information and Knowledge Management*, (McLean, Virginia, USA, 2002), 461-468.
- [26] Sheth, A.P., Aleman-Meza, B., Arpinar, I.B., Halaschek, C., Ramakrishnan, C., Bertram, C., Warke, Y., Avant, D., Arpinar, F.S., Anyanwu, K. and Kochut, K. Semantic Association Identification and Knowledge Discovery for National Security Applications. *Journal of Database Management*, 16 (1). 33-53.
- [27] Sheth, A.P., Arpinar, I.B. and Kashyap, V. Relationships at the Heart of Semantic Web: Modeling, Discovering and Exploiting Complex Semantic Relationships. In Nikraves, M., Azvin, B., Yager, R. and Zadeh, L.A. eds. *Enhancing the Power of the Internet Studies in Fuzziness and Soft Computing*, Springer-Verlag, 2003.
- [28] Sheth, A.P., Bertram, C., Avant, D., Hammond, B., Kochut, K. and Warke, Y. Managing Semantic Content for the Web. *IEEE Internet Computing*, 6 (4). 80-87.
- [29] Tartir, S., Arpinar, I.B., Moore, M., Sheth, A.P. and Aleman-Meza, B. OntoQA: Metric-Based Ontology Quality Analysis *IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*, Houston, TX, USA, 2005.
- [30] Townley, J. The Streaming Search Engine That Reads Your Mind *Streaming Media World*, 2000.
- [31] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A. and Ciravegna, F., MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. In *13th International Conference on Knowledge Engineering and Management (EKAW 2002)*, (Sigüenza, Spain, 2002), Springer Verlag.