On Boosting Semantic Web Data Access *

Li Ding
Department of CSEE, University of Maryland Baltimore County
1000 hilltop circle, Baltimore, MD 21250, USA
dingli1@cs.umbc.edu

Introduction

The Semantic Web is a promising framework for agents (especially software agents) to represent and share knowledge throughout the Web; hence its utility should be evaluated in three equally important aspects: **availability** – is there enough relevant data published in the Semantic Web, **accessibility** – can users discover and obtain the data they need effectively, and **quality** – can users evaluate the data's quality to select good information.

The first aspect *availability* depends on the adoption of semantic web technology. Our preliminary work – Swoogle (Ding *et al.* 2004) has discovered over 346,126 RDF documents in the Web covering different topics such as *personal profile* (FOAF), *news feeds* (RSS), *digital library* (Dublin Core), *copyright protection* (Creative Commons), *dictionary* (WordNet 1.6), and *scientific data sharing* (e.g. California Invasive Species Information Catalog). Steady growth rate of semantic web data is also observed in Swoogle Statistics[1]. We attribute such growth to industry adoptions which use (semi)automatic tools (Dill *et al.* 2003) to translate data from database or text into semantic web data. Based on these observations, we assume that *availability* issue is addressed and we will finally facing a very large semantic web on the Web (or within an enterprise's intranet).

The latter two important aspects (accessibility and quality), however, remain neglected by semantic web research. This dissertation pursues these two issues and proposes an ontology based approach: building ontologies and metadata for modeling the Semantic Web and its context (i.e. *the Web* that stores semantic web data and *the agents* who produce and consume semantic web data); and using the metadata and ontologies to address the following challenges:

- **semantic web vocabulary** is published in ontologies throughout the Web by many authors; however, its distributed and dialectic features (e.g. 'name' is defined under 560 distinctive namespaces) made it hard to even compose an effective RDF graph query. In addition, users may be interested in the definition and usage of URIrefs before using them in applications.
- **semantic web data access** remains in small scale, i.e. only in-memory or database data access tools are available; however scalability issues always exist when using the semantic web data from the entire web. Using metadata for RDF documents may leverage this situation.
- **semantic web data quality** becomes an important issue because anyone can publish any statements about any resource in RDF graph; however, related works remain in early stage – effective evaluation mechanisms (e.g. ranking RDF resources and judging RDF graph trustworthiness) are needed.

Research Overview

The Web Of Belief (WOB) ontology family models the Semantic Web and its context in three interactive worlds: the Web, the RDF graph world, and the agent world as shown in figure 1. *WOB-core* ontology has been developed to cover nodes and relations in figure 1 except trust. Ongoing work includes *WOB-graph* ontology for referencing an arbitrary RDF graph efficiently without extending RDF syntax[2]; *WOB-swoogle* for Swoogle's semantic web metadata; and *WOB-trust* ontologies for user's trust knowledge.

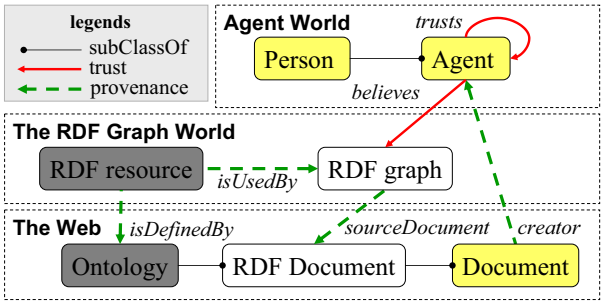

Figure 1: The Semantic Web and its Context

Figure 2 shows our proposed approach to semantic web data access activity: agent users first compose query by consulting data access service and then build local RDF graph

*Partial support for this research was provided by DARPA contract F30602-00-0591 and by NSF awards NSF-ITR-IIS-0326460 and NSF-ITR-IDM-0219649. Special thanks go to Dr. Tim Finin and Rong Pan for their great contribution to Swoogle development.

[1]http://swoogle.umbc.edu/modules.php?name=Swoogle_Statistics

[2]This feature distinct WOB-graph from *RDF reification* (Hayes 2004) and *Named Graphs* (Carroll *et al.* 2004)

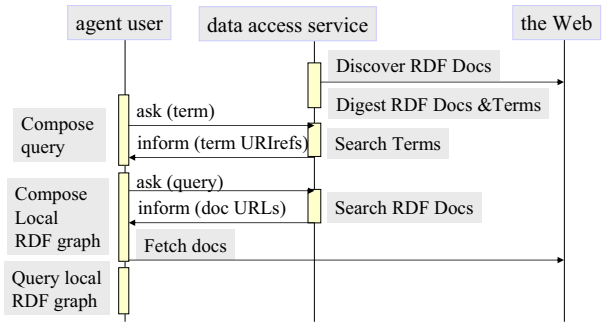

Figure 2: A semantic web data access process

by searching relevant RDF documents through data access service. In order to keep a global view about the semantic web, we have built crawler for discovering RDF document in the web and then collecting metadata for RDF documents and ontological terms. We also provide corresponding search services in Swoogle (Ding *et al.* 2004). Ongoing work is highlighted by the following: (i) *Ontology dictionary* for managing ontologies at term level users may construct/use ontologies by selecting relevant classes and properties without importing the entire ontologies that defines the term; and (ii) *semantic web navigation model* for capturing users' browsing behavior in the semantic web – users enter the semantic web world through document/term search and navigate in the semantic web using seven categories of arcs. An advanced feature of ontology dictionary is its support to ontology emergence through reverse engineering the domain/range relation between classes and properties. Besides, we also proposed *RDF sitemap* for publishing RDF documents and adding RDF document level links; and *provenance service* for tracking the provenance of each RDF graph at "molecule" level and adding document-resource relations. Since centralized service model is vulnerable to single point failure, future work will explore P2P architecture for maintaining the collected metadata and providing data access service.

The byproduct of semantic web data access service is to survey the semantic web in the Web and produce statistical reports. One metric is the amount of publicly available RDF document from web. An interesting metric is the statistics of last-modified time of collected RDF documents, which partially quantifies the deployment of semantic web. We also need to know the distribution of the content of semantic web data.

This dissertation also starts investigating data quality (Wang, Storey, & Firth 1995) issues in the Semantic Web from two aspects: (i) ontological aspect, i.e., studying the dimensions of data quality for different concepts in WOB and proposed ontologies for representing users quality judgments (esp. trust judgments) explicitly; (ii) computational aspect, i.e. evaluating data quality using heuristics, especially context analysis. Ongoing work is highlighted as the following: (i) ranking importance of RDF resource and RDF documents based on *semantic web navigation model*, which remarks the fact that RDF documents are always connected through the usage/definition of RDF resources but seldom connected by direct links; and (ii) evaluating trustworthiness of a given RDF graph, which is studied by content based approaches such as RDF graph difference (Berners-Lee & Connolly 2004) and context based approaches such as information security (Hyvonen 2002), trust network (Golbeck, Parsia, & Hendler 2003; Richardson, Agrawal, & Domingos 2003). A by-product of trust analysis is a trust based semantic web navigation model based on (Ding, Zhou, & Finin 2003). Currently, we only sketched some primitive thoughts, future work will first implement and evaluate the proposed algorithms, then investigate the applicability of other uncertainty reasoning methods, and draw a better picture of semantic web data quality issues.

References

Berners-Lee, T., and Connolly, D. 2004. Delta: an ontology for the distribution of differences between rdf graphs. http://www.w3.org/DesignIssues/Diff.

Carroll, J.; Bizer, C.; Hayes, P.; and Stickler, P. 2004. Named graphs, provenance and trust. Technical report, HP Lab.

Dill, S.; Eiron, N.; Gibson, D.; Gruhl, D.; Guha, R.; Jhingran, A.; Kanungo, T.; Rajagopalan, S.; Tomkins, A.; Tomlin, J. A.; and Zien, J. Y. 2003. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *The Twelfth International World Wide Web Conference (WWW2003)*.

Ding, L.; Finin, T.; Joshi, A.; Pan, R.; Cost, R. S.; Peng, Y.; Reddivari, P.; Doshi, V. C.; and Sachs, J. 2004. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the 13th CIKM*.

Ding, L.; Zhou, L.; and Finin, T. 2003. Trust based knowledge outsourcing for semantic web agents. In *IEEE/WIC 2003*.

Golbeck, J.; Parsia, B.; and Hendler, J. 2003. Trust networks on the semantic web. In *Proceedings of Cooperative Intelligent Agents*.

Hayes, P. 2004. Rdf semantics (w3c recommendation). http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.

Hyvonen, E. 2002. The semantic web – the new internet of meanings. In *Semantic Web Kick-Off in Finland: Vision, Technologies,Research, and Applications*.

Richardson, M.; Agrawal, R.; and Domingos, P. 2003. Trust management for the semantic web. In *the 2nd ISWC*.

Wang, R.; Storey, V.; and Firth, C. 1995. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering* 7(4):623–639.