Boosting Semantic Web Data Access Using Swoogle *

Li Ding and **Tim Finin**
Department of CSEE, University of Maryland Baltimore County
1000 hilltop circle, Baltimore, MD 21250, USA
{dingli1,finin}@cs.umbc.edu

Introduction

One of the unique advantages brought by the Semantic Web is that semantic web languages, such as RDF and OWL, offer a small but expressive set of common ontological constructs for agents to share knowledge on the Web. Instead of hard coding knowledge inside intelligent agents, the semantic web enables agents to publish and consume knowledge explicitly stored in web documents. The utility of the Semantic Web can be evaluated from three equally important aspects: *availability* – is there enough useful data, *accessibility* – can users find the data they need, and *quality* – can users evaluate data quality to select good information.

Survey of the Semantic Web

One of our ongoing researches is to survey the actual semantic web on the Web so as to estimate *data availability*. To this end, we are developing metrics for characterizing the semantic web deployment status, adaptive crawlers for discovering RDF documents in the Web, and tools for analyzing semantic web data.

The size of semantic web can be measured by the amount of web documents containing RDF graph and the amount of triples. (Eberhart 2002) reported 1,479 RDF documents with 254,783 triples out of 2,952,010 web documents. Recently, Swoogle (Ding *et al.* 2004) has discovered 346,126 RDF documents with 65,747,150 triples. Although this number is still trivial in comparison with 8,058,044,651 web pages indexed by Google, it is a big number of semantic web researchers (Guo, Pan, & Heflin 2004). The observed rapid growth rate of semantic web data is partially guaranteed by (semi)automatic tools (Dill *et al.* 2003) for translating data in database and text into semantic web data.

Large portion of semantic web data came from industry adoptions, e.g. FOAF personal profiles, RSS news feeds, the embedded RDF metadata in PDF files, Dublin Core metadata for digital library, Creative Commons' copyright statements, linux configuration (trustix.com), CIA world fact book and etc. Another important source of semantic web data is ontology, e.g. *upper ontologies* (OpenCyc, IEEE's SUMO), *dictionary and thesauri* (WordNet, SKOS), OWL-S web service ontology, and SWRL rule ontology. While industry focuses on a narrow spectrum of semantic web data, academic community proposes many ontologies for a great variaty of fields (Noy & Hafner 1997).

Web-Scale Semantic Web Data Access

The Semantic Web has reduced the dependency between publishers and consumers; hence *accessibility* issue rises when consumers need to find the data they want from the vast distributed semantic web. For example, how could a user translate her concept into URIrefs and thus compose a query. Even with the composed query, consumers may have no clue about the URLs of RDF documents that can be used to answer the query; hence effective *semantic web data access* services are in great need.

We have developed Swoogle, which discovers, digests and searches the Semantic Web in the Web to address the above issues. Swoogle differs from *ontology based annotation systems* such as SHOE (Luke *et al.* 1997), Ontobroker (Decker *et al.* 1999), EDUTELLA (Nejdl *et al.* 2002) and CREAM (Handschuh & Staab 2003) in its focus on creating metadata for online RDF documents and semantic web vocabulary; it differs from *ontology repositories* such as DAML Ontology Library (daml 2004) (indexed 282 ontologies) and Schema Web (schemaweb 2004)(indexed 202 ontologies) in its automated ontology discovery mechanism which has found over 4,000 ontologies from the Web; and it differs from *W3C's Ontaria* (ontaria 2004) in its rich search/navigation mechanisms which are highlighted by "swoogle search" and "ontology dictionary". Currently Swoogle manages its metadata using a centralized approach; P2P metadata management as in EDUTELLA could be a promising direction for future scalability and efficiency study.

Semantic Web Navigation Model and Ranking

Since the semantic web is composed of many RDF graphs distributed throughout the web, consumers may need to evaluate *data quality* (Wang, Storey, & Firth 1995) (e.g. relevance, accuracy, and trustworthiness) before using data. We

*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. Full version available at http://ebiquity.umbc.edu/v2.1/get/a/resource/76.pdf. Special thanks go to Rong Pan for his contribution to Swoogle development, especially *Ontology Rank* algorithm.

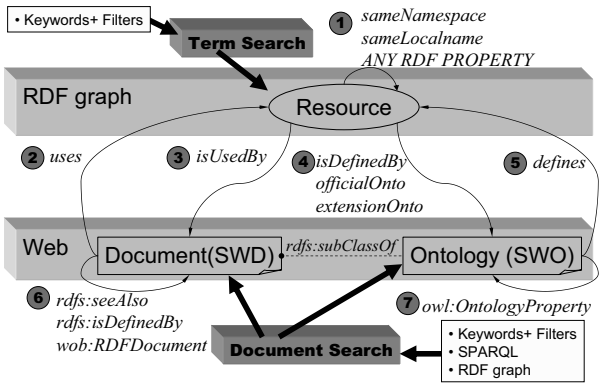

Figure 1: Semantic web search/navigation model

pursue one important branch of data quality analysis – context, i.e. *the Web* that stores semantic web data and *the agents* who produce and consume semantic web data.

To this end, we first build the Web Of Belief (WOB) ontology family that models the Semantic Web and its context with the entities and relations in three interactive worlds, namely the Web, the RDF graph world, and the agent world.

Based on WOB model and Swoogle services, we proposed a novel semantic web search/navigation model as shown in figure 1, where users can either enter the Semantic Web by document/term search or navigate the Semantic Web via seven categories of relations: (i) inter-resource relation that is derived from RDF graph or sharing namespace or local-name; (ii) usage and definition provenance that relate resources with documents; (iii) inter-document relations such as owl:imports.

This model acknowledges the fact that RDF documents are always connected through the usage/definition of RDF resources but seldom connected by direct links; hence, the corresponding ranking model differs from existing web document ranking models (e.g. PageRank, HITS) which use hyperlinks among web documents, and semantic web ranking models (Patel *et al.* 2003; Ding *et al.* 2004) which only consider document level relations. As the first step towards the ultimate ranking model, we suggest *TermRank* SWTs as shown in equation 1. The intuition is to split the rank of SWDs to their populated terms. The weight is computed proportional to the term frequency within the document (i.e., $cnt\_uses(d,t)$ which shows how many times the term t is used in the document d), and inverse term frequency (i.e. $|\{d|uses(d,t)\}|$ which shows how many documents have used the term t).

$$\begin{aligned} TermRank(t) &= \sum_{uses(d,t)} \frac{OntoRank(d) \times Weight(d,t)}{\sum_{uses(d,x)} Weight(d,x)} \\ Weight(d,t) &= cnt\_uses(d,t) \times |\{d|uses(d,t)\}| \end{aligned} \quad (1)$$

Conclusion

This work provides comprehensive ontologies as well as effective tools to boost accessibility and quality factors in semantic web data access. Its practical contributions lie in the deployment of Swoogle, which is among the first metadata and search services for the semantic web. Its theoretical contributions include the WOB ontology family which first model the Semantic Web and its context comprehensively, the novel semantic web navigation models and corresponding ranking mechanisms. This work is still in its preliminary stage and we will refine, complete, and evaluate prototypes in the future.

References

2004. Daml ontology library. http://www.daml.org/ontologies/.

Decker, S.; Erdmann, M.; Fensel, D.; and Studer, R. 1999. Ontobroker: Ontology based access to distributed and semi-structured information. In *DS-8*, 351–369.

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 *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 *CIKM2004*.

Eberhart, A. 2002. Survey of rdf data on the web. Technical report, International University in Germany.

Guo, Y.; Pan, Z.; and Heflin, J. 2004. An evaluation of knowledge base systems for large owl datasets. In *International Semantic Web Conference*, 274–288.

Handschuh, S., and Staab, S. 2003. Cream: Creating metadata for the semantic web. *Comput. Networks* 42(5).

Luke, S.; Spector, L.; Rager, D.; and Hendler, J. 1997. Ontology-based web agents. In *Proceedings of the First International Conference on Autonomous Agents*.

Nejdl, W.; Wolf, B.; Qu, C.; Decker, S.; Sintek, M.; Naeve, A.; Nilsson, M.; Palmer, M.; and Risch, T. 2002. Edutella: a p2p networking infrastructure based on rdf. In *WWW2002*, 604–615.

Noy, N. F., and Hafner, C. D. 1997. The state of the art in ontology design: A survey and comparative review. *AI Magazine* 18(3):53–74.

2004. Ontaria. http://www.w3.org/2004/ontaria/.

Patel, C.; Supekar, K.; Lee, Y.; and Park, E. K. 2003. Ontokhoj: a semantic web portal for ontology searching, ranking and classification. In *WIDM '03: Proceedings of the 5th ACM international workshop on Web information and data management*, 58–61.

2004. Schema web. http://www.schemaweb.info/.

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.