Swoogle: Searching for knowledge on the Semantic Web [*]

Tim Finin, Li Ding, Rong Pan, Anupam Joshi, Pranam Kolari, Akshay Java and Yun Peng
University of Maryland Baltimore County, Baltimore, MD

Introduction

Most knowledge on the Web is encoded as natural language text, which is convenient for human users but very difficult for software agents to understand. Even with increased use of XML-encoded information, software agents still need to process the tags and literal symbols using application dependent semantics. The Semantic Web offers an approach in which knowledge can be published by and shared among agents using symbols with a well defined, machine-interpretable semantics.

The Semantic Web is a "web of data" in that (i) both ontologies and instance data are published in a distributed fashion; (ii) symbols are either 'literals' or universally addressable 'resources' (URI references) each of which comes with unique semantics; and (iii) information is semi-structured. The Friend-of-a-Friend (FOAF) project (http://www.foaf-project.org/) is a good application of the Semantic Web in which users publish their personal profiles by instantiating the *foaf:Person* class and adding various properties drawn from any number of ontologies.

The Semantic Web's distributed nature raises significant data access problems – how can an agent discover, index, search and navigate knowledge on the Semantic Web? Swoogle (Ding *et al.* 2004) was developed to facilitate web-scale semantic web data access by providing these services to both human and software agents. It focuses on two levels of knowledge granularity: URI based *semantic web vocabulary* and *semantic web documents* (SWDs), i.e., RDF and OWL documents encoded in XML, NTriples or N3.

Figure 1 shows Swoogle's architecture. The **discovery** component automatically discovers and revisits SWDs using a set of integrated web crawlers. The **digest** component computes metadata for SWDs and *semantic web terms* (SWTs) as well as identifies relations among them, e.g., "an SWD instantiates an SWT class", and "an SWT class is the domain of an SWT property". The **analysis** component uses cached SWDs and their metadata to derive analytical reports, such as classifying ontologies among SWDs and ranking SWDs by their importance. The **service** component supports both human and software agents through conventional web interfaces and SOAP-based web service APIs. Two key services are (i) a *swoogle search* service that searches for SWDs by constraints on their URLs, the sites which host them, and the classes/properties used or defined by them and (ii) a *ontology dictionary* service that searches for SWTs and their relationships with other SWTs and SWDs.

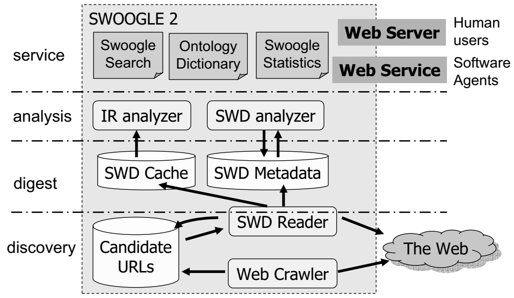

Figure 1: Swoogle has four components that discover, digest, analyze and serve semantic web data.

Discovering Semantic Web Documents

The size of the Semantic Web is measured by the number of discovered SWDs. (Eberhart 2002) reported finding 1,479 SWDs with about 255K triples out of nearly 3M web pages. As of May 2005, Swoogle has found over 368K SWDs with more than 70M triples. Although this number is far less than Google's eight billion web pages, it represents a non-trivial collection of semantic web data (Guo, Pan, & Heflin 2004).

The Semantic Web's content can be divided into two categories – program generated instance data and (mostly) hand crafted ontologies. The first category is the larger and includes FOAF personal profiles, RSS news feeds, RDF metadata embedded in PDF files, Dublin Core digital library metadata, Creative Commons' copyright statements, and assertions extracted from structured data sources such as WordNet and the CIA fact book. While some ontologies have been derived from structured sources, most appear to be designed by semantic web researchers. Although these ontology documents are far outnumbered by instance data documents, they are critically important since they convey symbol semantics.

[*] Research support was provided by DARPA contract F30602-00-0591 and NSF awards NSF-ITR-IIS-0326460 and NSF-ITR-IDM-0219649.

Navigating and Ranking SWDs and SWTs

Since semantic web data is highly distributed, facilitating *data access* and assessing *data quality* (Wang, Storey, & Firth 1995) are challenging. For example, how can users find relevant domain ontologies and then choose a popular and trustworthy one for use? To this end, we start with modeling navigational paths in the Semantic Web and then ranking the importance of objects in the Semantic Web.

Swoogle's services provide agents with the semantic web search and navigation framework modeled in figure 2. This model is defined by the links and paths within the Semantic Web and differs from conventional web navigation model in that it considers the interactions between two levels of abstraction: the RDF graph and the web of SWDs.

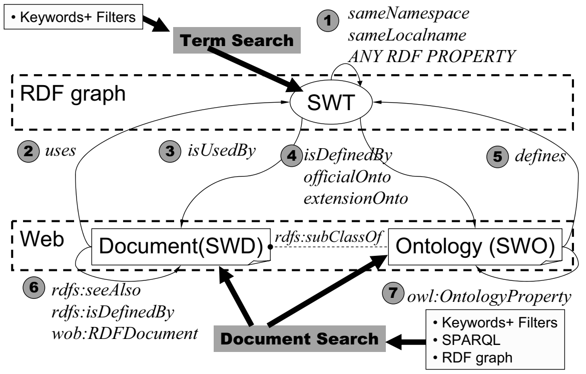

Figure 2: Agents access the Semantic Web using document/term search and navigate within it via three kinds of paths: *inter-resource* paths (1) enhance links between SWTs in RDF graph by additionally linking SWTs sharing a namespace or "local name"; *resource-document* paths (2,3,4,5) provide provenance (usage or definition) links between SWTs and SWDs; and *inter-document* paths (6,7) manifest explicit links between SWDs.

Our model gives rise to semantic web ranking metrics that differ from those used in web ranking (e.g., PageRank, HITS), which used only hyperlinks among web pages, and other semantic-aware ranking methods (Patel *et al.* 2003), which use a small set of document level semantic relations.

OntoRank is grounded on the *rational surfer model*, which is loosely derived from the *random surfer model* (Page *et al.* 1998). An agent navigates from one SWD to another with a constant probability or jumps to a random SWD. The surfing agent is also 'rational' in that it jumps non-uniformly according to link semantics. Moreover, on encountering an SWD D, the rational surfer **must** transitively import the "official" ontologies defining the terms (classes and properties) used by D in order to fully understand it. Intuitively, *OntoRank* estimates the probability of a *rational surfer* will visit an SWD with the bias that ontologies are more preferred to instance data. In equation 1, let $wPR(a)$ be a weighted PageRank variation, $f(a,b)$ be the sum of tag weight from SWD a to SWD b, d be a constant between 0 and 1, $link(a,l,b)$ be the semantic link from SWD a to SWD b using semantic tag l; $weight(l)$ be user's preference of choosing semantic links with tag l; $OTC(a)$ be a set of SWDs that (transitively) import a as ontology.

$$\begin{aligned}
&OntoRank(a) = wPR(a) + \sum_{x\in OTC(a)} wPR(x)\\
&wPR(a) = (1-d) + d \sum_{link(x,\_,a)} \frac{wPR(x)\times f(x,a)}{\sum_{link(x,\_,y)} f(x,y)} \qquad (1)\\
&f(a,b) = \sum_{link(a,l,b)} weight(l)
\end{aligned}$$

TermRank ranks the SWTs found on the Semantic Web and is defined by equation 2. Intuitively, we divide the rank of an SWD among the SWTs it uses. Given a term T and an SWD d, $TWeight(t,d)$ is computed from $cnt\_uses(d,t)$, which reflects how many times d uses t, and $|\{d|uses(d,t)\}|$, which shows how many discovered SWDs use t.

$$\begin{aligned}
&TermRank(t) = \sum_{uses(d,t)} \frac{OntoRank(d)\times TWeight(d,t)}{\sum_{uses(d,x)} TWeight(d,x)} \qquad (2)\\
&TWeight(d,t) = cnt\_uses(d,t) \times |\{d|uses(d,t)\}|
\end{aligned}$$

Conclusion

Swoogle is an implemented system that discovers, analyzes and indexes knowledge encoded in semantic web documents on the Web. Swoogle reasons about these documents and their constituent parts (e.g., terms and triples) and records meaningful metadata about them. Swoogle provides web-scale semantic web data access service, which helps human users and software systems to find relevant documents, terms and triples, via its search and navigation services. Swoogle also provides a customizable algorithm inspired by Google's PageRank algorithm but adapted to the semantics and use patterns found in semantic web documents.

References

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 Thirteenth ACM Conference on Information and Knowledge Management*.

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.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Database group.

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.

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.