7   Swoogle

Objective:

This chapter aims at exploring Swoogle which is a research level implementation of the Semantic Web.

7.1   Introduction

The Semantic Web is essentially a web-universe parallel to the web of online documents. Semantic Web Document (SWD) is well known for its semantic annotation and meaningful reference. Since no conventional search engines can take advantage of such features, a search engine customized for SWDs especially for ontologies, is needed by human users as well as software agents and services. At this stage, human users are expected to be Semantic Web researchers and developers who are interested in accessing, exploring and querying the RDF and OWL documents found on the web.

Download free eBooks at bookboon.com

7.2   Swoogle

Swoogle is a search engine for Semantic Web ontologies, documents, terms and data published on the Web. Swoogle employs a system of crawlers to discover RDF documents and HTML documents with embedded RDF content. Swoogle reasons about these documents and their constituent parts (e.g., terms and triples) and records and indexes meaningful metadata about them in its database. Swoogle is a crawler-based indexing and retrieval system for the Semantic Web, i.e., for Web documents in RDF or OWL. It extracts metadata for each discovered document and computes relations between documents. Discovered documents are also indexed by an information retrieval system which can use either character N-Gram or URI refs as keywords to find relevant documents and to compute the similarity among a set of documents. One of the interesting properties we compute is rank, a measure of the importance of a Semantic Web document.

Screenshot 7.1: User Interface of Swoogle.

7.3   History of Swoogle

Swoogle is a research project being carried out by the Ebiquity research group of the Computer Science and Electrical Engineering Department at the University of Maryland, Baltimore County (UMBC). Partial research support was provided by DARPA contract F30602-00-0591 and by NSF, by awards NSF-ITR-IIS-0326460 and NSF-ITR-IDM-0219649. Contributors include Tim Finin, Li Ding, Rong Pan, Anupam Joshi, Pavan Reddivari, Joel Sachs, Pranam Kolari, Akshay Java, Lushan Han, Yun Peng, R. Scott Cost, Sandor Dornbush and Vishal Doshi. Swoogle indexes only Semantic Web documents, currently including those written in RDF/XML, N-Triples, N3(RDF) and some documents that embed RDF/XML fragments. The data presented at Swoogle are collected from the public accessible World Wide Web. Swoogle has a privacy policy in English and a crawling policy.

7.4   Implementation of Swoogle

Swoogle adopts a hybrid approach to harvest the Semantic Web, including manual submission, Google-based meta-crawling, bounded HTML crawling and RDF crawling. Swoogle's statistic page is under re-construction now. Swoogle has over 1.4 million Semantic Web documents and 290 million triples indexed. Swoogle is mainly written in Java (JDK 1.4.2), and its web services are provided through Apache Tomcat server. The front-end of the Swoogle website is written in PHP. Currently, Swoogle stores its data in a MySQL database (mysql 4.1.16). The entire system runs on a Linux platform (Fedor Core 4). Swoogle is designed as a repository of URLs but not as a triple store, so it does not store all triples encountered. However, applications may be built on top of Swoogle that index a specific class of Semantic Web data and provide inference support using triple stores.

Swoogle tries to enhance the following features of the Semantic Web:

Finding appropriate ontologies:

Failing to find a proper ontology always leads to the creation of a new ontology, which is often customized to be reused. Swoogle helps users to find ontologies containing specified terms, and users may even qualify the type (class or property) of a term. Moreover, the ranking mechanism sorts ontologies by their popularity. This feature is believed to not only ease the burden of marking up data but also contribute to the emergence of canonical ontologies.

Finding instance data:

In order to help users integrate Semantic Web data distributed on the Web, Swoogle enables querying SWDs with constraints on the classes and properties used by them.

Characterizing the Semantic Web:

By collecting meta-data, especially inter-document relations, about the Semantic Web, Swoogle reveals interesting structural properties such as 'how the Semantic Web is connected?', 'how ontologies are referenced?', and 'how an ontology is modified externally?'.