Provenance-Aware Semantic Search Engines Based On Data Integration Systems

Domenico Beneventano, Dipartimento di Ingegneria “Enzo Ferrari”, Università degli Studi di Modena e Reggio Emilia, Modena, Italy

Sonia Bergamaschi, Dipartimento di Ingegneria “Enzo Ferrari”, Università degli Studi di Modena e Reggio Emilia, Modena, Italy

ABSTRACT

Search engines are common tools for virtually every user of the Internet and companies, such as Google and Yahoo!, have become household names. Semantic Search Engines try to augment and improve traditional Web Search Engines by using not just words, but concepts and logical relationships. Given the openness of the Web and the different sources involved, a Web Search Engine must evaluate quality and trustworthiness of the data; a common approach for such assessments is the analysis of the provenance of information. In this paper a relevant class of Provenance-aware Semantic Search Engines, based on a peer-to-peer, data integration mediator-based architecture is described. The architectural and functional features are an enhancement with provenance of the SEWASIE semantic search engine developed within the IST EU SEWASIE project, coordinated by the authors. The methodology to create a two level ontology and the query processing engine developed within the SEWASIE project, together with provenance extension are fully described.

Keywords: Data Integration, Data Provenance, Mediator System, Ontologies, Query Reformulation, Schema Based Peer-to-Peer Networks, Semantic Search Engines

INTRODUCTION

Commercial search engines are mainly based upon human-directed search. The human directed search engine technology utilizes a database of keyword, concepts, and references. The keyword searches are used to rank pages, but this simplistic method often leads to voluminous irrelevant and spurious results. Google with its over 2 billion of hits per day, and over 40 billion indexed Web pages, is undeniably the most popular commercial search engine used today, but even with Google, there are problems. For example, how can you find just the right bit of data that you need out of the ocean of irrelevant results provided? A well-know problem is that traditional Web search engines use keywords that are subject to the two well-known linguistic phenomena that strongly degrade a query’s precision and recall:

DOI: 10.4018/ijoci.2014040101
• Polysemy (one word might have several meanings) and
• Synonymy (several words or phrases, might designate the same concept).

Precision and recall are classical information retrieval evaluation metrics. Precision is the fraction of a search output that is relevant for a particular query, i.e., is the ratio of the number of relevant Web pages retrieved to the total number of irrelevant and relevant Web pages retrieved. The recall is the ability system to obtain all or most of the relevant pages, i.e., is the ratio of the number of relevant Web pages retrieved to the total number of relevant pages in the Web.

As Artificial Intelligence (AI) technologies become more powerful, it is reasonable to ask for better search capabilities which can truly respond to detailed requests. This is the intent of semantic-based search engines and agents. A semantic search engine seeks to find documents that have similar concepts not just similar words. In order for the Web to become a semantic network, it must provide more meaningful meta-data about its content, through the use of Resource Description Framework (RDF) (www.w3.org/RDF/) and Web Ontology Language (OWL) (www.w3.org/TR/owl-features) tags which will help to form the Web into a semantic network. In a semantic network, the meaning of content is better represented and logical connections are formed between related information.

Semantic search methods augment and improve traditional search results by using not just words, but concepts and logical relationships (Studer & Yong, 2011).

Several systems have been built based on the idea of annotating Web pages with Resource Description Framework (RDF) and Web Ontology Language (OWL) tags to represent semantics (see Related Work). However, the limitation of these systems is that they can only process Web pages that are already manually annotated with semantic tags and it seems unfeasible to annotate the enormous amount of Web pages. Furthermore, most semantic-based search engines suffer performance problems because of the scale of the very large semantic network. In order for the semantic search to be effective in finding responsive results, the network must contain a great deal of relevant information. At the same time, a large network creates difficulties in processing the many possible paths to a relevant solution.

The requirements for an intelligent search engine are given by a special class of users, small and medium-sized enterprises (SMEs) which are threatened by globalization. One of the keys to sustainability and success is being able to access information. This could be a cheaper supplier, an innovative working method, a new market, potential clients, partners, sponsors, and so on. Current Internet search tools are inadequate because even if they are not difficult to use, the search results are often of little use with their pages and pages of hits. Suppose an SME needs to find out about a topic -a product, a supplier, a fashion trend, a standard, etc. For example, a search is made for fabric dyeing processes for the purpose of finding out about the disposal of the dyeing waste material. A query to www.google.com for fabric dyeing listed more than 10 million hits at the time of writing, which related not only manufacturers of fabric dyeing equipment, but also the history of dyeing, the dyeing technology, and so on. Eventually, a useful contact may be found, and the search can continue for relevant laws and standards concerning waste disposal. But is it law or the interpretation of the law? What if the laws are of a different country where the practices and terminologies are different? Thus, intelligent tools to support the business of SMEs in the Internet age are necessary.

This article is a revised and extended version of the chapter “Semantic Search Engines Based on Data Integration Systems. In J. Cardoso (Ed.), Semantic Web Services: Theory, Tools and Applications (pp. 317-342).” In that chapter we showed that data integration systems, domain ontologies and peer-to-peer architectures are good ingredients for developing Semantic Search Engines with good performance. To this end, we described two projects, SEWASIE and WISDOM, which rely
Congestion Management Using Hybrid Particle Swarm Optimization Technique
Sujatha Balaraman and N. Kamaraj (2012). *Innovations and Developments of Swarm Intelligence Applications* (pp. 165-181).
www.igi-global.com/chapter/congestion-management-using-hybrid-particle/65812?camid=4v1a

Theoretical Derivations and Application Issues
E. Parsopoulos Konstantinos and N. Vrahatis Michael (2010). *Particle Swarm Optimization and Intelligence: Advances and Applications* (pp. 42-87).
www.igi-global.com/chapter/theoretical-derivations-application-issues/40630?camid=4v1a