ABSTRACT

As the use of the World Wide Web has become increasingly widespread, the business of commercial search engines has become a vital and lucrative part of the Web. Search engines are common place 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. In this chapter a relevant class of semantic search engines, based on a peer-to-peer, data integration mediator-based architecture is described. The architectural and functional features are presented with respect to two projects, SEWASIE and WISDOM, involving the authors. The methodology to create a two level ontology and query processing in the SEWASIE project are fully described.

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 400 million hits per day, and over 4 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-known 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:

- 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, that is, 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, that is, 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/2004/OWL/) 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 (Alesso, 2002; Boley, 2002).

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, and so forth. 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 44,600 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.
24 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/chapter/semantic-search-engines-based-data/28889?camid=4v1


www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

A Collaborative QoS-Aware Service Evaluation Method Among Multi-Users for a Shared Service
Wanchun Dou, Chao Lv, Xuyun Zhang and Jinjun Chen (2012). International Journal of Web Services Research (pp. 30-50).
www.igi-global.com/article/collaborative-qos-aware-service-evaluation/64222?camid=4v1a

Issues on the Compatibility of Web Service Contracts
www.igi-global.com/chapter/issues-compatibility-web-service-contracts/59923?camid=4v1a

XML Compression for Web Services on Resource-Constrained Devices
www.igi-global.com/article/xml-compression-web-services-resource/3123?camid=4v1a

Fully Automated Web Services Discovery and Composition Through Concept Covering and Concept Abduction
www.igi-global.com/article/fully-automated-web-services-discovery/3106?camid=4v1a