# Perspectives and Key Technologies of Semantic Web Search

**Konstantinos Kotis**
*University of the Aegean, Greece*

## INTRODUCTION

Current keyword-based Web search engines (e.g. Google[a]) provide access to thousands of people for billions of indexed Web pages. Although the amount of irrelevant results returned due to polysemy (one word with several meanings) and synonymy (several words with one meaning) linguistic phenomena tends to be reduced (e.g. by narrowing the search using human-directed topic hierarchies as in Yahoo[b]), still the uncontrolled publication of Web pages requires an alternative to the way Web information is authored and retrieved today. This alternative can be the technologies of the new era of the Semantic Web.

The Semantic Web, currently using OWL language to describe content, is an extension and an alternative at the same time to the traditional Web. A Semantic Web Document (SWD) describes its content with semantics, i.e. domain-specific tags related to a specific conceptualization of a domain, adding meaning to the document's (annotated) content. Ontologies play a key role to providing such description since they provide a standard way for explicit and formal conceptualizations of domains. Since traditional Web search engines cannot easily take advantage of documents' semantics, e.g. they cannot find documents that describe similar concepts and not just similar words, semantic search engines (e.g. SWOOGLE[c], OntoSearch[d]) and several other semantic search technologies have been proposed (e.g. Semantic Portals (Zhang et al, 2005), Semantic Wikis (Völkel *et al*, 2006), multi-agent P2P ontology-based semantic routing (of queries) systems (Tamma *et al*, 2004), and ontology mapping-based query/answering systems (Lopez *et al*, 2006; Kotis & Vouros, 2006, Bouquet *et al*, 2004). Within these technologies, queries can be placed as formally described (or annotated) content, and a semantic matching algorithm can provide the exact matching with SWDs that their semantics match the semantics of the query.

Although the Semantic Web technology contributes much in the retrieval of Web information, there are some open issues to be tackled. First of all, unstructured (traditional Web) documents must be semantically annotated with domain-specific tags (ontology-based annotation) in order to be utilized by semantic search technologies. This is not an easy task, and requires specific domain ontologies to be developed that will provide such semantics (tags). A fully automatic annotation process is still an open issue. On the other hand, SWDs can be semantically retrieved only by formal queries. The construction of a formal query is also a difficult and time-consuming task since a formal language must be learned. Techniques towards automating the transformation of a natural language query to a formal (structured) one are currently investigated. Nevertheless, more sophisticated technologies such as the mapping of several schemes to a formal query constructed in the form of an ontology must be investigated. The technology is proposed for retrieving heterogeneous and distributed SWDs, since their structure cannot be known a priory (in open environments like the Semantic Web).

This article aims to provide an insight on current technologies used in Semantic Web search, focusing on two issues: a) the automatic construction of a formal query (*query ontology*) and b) the querying of a collection of knowledge sources whose structure is not known a priory (distributed and semantically heterogeneous documents).

## BACKGROUND

A keyword-based Web search mainly concerns search techniques that are based on string (lexical) matching of the query terms to the terms contained in Web documents. Traditionally, keyword-based search is used for unstructured Web documents' (text with no semantics

attached) retrieval, where retrieval is obtained when query terms are matched to terms found in documents. Several techniques for keyword-based Web search have been introduced (Alesso, 2004), with the most popular being the simple Boolean search, i.e. combination of keywords based on Boolean operators AND, OR, NOT. Other techniques include

- wildcard and proximity search (syntactic analysis of documents or query terms),
- fuzzy search (handles misspelling and plural variations of keywords),
- contextual search (analyse the content of Web pages and return the subject of the page),
- keyword location-based search (keywords occurring in the title tags of the Web page are more important than those in the body),
- human(or topic)-directed search (use of topic hierarchies, manually created, to help users to narrow the search and make search results more relevant),
- thesaurus-based search (use specific relations such as synonym to help retrieve relevant information even if keyword is not present in a document),
- and finally statistics-based search such as Google's PageRank[e] technology.

Keyword-based search technology has been also used to retrieve SWDs by matching NL query terms to terms that lexicalize concepts of a SWD (e.g. an ontology concept). Such technology, when used in semantic search engines (e.g. SWOOGLE), do not utilize the semantics of the SWD in the matching algorithm. Matching is based on lexical techniques (string matching of keywords with terms that lexicalize concepts of a an ontology) although the retrieved content is semantically described (i.e. SWDs). Generally, semantic matching is performed in extension to the lexical one and the syntactic similarity between terms is not of interest. In fact, what is important is the similarity of the meaning of two terms. For instance, a match between a query-term "book" and a document-term "reserve" may be correctly identified if the sense of concept "book" is "the reservation of a ticket" (synonymy). On the other hand, a match between the term "book" found in a query and an identical term found in a Web document, may be incorrectly identified if their senses are completely different i.e. the query-term "book", meaning a publication, and the document-term "book", meaning a reservation (polysemy).

Semantic matching requires that the semantics of both the query and the document must be known or uncovered prior their matching. If the query is formally specified, the semantics of each term can be explicitly defined. Thus, if a query is represented as an ontology (query ontology), the semantics of each term that lexicalizes an ontology concept can be revealed by the semantic relations between this concept and the other concepts of the ontology (structure of its neighborhood). Such semantic relations are not only subsumption (is-a) relations, but also others such as "part-of", "meronym", "synonym", etc. On the other hand, if the query is informally specified, i.e. in natural language, the semantics of each term in the query must be somehow uncovered. The issue here is how a machine can "guess" what the intended meaning of an informal query is, in order to retrieve the document that is closer to this meaning and therefore, more interesting to the user. Intelligent search engines such as AskJeeves[f] (Teoma technology) try to tackle this issue by analysing the terms and their relations in a sophisticated way using natural language processing techniques or by refining the query in collaboration with the users. An alternative technique map each term of a query to its intended meaning (sense found in lexicon), using a combination of vector space indexing techniques such as LSI (Deerwester et al, 1990) and a lexicon such as WordNet (Miler, 1995). Furthermore, to be able to execute a semantic matching, the document (in addition to the query) must also provide its semantics. In case of a SWD, the semantics of the document are formally and explicitly specified in an ontology. In case of unstructured documents, advanced ontology learning techniques are required in order to extract their semantics and use them to annotate the related documents.

Further related work has been also carried out and presented (Karanastasi & Christodoulakis, 2007), where an ontology-driven semantic ranking methodology for ontology concepts is used for natural language disambiguation. This work has been proposed in the context of OntoNL framework (Karanastasi et al, 2007). The methodology uses domain specific ontologies for the semantic disambiguation. The disambiguation procedure is automatic and quite promising.

There are several other proposals concerning the retrieval of SWDs. The majority of them assume that