INTRODUCTION

Current keyword-based Web search engines (e.g. Google\textsuperscript{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\textsuperscript{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\textsuperscript{c}, OntoSearch\textsuperscript{d}) and several other semantic search technologies have been proposed (e.g. Semantic Portals (Zhang et al, 2005), Semantic Wikis (Völkel \textit{et al}, 2006), multi-agent P2P ontology-based semantic routing (of queries) systems (Tamma \textit{et al}, 2004), and ontology mapping-based query/answering systems (Lopez \textit{et al}, 2006; Kotis & Vouros, 2006, Bouquet \textit{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 (\textit{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
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
Related Content

Guided Sequence Alignment
www.irma-international.org/chapter/guided-sequence-alignment/10937/

Mining Data Streams
www.irma-international.org/chapter/mining-data-streams/10982/

Discovering Frequent Embedded Subtree Patterns from Large Databases of Unordered Labeled Trees
www.irma-international.org/chapter/discovering-frequent-embedded-subtree-patterns/7832/

Quantization of Continuous Data for Pattern Based Rule Extraction
www.irma-international.org/chapter/quantization-continuous-data-pattern-based/11039/

Comparative Genome Annotation Systems
www.irma-international.org/chapter/comparative-genome-annotation-systems/7731/