Semantic Web — Interoperability, Usability, Applicability 0 (0) 1
IOS Press

Semantic Search on the Web

Bettina Fazzinga [a] and Thomas Lukasiewicz [b]
[a] *Dipartimento di Elettronica, Informatica e Sistemistica, Università della Calabria, Italy*
E-mail: bfazzinga@deis.unical.it
[b] *Computing Laboratory, University of Oxford, UK*
E-mail: thomas.lukasiewicz@comlab.ox.ac.uk

Abstract. Web search is a key technology of the Web, since it is the primary way to access content on the Web. Current standard Web search is essentially based on a combination of textual keyword search with an importance ranking of the documents depending on the link structure of the Web. For this reason, it has many limitations, and there are a plethora of research activities towards more intelligent forms of search on the Web, called *semantic search on the Web*, or also *Semantic Web search*. In this paper, we give a brief overview of existing such approaches, including own ones, and sketch some possible future directions of research.

Keywords: semantic search on the Web, Semantic Web search, Web search, Semantic Web, ontologies.

1. Introduction

Web search is a key technology of the Web, which is essentially based on a combination of textual keyword search with an importance ranking of the documents depending on the link structure of the Web. For this reason, it has many limitations, and there are a plethora of research activities towards more intelligent Web search, called *semantic search on the Web*, or also *Semantic Web search*, which is currently one of the hottest research topics in both the Semantic Web and Web search (see [13] and [1], respectively).

There is no unique definition of the notion of semantic search on the Web. However, the most common use is the one as an improved form of search on the Web, where meaning and structure are extracted from both the user's Web search queries and different forms of Web content, and exploited during the Web search process. Such semantic search is often achieved by using Semantic Web technology for interpreting Web search queries and resources relative to one or more underlying ontologies, describing some background domain knowledge, in particular, by connecting the Web resources to semantic annotations, or by extracting semantic knowledge from Web resources. Such a search usually also aims at allowing for more complex Web search queries whose evaluation involves reasoning over the Web. Another common use of the notion of semantic search on the Web is the one as search in the large datasets of the Semantic Web as a future substitute of the current Web. This second use is closely related to the first one, since the above semantic annotation of Web resources, or alternatively the extraction of semantic knowledge from Web resources, actually corresponds to producing a knowledge base, which may be encoded using Semantic Web technology. That is, the latter semantic search on the Web can essentially be considered as a subproblem of the former one.

Another closely related use is the one as natural language search on the Web, where search queries are formulated in (written or even spoken) natural language. Many approaches try to translate such queries into formal queries in a structured query language, which are generally available in the above semantic search in the context of the Semantic Web. The answers to such natural language queries may be Web resources as usual, or they may also be structured or natural language results, towards more informative results, e.g., by showing structured information extracted from the resulting Web pages, and by additionally connecting the search result with Wikipedia articles. This is another meaning of semantic search, which is actually already a very simple form of question answering.

Frequently, the notion of semantic search also covers some other (often less) semantic ideas and concepts. For example, faceted search allows for exploring results according to a collection of predefined categories, called facets. Closely related is clustered search, where such facets are not predefined. A further example is the suggestion of related searches, such as the completion and correction of Web search queries, which are well-known from standard Web search en-

gines. A similar example is full-text similarity search, where blocks of text ranging from phrases to full documents, rather than few keywords, are submitted.

In this paper, we discuss especially the two initial interpretations of the notion of semantic search on the Web, which both refer to the context of the Semantic Web, as well as their generalizations towards natural language search on the Web. The rest of this paper is organized as follows. In Section 2, we describe some representative approaches to semantic search on the Web. Section 3 sketches our own such approach. In Section 4, we conclude and describe our vision for the future of semantic search on the Web.

2. Overview of Existing Approaches

State-of-the-art approaches to semantic search on the Web can be classified as follows:

1. approaches based on structured query languages, such as [5,9,12,14];
2. approaches for *naive* users, where no familiarity with ad-hoc query languages is required. In turn, these approaches can be divided into:
   - keyword-based approaches, such as [3,11,15, 19], where queries consist of lists of keywords;
   - natural-language-based approaches, such as [4,8,16,10], where users can express queries by means of the natural language.

In the following, we give an overview of the main approaches belonging to the above categories.

2.1. Approaches Based on Structured Languages

SHOE [12] is one of the first attempt to semantically query the Web. SHOE provides the following: a tool for annotating Web pages, allowing users to add SHOE markup to a page by selecting ontologies, classes, and properties from a list; a Web crawler, which searches for Web pages with SHOE markup and stores the information in a knowledge base (KB); an inference engine, which provides new markups by means of inference rules (basically, Horn clauses); several query tools, which allow users to pose structured queries against an ontology. One of the query tools allows users to draw a graph in which nodes represent constant or variable instances and arcs represent relations. To answer the query, the system retrieves subgraphs matching on the user graph. The SHOE search tool allows user to pose queries by first choosing an ontology from a drop-down list and next choosing classes and properties from another list. Finally, the system builds a conjunctive query, issues the query to the KB, and presents the results in a tabular form.

Subsequent approaches are [9,5], which mainly focus on RDF. Swoogle [9] is a crawler-based system for discovering, indexing, and querying RDF documents. Swoogle mainly provides a search for Semantic Web documents and terms (i.e., the URIs of classes and properties). It allows users to specify queries containing conditions on the document-level metadata (i.e., queries asking for documents having `.rdf` as the file extension), and it also allows users to search for Semantic Web documents using RDF/XML as the syntax language. Retrieved documents are ranked according to a ranking algorithm measuring the documents' importance on the Semantic Web.

The Corese system presented in [5] is an ontology-based search engine for the Semantic Web, which retrieves Web resources annotated in RDF(S) by using a query language based on RDF(S). Corese is able to *approximately* search the Semantic Web. Approximation is provided by employing inference rules and by computing the semantic distance of classes or properties in the ontology hierarchies. Specifically, Corese retrieves Web resources whose annotations are specializations of the query, and it also retrieves those resources whose annotations refer to concepts and relations that are hierarchically *close enough* to those of the query.

A more recent approach is [14], where the NAGA semantic search engine is presented, which provides a graph-based query language to query the underlying KB represented as a graph. The KB is built automatically by a tool for knowledge extraction from Web sources, which extends the approach proposed in [18]. The nodes and edges in the knowledge graph represent entities and relationships between entities, respectively. The NAGA query language extends SPARQL [20], allowing complex graph queries with regular expressions over relationships on edge labels. Answers to a query are subgraphs of the knowledge graph matching the query graph and are ranked using a specific scoring model for weighted labeled graphs.

2.2. Keyword-Based Approaches

As one of the first approaches, [11] focuses on augmenting the results of traditional keyword search with data retrieved from the Semantic Web. Query processing can be summarized as follows: when a user query

is issued, query terms (keywords) are mapped to Semantic Web nodes: in the case of multiple matching, some heuristics (for instance, taking into account the user profile, etc.) are employed to find the right one. Once nodes matching the search terms are found, the approach uses some heuristics to choose what part of the Semantic Web graph around these nodes, has to be returned as a result (i.e., the first N triples, where N is some threshold). Moreover, [11] proposes an approach to improve traditional keyword search by disambiguating the meaning of the terms in the query. To this end, an additional link next to each search result is added, so that, if the user clicks on this link, only Web documents having a content *semantically similar* to the document reachable from that link are shown.

More recent approaches for naive users based on keyword search are [15,19,3]. SemSearch [15] provides a Google-like query interface allowing users to specify queries without requiring any knowledge about ontologies or specific languages. User queries consist of two or more keywords, whose semantic meaning is taken into account to reformulate the queries themselves according to a formal query language syntax. Keywords are assigned a semantic meaning by matching them against a collection of classes, properties, and instances in semantic data repositories. Since each keyword can match a class, a property, or an instance, several combinations of semantic matchings of the keywords are considered. For instance, it can be the case that every keyword matches a class, or that the first keyword matches a class, while the second matches a property, and so on. All the combinations of matchings are taken into account in the reformulation process, and each combination leads to a distinguished formal query, obtained from a pre-determined set of query templates. After the reformulation, formal queries are exactly evaluated, and this yields results that are semantically related to all the user keywords.

In [19], a similar approach to [15], keyword queries are translated into conjunctive queries to be evaluated against an underlying KB. Here, the structure of the formal queries that are eventually evaluated does not conform to pre-determined templates. Formal queries are built exploiting a graph-based technique to find the connections between the entities in the user queries. Specifically, query translation consists of the following three steps. First, the keywords in the user query are mapped onto ontology elements. Then, relations among these ontology elements are examined, and subgraphs of the KB are extracted. Each subgraph represents a set of relations connecting all the considered elements, thus the set of these subgraphs represents all the possible relationships among user keywords that could not be explicitly specified by the user. Hence, these subgraphs correspond to the different queries that the user may be interested in. Finally, formal queries are generated by translating the subgraphs according to a proper language, and evaluated against the KB.

Falcons [3] is a keyword-based search engine for the Semantic Web, allowing concept and object search. Concept search is carried out by searching the classes and properties that match the query terms in the ontology selected by the user, and, furthermore, recommending other ontologies on the basis of a combination of the TF-IDF technique and the popularity of ontologies. Object search is performed in a similar way: besides returning the objects that match the query terms, the system also recommends other types of objects that the user is likely to be interested in.

2.3. Natural-Language-Based Approaches

Some of the most recent approaches focusing on natural language queries are [4,8]. In [4], the ORAKEL system is presented, where, before being evaluated, queries are first translated into a logical form, and then reformulated according to a target language, i.e., the language of the underlying KB. The translation from the logical form to the target language is described declaratively by a Prolog program. The overall approach is independent from the specific target language, since changing the ontology language only requires a declarative description of the transformation as a Prolog program, but no further change to the underlying system. The system relies on a specific kind of user, called lexicon engineer, who specifies how natural language expressions can be mapped onto predicates in the KB, i.e., how verbs, adjectives, and relational nouns can be mapped onto corresponding relations specified in the domain ontology.

The system presented in [8] supports (i) Semantic Web search over ontologies and (ii) semantic search over non-Semantic Web documents. As regards the first kind of search, answers to a natural language query are retrieved by exploiting a previous system, called PowerAqua [17], which works in the following way: first, the user query is translated from natural language into a structured format, called *linguistic triple*; second, the terms of the linguistic triple are mapped to semantically relevant ontology entities. Finally, the ontological entities that best represent the user query are selected and returned. PowerAqua ex-

tends the AquaLog system proposed in [16], which works in the presence of a single ontology only, to the case of multiple ontologies. The second kind of search in [8], namely, the semantic search over non-Semantic Web documents, is accomplished by extending the system proposed in [2]. Specifically, this relies on a new approach for annotating documents, consisting of the following steps: (i) extracting the textual representation of semantic entities, (ii) searching this textual representation in Web documents, and (iii) generating an annotation linking the semantic entities to each of the documents containing their textual representation. Furthermore, [8] deals with the problem of knowledge incompleteness, by switching to the traditional keyword search when no ontology satisfies the query.

The most recent approach belonging to the category of natural-language-based approaches is the newest version of Google [10]. Besides being a widely used keyword search engine, Google is now evolving to a natural-language-based search engine. In fact, it has been recently augmented with a new functionality, which provides more precise answers to queries: instead of returning Web page links as query results, Google now tries to build query answers, collecting information from several Web pages. As an example, the simple query "barack obama date of birth" gets the answer "4 August, 1961 ". Next to the answer, the link *Show sources* is shown, that leads to the Web pages from which the answer has been obtained.

3. The FGGL Approach

We now describe our approach to semantic search on the Web presented in [7], which is based on a structured query language that allows to formulate complex ontology-based (conjunctive) search queries.

More specifically, an ontologically enriched Web along with complex ontology-based search on the Web are achieved on top of the existing Web and using existing Web search engines. Intuitively, rather than being interpreted in a keyword-based syntactic fashion, the pieces of data on existing Web pages are connected to (and via) some ontological KB (in a lightweight ontology language) and then interpreted relative to this KB. That is, the pieces of data on Web pages are connected to (and via) a much more precise semantic and contextual meaning. More concretely, we are actually mapping the Web into an ontological KB, which then allows for Semantic Web search relative to the underlying ontology. Intuitively, such a KB can be considered as an ontological index over the Web, against which ontological Web search queries can be answered. This allows for answering Web search queries in a much more precise way, taking into account the meaning of Web search queries and pages, and it also allows for more complex ontology-based Web search queries that involve reasoning over the Web, which are also much closer to complex natural language search queries than current Boolean keyword-based search queries.

Query processing is based on new techniques (i) for pre-compiling the ontological knowledge using standard ontology reasoning techniques and (ii) for translating complex ontology-based Web queries into (sequences of) standard Web queries that are answered by standard Web search. That is, essential parts of ontological search on the Web are actually reduced to state-of-the-art search engines. As important advantages, this approach can immediately be applied to the whole existing Web, and it can be done with existing Web search technology (and so does not require completely new technologies). Such a line of research aims at adding ontology-based structure and semantics (and thus in a sense also intelligence) to current search engines for the existing Web by combining existing Web pages and queries with ontological knowledge.

The ontological knowledge and annotations that are underlying our semantic search on the Web can be classified according to its origin and contents. As for the origin, they may be either (a) explicitly defined by experts, or (b) automatically extracted from the Web, eventually coming along with existing pieces of ontological knowledge and annotations (e.g., from existing ontologies or ontology fragments, and/or from existing annotations of Web pages in microformats or RDFa). In the latter case, generating, maintaining, and updating the ontological knowledge and annotations is done automatically and much less cost-intensive than in the former case. As for the contents, (a) the ontological knowledge and annotations may either describe fully general knowledge (such as the knowledge encoded in Wikipedia) for general ontology-based search on the Web, or (b) they may describe some specific knowledge (such as biomedical knowledge) for vertical ontology-based search on the Web. The former results into a general ontology-based interface to the Web similar to Google, while the latter produces different vertical ontology-based interfaces.

In [6], a variant of the above approach is explored, which uses inductive reasoning techniques rather than

deductive ones. This adds especially the ability to handle inconsistencies, noise, and incompleteness.

4. Conclusion and Vision for the Future

We have given a brief overview of approaches to *semantic search on the Web* (also called *Semantic Web search*), which is currently one of the hottest research topics in both the Semantic Web and the Web search community. In semantic search on the Web, the current strong research activities of the former to realize search on the Semantic Web are merged with the current strong research activities of the latter to add semantics to Web queries and content when performing Web search. It is through this integration that the reasoning capabilities envisioned in Semantic Web technologies are coming to Web search and the Web. As we have seen, the formulation of queries and their results in semantic search on the Web is ultimately directed by a third area, namely, the one of question answering systems, which is based on natural language processing.

Although many approaches and systems to semantic search on the Web already exist, the research in this area is still at the very beginning, and many open research problems still persist. Some of the most pressing research issues are maybe (i) how to automatically translate natural language queries into formal ontological queries and (ii) how to automatically add semantic annotations to Web content, or alternatively how to automatically extract knowledge from Web content.

Performing Web search in the form of returning simple answers to simple questions in natural language is still science fiction, let alone performing Web search in the form of query answering relative to some concrete domain or even general query answering. However, with the current activities towards semantic search on the Web, we are moving one step closer to making such science fiction become true, which ultimately aims at a human-like interface to the knowledge, information, services, and other resources available on the Web.

Acknowledgments. Thomas Lukasiewicz's work has been supported by a Yahoo! Research Fellowship.

References

[1] R. A. Baeza-Yates and P. Raghavan. Next generation Web search. In S. Ceri and M. Brambilla, editors, *Search Computing*, *LNCS* 5950, pp. 11–23. Springer, 2010.

[2] P. Castells, M. Fernández, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng.*, 19(2):261–272, 2007.

[3] G. Cheng, W. Ge, and Y. Qu. Falcons: Searching and browsing entities on the Semantic Web. In *Proc. WWW-2008*, pp. 1101–1102. ACM Press, 2008.

[4] P. Cimiano, P. Haase, J. Heizmann, M. Mantel, and R. Studer. Towards portable natural language interfaces to knowledge bases — The case of the ORAKEL system. *Data Knowl. Eng.*, 65(2):325–354, 2008.

[5] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the Semantic Web with Corese search engine. In *Proc. ECAI-2004*, pp. 705–709. IOS Press, 2004.

[6] C. d'Amato, F. Esposito, N. Fanizzi, B. Fazzinga, G. Gottlob, and T. Lukasiewicz. Inductive reasoning and Semantic Web search. In *Proc. SAC-2010*, pp. 1446–1447. ACM Press, 2010.

[7] B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz. Semantic Web search based on ontological conjunctive queries. In *Proc. FoIKS-2010*, *LNCS* 5956, pp. 153–172. Springer, 2010.

[8] M. Fernández, V. Lopez, M. Sabou, V. S. Uren, D. Vallet, E. Motta, and P. Castells. Semantic search meets the Web. In *Proc. ICSC-2008*, pp. 253–260. IEEE Computer Society, 2008.

[9] T. W. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the Semantic Web. In *Proc. AAAI-2005*, pp. 1682–1683. AAAI Press / MIT Press, 2005.

[10] Google. `http://www.google.com`.

[11] R. V. Guha, R. McCool, and E. Miller. Semantic search. In *Proc. WWW-2003*, pp. 700–709. ACM Press, 2003.

[12] J. Heflin, J. A. Hendler, and S. Luke. SHOE: A blueprint for the Semantic Web. In D. Fensel, W. Wahlster, and H. Lieberman, editors, *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, pp. 29–63. MIT Press, 2003.

[13] J. Hendler. Web 3.0: The dawn of semantic search. *Computer*, 43(1):77–80, 2010.

[14] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In *Proc. ICDE-2008*, pp. 953–962. IEEE Computer Society, 2008.

[15] Y. Lei, V. S. Uren, and E. Motta. SemSearch: A search engine for the Semantic Web. In *Proc. EKAW-2006*, *LNCS* 4248, pp. 238–245. Springer, 2006.

[16] V. Lopez, M. Pasin, and E. Motta. AquaLog: An ontology-portable question answering system for the Semantic Web. In *Proc. ESWC-2005*, *LNCS* 3532, pp. 546–562. Springer, 2005.

[17] V. Lopez, M. Sabou, and E. Motta. PowerMap: Mapping the real Semantic Web on the fly. In *Proc. ISWC-2006*, *LNCS* 4273, pp. 414–427. Springer, 2006.

[18] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proc. WWW-2007*, pp. 697–706. ACM Press, 2007.

[19] T. Tran, P. Cimiano, S. Rudolph, and R. Studer. Ontology-based interpretation of keywords for semantic search. In *Proc. ISWC/ASWC-2007*, *LNCS* 4825, pp. 523–536. Springer, 2007.

[20] W3C. SPARQL Query Language for RDF, 2008. W3C Recommendation (15 January 2008). Available at `http://www.w3.org/TR/rdf-sparql-query/`.