International Journal of Engineering, Applied and Management Sciences Paradigms, Vol. 01, Issue 01, February 2013
ISSN (Online): 2320-6608
www.ijeam.com

Semantic Web Search Engine: A Review

Sachin Minocha

M. Tech Student, Vaish College of Engineering, Rohtak, Haryana (India)

Abstract

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. 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. In this paper we present survey on the web semantic, their architecture and also explain some web search engine.

Keywords: *Web-semantic, Ontology, Search engine.*

1. Web Semantic

Web comprises a huge amount of heterogeneous data (structured data, semi-structured data, textual data and multimedia data), dedicated to human users of the Web. The Semantic Web [1] aims at enabling the semantic contents of Web resources to be also processed by automated tools. It relies on rich metadata, also called semantic annotations, offering explicit semantic descriptions of Web resources and built on domain ontologies [2].

Semantic Web is an extension of the current Web [2] that allows the meaning of information to be precisely described in terms of well-defined vocabularies that are understood by people and computers. On the Semantic Web information is described using a new W3C standard called the Resource Description Framework (RDF). Semantic Web Search is a search engine for the Semantic Web. Current Web sites can be used by both people and computers to precisely locate and gather information published on the Semantic Web [3]. In a row with the extraordinary growth of web, there are many search engines come out to help the users on finding their need, but search engines find it increasingly difficult to provide useful results.[4] To manage this explosively large number of web documents, automatic clustering of documents and organizing them into domain dependent directories became very popular. [5]

All search engines consist of three parts [5]:

1. A database of web documents,
2. A search engine operating on that database and
3. A series of programs that determine how search results are displayed [7]. Part of the search engines' success might be due to their simplicity: you enter some words and the results are then output in form of a ranked list, in which the search engine estimates the relevance of each indexed website to your query. [6]

The utility of any search engine depends on two parts: the quality of the system, and content, which in our case is provided by a large number of contributors (personal and corporate web sites, for example). Importantly, content suppliers have to agree on a social contract (as anywhere on the Internet) on how to provide and publish data [5].

Currently, the Semantic Web, (i.e. online documents written in RDF or OWL), is essentially a web universe parallel to the web of HTML documents. Semantic Web documents (SWDs) are characterized by semantic annotation and meaningful references to other SWDs [2]. Since conventional search engines do not take advantage of these features, a search engine customized for SWDs, especially for ontologies, is needed by human users as well as by software agents and services. At this stage, human users are expected to be semantic web researchers and developers who are interested in accessing, exploring and querying RDF and OWL documents found on the web [8].

2. Semantic Web Technologies [10]

Semantic web [9] is the next generation of the web which evolves toward semantic knowledge representations and intelligent services (e.g., information brokers, search agents) where information can be processed by machines. To fully realize this goal, standards for exchanging machine-understandable information have to be established. These standards define not only the syntactic representation of information, but also their semantic content.

A technology stack, suggested by the W3C, that we use in our work consists of Resource Description Framework (RDF), which provides data model specification and an XML-based serialization syntax; ontologies, which enable the definition and sharing of domain vocabularies; and rules, which allow declarative processing of data [10].

2.1 The Resource Description Framework [11]

At present, services on the Web are single islands. Common data models and data exchange standards are required in order to enable the fast integration of different datasources and to bridge semantic differences. The Web community has proposed the Resource Description Framework (RDF) [11] as a data model suitable for information integration tasks. The data model serves as a foundation for ontology languages.

2.2 Ontologies[10]

Ontology is a specification of a conceptualization [12]. In this context, specification refers to an explicit representation by some syntactic means. In contrast to schema languages (like XML Schema or DTDs) ontologies try to capture the semantics of a domain by deploying knowledge representation primitives, enabling a machine to (partially) understand the relationships between concepts in a domain. Additional knowledge can be captured by axioms or rules. In the Web context, RDF-Schema and OWL1 are recommendations from the W3C for ontology modeling languages.

2.3 Rules

Rules, in combination with RDF and ontologies, are an active field of research. Rules can be used to capture domain knowledge.

2.3.1 Mapping between keywords and concepts [13]

It is a common approach in the semantic search engine. It has got several advantages thatare generally the available data may not be formally encoded .For example in fuzzy keyword to concept mapping deals about the mapping of the textual materials to the well-defined information. In other way is the natural language system, natural languages are in the form of expressions which is mostly acceptable by the humans. As far as considering the mapping patterns as we map from graph to sentence the visual query tool like SEWWAISE gives most accurate picture to the humans about the relationships. The humans also will be more comfortable with the natural language since that is more comfortable to them [13].

2.3.2 Graph patterns

It is an important concept in semantic search. In semantic search it is used in multiple varying roles. It is used to solve or encode the complex constraint queries given by the user, it is solved by locating the corresponding graph in the RDF network. The common RDF pattern Anyanwu and sheth is also used to have connection between the paths and the common names. Graphs patterns also gives the idea that where to collect or to fetch the data for particular item [13].

2.3.3 Logics

It is internally very much tied with semantic web. Even the standard web ontology language (OWL) is based on the description logics. Still the application built on the top of the logical frameworks with the wine agent is an exception than the common ordinary example. In most of the cases the applications take few entailments for the base and create functionality according to their requirements over that. The examples GRQL, ODESeW,SHOE, and SEWASIE do the same [13].

2.3.4 Combining uncertainty with logics

While augmenting the text search with the ontology technique, to combine the uncertain annotation some formalization could be needed. For this probability logics or the fuzzy logic experiments are under taken in this field [13].

3. Architecture of Semantic Web Components

The term "Semantic Web" is often used more specifically to refer to the formats and technologies that enable it.[14] The collection, structuring and recovery of linked data are enabled by technologies that provide a formal description of concepts, terms, and relationships within a given knowledge domain. These technologies are specified as W3C standards and include:

1. Resource Description Framework (RDF), a general method for describing information
2. RDF Schema (RDFS)
3. Simple Knowledge Organization System (SKOS)
4. SPARQL, an RDF query language
5. Notation3 (N3), designed with human-readability in mind
6. N-Triples, a format for storing and transmitting data
7. Turtle (Terse RDF Triple Language)
8. Web Ontology Language (OWL), a family of knowledge representation languages

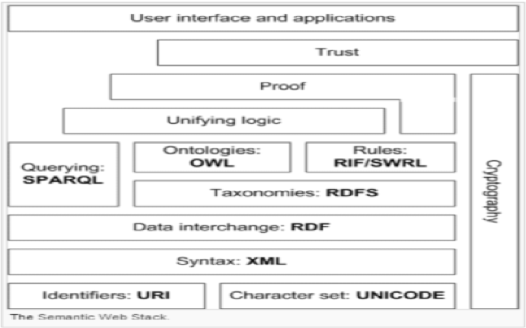

Figure 1: Web Semantic Components

The Semantic Web Stack illustrates the architecture of the Semantic Web. The functions and relationships of the components can be summarized as follows:[15]. XML provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within. XML is not at present a necessary component of Semantic Web technologies in most cases, as alternative syntaxes exists, such as Turtle. Turtle is a de-facto standard, but has not been through a formal standardization process.

- XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents.
- RDF is a simple language for expressing data models, which refer to objects ("web resources") and their relationships. An RDF-based model can be represented in a variety of syntaxes, e.g., RDF/XML, N3, Turtle, and RDF.[16] RDF is a fundamental standard of the Semantic Web.[17]
- RDF Schema extends RDF and is a vocabulary for describing properties and classes of RDF-based resources, with semantics for generalized-hierarchies of such properties and classes.
- OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties and characteristics of properties (e.g. symmetry), and enumerated classes.
- SPARQL is a protocol and query language for semantic web data sources [17].

4. Web Search Engine Design

The term "search engine" is often used generically to describe both crawler-based search engines and human powered directories [8]. These two types of search engines gather their listings in radically different ways. Crawler-based search engines, such as Google, create their listings automatically. They "crawl" or "spider" the web, then people search through what they have found. A human powered directory, such as the Open Directory, depends on humans for its listings. When we submit a short description to the directory for your entire site or editors write one for sites they review. A search looks for matches only in the descriptions submitted [8].

There are several components in search engine, such as crawler, indexer, sorter, analyzer, and searcher. Crawler is a program that is created to visit each webpage periodically and collect the important information and store them into database. The Web Crawler Application is divided into three main modules [5].

a. **Controller Module -** This module focuses on the Graphical User Interface (GUI) designed for the web crawler and is responsible for controlling the operations of the crawler. The GUI enables the user to enter the start URL, enter the maximum number of URL’s to crawl, view the URL’s that are being fetched. It controls the Fetcher and Parser.
b. **Fetcher Module -** This module starts by fetching the page according to the start URL specified by the user. The fetcher module also retrieves all the links in a particular page and continues doing that until the maximum number of URL’s is reached.
c. **Parser Module -** This module parses the URL’s fetched by the Fetcher module and saves the contents of those pages to the disk. [18]

After that indexer create index in the database to organize the data by categorize them. The indexer extracts all the information from each and every document and stores it in a database. All high-quality search engines index each and every word in the documents and give a unique word Id. Then the word occurrences, which some search engines call “hits,” are checked, recording all the words, including their placement in the document, their font size and capitalization. [19]

The typical crawler-based search engine has several major elements. First is the spider, also called the crawler. The spider visits a Web page, reads it, and then follows links to other pages within the site [8]. This is what it means when someone refers to a site being "spidered" or "crawled." The spider returns to the site on a regular basis, such as every month or two, to look for changes. Everything the spider finds goes into the second part of the search engine, the index. The index, sometimes called the catalogue, is like a giant book containing a copy of every Web page that the spider finds. If a Web page changes, then this book is updated with new information. Search engine software is the third part of a search engine. This is the program that shifts through the millions of pages recorded in the index to find matches to a search and rank them in order of what it believes is most relevant [8].

A search engine cannot work without a proper index where possible searched pages are stored, usually in a compressed format. This index is created by specialized robots, which crawl the Web for new/modified pages (the actual crawlers, or spiders).

"Smart" crawling technology is used to crawl 'valuable' sites more frequently or more deeply. The measure of 'value' is, of course, itself an important research topic. Smart crawling is also used to estimate the rate of change of web pages and adjust the crawling algorithm to maximize the freshness of the pages in the repository [8].

5. Semantic Web Process Lifecycle [25]

Semantic Web services will allow the semi-automatic and automatic annotation, advertisement; discovery, selection, composition, and execution of inter-organization business logic, making the Internet become a global common platform where organizations and individuals communicate among each other to carry out various commercial activities and to provide value-added services. In order to fully harness the power of Web services, their functionality must be combined to create Web processes. Web processes allow representing complex inter-actions among organizations, representing the evolution of workflow technology. Semantics can play an important role in all stages of Web process lifecycle [25].

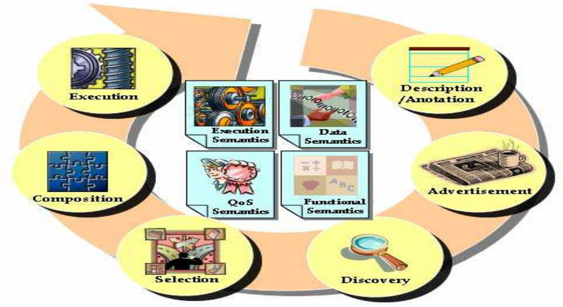

Figure 2: Web process lifecycle and semantics.[25]

The lifecycle of semantic Web processes includes the description/annotation, the advertisement, the discovery, the selection, the composition of Web services that makeup Web processes, and the execution of Web processes. All these stages are significant for the Web process lifecycle and their success. [25]

5.2.1 Semantics and Ontologies

There is a growing consensus that Web services alone will not be sufficient to de-velop valuable and sophisticated Web processes due the degree of heterogeneity, autonomy, and distribution of the Web. Several researchers agree that it is essential for Web services to be machine understandable in order to allow the full deployment of efficient solutions supporting all the phases of the lifecycle of Web processes.

The idea and vision of the "Semantic Web" [26] catches on and researchers as well as companies have already realized the benefits of this great vision. Ontologies [27] are considered the basic building block of the Semantic Web as they allow machine supported data interpretation reducing human involvement in data and process integration.

Ontology "is a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine readable. Shared reflects that ontology should capture consensual knowledge accepted by the communities" [28].

When the knowledge about a domain is represented in a declarative language, the set of objects that can be represented is called the universe of discourse. We can de-scribe the ontology of a program by defining a set of representational terms. Definitions associate the names of entities in the universe of discourse (e.g. classes, relations, functions or other objects) with human-readable text describing what the names mean and formal axioms that constrain the interpretation and well-formed use of these terms [25].

A set of Web services that share the same ontology will be able to communicate about a domain of discourse. We say that a Web service commits to ontology if its observable actions are consistent with the definitions in the ontology.

5.2.2 Semantics for Web Services [25]

In Web services domain, semantics can be classified into the following types [8] illustrated in Figure 2:

- Functional Semantics
- Data Semantics
- QoS Semantics and
- Execution Semantics

These different types of semantics can be used to represent the capabilities, requirements, effects and execution of a Web service. In this section we describe the nature of Web services and the need for different kind of semantics for them.

5.2.2.1 Functional Semantics: The power of Web services can be realized only when appropriate services are discovered based on the functional requirements. It has been assumed in several semantic Web service discovery algorithms [29] that the functionality of the services is characterized by their inputs and outputs. Hence these algorithms look for semantic matching between inputs and outputs of the services and the inputs and outputs of the requirements. This kind of semantic matching may not always retrieve an appropriate set of services that satisfy functional requirements. Though semantic matching of inputs and outputs are required, they are not sufficient for discovering relevant services.

5.2.2.2 **Data Semantics:** All the Web services take a set of inputs and produce a set of out-puts. These are represented in the signature of the operations in a specification file. However the signature of an operation provides only the syntactic and structural de-tails of the input/output data. These details (like data types, schema of a XML complex type) are used for service invocation. To effectively perform discovery of services, semantics of the input/output data has to be taken into account. Hence, if the data involved in Web service operation is annotated using ontology, then the added data semantics can be used in matching the semantics of the input/output data of the Web service with the semantics of the input/output data of the requirements. Semantic discovery algorithm proposed in [29] uses the semantics of the operational data.

5.2.2.3 **QoS Semantics**: After discovering Web services whose semantics match the semantics of the requirements, the next step is to select the most suitable service. Each ser-vice can have different quality aspect and hence service selection involves locating the service that provides the best quality criteria match. Service selection is also an important activity in web service composition [30]. This demands management of QoS metrics for Web services. Web services in different domains can have different quality aspects [25].

5.2.2.4 **Execution Semantics:** Execution semantics of a Web service encompasses the ideas of message sequence, conversation pattern of Web service execution, flow of actions, preconditions and effects of Web service invocation, etc. Some of these details may not be meant for sharing and some may be, depending on the organization and the application that is exposed as a Web service. In any case, the execution semantics of these services are not the same for all services and hence before executing or invoking a service, the execution semantics or requirements of the service should be verified [25].

6. Some Semantic Search Engines

Currently many of semantic search engines are developed and implemented in different working environments, and these mechanisms can be put into use to realize present search engines.

Alcides Calsavara and Glauco Schmidt proposed and defines a novel kind of service for the semantic search engine. A semantic search engine stores semantic information about Web resources and is able to solve complex queries, considering as well the context where the Web resource is targeted, and how a semantic search engine may be employed in order to permit clients obtain information about commercial products and services, as well as about sellers and service providers which can be hierarchically organized [20]. Semantic search engines may seriously contribute to the development of electronic business applications since it is based on strong theory and widely accepted standards.

Sara Cohen Jonathan Mamou et. al. presented a semantic search engine for XML (XSEarch) [21].It has a simple query language, suitable for a naïve user. It returns semantically related document fragments that satisfy the user's query. Query answers are ranked using extended information-retrieval techniques and are generated in an order similar to the ranking. Advanced indexing techniques were developed to facilitate efficient implementation of XSEarch. The performance of the different techniques as well as the recall and the precision were measured experimentally. These experiments indicate that XSEarch is efficient, scalable and ranks quality results highly.

Bhagwat and Polyzotis proposed a Semantic-based file system search engine- Eureka, which uses an inference model to build the links between files and a File Rank metric to rank the files according to their semantic importance [22]. Eureka has two main parts: a) crawler which extracts file from file system and generates two kinds of indices: keywords' indices that record the keywords from crawled files, and rank index that records the File Rank metrics of the files; b) when search terms are entered, the query engine will match the search terms with keywords' indices, and determine the matched file sets and their ranking order by an information retrieval based metrics and File Rank metrics.

Wang et al. project a semantic search methodology to retrieve information from normal tables, which has three main steps: identifying semantic relationships between table cells; converting tables into data in the form of database; retrieving objective data by query languages [23]. The research objective defined by the authors is how to use a given table and a given domain knowledge to convert a table into a database table with semantics. The authors' approach is to denote the layout by layout syntax grammar and match these denotation with given templates which can be used to analyze the semantics of table cells. Then semantic preserving transformation is used to transform tables to database format.

7. Conclusion

World Wide Web (WWW) has rapidly evolved into a huge mine of global information and it is growing in size everyday. The presence of huge amount of resources on the Web thus poses a serious problem of accurate search. This is mainly because today's Web is a human-readable Web where information cannot be easily processed by machine. Highly sophisticated, efficient keyword based search engines that have evolved today have not been able to bridge this gap. So comes up the concept of the Semantic Web which is envisioned by Tim Berners-Lee as the Web of machine interpretable information to make a machine process able form for expressing information. Based on the semantic Web technologies we present in this paper the components and their architecture.

References

[1] Olivier Corby et. al., "Querying the Semantic Web with Corese Search Engine",

[2] T. Berners-Lee, J. Handler, O. Lassila. The Semantic Web, Scientific American, May, 2001.

[3] G. Madhu et. al., "Intelligent Semantic Web Search Engines: A Brief Survey", International journal of Web & Semantic Technology (IJWesT) Vol.2, No.1, January 2011.

[4] Debnath, Sandip, et al. "Knowledge discovery in web-directories: Finding term-relations to build a business ontology." E-Commerce and Web Technologies. Springer Berlin Heidelberg, 188-197, 2005.

[5] Junaidah Mohamed Kassim et. al., "Introduction to Semantic Search Engine", 2009 International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia.

[6] Bifet, Albert., et al., "An Analysis of Factors Used in Search Engine Ranking". Technical University of Catalonia, 2005.

[7] Barker, Joe., "What Makes a Search Engine Good?",Available: http://www.lib.berkeley.edu/TeachingLib/Guides /Inter net/SrchEngCriteria.pdf, 2003.

[8] Ms. S.Latha Shanmugavadivu et. al., "Semantic Based Multiple Web Search Engine", IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1722-1728.

[9] Tim Berners-Lee. Weaving the Web. Texere Publishing, NA, 2000.

[10] Tangmunarunkit, Hongsuda, Stefan Decker, and Carl Kesselman. "Ontology-based resource matching in the Grid–the Grid meets the semantic web." The Semantic Web-ISWC 2003. Springer Berlin Heidelberg, 2003. 706-721.

[11] O. Lassila and R. R. Swick. Resource description framework (rdf) model and syntax specification. In W3C Recommendation, World Wide Web Consortium. February 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222.

[12] T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220, 1993.

[13] Anusree.ramachandran et. al. , “Semantic search engine: A survey”, Int. J. Comp. Tech. Appl., Vol 2 (6), 1806-1811, ISSN:2229-6093, IJCTA | NOV-DEC 2011.

[14] "W3C Semantic Web Activity", (2011) World Wide Web Consortium (W3C).

[15] "OWL Web Ontology Language Overview", (2011) World Wide Web Consortium (W3C).

[16] Sikos, L., (2011)"RDF tutorial".

[17] Allemang, D., Hendler, J. (2011). "RDF –The basis of the Semantic Web”, In: Semantic Web for the Working Ontologist (2nd Ed.)". Morgan Kaufmann. doi:10.1016/B978-0-12-385965-5.10003-2.

[18] Web Crawler Application Design. Available: http://searchenginecrawler.blogspot.com/2007/09/webcrawler-application-design.html

[19] The Anatomy Of An Automated Search Engine! Available: http://www.templatesfactory.net/articles/the-anatomyof-an-automated-search-engine.html

[20] F. F. Ramos, H. Unger, V. Larios (Eds.): LNCS 3061, pp. 145–157, Springer-Verlag Berlin Heidelberg 2004.

[21] Cohen, S. Mamou, J. Kanza, Y. Sagiv, Y “XSEarch: A Semantic Search Engine for XML” proceedings of the international conference on very large databases, pages 45-56, 2003.

[22] D. Bhagwat and N. Polyzotis, "Searching a file system using inferred semantic links," in Proceedings of HYPERTEXT '05 Salzburg, 2005, pp. 85-87.

[23] H. L. Wang, S. H. Wu, I. C. Wang, C. L. Sung, W. L. Hsu, and W. K. Shih, "Semantic search on Internet tabular information extraction for answering queries," in Proceedings of CIKM '00 McLean, 2000, pp.243-249.

[24] Jorge Cardoso et. al., “Introduction to Semantic Web Services and Web Process Composition”

[25] Cardoso, Jorge, and Amit Sheth, “Introduction to semantic web services and web process composition”. Springer Berlin Heidelberg, 2005.

[26] W3C, W3C Semantic Web Activity. http://www.w3.org/2001/sw/. 2004.

[27] Uschold, M. and M. Gruninger, Ontologies: Principles, methods and applications. Knowledge Engineering Review, 1996. 11(2): p. 93-155.

[28] Gruber, T.R., toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 1995. 43(5-6): p. 907-928.

[29] Paolucci, M., et al. Importing the Semantic Web in UDDI. in Proceedings Web Services, E-Business and Semantic Web Workshop, CAiSE 2002. 2002. Toronto, Canada.

[30] Cardoso, J. and A. Sheth, Semantic e-Workflow Composition. Journal of Intelligent Infor-mation Systems (JIIS). 2003. 21(3): p. 191-225.