OVERVIEW OF APPROACHES TO SEMANTIC WEB SEARCH

Meena Unni[1] **and K. Baskaran**[2]

[1]Computer Science, Karpagam University, Coimbatore, India.
E-mail: meena_ukh@yahoo.com
[2]Asst. Professor (RD), Dept. of CSE and IT, Govt. College of Technology, Coimbatore, Tamilnadu, India.
E-mail: baski_101@yahoo.com

ABSTRACT

Web search is the hot topic of research today, since it is the one of the main ways of accessing content from the Web. Nowadays a search engine employs a special software robot to search for information on the millions of pages on the web. The software is also called as spiders or bots. Spiders build lists of the words found on Web sites. The spider is a program which follows links from one page to another and from one site to another.

The spider will start from a popular site, adding the words found on this site to the search engine index and then it follows all links found within this site. In this way, the spidering system rapidly moves through different sites covering most of the sites on the web. When you enter a query at the search engine site to search for information, your input is actually checked against the search engine's index which is created by analyzing all the web sites. Infact you are not searching the web. These indices are large databases of information that is collected and stored by the search engine site and subsequently searched when user inputs query.

To make results more meaningful, most search engines rank the pages according to relevance. Such type of search has many limitations, and a number of research activities is being carried out in this direction so that end user will be able to retrieve intelligent information from the web. Such intelligent information retrieval from the web is called semantic search or Semantic Web search. Semantic search augments the traditional web search. In this paper, we will make a preliminary survey over the existing literature regarding semantic search engine, and describe some possible future directions of research.

General Terms: Semantic Search, Web Search, Keyword Based Search, Structured Query Search, Natural Language Search.

Keywords: Semantic Web, Web Search, Semantic Web Search, Semantic Search on the Web, Ontologies.

1. INTRODUCTION

There are billions of web pages on the world wide web. Thus it becomes difficult to search and find useful information from such large amounts of data. Web search was born to facilitate the fast and relevant retrieval of information.

Web search is an important concept of the Web, where in keyword search is used to retrieve information from the web and the retrieved information is ranked according to the link structure of the web. For example if the user gives a query like "cheapest hotels in Dubai" the search engine will start searching for the document which contains all the four words. If we go through the list of web pages, we will find that most of the pages retrieved are not useful. Such searching technique has many short comings. These search engines cannot answer intelligent queries from the user. This has resulted in a number of research activities for retrieving intelligent information from the web, called semantic search on the Web. One of the intense topics of research [2] today is semantic search. Semantic search systems are information retrieval systems that employ semantic technologies to enhance different parts of information retrieval by using structured query languages, keywords or natural languages.

The aim of semantic search technology is to enable the retrieval of accurate information by matching the concept or meaning. Semantic search doesn't replace the traditional web but has the power to enhance it. However, the great hope with semantic search is that it may one day be able to go beyond the keywords that we type, and find exactly what we mean.

The research in this area is important for two reasons. It enhances the traditional keyword based information retrieval by providing search services based on entities, relations, and knowledge. And secondly the semantic web development demands better search paradigms for acquisition, processing, storage, and retrieval of the information.

Such a search permits search queries which are complex and which require reasoning to retrieve information over the web.

This is also very important if the user uses natural language to search for information on the web.

These queries in natural language will have to be transformed into formal queries in a structured query language. Such structured query language is available in the context of semantic search.

Semantic search also includes faceted search where results are derived from facets, which is a collection of predefined categories. If facets are not predefined, then it is called as clustered search. Semantic search also involves searches based on relations or similarities or ontologies.

In this paper, we explain understandings about semantic search on the Web. And how it can be used if the end user wants to search the web using natural language, We provide a review focusing on methodologies, and most distinctive features.

The remainder of this paper is organized as follows. In Section 2, we describe overview of some of the existing approaches to semantic search on the Web. Section 3 illustrates our own proposed approach. In Section 4, we conclude and present a detail account of our vision for the future of semantic search on the Web.

2. OVERVIEW OF EXISTING APPROACHES

Varied semantic search approaches on the Web can be broadly categorised as Approaches which are based on structured query languages and Approaches for novice users, who may have no familiarity with query languages. Approaches for naïve users can be further classified into:

(i) Keyword based approaches where queries consist of lists of keywords;

(ii) Natural language based approaches, where users use natural language to write queries.

Below we give a general review of the main approaches belonging to the above categories.

2.1 Approaches Which are Based on Structured Languages

SHOE is a prototype ontology language for the Web. SHOE can be embedded directly in HTML documents or used in XML documents. SHOE ontology has both an identifier and a version number which avoids the problem to list all versions of the ontology. The included ontology is specified by both identifier and version number. SHOE also provides additional axioms called inference rules. SHOE declares instances to associate knowledge with resources. Every instance points to one ontology, which defines the categories and relations used. To use SHOE, appropriate ontologies must be defined first. Then these ontologies are placed on the internet so that different search engines will be able to access them. Then the information is added to the web pages which are called as annotation. Once the ontologies and instances are placed on the web, Shoe agents and search engines access this information either by direct access approach or repository-based access approach. Finally this information is used by an intelligent web agent to locate useful documents.

Another structured language approach is Swoogle. It is a search engine for Semantic Web documents. Since the documents in semantic web are encoded in RDF AND OWL, Swoogle[4] employs crawlers which are a script to browse the World Wide Web in a methodical, automated manner to locate RDF documents and HTML documents embedded with RDF. Swoogle is designed to support both human users as well as software agents. Swoogle system consist of a database that stores metadata about semantic web documents. Crawlers locate these metadata and compute relationships amoung them. It has a simple user interface for querying and locates semantic web documents which are then indexed and metadata is generated, generated metadata is analyzed and ranked and finally agent based services allows access to the metadata and navigates the Semantic Web.

The corese works internally on conceptual graphs. In order to match a query with annotation according to a shared ontology, corese translates the query, annotation and ontology into conceptual graph model. Ontology in corese is represented using RDF Schema. It also handles OWL Lite. Corese's query language is extended so that mismatches between user query and ontological concepts can be taken into consideration. Corese[11] will return result even if exact answer doesn't exist by approximating the query's structure, semantics or both. Corese will calculate the semantic distances between ontological types during approximation. On this basis, it returns web resources which has the same annotations like query or those whose annotations are semantically close. Corese measures similarity of retrieved annotations with the query and only those results are returned to the user whose similarity doesn't exceed a given value. The result is sorted in descending order of similarity. Corese is able to retrieve result if there is a difference between annotation structure and query structure. Such approximations are supported through the path graph feature. Here the resources are searched through the path graph feature.

Another approach [13] queries RDF repositories for approximate answers. This method approximately queries RDF datasets with SPARQL. SPARQL query is encoded as a set of triple constraints with variables and finding approximate solution is through standard evolutionary methods guided by the number of satisfies constraints. This method approximates the dataset. In this method, evolution of the result set can be stopped at any point without satisfying all the constraints. This could result in some incorrect result and could also be incomplete as some possibilities would not have been explored.

One of the more recent approaches based on structured languages is ONTOSEARCH2.

ONTOSEARCH2[5] is an ontology based search engine. It has two components, an ontology repository and a query engine. It queries the repository using SPARQL. Additional ontology can be added to indexby RDF data.

The Coraal[8] system is a knowledge based search engine for biomedical publication. Meta data along with the knowledge present in the text is extracted and the content is integrated with the domain knowledge. Texts and meta data are represented as RDF graphs in a triple store. Natural language processing heuristics is used for processing the texts. This way coraal exploits large scale knowledge associated with the text.

NAGA[2] is a graph based query language in which the nodes represent entities and the edges represent the relationships between the entities. A fact in NAGA semantic search engine is an edge in the graph. Facts are derived from various web data sources. A confidence value is attached to each fact and this value reflects the authority of source and certainty of extraction process. The user can formulate queries providing information regarding the requirements using NAGA's query language. NAGA may return multiple answers for the user's query. The query results are ranked according to a scoring mechanism taking into consideration the confidence values.

2.2 Approaches Which are Based on Keywords

Below we propose two preliminary approaches to the problem of semantic search. OntoSelect is an ontology search engine which centers on issues dealing with ontology search. The users can specify the ontology title or the topic of interest in order to search for ontologies.

OntoSelect[6] is a dynamic web based ontology library used in knowledge markup, which integrates, analyzes and selects an ontology on the semantic web. Ontoselect not only allows users to search for ontologies on any domain but also checks the web for new ontologies in the representation format (RDFS,DAML, OWL). Ontologies are analyzed using the OWL and extracts the structure and content. Ontologies are organized according to class and property names or class and property labels. After appropriate ontology is selected, a relevant document will be marked with the knowledge.

Below we outline more recent approaches for novice users based on keyword search.

Semsearch[12] is a keyword-based semantic search engine. It hides the complexity of semantic search from the end users who may not be familiar with the problem domain or with the specified query language. Semsearch is a layered architecture which segregates end users from complex data repositories. It consists of five layers: (i) Query interface layer like google where in end users can specify queries using multiple keywords; (ii) a Text search layer, which finds out the semantic meaning of the user keywords using either ontology or semantic data repositories. This layer is composed of index engine and search engine. Index engine indexes documents and their associated semantic entities and search engine which searches for semantic entity matches for the given user keywords (iii) semantic query layer, which translates entered by user into formal queries. This layer consists of a formal query construction engine (iv) a formal query layer, which retrieves results from the semantic data repositories. It consist of ranking engine which ranks the searched results in terms of the relevance of user queries and (v) a semantic data layer, which contains ontologies and semantic metadata of the problem domain that are gathered from heterogeneous data sources.

Falcons extend the keyword-based search system by using ontology driven approach. It is a search system for objects and concepts (classes and properties), on the Semantic Web. Ontology and a knowledge base for the domain is built by crawling the world wide web and indexing them on the objects in the knowledge base. Falcons[1] search is based on the associations between objects. Associations between entities are meaningful paths consisting of one or more properties which are predefined. When user executes a query, the user starts with an individual in the domain to start exploring. Because of the semantic association between objects, user browses all the linked objects and then corresponding objects will be returned as results. Falcons exhibit two types of associations. One is the basic associations which are defined as paths between classes and another is compound associations which are composed with the pre-defined basic associations. Falcons provide the end users with the flexibility to customize their own associations as queries. These associations are converted into RDF query statements, which are used to search the knowledge base.

Two recent semantic search engines SWSE [9] and Sig.Ma [10] allows users to locate RDF entities using keyword search.

The SemanticWeb Search Engine (SWSE) contains components for crawling, ranking and indexing data. SWSE also handles RDF data. The crawler retrieves a large set of RDF data from the web and finds synonym identifiers in the data. Links based analysis is performed over the crawled data and scores are given to individual elements by the ranking component. The reasoning component evaluates the trustworthiness of data and materializes new data depending on the semantics. Indexing component prepares an index which supports the user while retrieving information.

The core of SWSE is the Semantic Search and Query Engine is YARS2. It is a semantic search engine that retrieves query results using graph structured data model. It can scale to keep up the growing data volume in the web. Distributed architecture is used for indexing and querying. YARS2[3] uses multicrawler, which transforms data from multiple sources into RDF. RDF extends triple with context as a quadruple or quad .URIs is used to uniquely identify entities within RDF. It is provided with a local index creation and management system. On text it manages indices on keyword, on graph structure it manages statement indices such as quad. For combinations of data values it manages join indices. The query processor along with index management system executes the query over the network in a parallel multithreaded fashion and returns result. Evaluation of queries is done in SPARQL format. Returned results are ranked based on the relevance using ReConRank. ReConRank derives ranks of entities and data sources which is useful in prioritizing the results.

Like SWSE, Sig.Ma combines results from several sources. An aggregate view of information is derived from these sources and is provided to the user along with the sources from where it is derived.

Sig.ma starts with a search based on a keyword or simple structured query. It provides an aggregated view of the heterogeneous data sources returned by a query into a single entity profile. Entity profile is a summary of an entity that is presented in a visual interface to the end user or returned as a RDF document. Sig.ma also provides links from one entity profile to another. The returned results are ranked using ranking matric. Sig.ma offers tools to the user to modify the source list in order to refine the presented entity profile. It also offers many interactive data cleaning mechanisms. This step generates a user driven expansion refinement loop.

2.3 Natural-language-based Approaches

In this section we discuss some of the well known approaches based on natural language queries.

Orakel[7] is an ontology based natural language interface wherein it accepts questions in natural language and obtains response on the basis of a given knowledge base.

The user asks questions beginning with wh pronouns such as where, which, what, who etc which is interpreted by the query interpreter. Query interpreter parses and translates into a logical form. This logical form is translated into knowledge representation language of the knowledge base by query converter component. The answer generation component evaluates the query and returns answers to the user. The only resources used to answer the queries are the general and the domain specific lexicon sources created by the domain experts. ORAKEL works with the knowledge representation language *F*-Logic and query language is as implemented in Ontobroker system.

Sig.Ma[10] supports both Semantic Web search over ontologies and over non Semantic Web documents. As regards to semantic web search over ontologies, PowerAqua, which is a previous system is used to answer the natural language query.

Unlike other semantic search engines, PowerAqua is not limited by single ontology. It uses multiple ontology based question answering system. Information in the semantic web is distributed across heterogeneous resources. The end user inputs queries in natural language. The element mapping component identifies the ontology relevant to the query. In case there is an ambiguity regarding the ontology discovery, word sense disambiguation technique is used to disambiguate between different interpretations across ontologies. Triple mapping component determines the most likely interpretation of the query as a whole. Merging and ranking component return results after applying a set of ranking criteria to sort the results, by composing information from multiple heterogeneous sources of different domains.

Google's latest version uses natural language based approaches for its query. Google has recently enhanced its search capacity by adding a new functionality, which gives more accurate answers to queries than before. It correctly answers simple queries like "barack Obama date of birth". It returns the answer and the link from where the information is extracted is also shown.

3. OTHER DISCUSSION

Users natural language query request is processed by the ontology based component. The user query is translated into the terms of the available ontologies and retrieves a list of available ontological entities as a response. These ontologies are indexed a priori. There are a number of ontologies which cover various domains.and in semantic search, response is obtained from these unlimited number of ontologies. E.g given the query "How to play bridge?" and the two ontologies covering the term "Bridge" (one about bridge on river and one about card game bridge). Here it should be able to select appropriate ontology after disambiguating the query using its context and available information and obtain an answer in the form of ontological entities.

Once the exact answer to the user's query has been retrieved, the system performs operations to retrieve Web documents. The retrieved document should be ranked according to relevance and displayed to the user.

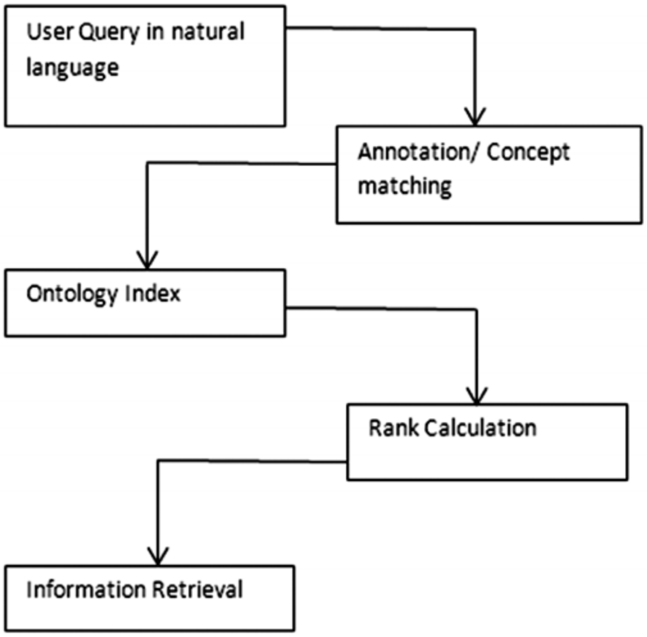

Figure 1: Framework for Ontology Based Information Retrieval

4. CONCLUSION

In this paper, we have given a general review of approaches to semantic search on the web. We have categorised approaches based on structured query language, keyword based or natural language based. Although many approaches for semantic search exist, research in this area is at its initial stages. Very little progress has been done in the last few years. It is likely to change in the near future.

One area where the research needs urgent attention is to find out is how to automatically extract knowledge from the web content. Another area where immediate attention required is to find out how to translate query in natural language into formal ontological queries. In that case how to create and maintain the underlying ontologies. Should this be done manually by experts as done in Wikipedia like manner or it should be done automatically by extracting from the existing pieces of ontological knowledge and annotations in the web or semiautomatically i.e combining the the above two methods.

With the increase in size of automation, bigger will be the size of ontologies and subsequently the cost and efforts for generating and maintaining them will be lesser.

One more point to note is semantic search mainly focuses on the trust and the quality of knowledge which varies from source to source.

Getting answers from the web for simple questions given by users in natural language is still a science fiction. However with the advancement of research towards semantic web search, this could become a possibility in the near future.

ACKNOWLEDGEMENTS

Our thanks to Karpagam university for providing the needed infrastructure to carry out the study.

REFERENCES

[1] Honghan Wu, Gong Cheng, and YuzhongQu, “Falcon-S: An Ontology-Based Approach to Searching Objects and Images in the Soccer Domain.

[2] GjergjiKasneci, Fabian M., Suchanek, Georgiana Ifrim, Maya, Ramanath, and Gerhard Weikum, “NAGA: Searching and Ranking Knowledge”, MPI–I–2007–5–001 March 2007.

[3] Andreas Harth JüurgenUmbrichAidan Hogan Stefan Decker, “Yars2 : A Federated Repository for Searching and Querying Graph Structured Data”, DERI Technical Report.

[4] Li Ding, Tim Finin, Anupam Joshi, Rana Pan, R. Scott Cost Yun Pena, Pavan Reddivari, Vishal Doshi and Joel Sachs, “Swoogle: A Search and Metadata Engine for the Semantic Web”, ACM 2004.

[5] http://www.foaf-project.org.

[6] http://www.daml.org/ontologies/

[7] Philip Cimiano, Peter Haase, J. ORGHeizmann, Matthias Mantel, “ORAKEL: A Portable Natural Language Interface to Knowledge Bases”, March 1, 2007.

[8] VitNovacek_ and Tudor Groza and Siegfried Handschuh and Stefan Decker, CORAAL; Dive into Publications, Bathe in the Knowledge, Digital Enterprise Research Institute, National University of Ireland Galway, IDA Business Park, Dangan, Galway, Ireland.

[9] Andreas Harth, Aidan Hogan, Renaud Delbru, J. UrgenUmbrich, Sean O’ Riain and Stefan Decker, National university of Ireland, Galway, DERI.

[10] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, SzymonDanielczyk, Renaud Delbru, Stefan Decker, Sig.ma: Live Views on the Web of Data, DERI, Galway, Ireland.

[11] http://www-sop.inria.fr/edelweiss/

[12] Yuangui Lei, Victoria Uren, and Enrico Motta, *Knowledge Media Institute*, the Open University, Milton Keynes.

[13] Eyal Oren, Christophe Gúeret, and Stefan Schlobach, Any time Query Answering in RDF through Evolutionary Algorithms, ISWC 2008.