(Draft)
International Symposium on Distributed Computing and Artificial Intelligence 2010 (DCAI-2010).
Advances in Intelligent and Soft-Computing. Springer, 2010, vol. 79, p. 613-620.

Exploring the Advances in Semantic Search Engines

Walter Renteria-Agualimpia[1], Francisco J. López-Pellicer[1], Pedro R. Muro-Medrano[1], Javier Nogueras-Iso[1], and F.Javier Zarazaga-Soria[1]

[1] Computer Science and Systems Engineering Department, University of Zaragoza.
Zaragoza, Spain
{walterra, fjlopez, prmuro, jnog, javy}@unizar.es

Abstract. With the vertiginous volume information growing, the amount of answers provided by traditional search engines and satisfying syntactically the user queries has enlarged directly. In order to reduce this problem the race to develop Semantic Search Engines (SSE) is increasingly popular. Currently, there are multiple proposals for Semantic Search Engines, and they are using a wide range of methods for matching the semantics behind user queries and the indexed collection of resources. In this work we survey the semantic search engines domain, and present a miscellaneous of perspectives about the different classification of approaches. We have created a comparative scheme and identified the prevalent research directions in SSE.

Keywords: Semantic Search Engines, Semantic Web, Information Retrieval.

1 Introduction

With the vertiginous increase of volume information on the Web, the results provided by traditional search engines in response to user queries do no longer satisfy the needs of specific communities of users. There is an increasing amount of answers that satisfies the terms contained in user queries. However, these answers are not precise enough for some users demanding a more refined list of results according to the semantics of their queries. This open problem has motivated a new era of search systems that have received the name of Semantic Search Engines.

A Semantic Search Engine (SSE) can be understood as a semantic Web application that can answer questions based on the meaning of users query specification, resources in the repositories and in many cases it is based on predefined do-

(Draft)
International Symposium on Distributed Computing and Artificial Intelligence 2010 (DCAI-2010).
Advances in Intelligent and Soft-Computing. Springer, 2010, vol. 79, p. 613-620.

main semantics or a knowledge model. SSE can return relevant results on your topics that do not necessarily mention the word you searched for explicitly.

The goal of this work is to study and discuss various widespread research directions in semantic search engines, as well as identifying common features and main approaches used in them. In this work we can get an overview of current approaches to semantic search and its state of development, but not an exhaustive review of all implemented systems.

The rest of the paper is structured as follows. The next section shows several schemes of classification approaches, in order to identify what kind of semantic search approach is the base for each SSE. And then we analyze the trend and prevalent research directions in the SSE domain. Section 3 provides a comprehensive analysis based on a survey of more than 30 SSE. We discuss and compare the classifications of approaches. All results of this analysis are available online as linked data (see sect. 3). Finally we summarize our main conclusions in section 4.

2 Different schemes for comparison and classification of SSE

Currently, there are multiple proposals for Semantic Search Engines, and they are using a wide range of methods for matching the semantics behind user queries and the indexed collection of resources. Several authors have studied the current status of semantic search engines from different viewpoints. We present a review of approaches from a research perspective.

A first review was presented by Miller et al. [5]. They present a classification based on the intention of users. That is, if the users want to navigate to a particular intended document, this approach is called: *Navigational Searches.* On the other hand, there maybe users trying to locate a number of documents, which together will give them the information they are trying to find. This is *Research Searches*.

Mangold [2] presents a categorization scheme that he uses to classify different approaches for semantic search along several dimensions. His classification is based on the next criteria: architecture, coupling, transparency, user context, query modification, ontology structure and technology. The analyzed approaches were implemented by the next technologies: SHOE, Inquirus2, TAP, Hybrid spreading Activation, ISRA, Librarian agent, SCORE, TRUST, Audio, and Ontogator.

The next author viewpoints are more focused in the semantic processing method to resolve queries. Mäkelä identifies five distinct research directions emerged and prevalent research directions in semantic search, based on similarity of research goals[1]. He observed that sometimes the categories do not differ much in methodology, but they seem sufficiently separate. His classification is:

- *Augmenting Traditional Keyword Search with Semantic Techniques:* more specific ontological techniques are used. i.e, Terms are expanded to their synonym and meronym sets [9]. Direct ontological Browning is supported. The intention is to find related concepts as the writer of the document [15].

- *Basic Concept Location*: The main goal is to locate instances of the core semantic web formed by concepts, instances and relationships. Users can choose the class of instances by means of ontological navigation [7].
- *Complex Constraint Queries*: Many SSE with this approach are based on navigating the ontology as the last approach [13]. One way is based on a global intersection of distinct selectors, constraining do not need to be ontological.
- *Problem Solving*: The SSE use ontological knowledge to solve a problem; searching for solutions by inference and other reasoning techniques [14, 16].
- *Connecting Path Discovery:* The SSE are based on the ideas of a vast amount of varied semantic data will be available to be mined for semantic connections. The major technical problems are the locating complex and hidden relations.

Hildebrand et al. [3] systematically scanned proceedings about Web Semantics to compile a list of end-user applications described or referred to. For each system they collected basic characteristics such as the intended purpose, intended users, the scope, the triple store and the technique or software used for literal indexing, giving a total of 35 systems. Based on the data resulting from the survey they perform a more thorough analysis of the three individual phases in the search process: *query construction in section*, *search algorithms*, *presentation of the results.*

Now, in the search algorithms stage, we can find the semantic component, that is, the main interest in this work. The cores of SS approaches identified are the:

- *Graph Traversal*: Takes only the structure of the graph into account. It uses weighted graph search algorithm. Weights reflect the importance of relations.
- *Query Expansion*: Thesaurus relations are used for query expansion. Semantic matching with hierarchical broader, narrower and the associative related term.
- *Spread Activation*: It uses weights as well as the number of incoming links.
- *RDFS/OWL Reasoning*: Has the ability to influence the search results. RDFS. Some SSE support OWL reasoning based on logic programming or rules [12].

Dietze and Schroeder [6] suggest a new classification approaches. They developed an interesting study about 27 SSE and use a classification based on 9 criteria: structured/unstructured file, ontologies, text mining type, number of documents, type of documents, clustering, result type, highlighting, scientifically evaluated.

Dong et al. [4] present a extended classification: Semantic Search (SS) Algorithm based on the Graph, SS Methodology on Distributed Hash Tables (DHT), Logics (DL)-based Information retrieval (IR) Thesaurus-DL form Knowledge Base (TK), DAML+OIL-based Semantic Search, Keyword-based Search Engines combined with Semantic Techniques, SSE based on Ontology Annotations, Agent-based SSE, SS Engine and XML Objects, Semantic Multi-media SE.

Finally, Grimes [11] presents an extensive classification of approaches:

- *Related searches/queries:* The SSE recommends searches that are in some "sense" similar to the user search.
- *Reference results*: SSE is responding with resources that define the search terms, via a dictionary look-up, or elaborately, pulling Wikipedia pages.
- *Semantically annotated results*: SSE returns pages or documents with highlighting of text features, especially named or pattern-defined entities.

- *Full-text similarity search:* SSE use a block of text ranging submitted from a phrase to a full document, rather than a few keywords.
- *Search on semantic/syntactic annotations*. Users define the semantic of search by means of indicate the syntactic role the term play.
- *Concept search*: The SSE identifies specific concept to seek the original and their equivalent concepts semantically.
- *Ontology-based search:* SSE can understand hierarchical relationships of entities and concepts as in taxonomy, and more complex inter-entity relations.
- *Semantic Web Search*: SSE capture data relationships and make the resulting "Web of data" queryable.
- *Faceted search:* It provides a means of exploring results according to a set of predefined, high-level categories called facets.
- *Clustered search:* It is like faceted search, but without the predefined categories. Meaning is inferred from topics extracted from the search results.
- *Natural language search:* The SSE understands the semantic behind the questions, and present answers in natural language.

A summary about the classifications is presented in Table 1. The 5th column shows the final classification based on Grimes, because this is more extensive than other perspectives. The goal is identify the main active areas in SSE domain.

Table 1. Comparison of semantic search Approaches

Mäkelä	Hildebrand	Dong et al.	Grimes	Proposed final classification
Connecting Path Discovery	Graph Traversal	SS Algorithm based on the Graph	-	SS based on Graphs
Augmenting Traditional Keyword Search with Semantic Techniques	Query Expansion	Keyword-based SE with Semantic Techniques	Related Searches/Queries	Related Searches/Queries
	Spread Activation	SSE based on Ontology Annotations	Search on Semantic/Syntactic Annot.	Search on Semantic/Syntactic Annotations
			Semant. Annot. Results	Semant. Annot. R.
Problem Solving Complex Constraint Queries	RDFS/OWL Reasoning	Agent-based SSE (DL)-based on IR TK DAML+OIL-based SS	Ontology-based Search	Ontology-based Search
-	-	SSE and XMLObjects Semantic Multimedia SE	Semantic Web Search	Semantic Web Search
-	-	SS Methodology on DHT	Reference Results	Reference Results
-	-	-	Full-Text Similarity S	Full-Text Similarity S
Basic Concept Location	-	-	Concept Search	Concept Search
			Faceted Search	Faceted Search
			Clustered Search	Clustered Search
-	-	-	Natural Language Search	Natural Language S.

(Draft)
International Symposium on Distributed Computing and Artificial Intelligence 2010 (DCAI-2010).
Advances in Intelligent and Soft-Computing. Springer, 2010, vol. 79, p. 613-620.

3 Analysis of current SSE

The methodology used in this work was based on 4 steps. The first step was to review the applications available in the Web, publications and projects in the state of art. Then, we evaluated a series of parameters (see below for the list of parameters and their description) for each semantic search engine. The third step was to contact some authors because the available information for some engines was uncomplete. However, in some cases we could not contact some of the authors. Then we complement the analysis with the several previous works and their operation mode and results [1, 2, 4]. Finally the complete research results are detailed in an updatable technical report and all results are published in the research group portal IAAA[1] by means of a RDF file, and linked data in order to obtain more feedback.

We have studied different scheme classifications and the several author viewpoints about main semantic search approaches. Numerous criteria and parameters have been used in this purpose. Our objective is not to reward or dismiss those proposals; but to identify the predominant or prevalent active areas or approaches by means of exploring many semantic search engines at present.

Researchers and developers are aware of the need to improve traditional engines, including features like: *user feedback; results explanation* and compressive presentation of results; and more dialogue with the users about possible problem with their request, e.g *ambiguity advertisement.*

Many of these aspects are related to human understanding, but it is important to study the *interoperability*, that is, to analyze what kind of interoperability do SSE present? Is the SSE a machine or informatic agent queryable? We have summarised the SSE exploration in Table 2, which show the following 8 parameters:

- *Main approach(es):* This field identifies the type of approach used by each SSE. The type of approach was presented in Table 1; it was obtained by means of unifying the Grimes classifications with the other approaches unmentioned. The complete results are available in the RDF file cite above.
- *Features:* It is a description about the main SSE qualities.
- *Type of Result*: It specifies the query result: summary, link, free text or other.
- *User feedback:* This is useful when there are multiple controlled terms that match with the free text input semantically. There are two ways. The first one is called “pre-query disambiguation”, allow us to select the intended term before it is processed by the search algorithm [10]. The second way is called “post-query disambiguation”; feedback is taking into account on the results.
- *Multilingual*: Multiple language support.
- *Interoperability*: It evaluates if the SSE is able to exchange machine-understable content by mean of a standard protocol.
- *Result explanation*: Here we recognize if the SSE argue the query answer, justifying by means a graph, conceptual structure or other.

[1] http://iaaa.cps.unizar.es/openknowledge/papers/2010/dcai/sse/

- *Ambiguity alarm*: In many cases, there are results that match with the query. SSE must advert to user about the different senses that satisfy the query.

Additional we present two features available in online version, as following:

- *Geospatial component*: It allows evaluate as if the SSE takes into account additional richness aspects, such as *geospatial location information* when is required to complement or clarify the semantic or to confirm the result sense. i.e Washington state instead of Washington president (see RDF online).
- *Availability*: It examines if the Web application is available now (see RDF).

It is worth noting that we had to face the problem that some systems, Web applications and publications describe their approaches from a very abstract viewpoint. For this reason we relied on the given information without knowing the deep details, but assigning and classifying the SSE according to their external description and comparing with similar semantic search engines.

Table 2. Comparison of semantic search Engines

Engine	Main Approach(es)	Features	Type of Result	M [a]	Interoperability	RE [b]	AA [c]
SenseBot	Concept Search	Text mining	Summary	Yes	SOAP, REST	No	No
BotPowerseet	Natural Language Processing (NLP)	Free text input, disambiguate.	Summary	Yes	-	Yes	Yes
DeepDyve	Semantic/Syntactic Annot., Reference results	Analysis across large amounts of data	Summary	Yes	-	No	No
Cognition	NLP	Business, APIs	Link	Yes	API	Yes	Yes
Hakia	Related searches, NLP	Excellent resumes	Link & Free text	Yes	Yes	Yes	No
TrueKnowledge	Ontology-based search, Semantically annot. results	Questions – answering	Summary and classification	No	Direct Answer API, Query API	Yes	Yes
Open Mind Common Sense	NLP, concepts search	Learn general knowledge	Free text	No	-	No	No
Swoogle	Semantic Web search	Semantic Web documents.	OWL, RDF	No	REST web service	No	No
TrueVert	Concept search, NLP and Clustered results	model of word relations in context	Free text	Yes	-	No	No
Wolfram Alpha	Reference results, Ontology-based search, Clustered search	Web, parallel computing, mathematical, grid knowledge	Taxonomy, graph	Yes	REST API	Yes	No
Duck Duck Go	Clustered search, NLP	Zero-click Info above links, Disambiguation	Summary	Yes	XML-based API	-	Yes

[a] Multilingual, [b] Result explanation, [c] Ambiguity alarm

(Draft)
International Symposium on Distributed Computing and Artificial Intelligence 2010 (DCAI-2010).
Advances in Intelligent and Soft-Computing. Springer, 2010, vol. 79, p. 613-620.

In table, the symbol "-" represents unknown information. The main parameter of comparison in the table is the second column "*Main approach(es)*". It allows us to identify the research areas with the more intense activity in the semantic search.

Which is the prevalent Semantic Search Approach? Have the Semantic Search Engines analized a unique approach? Taking into account these questions and the information provided in Tables 2 and 3, we can see five main groups with major activity, i.e. a significant number of SSE using that approaches. Those groups are: in first place, Concept Search, Faceted Search, Clustered Search; then Search Engines based on NLP; in third place, SE based on Related Searches/Queries, Search on Semantic/Syntactic Annotations and Semantically Annotated; then Ontology-based Search; and finally Semantic Web Search. We have analized several SSE implementing different approaches, and based on combinations of the last groups mentioned. Probably these research directions will be the dominant approaches.

Table 3. Summary of prevalent research directions in SSE

Group	Approach(es)	Number of SSE using this approach
Group 1	SS based on Graphs	0
Group 2	Related Searches/Queries, Search on Semantic/Syntactic Annotations, Semantically Annotated Results	9
Group 3	Ontology-based Search	9
Group 4	Semantic Web Search	8
Group 5	Reference Results	4
Group 6	Full-Text Similarity Search	0
Group 7	Concept Search, Faceted Search, Clustered Search	11
Group 8	Natural Language Search	10

4 Conclusions

There is one common idea in the majority of approaches, that is, the machines must understand the meaning behind the Query and Data sources in order to return answers based on the meaning. Maybe this is the main requirement for a SSE. Intuitively we can say that many SSE will be based on a similar core, including conceptual structures such as ontologies, and founded on main components to process queries in form of natural language. A direct consequence is the need to develop SSE allowing the users to play a part of the answers, before and after the query, that is, pre-query disambiguation, advertisement of ambiguity presence, and feedback to improve futures answers.

Another aspect related to the semantic legibility intrinsically is the system ability to explain results, that is, what was the form to generate one or other result? In this aspect, many systems are working to improve the visualization and interpreta-

tion form strongly, e.g. some SSE such as *Wolfram*[8], *Google*[2], and Kolline [6] provide visualization of results by means of concept connection graph or surfable graph.

Acknowledgments. This work has been partially funded by the Spanish government through the projects "España Virtual" (ref. CENIT 2008-1030) and TIN2009-10971, and the Government of Aragon through the project PI075/08.

References

1 Mäkelä E. Survey of semantic search research. In Proc. of the Seminar on Knowledge Management on the Semantic Web, 2005.
2 Mangold C. A survey and classification of semantic search approaches. International Journal of Metadata, Semantics and Ontologies, 2(1), 2007, pp. 23–34
3 Hildebrand, M., Ossenbruggen, J., and Van Hardman. L., An analysis of search-based user interaction on the semantic web. Report, CWI, Amsterdam, Holland, 2007.
4 Dong, H., Hussain, FK., and Chang, E., A survey in semantic search technologies, Second IEEE International Conference on Digital Ecosystems and Technologies, 2008.
5 Guha, R., McCool, R., and Miller, E. Semantic Search. Proceedings of the WWW'03, Budapest, 2003.
6 Figueira, F., Porto de Albuquerque, J., Resende, A.; Geus, Lício de Geus, P., Olso, G., *A visualization interface for interactive search refinement*. In: Proc. 3rd Annual Workshop on Human-Computer Interaction and IR, pp. 46-49, Washington DC, 2009.
7 Mäkelä, E., Hyvönen, E., Saarela, S., Viljanen, K.: OntoViews - A Tool for Creating Semantic Web Portals. In: Proc. of the 3rd International Semantic Web Conf., 2004.
8 Wolfram alpha, system available at, http://www.wolframalpha.com/ (Mach/2010)
9 Buscaldi, D., Rosso, P., Arnal, E.S.: A wordnet-based query expansion method for geographical information retrieval. In: Working Notes for the CLEF Workshop, 2005.
10 Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A Browser for Heterogeneous Semantic Web Repositories. In: The Semantic Web – ISWC, 2006, pp. 272–285.
11 Informationweek,http://intelligent-nterprise.informationweek.com/channels/information_management/showArticle.jhtml;jsessionid=QTPB04LAEZJMHQE1GHOSKHWATMY32J VN?articleID=222400100 (March-2010).
12 UMBC: F-OWL: An OWL Inference Engine in Flora-2 http://fowl.sourceforge.net/, (Ap/10).
13 Zhang, L., Yu, Y. Yang, Y., Zhou, J., and Lin, C., "An Enhanced Model for Searching in Semantic Portals," in In WWW'05: Proceedings of the 14th international conference on World Wide Web. ACM Press, 2005, pp. 453–462.
14 Duke, A., Glover, T., Davies, J.: Squirrel: An Advanced Semantic Search and Browse Facility. In: Proc. ESWC, Innsbruck, Austria, 2007.
15 Rocha, C., Schwabe, D., de Aragao, M.P.: A hybrid approach for searching in the semantic web. In: Proc. of the 13th international conf. on World Wide Web, 2004, pp. 374–383.
16 Wine Agent 1.0, http://onto.stanford.edu:8080/wino/index.jsp, (Mach 2010)
17 Guha, R., McCool, R., Miller, E.: Semantic search. In: WWW'03: Proc. of the 12th international conference on World Wide Web, ACM Press, 2003, pp 700–709

[2] At the time of writing this paper, one could reach the Google tool by selecting "Show options" in the main page of Google.