A Categorization Scheme for Semantic Web SearchEngines

Prof. Haresh R. Parmar

Assistant Professor in Computer Engineering Department
Silver Oak College of Engineering & Technology, Ahmedabad.

Abstract—Semantic web search engines are evolving and manyprototype systems and some implementation have been developed.However, there are some different views on what a semanticsearch engine should do. In this paper, a categorization schemefor semantic web search engines are introduced and elaborated.For each category, its components are described according to a proposed general architecture and various approaches employedin these components are discussed. We also propose some factorsto evaluate systems in each category**.**

I. INTRODUCTION

This paper tries to analysis semantic search engines andprovides a rational categorization scheme for them. To the bestof our knowledge, in this regards there is no work reported.However, there is only a short explanation in [19].According to [5], semantic web (hereafter is referred toas SW) has some stinguishing features that affects searchprocess:

• Instead of web documents, in the SW, all objects of thereal world are involved in the search.

• Information in SW is understandable by machines as wellas human.

• SW languages are more advanced than html. It is possible to distribute information about a singleconcept in SW.Therefore semantic search engines have following fundamentaldifferences to the traditional search engines:

Using a logical framework lets more intelligent retrievalpossible.

• There are more complex relations in documents resultingin the importance of the problem of meta-data maintenance,update and more complex ranking.

• Specifying relationships among objects explicitly highlightsthe need for better visualization techniques for theresults of a search.One important aspect of SW search is the usage of ontologyand meta-data. Ontology provides explicit conceptualizationfor entities in a specific domain. Another important aspectis the annotation for the current web pages. Annotations aremeta-data useful for machines to understand the content of aweb page. In such meta-data concepts are pointers to alreadydefined ontologies [24].

Respecting the kind of search in SW,it is possible to categorize users to two groups. One group areordinary users that do searching as like the current web butdemand more accurate and complete results than traditionalsearch engines. Second group are application developers inthe SW where their primary goal is to search and retrieve SWdocuments. According to these two categories of users, we cancategorize SW search engines to the following two categories:

• Engines specific to the SW documents: they search onlydocuments that are represented in one of the languagesspecific to SW.

• Engines that tries to improve search results using SWstandards and languages. Using context information (representedby domain ontology and metdata) is one of the
key aspects for these engines.
The paper is organized as follows. In section 2 different typesof annotation and methods of generating them are discussed. Section 3 provides explanations about search engines fordevelopers and advanced users in SW. Search engines that use
SW concepts to provide better search results are discussedand further categorized in section 4. For each category, insection 5, we have a brief analysis and propose some points forevaluating them. Section 6 concludes the paper by providing
a summary of all reported works in SW search.

II. ANNOTATION METHODS

One of the major problems facing researchers in SW isannotation which is a prerequisite for SW search engines. Toadopt current web pages for SW search engines they should beannotated by finding appropriate meta-data to be added to each
one. There are different approaches which spawn in a broadspectrum from complete manual to full automatic methods.Selection of an appropriate method depends on the domainof interest [24]. In general meta-data generation for structureddata is simpler [24].Annotations can be categorized based on following aspects:
• Type of meta-data: According to [25] meta-data canbe divided to two types of Structural and Semantic. Inthe former, non contextual information about content isexpressed (e.g. language and format). In the later, themain concern is on the detailed content of informationand usually is stored as RDF triples.
• Generation approach: a simple approach is to generatemeta-data without considering the overall theme of thepage and only using structural information of a pagetogether with natural language processing techniques. Abetter approach is to use an ontology in the generationprocess. In this case it is possible to use clusteringmethods to distinguish the general type of a page [25].Then using a previously specified ontology for that type,
generate meta-data that instantiates concepts and relationsof ontology for that page. The main advantage of thismethod is the usage of contextual information.• Source of generation. The ordinary source of meta-datageneration is a page itself but sometimes it is beneficialto use other complementary sources. For example [1] and[5] discusses about using network available resources foraccumulating more information for a page. For examplefor a movie it might be possible to use IMDB to extractadditional information like director, genre, etc.
Although there is no complete reference about meta-data but[18] provides a rather complete list of systems that generatemeta-data.

II. ONTOLOGY SEARCH ENGINES

According to [1] and [8], for the following reasons, it is notpossible to use current search engines for SW documents:
• Current techniques does not let to index and retrievesemantic tags.

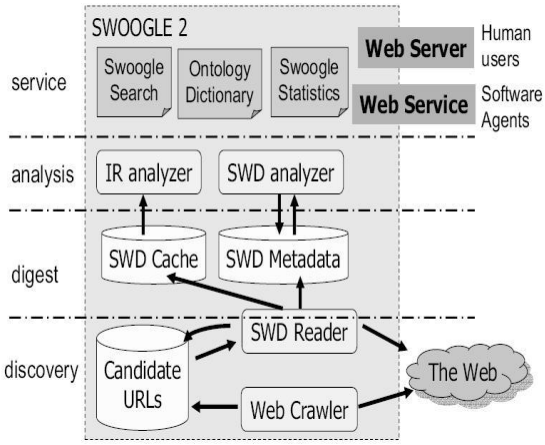
• They don't use the meaning of tags
• Can't display results in visual form
• Ontologies are not separated entities which usually havecross references that current engines don't process.In general there are two approaches to handle these documents:using current search engines with some modificationsor creating a special search engine. In what follows each ofthese approaches is further elaborated.

A. Ontology Meta Search Engines

This group do retrieval by putting a system on top of acurrent search engine. There are two types of this systems.

In the first type, there is a search engine that only searchesspecific file types (e.g. RSS, RDF, OWL). The main concernsof such systems are on the visualization and browsing ofresults. For example in [8] an engine forwards a users request for a specific file type to Google search engine and then usinga visualization tool lets user to navigate and display results.

In the second type there is possible to search on semantictags. But since those tags are ignored by the underlyingsearch engine, an intermediate format for documents anduser queries are used. In [2] a technique named Swangle issued for this purpose. With this technique RDF triples aretranslated into strings suitable for underlying search engine.For example consider the following triple which is in n3 Notation.



Each of these terms are converted to a string and added tothe document for indexing. On the other side, this translationis done for user queries too.

B. Crawler Based Ontology Search Engines

These engines uses a specific crawler for SW documents.

In the figure 1 one complete system is shown [4]: Architectureof a SW document specific search engineHere, based on the four sections specified in the architecture, Analysis of them is given.*1) Discovery:* Crawling of SW documents are different fromhtml documents. Actually they are knowledge crawlers whichare more complex than traditional ones [27]. In SW we expressknowledge using URI in RDF triples. Unlike html hyperlinks,URIs in RDF may point to a non existing entity. Also RDFmay be embedded in html documents or be stored in a separatefile.

Such crawlers should have the following properties [27]:

• Should crawl on heterogeneous web resources (owl, oil,
daml, rdf, xml, html) .
• Avoid circular links
• Completing RDF holes
• Finding new semantic web documents from informationin the currently under

process document (e.g. Extend andImport specifications).

Priority	Relationship	Language Specific
1	instantiation	rdf:type
2	subClass	rdfs:subClass, daml:subClass
3	domain/range	rdfs:domain, daml:range

TABLE I
RELATIONSHIP TYPES

In [26] crawling of SW documents is explained in detail.According to [27] it is possible to categorize the derivedontologies based on a clustering method. For indexing andretrieval of SW documents N-grams and bag of URI refs areproposed. More explanation and comparison of them are givenin [3].Three types of meta data are used in these systems [3]:

• Language attributes
• Relationships between SW documents (prior version,imports, extends, etc.)
• Meta data resulting from analysis

2) Analysis: For offline ranking it is possible to use thereferences idea of PageRank. But the main point is that count,
type and meaning of relations in SW is more complete thanthe current web. In the table I three types of relation and theircorresponding values are specified[27]:In [4] Onto Rank values for each ontology is calculated verysimilar to PageRank in traditional search engines like google.

3) Digest: In a recent version this algorithm is similar towhat is given in [27] which uses a sum of rank and priorityof concepts in a SW document to calculate the overall rank ofa document.*4) Service:* In addition to user interface in this section,services to application systems are provided too.

III. SEMANTIC SEARCH ENGINES

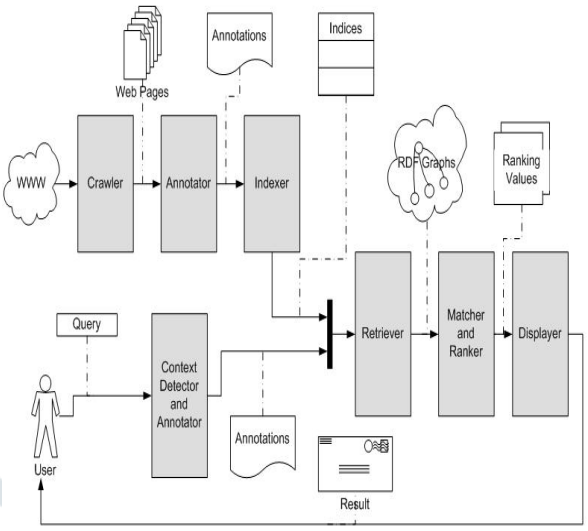
Searching in the web is done either using search enginesor web directories, each having respected restrictions [5]. Oneinteresting example is shown in [5]: if we search for ”Matrix”,non homogeneous results ranging from mathematical matrix,Matrix movie and so on is returned. Semantic web is introducedto overcome such problems.The most important tool in semantic web for improvingsearch results is context concept and its correspondence withOntologies. This type of search engines uses such ontologicaldefinitions.
It is possible to categorize this type of search engines tothree groups. In the first group which is the largest one,aim is to add semantic operations for better results. In the
second group, using facilities of semantic web the goal isto accumulate information on a topic we are researching on.Search engines in the third group try to find semantic relationsbetween two or more terms.

A. Context based Search Engines

Figure 2 shows an overall framework for this kind ofengines. It should be emphasized that very limited number ofengines have all of the functionalities specified in the figure.

1) Crawling the semantic web: There is not much differencebetween these crawlers and ordinary web crawlers and infact many of the implemented systems uses an existing webcrawler as underlying system. For example in [1] haircut isused as underlying system and also [15] uses

one that understandsspecial semantic tags. One of the important features oftheses crawlers should be the exploration of ontologies that
are referred from existing web pages.*2) Metadata generation:* According to discussion in section3, there are different ways for metadata generation. Forexample [1] and [5] use external metadata. [1] Uses Aero Textto extract names and expressions and then generates metadatain RDF format. One of the important problems in this regardsis semantic normalization [25] meaning to generate metadatafor different resources in same form. For example [12] is anon-standard example in which metadata is represented in adhoc model.
In semantic portal [23] producers should generate annotationsand there is a summarization and collection of metadata
in the central server.As explained before, metadata generation is simpler andmore accurate when the theme of a page is known. For examplein [15] using a tool named Knowledge Annotator termsof ontology is used to describe information in a given page.Also [18] proposed a method for generating and managingmetadata according to already defined ontologies.
But if ontology for a page is not known in advance, itis possible to use clustering techniques like what explainedin [25] to find an appropriate ontology. Knowledge Parser[24] is a kind of complete system using important techniquesfrom different areas like NLP, Text Engineering, DocumentStructure Processing, and Layout Processing. Its operation isshown in the figure.3.*3) Indexing:* Most of the engines does not provide any specialfunctionality regarding indexing. OWLIR [1] uses Swanglingexplained earlier. [6] Introduces Ontological Indexing inwhich indexing is done based



on a reference ontology. Also in[18] possibility of dividing documents to smaller parts is used
to improve indexing performance. Also in p2p architecture of[22] for each of concepts in the reference ontology there existan agent that maintains information corresponding to it.*4) Accepting user's requests:* There are two different approaches:term-based and form-based. In term-based approachused in [5], [23], and [24], it is tried to find the search context
from entered keywords. In the form-based approach used in[1], [15], [23], and [24], user interface is generated accordingto the ontology selected by user.*5) Generating meta data for user requests:* This operation
is very similar to generating metadata for documents. Forexample in [18] the same Semantic Mapper is used forgenerating metadata both for documents and user requests.Often Wordnet is used to expand user requests. For examplein [20] for termed entered by a user, using Wordnet, synonymsis found and used to expand the query.*6) Retrieval and ranking model:* Usually an ordinary VSMmodel [30] is used and then based on RDF graph matching

Fig. 2. Semantic Search Engines' Architecture

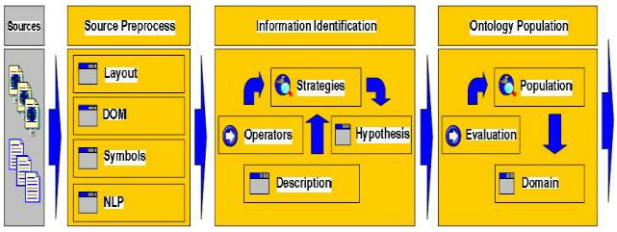


Fig. 3. Annotation Generation Steps in KnowledgeParser

results are pruned. In [9] from the equivalence of RDF graphsand Conceptual Graphs (CG), already existing operations onCGs is used to match user request and documents.

Semantic Distance concept is often used to estimate similarityof concepts in a matching process. In [21] this measureis defined for different elements in graphical representations.It is also possible to use graph similarity for ranking results.However, in [7] a fuzzy approach is used for this purpose.*7) Display of results:* A major difference of semanticsearch engines and ordinary ones is the display of results.One of the primary tasks is to filter the results (for examplefor eliminating repetitions). In [6] in addition to normal displayof results, a number of classes is displayed and when a userselects one, only those results having instances of the classes

is shown. In [23] display is a kind of hierarchy in which topconcepts of ontology is shown and by selecting one, detail ofit according to the ontology is displayed.

B. Evolutionary Search Engines

The advanced type of search is some thing like research; infact as mentioned in [5] this kind of searches aim at gatheringsome information about specific topic. For example if wegive the name of a singer to the search engine it should beable to find some related data to this singer like biography,posters, albums and so on. These engines usually use oneof the commercial search engines as their base componentfor searching and then augment returned result by these baseengines. This augmented information is gathered from somedata-insensitive web resources. In figure 4 we showed overallarchitecture for such engines.As it can be deduced from the figure this architecture hassome similarities with what we discussed in previous subsection;here we crawl and generate annotation just for somewell known informational web pages i.e. CDNow, Amazon,IMDB as mentioned in [2] and [5]. After this phase we collectannotations in a repository. Whenever a sample user poseda query two processes must be performed: first, we shouldgive this query to a usual search engine (usually Google) toobtaining raw results. Second, system will attempt to detectthe context and its corresponding ontology for the user'srequest in order to extract some key concepts. Later we usethese concepts to fetch some information from our metadatarepository. The last step in this architecture is combining anddisplaying results. Main problems and challenge in these typesof engines are [5]:

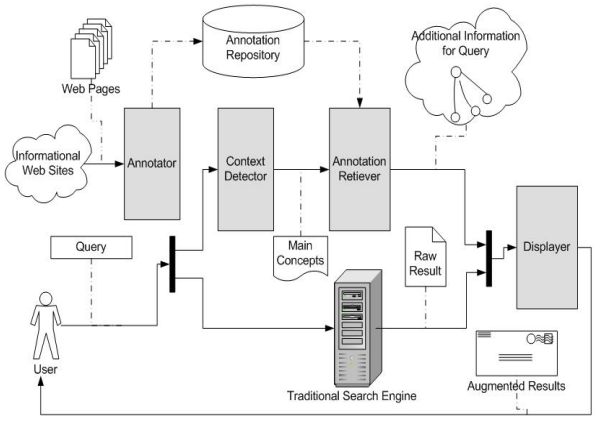


Fig. 4. Evolutionary Search Engines' Architecture

Concept extraction from user's request: there are someproblems that lead to misunderstanding of input query bysystem; for example inherent ambiguity in query specifiedby user or complex terms that must be decomposed to beunderstood.• Selecting proper annotation to display and their order:often we find a huge number of potential metadata relatedto the initial request and we should choose those ones thatare more useful for user. A simple approach is using otherconcepts around our core concept (which we extracted itbefore) in base ontology and if we have more than onecore concept we must focus on those concepts that areon the path between those core concepts.

*C. Semantic Association Discovery Search Engines*Usually one of the user's interests is finding semanticrelations between two input terms. Old search engines handledthese request using learning and statistical methods [25], butsemantic web standards and languages have provided moreeffective and precise methods. SemDis [10],[14] is a realsample for these systems, its goals is finding and ranking semanticassociations. Overall architecture of SemDis is shownin figure 5. There are different types of semantic associationbut most known of them is a sequences of classes and relationsbetween two classes. In fact we talk about just two termsbecause as said in [13] average length for users' queries is 2.3term. With respect to our definition for semantic association,two terms may have one of these association: Null (both ofthem are instances of one concept), Direct (when there is adirect relation between them) and Indirect (chain of relationsinstead of single direct one). In the [13] Bayesian networkswas applied in order to discover semantic association. Our
reference ontology forms the graph of this network and logsof user's queries are used to computing its parameters. Ingeneral manner, for finding semantic association betweenmore than two terms some techniques have been proposed,for example in [16] Spread Activation Technique is used toexpand an initial set of instances to contain most relativeinstances to them. The initial set is populated by extractingimportant terms from user's query, then with respect to themetadata repository corresponding instances is retrieved andafter expanding them an instances graph is produced whicheach of its edges has correctness weight in addition to usualsemantic label. Technically speaking, after discovery phaseoften we have numerous semantic association, therefore aranking policy must be used. In [10] some criteria for theseranking algorithms are introduced:
Context: special part of reference ontology that is interestedby user

- Subsumption: low level classes in hierarchy have moreinformation then their parents
- Path Length: having a shorter path between two termsindicates that they have near meaning

A. Ontology Search Engines

In contrast to usefulness of meta-search engines for regularpages in traditional web, it seems that they are not so good forontologies. In fact we can not collect the all ontologies in theweb just but using filetype command within commercial searchengines. In addition swangling operation has a huge amountof overhead, therefore it's much better to use crawler-basedontology search engines (2nd category) rather than ontologymeta-search engines (1st category). In order to evaluatingperformance of this kind of search engines there is no standardtest collection, but we can simply test them by searching forontologies using name of ontologies, classes and propertiesand judge their results according to the precision measure
(portion of relevant result from all result returned)
• Trust: obtained results from trusted resources is morevaluable in final ranking results

B. Semantic Search Engines

1) Context-Based Semantic Search Engines: The mainstrangeness of these engines is their simplicity. In fact becausethey tried to be as simple as textbox search engines (likegoogle) they are most popular search engines in the semanticweb. Here quality of results heavily depends on power ofits annotation module. The biggest problem of these searchengines is that they are limited to the special contexts. It is
much better if we can develop a multi-contextual semanticsearch engine. Fortunately we can apply standard measures(i.e. Precision and Recall) and standard test collections (i.e.TREC tracks) of traditional information retrieval to evaluatethis kind of semantic web search engines. It should be notedthat if we can prepare ontology for test documents, the resultswill show much improvements.

VI. CONCLUSION

Compared toother categories, the semantic association discovery enginesare related to higher layers of semantic web cake (logic andproof). Result of these engines is very depending on theirontology repository. For evaluating them we can use an upperontology like WordNet, after selecting two concepts randomly,the correctness and speed of discovering paths between themare two useful measures for performance evaluation.

REFERENCES

[1] J. Mayfield, T. Finin, and B. County, "Information retrieval on thesemantic web: Integrating inference and retrieval," in *SIGIR Workshopon the Semantic Web*, Toronto, Canada, 2004.
[2] T. Finin, J. Mayfield, C. Fink, A. Joshi, and R. S. Cost, "Informationretrieval and the semantic web," in *Proceedings of the 38th InternationalConference on System Sciences*, Hawaii, United States of America, 2005.
[3] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari,V. C. Doshi, and J. Sachs, "Swoogle: A search and metadata engine forthe semantic web," in *Proceedings of the Thirteenth ACM Conferenceon Information and Knowledge Management*, 2004.
[4] T. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng,"Swoogle:

Searching for knowledge on the semantic web," in *Proceedingsof the AAAI 05*, 2005.
[5] R. Guha, R. McCool, and E. Miller, "Semantic search," in *Proc. of the12th international conference on World Wide Web*, New Orleans, 2003,pp. 700–709.
[6] J. Davies, R. Weeks, and U. Krohn, "Quizrdf: Search technology forthe semantic web," in *WWW2002 Workshop on RDF and Semantic WebApplications*, 2002.
[7] T. Priebe, C. Schlaeger, and G. Pernul, "A search engine for RDFmetadata," in *Proc. of the DEXA 2004 Workshop on Web Semantics*,2004.
[8] Y. Zhang, W. Vasconcelos, and D. Sleeman, "OntoSearch: An ontologysearch engine," in *The Twenty-fourth SGAI International Conferenceon Innovative Techniques and Applications of Artificial Intelligence*,Cambridge, 2004.
[9] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker, "Querying the semanticweb with the corese search engine," in *Proc. 15th ECAI/PAIS*, Valencia,Spain, 2004.
[10] C. Halaschek, B. Aleman-Meza, I. Arpinar, and A. Sheth, "Discoveringand ranking semantic associations over a large RDF metabase," in *30th International Conference on Very Large Data Bases(VLDB)*, Toronto,Canada, 2004.
[11] H. Yu, T. Mine, and M. Amamiya, "An architecture for personal semanticweb information retrieval system," in *14th international conference onWorld Wide Web table of contents*, Chiba, Japan, 2005.
[12] B. Sigrist and P. Schubert, "From full text search to semantic web: TheInfofox project," in *Proceedings of the Tenth Research Symposium onEmerging Electronic Markets*, 2003, pp. 11–22.