A Survey on Semantic Search Engines and Their Behooves and Limits

Javad Mohammadi Madavani
Software Engineering Department
Islamic Azad University of Larestan,Larestan,Iran

Abbas Akkasi
Software Engineering Department
Islamic Azad University of Larestan,Larestan,Iran

Fazel Rezaie
Software Engineering Department
Islamic Azad University of Larestan,Larestan,Iran

Hamed Mohebi Mahdi Abadi
Software Engineering Department
Islamic Azad University of Larestan,Larestan,Iran

ABSTRACT
Current age is age of information explosion. Ever-expanding of the World Wide Web makes the finding required information difficult. Search engines play a principle role in finding info and a high volume of internet traffic related to them. Despite the remarkable progress in search engines, the results still are not satisfactory. Semantic search engines by using facilities at the semantic can add a lot of quality to search results. More specifically semantic web provides clear and intelligible meanings for search engine so it can generate more desired results. What is in this paper is a review of some efforts have done in this field and results achieved by these efforts. At last it will be realized that despite all these efforts, still semantic web cannot be an alternative to current search engines.

Keywords
Search engines, Semantic search engine, information retrieval, Ontology

1. INTRODUCTION
Nowadays the World Wide Web impresses human life in many aspects. Such a huge information resource that grows progressively is one of human's greatest achievement. Nowadays whatever comes to mind could be found on the web. So web is a great resource of information that everybody can access to and take advantages of it as required. But, the big deal is how to find required information in this immense ocean of data? It's just like looking for a needle in a haystack. So the effective solution for this problem is nothing but the search engines.

A search engine is an information retrieval system that by receiving a query from a user, finds relevant information, and then sorts them by relevancy to query, using some algorithms and return the result to user.

Low precision is the deficiency of current search engines, so in addition to relevant information, they retrieve large amount of irrelevant results too. This problem is due to dependency of search engines to key words. They retrieve all documents containing key words; whereas a word may have several different meanings. In other words, there's no semantic perception of user demands.

Semantic web is a tool to make information understandable for search engines. If search engines get the meaning of queries, they will do the search more accurate and desired. That is what semantic search engines are looking for. Although, several search engines with various technologies have been developed to get this.

Rest of this paper, is an overview of some of these approaches and their practical results.

In this section the overall structure of paper will be described. In the following section we'll review current search engines performance and their drawbacks.

The third section briefly introduces the Semantic Web and Ontology. In section forth, a set of criteria will be presented to compare semantic search engines and finally conclusion part will come along.

2. SEARCH ENGINES AND THEIR CONSTRAINS
In this section search engines operations and their constraints have been discussed. In next parts semantic web solutions to eliminate these constraints will be explained.

2.1 How Search Engines Work
Figure 1 shows how search engines work. As can be seen each search engine has a software module called spider, which has the duty to follow links and read web pages. After reading web page, Pre-processing step will be done on data to extract key words and other useful information for indexing the page. After extraction, categorized data will be stored in the database of search engine until the user enters a query. Search engine extracts a set of pages from database based on keywords, which contained specified keywords.

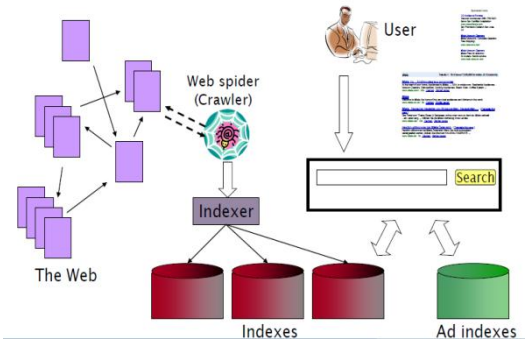

Figure 1 – A typical search engine's schematic

Then a ranking algorithm is applied to extracted pages to sort them by relevance to the query. In this way, more relevant pages will be at the top of search result list.

Most popular ranking algorithm is "PageRank", which is used in the Google search engine. It gives each page a numerical index based on input and output links of the page and the thematic relationship between linked pages. This number is used to rank results [1]. The purpose of a search engine is to maximize two Precision and Recall criteria [2] and of course the maximum value of both is 1.

$$precision = \frac{number\_of\_retrieved\_relevant\_documents}{number\_of\_retrieved\_documents}$$

$$recall = \frac{number\_of\_relevant\_retrieved\_documents}{number\_of\_retrieved\_documents}$$

2.2 Constraints Of Current Search Engines And Web

Current World Wide Web is a public database with absence of a semantic structure. So realizing information entered by user in form of a string would be hard for search engines. That's why the engines return ambiguous or partially obscured data as a result [3].

As it was mentioned World Wide Web has no specific structure to display information. Html codes are only markup for browsers to display information in a specific format, regardless of the meanings. So search engine cannot comprehend the concept of pages and relation between them.

One issue with current search engines is high Recall and low Precision. Although a lot of relevant information is retrieved but this amount of information will be lost in a large volume of irrelevant data. So it leads to results in poor quality (Figure 2).

Another problem in current search engines is hypersensitivity to words in the query. Whereas the same word maybe have different connotations in each sentence or in combination with other words. In another case there may be different words with the same meaning in a web page. Actually finding pages of user interest according to exact words in query and disregarding their meanings definitely lowers precision.

Another issue with current search engines is ignoring implications in query while comprehend them can help to have high-quality results.

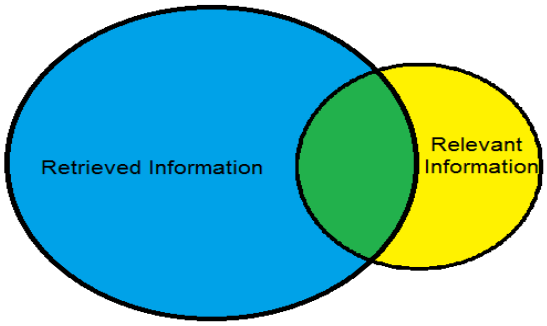

Figure 2 – Precision and Recall ratio

3. SEMANTIC WEB AND ONTOLOGY

This section is a brief description of the Semantic Web, and ontology as its main constituent. As mentioned before current web, suffers the absence of a specific structure to display information. Web info can be understood by human but search engine finds data as a string of bits and has no perception of concepts. Semantic Web is an attempt to solve this issue, which aims to display information in a specified format and providing the meaning for them till both human and engine can comprehend the concepts and relationship between them.

Figure 3 demonstrates different layers of Semantic Web called Semantic Cake.

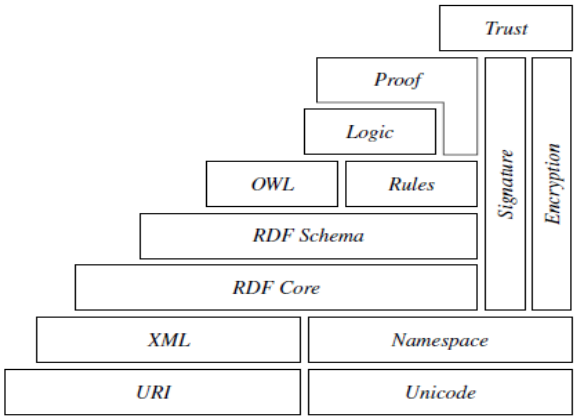

Figure 3 - Semantic Cake

Lowest layer contain information about web structure. XML is located on this level. XML is an attempt for structuring information and it doesn't make much meaning. On top of this layer is ontology layer that is created by one of ontology description languages such as RDF. That's one of most fundamental layers in semantic cake which develops a vocabulary, and describes entities and relations. For an example in this layer, it can be said human beings are divided into two kinds of men and women and they are disjoint or it can be defined filiations or sibling relationship between two people. Due to constraints in RDF, RDF-schema has been created to add new functionalities to RDF. "OWL", has been made on the basis of RDFS intended to overcome the restrictions by adding new features to it. Other upper layers utilize of ontology layer for knowledge-based inference. E.g. if there are two statements: "Barney is Marshall's father" and "Ted is Barney's brother" then it can be interfered that "Ted is Marshall's uncle". To learn more about the various layers of semantic cake see [4]. The purpose of this section was just getting to know the terms used in the following.

4. CLASSIFICATION CRITERIA FOR SEMANTIC SEARCH ENGINES

In this section some criteria will be introduced, that used to classification of semantic search engines as follows: Architecture, Coupling, Transparency, User Context, Query Modification, Ontology Structure, and Ontology Technology [2].

4.1 Architecture

There are two major architectures for search engines:

First, standalone search engines: in these structures search engine as a full searcher does have all presented parts in section 3-1 without relying on other search engines. So search engine is responsible for collecting, storing and analysis of fetched data by itself.

Second, Meta search engine: in this structure search engine send user's query to other self-determining engines, then collects their answers, combines them and render the results to users.

4.2 Coupling

This criterion determines the coupling level of documents and ontologies which is divided into two categories:

Tight coupling: in this category metadata explicitly refer to the concepts in a specific ontology and vice versa. Sometimes documents themselves are considered as a sample of a concept in ontology. In this method, similarity and polysemic issues would be resolved by choosing the proper concept in ontology, although it will cost semantic annotation amidst documents information.

Loosely coupling: in this category, documents are independent from a specific ontology. Indeed picking out an appropriate ontology for query domain is an issue itself. So systems in this category have a low semantic power. For instance, the similarity and polysemic issues cannot be solved easily in such systems. Since in current World Wide Web there are a few documents with semantic annotations, this solution seems realistic. Systems in this category are easy to implement by Meta search engines.

4.3 Transparency

This criterion considers user involvement in semantic aspects of the system, in the following categories:

- Fully transparent: in these systems semantic aspect is completely hidden from user's view and to users these search engines seem as a typical one. Also the search engine does not request any additional information from user (e.g. user won't be asked about polysemic issue, for transparency).
- Interactive: This type of systems ask user to clarify their purpose and make suggestion to change the query. Sometimes they are called Recommender Systems.
- Hybrid: These systems are a combination of two previous one. They seem as typical search engines to users, and user assistance is not used except in very specific cases.

Transparency is like a double-edged sword. On one hand, transparent systems save users from menus and lots of questions; on the other hand, there would be not possibility of involving user's opinion in semantic aspects of searching process; so it leads to low precision.

4.4 User's Context

The relevance of any document is always related to its context. Most of semantic search engines utilize user's context in order to find his required information. From this point of view there are the following categories:

- Learning: In this category history of user dynamically achieved based on his interactions with system. System tries to reevaluate user's possible intended results, based on his previous queries, history of transparency and applying improvement to his query. If terms used in queries were about the same semantic context, then the system can solve polysemic issue.
- Hard Coded: In this category queries divided into groups called "the question categories", to determine user's need for information. The system provides a certain number of question categories that will be used during the query assessment. E.g. questions like "location of …" or "general resources for …" determines user requirements category. So the context could be inferred explicitly via user's words in query or implicitly via user's category or query analysis by engine. E.g. Figure 4 demonstrates "Bing" search engine which use this feature. It suggests queries to users that could improve quality of results.

Figure 4 - Bing search engine

4.5 Query Conversion

As mentioned before a search engines tend to increase two major parameters: Recall and Precision. Increasing Precision also is called Query Clarification.

When there is an ontology (general ontology specially) increase in Recall value is just in hand by using more general words in query. For example, while it is made use of synonyms in query, more results will be generated. On the other hand increasing precision is really a hard task that includes solving synonymy issue and hierarchical structure of words. As an example, when a user enters a query related to a concept in ontology, this can increase precision by using a sub concept in the ontology instead of main concept itself.

Figure 5 depict variety of query conversion methods explained briefly in the following.

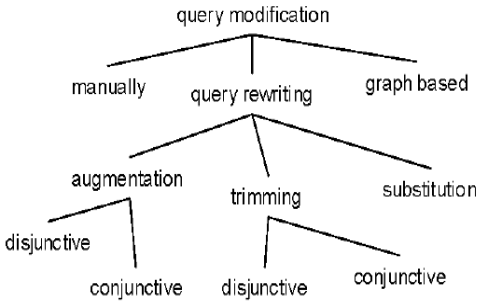

Figure 5 - Variety of query conversion methods

Manual conversion: in this method which is simplest one, search engine returns part of ontology in addition to search results. User can improve his query by tracing the ontology. Knowledge graph of Google is an instance, depicted in Figure 6.

Figure 6 - Google's Knowledge Base

Query rewriting conversion: In this method, system improves query's precision, using some of these ideas: Augmentation (adding more concepts based on ontology, for example "Theory of relativity" can be added to "Einstein" query), Trimming (Reverse of Augmentation, removing words in

query and comparing quality of results after removing. e.g. If a query contains "Theory of relativity", "Einstein" and "Black Holes" did not make desired results then search engine can find better results by removing "Einstein". Augmentation and trimming can be in two forms: Conjunction (outcome of several words conjunction, lead to more specific result) or Disjunction (outcome of several words disjunction, lead to more general result). Indeed conjunctive long queries lead to increasing Precision and disjunctive long queries lead to increasing Recall. And Substitution (replacement of query words by their synonyms or more specific or more general words).

Graph-driven conversion: this method requires tight coupling between documents and ontology. In this way, concepts in ontology and documents both appear in the form of nodes of a graph. Then words of query will be used to find relevant nodes. Based on these nodes, an algorithm traces the graph to find relevant semantic documents. In this method, user's query doesn't be changed but find relevant documents directly.

4.6 Ontology structure

Ontology is a collection of concepts, properties, constraints and axioms. In practice, search engines pay more attention to properties and divided them to several categories:

Anonymous properties: in this case, the name and meaning of properties are not considerable and only relationship between two properties indicates that they have a shared concept.

Standard properties: in this case only properties like "synonym with…", "more general than…","part of …"," example of…","reverse of…" are used. Applying these properties increase power of semantic search (e.g. using standard ontology in query conversion). To learn more about how standard properties impress on semantic search refer to [5].

Domain specific properties: a system can use domain specific properties in addition to standard properties, as type of cloth in an information retrieval system of fashion journal.

A combination of these three properties may be used in some systems. E.g. only applying some standard properties and using the others as anonymous. This criterion determines level of flexibility in search engines to reuse ontology.

4.7 Ontology Technology

To demonstrate ontology an ontology description language should be used. Ontology structure determines its semantic reusability but ontology technology concentrate on reusability from technical aspect and diversity of cases of reusability. Most common Languages used in this case are:

- F-Logic [6]
- RDF [7]
- DAML(+OIL) [8]
- OWL [9]

5. INTRODUCTION TO SOME POPULAR SEMANTIC SEARCH ENGINES

In this section we'll introduce some popular semantic search engines, and we'll compare them on previously mentioned aspects. There are many search engines that take advantages from their own unique technology and architecture. Those which introduced here are some of many, in this area.

5.1 SHOE Search Engine

This search engine has designed by Jeff Heflin and James Hendler in University of Maryland in 2000. Shoe works, based on a domain ontology in which each entity in a document will be mapped to an existing concept in ontology for comparison. For example, for a webpage of a university the ontology may contain webpages about facilities, students or scientific projects, and information about which student works on which project.

All the concepts described by "SHOE" annotation language which cannot be understood by web browsers but can be analyzed by semantic search engines. In SHOE every webpage called an "Individual" in anthology. For example, webpage of project "P" in an instance of "projects pages" which inherited from webpages entity, and there are some attributes that connect it to student's page "S".

To search in the ontology, first user picks a concept in a presented list and then search engine return a set of attributes related to that entity then user make a conjunctive query among entity and presented values and the query will be applied to SHOE database.

This search engine needs a tight coupling among webpages and ontology concepts. But it's architecturally independent from the cases and has its own knowledge base, consequently the system does not keep track of history of user's activities and queries only depend on concepts and attributes. Figure 7 represents SHOE architecture and complete information about that is available in [10].

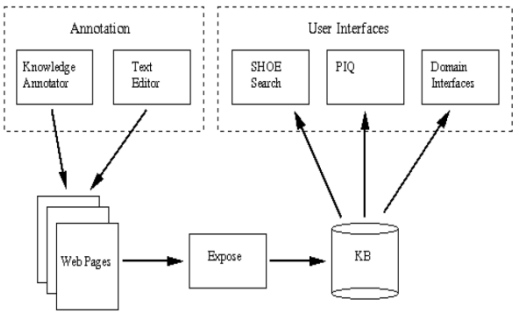

Figure 7 - SHOE system architecture

5.2 TAP Search Engine

This search engine is a combination of normal search engines and semantic search engines. In TAP, semantic search process considered as an add-on to normal searching process. In this method semantic web is sort of RDF ontology which is independent from web pages and their ranking methods. So, every result that comes from queries includes two parts: first part contains results of web search engines and second part contains RDF trilogy resulted from applied query to semantic web. To remove ambiguities from results while retrieving results from ontology, TAP use three methods: First, Measuring semantic distance between words in the query and resulted RDF graph. Second, using subjective background of user's conceptual context. Third, measuring of word's popularity in documents base.

Because of independency in outcome of semantic web search and web search engines, TAP categorizes as loosely coupled search engines. And from architectural aspect it's known as a "hyper search engine".

5.3 Inquirus2 Search Engine

This search engine known as a "transparent hyper search engine" and it uses batch queries to retrieve results from a knowledge base.

The search engine receives a query from user and a search group and optimize user's query according to presented search group, it happens by adding some words to the query, and using a proper sub search engines. Then resulted will be collected from search engines and will be sorted by relevance. Choosing a proper search group is key factor in the process.

5.4 ISRA Search Engine

This search engine benefits from manipulating user's query by using semantic networks, and making improvement in query's precision. The semantic networks build by synonyms, general and also opposite words which retrieved from "WordNet" and DAML hierarchical concepts.

According to generated semantic network, ISRA tries to guess the meaning of words and eliminate ambiguities and improve user's query. And finally resulted query will be sent to other search engines to retrieve results. ISRA categorized as "Rewriting user's query" methods. Because it only changes query, and documents do not be retrieved directly from user's query.

From architectural aspect ISRA is sort of hyper search engines which is basically a loosely coupled search engine and from transparency aspect it's a hybrid method because it uses user's feedback in case that searches does not include satisfying results [14].

5.5 Librarian Agent Search Engine

This search engine act like a librarian and users improve their queries in an interactive environment, this method use ontology to eliminate ambiguities, on the other hand the method uses user's search history to guess the meaning of words in the query and uses documents base to estimate number of retrieving results.

For example, a typical query that contain "Einstein", "Relativity theory" may lead to 20 results, and system says there is 190 results for "Einstein" and 220 results for "Relativity theory" independently, and also it says there is 11 results for "Einstein", "Relativity theory", "Special" , and let user choose which query would be appropriate. This method only support conjunctive queries and does not depend to any anthology-based structures.

6. CONCLUSION AND FUTURE WORK

In this paper we did a review on some semantic search engine and a brief about how they work. Since using semantic web isn't vastly used by World Wide Web and there are a few documents which annotated semantically, it's not expected that current search engines be replaced by semantic search engines in near future and Google remain a powerful and popular search engine for web users, but Google implicitly develop semantic web concepts and apply them to their searching methods to improve results. But migrating toward using semantic web, seems inevitable.

7. ACKNOWLEDGMENTS

special gratitude to Mr.Mehdi Sabbari from Software Engineering Department of Islamic Azad University of Qazvin. And Thanks to the experts who have contributed towards development of this paper.

8. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. .

[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

[4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[9] McGuiness, D.L. and van Harmelen, F. (2004) 'OWL web ontology language – overview – W3C recommendation 10 February 2004', Technical Report, W3C, http://www.w3.org/TR/owl-features/ (2004-7-15).

[10] Heflin, J., & Hendler, J. (2000). Searching the Web with SHOE. In *Artificial Intelligence for Web Search* (pp. 35–40).

[11] Finin, T., Reddivari, P., Cost, R. S., & Sachs, J. (2004). Swoogle : A Search and Metadata Engine for the Semantic Web. In *ACM conference on Information and knowledge management* (pp. 652–659). doi:10.1145/1031171.1031289

[12] Guha, R., McCool, R. and Miller, E. (2003) 'Semantic search', WWW '03: Proceedings of the Twelfth International Conference on World Wide Web, May, Budapest, Hungary.

[13] Glover, E.J., Lawrence, S., Gordon, M.D., Birmingham, W.P. and Giles, C.L. (2001) 'Web search – your way', Communications of the ACM, Vol. 44, No. 12, December, pp.97–102.

[14] Burton-Jones, A., Storey, V.C., Sugumaran, V. and Purao, S. (2003) 'A heuristic-based methodology for semantic augmentation of user queries on the web', Conceptual Modeling – ER 2003, 22nd International Conference on Conceptual Modeling, Chicago, IL, USA, October 13–16, Proceedings, pp.476–489.

[15] Stojanovic, N. (2003) 'On analysing query ambiguity for query refinement: the librarian agent approach', Conceptual Modeling – ER 2003, 22nd International Conference on Conceptual Modeling, Chicago, IL, USA, October 13–16, Proceedings, pp.490–505.