Comparative Analysis of Semantic Search Engines Based on Requirement Space Pyramid

Maliha Majid Qureshi, Bibi Asma, and Hikmat Ullah Khan

Abstract—**Semantic Web promises to add metadata to web content to make it understandable to computers. Search is the most widely used activity on web. Semantic search engines have already changed the way we search the data on web. Uren. V, Yuanguilei Uren et al., proposed requirement space pyramid arguing that iterative and exploratory search modes are important to the usability of search engines. It identified the types of semantic queries the users need to make, the issues concerning the search development and the problems intrinsic to semantic search in particular. We have extensively examined the semantic search engines and have done broad survey to analyse the semantic search engines. Comparative analysis of the semantic search engines have been done on the basis of factors cited in the pyramid. The research provides deep understanding of five main semantic search engines based on comparative analysis that may help for future work for semantic web in general and for semantic search engines in particular.**

Index Terms—**Semantic web, semantic search engine, requirement space pyramid.**

I. Introduction

Semantic Web [1], the future of web, promises to provide content that will be understandable to computer. Search is the most important and popular applications of the web. Semantic search [2] is one of the most significant applications of semantic web. Recently a number of semantic search engines have been introduced. Semantic Search engine is a tool that produces precise results to user queries by retrieving data semantically [3]. The purpose of this paper is to compare semantic search engines along with their salient characteristic approaches. This paper focuses mainly on five semantic search engines including Hakia, Sensebot, Powerset, Lexxe and Cognition. There are many benchmarks and standards used in every field and computer science is not an exception. Google and other search engines have revolutionized the world with their services but with the idea of semantic search, the rise of semantic search engines has changed the requirements and issues related with search engines. A pyramid scheme has been used as model for discussion to present the requirements, issues and challenges for semantic search engines. The remaining paper has been formatted as follows that related work have discussed in section II. Comparative evaluation of the semantic search engines on the basis of the pyramid has been described in section III. The semantic search engines along with their GUI's have been discussed in section IV while finally the concluding remarks include the issues and challenges of semantic search engines.

II. Related Works

Semantic search engine is new field and realtively less work has been done in this domain. Hakia has been compared with google, yahoo and msn in [4] and that of Hakia with dogpile. A relation-based page rank algorithm has been proposed to be used as semantic web search engine. In this work, relevance is measured as probability of finding connections completed by the user at the time when the query was carried out and the information contained in the base knowledge of the semantic web environment [5]. Semantic based approach for the evaluation of information retrieval systems has been proposed. The purpose is to increase the selectivity of search tools and to improve how these tools are evaluated [6]. Another paper proposes a novel approach to construct the snippets based on a semantic evaluation of the segments in the page. The target segments are identified by applying a model to evaluate segments present in the page and selecting the segments with top scores [7]. WebOWL consists of a community of intelligent agents, performing as crawlers and are able to discover and study the locations of semantic web neighborhoods on the Web, a semantic database to pile data from diverse ontologies, a query mechanism that maintenances semantic queries in OWL, and a ranking algorithm that defines the order of the returned results based on the semantic relationships of classes as well as individuals [8]. An advanced approach for semantic service search is proposed. The proposed approach comprises of three main stages. Firstly, the crawling phase, during which semantic service descriptions that are online are retrieved and stored locally. Secondly, the homogenization phase when the semantics of every description are mapped to a reference service model. And lastly, the search phase when the users are enabled to query the underlying repository and find online services [9]. A specialized and dedicated semantic search engine known as GeneView has been developed that targets only the medical domains [10]. To accomplish this unique analysis, our recent works in knowledge management [11] was very helpful and also our ability in context driven algorithms [12] and use of Suffix Array based RDF Indexing using RDQL Queries [13]. Semantic has been used for various other reasons like semantic based dynamic modeling has been carried out to find out the research interests of the authors [14] using topic models [15]. A recent survey has been done covering

Manuscript received November 8, 2012; revised January 28, 2013.

The authors are with the Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan (e-mail: Hikmatullah@comsats.edu.pk, maliha_red@hotmail.com, asmaasmi@live.com).

DOI: 10.7763/IJFCC.2013.V2.227

different promising features of some semantic search engines. Significance of work and many approaches for semantic search have been discussed. The most prominent part is that how the semantic search engines differ from the traditional searches and their results are shown by giving a sample query as input [16].

III. Pyramid Based Comparative Analysis

The comparison is based on a pyramid scheme representing the requirements space for semantic search systems. To the best of our knowledge, it is the first such standard for the new domain of semantic search engine. As the domain of the semantic web is still far from its promises and social web is taking its grip more on the world wide web as well as on the research horizon, thus very less research publications can be found in this comparative domain. The requirement space pyramid [17], as evident from the Fig. 1, consists of four main categories, i.e., Search Environment, Query Type, Iterative and Exploratory and Intrinsic Problems. Search Environment discusses the requirement covering scale which tells us at what scale the particular semantic engine can operate. How different topics can be incorporated and also what is the ontological depth and breadth. Another point of concern is heterogeneity and portability. Query type includes parameterized search, relation search and entity search. As name suggests, user opts for entity relation when know direct topic or keyword, else opts to retrieve from relative terms or entities so user looks for relation search and lastly if user has better knowledge then parameterized query is best option. Iterative exploratory requirement has been devised based on earlier learn experiences and with respect to refinement, recommendation and reusability. Refinement allows the user to alter the query, recommendation is done on the basis on previous experiences of the various users and reusability is based on the queries used already carried out on that particular semantic search engine. .Intrinsic problems include understanding, ranking of entities and matching of different concepts.

IV. Search Environment

A. Large Scale Search

It requires that semantic search systems should be able to support large scale Search Environment [17]. Hakia is Ontological Semantic i.e., it focuses on the depth of concept. Hakia works on Commercial Ontology that is a reality of Web. Most of the Search queries are exclusively relevant to Commercial Topics and it also provides two online services of Enterprise search and Sense News. It contains two important innovations of the development of concepts, lexicons and Sequence Approach. It classifies result in diverse categories like news, blogs, video, images etc [4].Sensebot represents summary of multiple documents in result of user query. It parses the top result from web and puts them into summary rather than web pages link. The summary is the main result of your search [16]. It searches on the basis of text mining and NLP and tries to match the human made summary but it has not yet been accomplished. It summarizes up to 100 pages at a time. Powerset is a better way to search Wikipedia articles. After being merged with Microsoft the changes are visible in live search and related search and in content of Wikipedia [16]. Lexxe answers the questions and gives better results on standard keyword searching that are much relevant to leading search engine experience, and Lexxe achieves this by implementing NLP Technology [6]. Cognition is linguistic search engine that supports ontology, morphology and synonym, tapping one of the world’s largest computational dictionaries. Cognition involves concepts such as knowledge, desire and preference and works on recently launched three websites by cognition with high value deep content in the domains of health, law and consumer information [18].

B. Heterogeneity

This requirement is the ability to support heterogeneity; the system must be able to search between several different search ontologies at the same time. Hakia is based on language independent ontology of thousands of interrelated concepts; ontology based English lexicons of 100,000 words senses. If a sentence of 6 significant words generates over a million sequences while only a few makes sense. The challenge is to reduce these possibilities down to few dozens and fuse the possible knowledge into different ontologies together. This is accomplished using commercial ontology [4]. Sensebot takes semantic result from Google, Yahoo and Live and summarizes them into one concise digest on the topic of query [16]. Powerset gives accurate results in the response of query and often answers question directly, and aggregates information across different multiple articles. Results are based on NLP and parsing but with very little semantic knowledge. In lexxe, it is based on the semantic and meaning of the query, and does not depend on the preset keyword grouping or inbound link measurement algorithms. NLP search delivers results in clusters with related topics. Cognition has parsing and semantic NLP techniques. Other aspects of parsing technology has ability to analyze the grammatical structure of sentence and NLP enables the computer to recognize the phrases and their relationship with other words and phrases [18]. Cognition retrieves relevant results with the help of these two techniques.

C. Portability

It requires that the system has to move between ontologies without any need for domain-specific reconfiguration. The important thing about Hakia is that it is built on well-developed taxonomy and constantly improving which ensures the homogeneity of the ontological concept and lexical entries by proven portability and extensibility of the resources to new domains. Hakia developed QDEX Algorithm that extracts all possible queries that can be asked in the content. These queries become gateway to originating documents, paragraphs and sentences during retrieval mode. Sensebot needs to reconfigure each query time by choosing the domain. As Powerset goes through different articles and aggregates the information and sometime gives exact answer in response of a query and it gives Wikipedia article to digest the content and information easily. Lexxe, using NLP technique, extracts candidate answer by routing through different documents, combining answers that are logically related and same, estimating which answer is more credible and then representing it to user. Cognition user sets preferences by specifying the domain and region for better understand ability of query structure and meaning to retrieve the relevant result and the review tool used in cognition for legal domain update the index on the fly [18].

V. Query Type

A. Parameterized Search

According to pyramid, this requirement serves where user has very precise definition of topic to be searched. In Hakia QDEX algorithm [4] decomposes query into meaningful sequence without getting lost into useless combinations and produces dozens of useful combinations where only few make sense at native user level. On other hand, in Sensebot every sentence has a link to its source page and then on the basis of text mining, it serves highly relevant results in response of long queries [16]. Powerset serves faster return of results, more accurate results for less keyword and more relevance and awareness on the user side [8].In Lexxe user can specify several parameters in the form of question, then it automatically extracts answer from internet web pages which means it can answer greater variety of questions. It is not like traditional or typical search engines which gives answers from manually prepared database. It present results in three categories of answers, cluster and web page snippets. Cognition provides results where the data conceptually match, first it displays the results which contain exact words then the next group of results contains those documents which match to query terms conceptually. After submitting the query cognition wants user to set the meaning of query and context of word to understand query much like the way user understands it and also to remove ambiguity [19].

B. Relational Search

The pyramid states that relational search look for the relation between the entities. Hakia classifies the related search in ontologies where it has related concept in same documents. Hakia classifies results in different categories and each category focuses on query phrase and then on synonyms [4]. Sensebot uses text mining and summarizes multiple documents to extract the sense from web, and then places related documents in semantic cloud [16] on the basis of related concept in same document. While Powerset aggregates results so well from multiple articles but the semantic technology of Powerset set is not well developed and still in process with Microsoft Bing, so when it comes to retrieve related concept from web it is not regarded highly [16]. Lexxe makes search queries more focused after processing them with linguistic support. By offering clustering Lexxe organizes information to group search results in semantically related clusters, and it does so on-the-fly. Cognition works on predefined relationships between words and phrases especially paraphrases and taxonomy [20].

C. Entity Search

Entity search is a most common provided type of search. It serves the case where user needs information about a particular kind of things. All of these five semantic search engines provide keyword searching to provide end-user with a straight forward way to specify query. Hakia works on OntoSem to retrieve relevant results efficiently [4]. Sensebot works on text mining [16] and Powerset, Lexxe and cognition work on strong NLP to retrieve results on-the-fly and human desired result [6], [7].

VI. Iterative and Exploratory Search

A. Reuse

It enables users to reuse query or the part of the query they have defined previously [17]. In Hakia ontology enables reuse of the domain to make query reusable. Sensebot facilitates user to add new term in his previous query to form new query. Powerset remembers the previous queries and learns earlier behavior of computer and facilitates user with previous queries along with changes in them. Lexxe is bit tricky on that part that it follows the proper format of question and it is complex to understand for novice user how to query in Lexxe and it retrieves complex results to understand for ambiguous query. So reuse of query is bit complicated and tricky in Lexxe. Cognition reuse of query is improved and extended to include different words with the same meaning as the query term. But the query is not domain specific but also works for other domains.

B. Recommendation

Query Recommendation suggests queries to users based on information system has learnt from their past search behavior [17]. Hakia, unlike other search engines, does not track the search history or monitors the search behavior. Hakia provides user privacy and place the cookie with user permission [4]. It utilizes work on search engine instead of personalized history. Sensebot tracks the users past search behavior and recommends search options during query submission. Sensebot recommendations are for books, ads and services and recommendation is provided by link sensor which is used to increase page views but at the same time it narrows the search focus. Powerset learn the behavior of search along with suggestion for users' queries. Factz is a box often appears in search page and provides several suggestions and references related to query. It also provides the source of the results as well as set of relevant web page links. Cognition helps advertisement companies to serve relevant ads to user activity on web for higher click through..

C. Refinement

Refinement allows users to modify previous queries to capture their information needs in better way. Search operations at Hakia are question type detection and relevance, categories, Qdexing (content characterization) and disambiguation [17]. All these help to get higher refined query. Sensebot summarizes multiple documents and gives coherent summary and allows modification in the summary according to user s' need by eliminating some of the results and the summary can be saved. Powerset refines query by giving advance options such as sort by articles and sort by sentence to get desired results. It provides refined results by presenting them in forms of clusters, answers and web page snippet [6]. Cognition advance syntactic parser provides result by making use of syntactic relation between words [18]. (Table I)

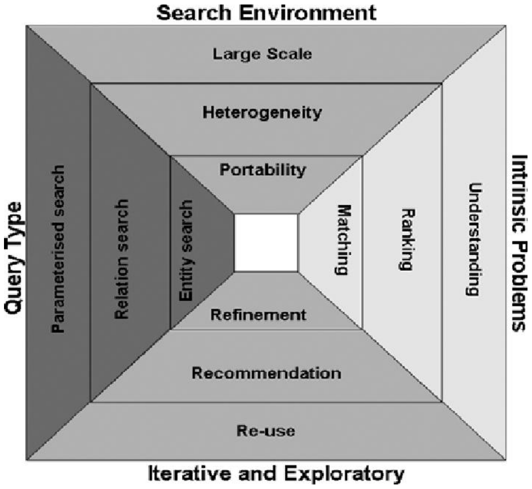

Fig. 1. Requirement Space Pyramid

TABLE I: COMPARATIVE ANALYSIS OF SEMANTIC SEARCH ENGINES ON THREE SCALES [LOW, MEDIUM, HIGH]

Requirements	Semantic Search Engines				
	Hakia	Sensebot	Powerset	Cognition	Lexxe
Search Environment					
Large Scale	High	High	High	High	High
Heterogeneity	High	Medium	High	Low	High
Portability	High	Low	Medium	Low	High
Query Type					
Parameterized Search	Medium	High	Medium	High	High
Relation Search	High	Medium	Medium	Medium	Medium
Entity Search	High	High	Medium	High	Medium
Iterative and Exploratory					
Reuse	High	High	High	High	High
Recommendation	Low	High	High	High	High
Refinement	High	High	High	High	High
Intrinsic Problem					
Understanding	High	High	High	High	High
Requirement	High	High	Medium	High	High
Matching	High	High	Medium	High	High

VII. INTRINSIC PROBLEMS

A. Understandability

Refinement allows users to modify previous queries to capture their information needs in better way. Search operations at Hakia are question type detection and relevance, categories, Qdexing (content characterization) and disambiguation [17]. All these help to get higher refined query. Sensebot summarizes multiple documents and gives coherent summary and allows modification in the summary according to user s' need by eliminating some of the results and the summary can be saved [19]. Powerset refines query by giving advance options such as sort by articles and sort by sentence to get desired results. Lexxe provides refined results by presenting them in forms of clusters, answers and web page snippet [6]. Cognition advance syntactic parser provides refine result by making fuller use of syntactic relation between words [19].

B. Ranking

This problem refers to find the way of returning results that are presented according to how well they satisfy the user query [9]. Hakia performs morphological and syntactic analysis to retrieve exact user requirement by filtering unwanted results [4]. Sensebot focuses on the particular requirement of user and gathers data from Google, Yahoo and rest of web in the coherent and digest form [18]. Powerset does semantic analysis on documents and thus queries create a better relevance which captures the user requirement from gathered information and provides the automatic summary of particular concepts. Lexxe extracts the most concise form of most credible answer. Sometime it provides the answer directly from database but it usually answers the question from web links [6]. Cognition is domain specific search engine so its relevancy is higher in comparison of other search engines [19] and it provides about 90% more relevant results according to users' need.

C. Matching

Matching refers how to do semantic matching of search term to entities. The final relevancy is determined by semantic rank algorithm based on advanced sentence analysis and concept match between the query and the sentence of each paragraph. Sensebot attempts to understand what the web pages are about and extracts the key phrases from those documents through text mining [18]. Powerset runs the content from semantic pipeline to reproduce another representation of content to identify the key concept from WebPages. Lexxe matches the query with context and returns the context in context. Cognition query is related to its domain and it retrieves results within the domain. One word can have more than one meaning, cognition runs its English semantic technology to understand context and structure of query [17].

VIII. DISCUSSION

Sense News, a new service of Hakia, focuses on real-time information and offers visualization and personalization tools helping users make sense of information from over 30,000 news sources, twitter and blogs. User can also personalize the charts according to current trends known as Sense Charts [20]. Hakia retrieves results including text as well as images but the image results do not provide direct navigation to source page. We find no search box but a lot of options on the search page of Sensebot which makes query more efficient but may cause difficulty for novice user. Sensebot provides Sentiment API to evaluate current trends with three basic sentiments of positive, negative and neutral by displaying pie charts or graphs. Powerset only covers domain knowledge of wikipedia thus shortcoming in knowledge of wikipedia also becomes that of Powerset. Lexxe needs a proper format while accessing natural language queries/questions whereas Cognition uses its semantic English dictionary to refine the search which may cause unease to novice user.

IX. CONCLUSION

The research has mainly been carried out to explore the different dimensions of semantic search. The pyramid used is an excellent benchmark to check the various domains of study about semantic search engines. The field of semantic search engines is still in its evolutionary mode, the main references have been taken from blogs and few from research papers as less number of publications are there in the emerging domain of semantic search engines. Also, much exploratory research has been carried out by querying a lot of diverse queries to each of the semantic search engine and various observations have been stored in recorded format. All such relevant material can be obtained from the corresponding author. Also the Pyramid for each semantic search engine has also been constructed. The exploratory research gives comparative analysis of main emerging search engines based on pyramid explaining the requirements for semantic search engines.

REFERENCES

[1] T. B. Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, May 2001.

[2] R. Guha, Mccool, R. Miller, "Semantic search," in *Proc. of the 12th international conferences on the World Wide Web*, ACM press (2003) , pp. 700-709.

[3] Y. Lei, V. Uren, and E. Motta, "SemSearch a search engine for the semantic web," in *Proc. 5th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks*, Lect. Notes in Comp. Sci., Springer, Podebrady, Czech Republic, 2010.

[4] D. Tumer, M. A. Shah, and Y. Bitirim, "An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, Yahoo, Msn and Hakia," in *Proc. Fourth International Conference on Internet Monitoring and Protection*, ICIMP' 09, pp. 51–55, 24-28 May 2009.

[5] M. Rojas, "A semantic association page rank algorithm for web search engines," *Journal of Computing Research Repository*, pp. 129-138, 2012.

[6] A. Bouramoul, M. K. Kholladi, B. L. Doan, "An ontology-based approach for semantics ranking of the web search engines results," *Journal of Computing Research Repository*, pp. 23-31, 2012.

[7] K. S. Kuppusamy and G. Aghila, "Semantic snippet construction for search engine results based on segment evaluation," *Journal of Computing Research Repository*, pp. 201-212, 2012.

[8] A. Batzios and P. A. Mitkas, "WebOWL: A Semantic Web search engine development experiment," *Journal of Expert System Applications (ESWA)*, vol. 39, pp. 5052-5060, 2012.

[9] N. Loutas, V. Peristeras, D. Zeginis, and K. A. Tarabanis, "The Semantic Service Search Engine (S3E)," *Journal of Intelligent Information System (JIIS)*, vol. 38, no. 3, pp. 645-668, 2012.

[10] W. Y. Ma, "Semantic search and a new moore's law effect in knowledge engineering," *Knowledge and Data Discovery*, pp. 98-109, 2012.

[11] T. A. Malik and H. U. Khan, "Perforamnce measurement using distributed perforamnce knowledge management system: Empirical case study of Coca Cola Enterprises," *International Review of Business Research Papers*, vol. 6, no. 1, pp. 250-282, Feburary 2010.

[12] T. A. Malik, H. U. Khan, and S. Sadiq, "Dynamic Time Table generation conforming constraints a novel approach," presented at International Conference on Computing and Information Technology (ICCIT 2012) 1st Taibah University International Conference held in Al-Madinah Al-Munawwarah, Saudi Arabia on 12-14 March 2012.

[13] H. U. Khan and T. A. Malik, "Finding resources from middle of RDF graph and at sub-query level in suffix array based RDF indexing using RDQL queries," *International Journal of Comuter Theory and Engineering*, vol. 4, no. 3, June 2012.

[14] A. Daud, "Using time topic modeling for semantics-based dynamic research interest finding," *Knowledge-Based Systems*, vol. 26, pp. 154–163, 2012.

[15] A. Daud, J. Li, L. Z. Zhou, and F. Muhammad, "Knowledge discovery through directed probabilistic topic models—A survey," *Journal of Frontiers of Computer Science in China (FCS)*, vol. 4, no. 2, pp. 280-301, June, 2010.

[16] G. Sudeepthi, G. Anuradha, and M. Surendra Prasad Babu, "A survey on seamantic web search engines," *International Journal of Computer Science Issues*, vol. 9, no. 1, March 2012.

[17] V. Uren, G. L. Yuan, V. Lopezi, M. L. Hai, E. Motta, and M. Giordanino, "The usability of semantic search tools: A review," in *Proc. The Knowledge Engineering Review*, Cambridge University Press, pp. 361-377, 2007.

[18] Technical Overview of Cognition's Semantic NLP™ (as applied to Search).white paper, Cognition Giving Technologies New Meaning, 2007.

[19] K. Dahlgren and D. Albro,"Cognition Technology Resources Overview Semantic Map, System Architecture and Tools at Cognition giving new meaning," white paper, Cognition Giving Technologies New Meaning, 2008.

[20] P. J. Hane, Cognition giving new meaning, "The Most Advanced Commercially Available Semantic," *Journal of Natural Language Processing*, pp. 90-99, 2010.

Hikmat Ullah Khan is serving as assistant professor, Department of Computer Science, Attock, Pakistan. He has done his Master of Computer Science and Master in Science (Computer Science) from Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan. He has already published many papers in international conferences and journals. He has served as reviewer in many international conferences including 2011 4th IEEE International Conference on Computer Science and Information Technology (IEEE ICCSIT 2011) held in China on June 10 - 12, 2011. He is member of the technical committee and reviewer of International Arab Journal of E-Technology. He has keen interests in research fields of Information Retrieval, Semantic Web and Semantic Cache, Web Mining, Social Network mining etc.

Maliha Majid Qureshi has done her Bachelor of Science(Computer Science) from Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan. She has keen interest in research domains like Semantic, Semantic Web, Artificial Intelligence, Machine Learning etc. Currently, she is a regular student of Master of Science (Computer Science) in Department of Computer Science, COMSATS Institute of Information Technology, Islamabad campus, Pakistan. In the deparmtnet, she is also serving as Teaching Assistant (email: maliha_red@hotmail.com)

Bibi Asma has done her Bachelor of Science (Computer Science) from Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan. (email: asmaasmi@live.com)