Studying and investigation of the Semantic Agent Case Study (*SemanSearch*)

محرك بحث باستخدام الويب الدلالي "SemanSearch"

- أسماء سالم اعبيد . محاضر بقسم الحاسوب كلية التربية قمينس. جامعة بنغازى.
- سعاد عوض البدرى مساعد محاضر بقسم الحاسوب كلية التربية قمينس. جامعة بنغازي.
- حنان فرج نصيب . محاضر بقسم الحاسوب كلية التربية قمينس. جامعة بنغازي.

MA: Asma S. Obaid. Lecturer, Computer Department, Faculty of Education, Qamens. Benghazi University.
Email: hasob23@gamil.com.
MA: Souad A. Al-Badri, Assistant Lecturer, Computer Department, Faculty of Education, Qamens. Benghazi University.
Email: suaadAwad@gmail.com.
MA: Hanan F. Nassib. Lecturer, Computer Department, Faculty of Education, Qamens. Benghazi University.
Email:hananfarj@gmail.com.

المخلص: تعتبر عمليات البحث وإنتاج الابحاث العلمية أكبر النشاطات على شبكة الإنترنت ومع التطور الهائل في مجال التكنولوجيا دخل الإنترنت إلى أغلب المجالات ,حيث تعمل محركات البحث عن طريق تخزين المعلومات عن عدد كبير من صفحات الويب حيث أن معظم محتويات الويب التقليدية هي مناسبة لاستخدامات المستخدم وليس للآلات ونظرا للكمية الهائلة للمعلومات على شبكة الإنترنت ,لذلك من الضروري لشبكة الإنترنت أن تكون مفهومة بالنسبة للآلة .وتعتبر محركات البحث على أساس الكلمات الرئيسية ,مثل (AltaVista, Yahoo and Google) هي الأدوات الرئيسية للبحث عن المعلومات على شبكة الإنترنت .المشكلة الرئيسية في تلك محركات البحث هو فقدانها لمعاني الكلمات فذلك أدى إلى فقدان كميات كبيرة من المعلومات وكذلك ظهور كمية كبيرة جدا من المعلومات الغير مرغوب فيها .الويب الدلالي Semantic Web هو النظام الذي يسمح للآلات بالفهم و الاستجابة لاستعلامات المستخدم بناءاً على معاني الكلمات فذلك يعطي إثراء لمعالجة البيانات مما يجعلنا نتغلب على المشكلة الرئيسة من محركات البحث التي تعتمد على الكلمات. هذا العمل يصف نموذجا أوليا يدعي SemanSearch.يعتبر هذا النموذج دلاليا ليساعد المستخدم على الحصول على نتائج أكثر أهمية عند البحث عن المعلومات باستخدام محرك البحث القائم على أساس الكلمات الرئيسية وكذلك محركات البحث التي تعتمد على المعني أيضا . وينقسم بناء SemanSearch إلي مرحلتين أساسيتين هما :بناء الأنتولوجيا وبناء النموذج .تنفيذ SemanSearch ركز على تعديل جملة الاستعلام نفسه. وتستخدم ترجمة المفاهيم وإيجاد المرادفات الدلالية لاستخراج دلالات الكلمات ثم تدرج في جملة الاستعلام وكذلك المرادفات للكلمات الرئيسية والمرادفات العربية وذلك عن طريق استخراجها من الأنتولوجيا المطورة.

الكلمات الداله: دبلن كور , الانتولوجي , محرك البحث Hakia ,الويب الدلالي , النحوية او المعني , محرك البحث SemanSearch , Swoogle.

Abstract:

Semantic Web can be defined on the Internet that it was able to describe things in a certain way to allow all computers understand it. Tim Berners-Lee, inventor of the World Wide Web, defines the Semantic Web as "An extension of the current Web in which information is given well-defined meaning, enabling computers and people to work in better cooperation"

This paper describes both architecture and a prototype of SemanSearch, a semantic agent that helps user to get more relevant results when searching for information using a keyword-based search engine. SemanSearch is implemented using Jena (a java frame work) with the help of ontology that developed for education domain. SemanSearch also includes the Arabic meaning of concepts to get documents that contain the needed meaning but in Arabic. A comparative study compares keyword-based search via Google with semantics-based search via the SemanSearch prototype is used for evaluation.

Keywords: Dublin Core, Hakia, Ontology, Semantic web, Syntactic, Swoogle, SemanSearch

Introduction

The semantic web designed to help machines to understand more information on the web so that it can support richer discovery, data integration, navigation, and automation of task. He Semantic Web will only be possible once further levels of interoperability have been established. Standards must be defined not only for the syntactic form of documents, but also for their semantic content
(T. Berners Lee, l td. 2001).

Internet search engines have popularized keyword based search in which users can submit keywords to the search engine and a ranked list of documents is returned to the user (Sanjay. A, Surajit,l td,2002). He big problem of keyword based search engine such as Yahoo and Google is the loss of keyword semantics which gives words or multi-word phrases as atomic elements in document and query representations.

The search procedure is essentially based on the syntactic matching of document and query representations. The solution of this problem is the semantic search. Semantic search is based on retrieving documents based on semantic analysis of their contents using natural language processing (Fausto. G, l td, 2009) .he idea is that, differently from syntactic search, semantic search exploits the meaning of words, thus avoiding many of the well known problems of syntactic search as discussed in Semantic Search (Stephan . B, l td, 2008) .but all still lacks the use of Arabic meanings in searching query which prevents many relevant pages to be retrieved. In this paper we discuss both architecture and implementation of SemanSearch. In Section 2, we highlight some of the top ranked semantic search engines for semantic search on the Web. Section 3 states the problem statement and the benefits of our approach. Section 4, 5 and 6 sketches our own such approach. Section 7 shows our experimental results. Finally in Section 8 **.**

Materials and Methods

Compared to the other search engines the semantic search engines helps to find results for user queries very fast and accurately rather than the keyword matching, it gives the more relevant data and their reference links.

A way to represent the difference between the traditional search engines and the semantic search engines is to compare the results of the same query by both of them. Description of some of the best semantic search engines are given below.

Swoogle: Swoogleis a crawler-based search engine for the Semantic Web. It. Swoogle uses a set of crawlers to discover RDF documents and HTML documents with embedded RDF content. Swoogle reasons about these documents and their constituent parts (e.g., terms and triples) and records and indexes meaningful metadata about them to produces additional facts, constraints and metadata. Swoogle provides also web scale semantic web data access service, which helps human users and software systems to find relevant documents, terms and triples, via its search and navigation services. (Tim. F, l td 2004**)**

Problem statement and approach: Traditional web search is essentially based on a combination of textual keyword search. Most of research activities goes towards a more intelligent web search called *Semanticsearch*, which is currently one of the hottest research topics in both the Semantic Web and Web search but it does not consider Arabic concepts (or results for the corresponding Arabic concepts) so many results are ignored even they may have important information related to user query. We reused an university ontology for benchmark tests with some modifications including adding Arabic terms and Arabic synonyms for terms defined in the ontology.

Ontology: Ontology is an explicit specification of a conceptualization and specifies the primary concepts and the relationships among the concepts in a particular domain (Thomas.G, 1995). Computer science uses ontologies to describe specific conceptual terms and relationships in a specific domain in a standardized machine readable format. Ontologies are used for organizing knowledge in a structured way in many areas from philosophy to knowledge management and the Semantic Web (John .D, 2006) .Machine readable ontologies require a computer language to define the concepts and associated relationships. One of the

standard languages is the Web Ontology Language (OWL) developed by the World Wide Web Consortium (W3C). An example of a small part of our OWL ontology (in education domain)

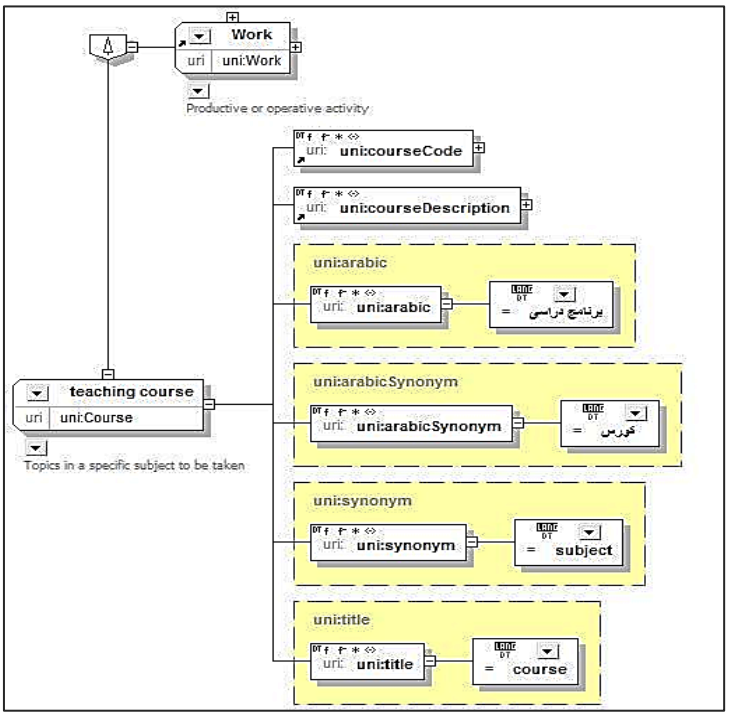

Figure 1 (Graphical representation of the class "*Course*")

Ontology implementation is a very difficult task because it‟s very difficult to collect all concepts in a domain and the relationships between them. The myriad of technical standards and specifications only address the formats of ontology. For example, RDF and OWL specify the syntax for how certain concepts and relationships should be represented but do not tell us whether „rock & roll‟ and „music‟ are related through a relationship called „genre‟ **Figure (1).** Graphical representation of the class "*Course*"

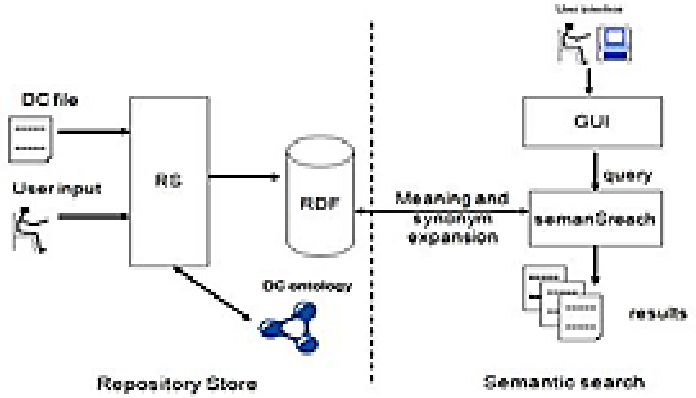

Figure 2 (SemanSearch work flow model).

System Overview: SemanSearch was designed as a plug-in for any kind of search engine. Thus, implementation of the SemanSearch focused on the modification of the query string itself, instead of modifying the search engine directly which is easier. As illustrated in Figure2, The key components of the agent are Repository store manage digital objects and other information. The repository store stores both the content and the metadata of the digital objects. Figure.3 shows a detailed architecture of the agent.

- **Preprocessing:** preprocessing of query string.
- **Produce Different Meanings of Keywords:** concept mapping and semantic matching is used to extract keywords semantics if a keyword hasmore than one meaning then different meanings produced to users to select the needed meaning.
- **Determine User Needed Meaning:** user selects the needed meaning.
- **Synonym Expansion:** extract.

Figure 2 SemanSearch work flow model.
Figure 3 Detailed architecture of the agent.

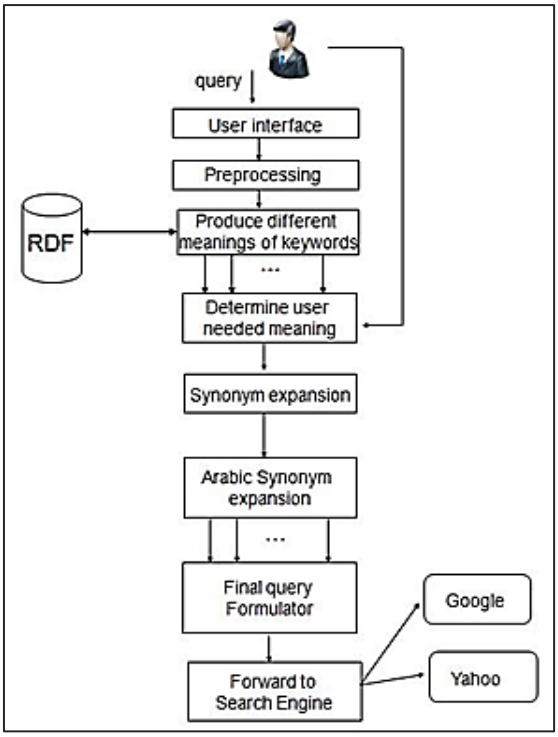

Figure 3 (Detailed architecture of the agent)

Semantic Search: Semantic search uses data model (RDF Model) and data structures of syntactic search with the only difference is that now keywords are substituted with

$$\text{Precision} = \frac{\text{Sum of the scores of sites retrieved by a search engine}}{\text{Total number of sites retrieved}}$$

concepts in ontology and syntactic matching of keywords is extended to semantic matching of concepts. This idea is schematically represented in the equation below:

Keyword search +Concept Mapping +Semantic Matching → Semantic Search

Let us consider in details how the keywords in the query are converted into the concepts in the ontology and also how the semantic matching is implemented.

Concept Mapping: Keyword search does not take into account concepts which are semantically related to the query concepts. For instance, a user looking for *"chair"* might

not be interested in documents which talk about the furniture word (seat) but in documents which talk about

the work position word (supervisor).

Semantic Matching: In semantic Search, the search process is done using concepts that are semantically related to query concepts. We assume that, when a user is searching for a concept, he is also interested in synonyms of that concept. For example, the *"executive"*

)f the *irector*" are
tor", " *and*

Therefore, documents describing the concept should be returned as an answer to the query describing the synonyms of the concept. Formally a query answer A (C_q, T) is defined as follows:

Where C^q is a query concept extracted from the query *q*, C^d is a document concept extracted from the document *d*, and T is a terminological knowledge base (the developed

$$A(C^q, T) = \{d \mid \exists C^d \in d, \text{s.t. } \mathcal{T} \models C^d \sqsubseteq C^q\}$$

(Enrico. F,l td 2010).

ontology) which is used in order to check if C^d is a synonym for C^q. Equation 2 states that the answer to a query concept C^q is the set of all documents *d*, such that, there exists concepts C^d in d which is a synonym for the query concept C^q.

During query processing, A (C^q; T) must be computed for every query concept C^q in the query. One approach is to sequentially iterate through each concept C^d, compare it to the query concept C^q using semantic matching.

Results

In a huge search results, the user is sometimes able to retrieve relevant information and sometimes able to retrieve irrelevant information. The quality of searching the right information accurately would be the precision value of the search engine .n the present study, the search results which were retrieved by Google were categorized as "more relevant", "less relevant" and "irrelevant" on the basis of the following criteria :

•If the content of the web page closely matched the subject query, then it was categorized as *'more relevant'* and it was given a score of 2.

•If the content of the web page not closely related to the subject of the search query, then it categorized as *'less relevant'* and it was given a score of 1.

•If the content of the web page is not related to the subject of the search query ,then it was categorized as *'irrelevant'* and it was given a score of 0 Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage Table1 shows a comparison between results retrieved by Google search for origin query and final query by SemanSearch for single-word queries and table 2 for multi- word query. The precision of Google was calculated using equation.

Table 1 Comparison between results retrieved by Google search for origin query and final query for single-word queries.

	More relevant	Less relevant	Irrelevant	Total
	sites (%)	sites (%)	sites (%)	Precision

Origin query	45	36	19	54
Final query	55	31	14	70

Table 2 Comparison between results retrieved by Google search for origin query and final query for multi-word queries.

	More relevant sites (%)	Less relevant sites (%)	Irrelevant sites (%)	Total Precision
Origin query	56	15	29	63.5
Final query	56	29	15	70.5

As seen Table 1: 45% sites were more relevant using origin query and the percentage is increased to 55% when using the final query produced by SemanSearch , It was also observed that 19% of the sites were irrelevant using origin query and that percentage is decreased to 14% when using the final query produced by SemanSearch.

As seen in Table 2: 56% sites were more relevant using origin query and that percentage is increased to 56% when using the final query produced by SemanSearch .It was also observed that 29% of the sites were irrelevant using origin query and percentage is decreased to 15 %when using the final query produced by SemanSearch.

Conclusion and Future Work

This comparison study showed that the Google gave better search results with more precision for final query produced by SemanSearch for simple one word and multi-word queries compare to precision of the origin query itself. Over all precision of final query results was higher than of origin query. This means that Google search is improved by getting more relevant results than submitting the origin query. Natural language processing tools for semantic and syntactic analysis over user queries will be needed to find corresponding concepts in the ontology. Exact string matching is not enough, since user queries are not only simple but rather contain complex phrases. Therefore, a matching technology based on case-based reasoning should be used, since complex queries consist of one or more phrases.

References

- Enrico. F, Simonetta. M, Wim. P, Daniela. T, 2010, Semantic Processing of Legal Texts: where the Language of Law Meets the Law of Language.
- Fausto. G, Uladzimir. K, Ilya. Z, 2009, The Semantic Web: Research and Applications.
- John. D, Rudi. S, Paul. W, 2006. Semantic Web Technologies: Trends and Research in Ontology-based Systems, England.
- Sanjay. A, Surajit. C and Gautam. D, 2002, DBX plorer: A System for Keyword-Based Search over Relational Databases.
- Stephan. B, Marko. G, Peter and Thanh. T, 2008, Semantic Search (SemSearch 2008), International Workshop located at the 5th European Semantic Web Conference Tenerife, Spain.
- The World Wide Web Consortium (W3C) website: http://www.w3.org/
- Thomas. G, 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human and Computer Studies.
- Tim F, Yun P,Scott . C, Joel .S, Anupam. J, Pavan. R, Pan Vishal D, Li. D, 2004, Swoogle: a search and metadata engine for the Semantic Web. Proceedings of 13th ACM Conference on Information and Knowledge Management.
- Tim. B, James .H and OraL, 2001. The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific America.