2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies

New Semantic Indexing and Search System based on Ontology

Fraihat Salam
Faculty of Information Technology
AL-Ahliyya Amman University
Amman, Jordan
Fraihat.salam@gmail.com

***Abstract*—Information retrieval becomes a very complex process for search engines on the Web, this is due to, first, the staggering growth speed of the number of web site and, in the other hand, the search algorithms by keywords (terms) used currently are not suitable to better exploit this huge information quantity. These elements make the information retrieval by the current search engines very difficult and does not meet the users' needs. In this paper, we propose a new method of semantic information retrieval based on ontology. Our method allows the indexing and searching engine to take in consideration the documents semantic level, which significantly improves the quality of search results.**

We have implemented our approach on an indexing and searching engine based on ontology (called MIRO Moteur d'Indexation et de Recherche base sur les Ontologies). MIRO offers a multilingual semantic search of documents using concept instead of term. Additionally, MIRO offers a guided search tool, and a tool for an automatic enrichment of ontology. Moreover a comparison of results between MIRO and PhpDig (an open source search engine) is presented.

Keywords-component; Ontology; Semantic Indexing System; Searching Engine; Semantic Web Information retrieval System; Ontology modeling; reuse, extraction, and evolution; Semantic Query Processing.

I. INTRODUCTION

Today's Web is a major source of information, and the richness of it is largely underexploited. Indeed, if its gigantism unanimous, it is different from its ability to meet our information needs. The question today, for the search engines is not how many pages will you find? But, how many relevant web pages will you give me? We lose a lot of time looking for our needs in the pages retrieved by the search engines, and often we are forced to change our search queries. These systems use a centralized database for indexing information. They are based on queries from simple keywords. The recall rate is high, but the accuracy is low. This is due to the disambiguation, wrong context, the use of different words (ex synonyms), more specific words, or more general (hypo-hyperonymic). These systems rarely take into consideration the semantic content of the document to the index. The approach allows taking into consideration the semantics of the document focuses on techniques of information retrieval based on ontologies. For these types of systems, documents are indexed according to the ontology concepts.

In this paper, we present a new indexing and retrieval technique of web documents based on ontologies. Latter consists of retrieves the information contained in a document from the used ontology concepts, in order to take into account the semantic content of documents.

This paper is structured as follows. First, we introduce the Web semantic then we present the ontologies and their uses. After that, we present the technique of indexing and searching websites based on ontology. Finally, a comparison between MIRO our search engine and PHPDig [1] (open source search engine by term) is presented followed by the conclusion and future works.

II. SEMANTIC WEB

As defined by Tim Berners-Lee (creator of W3C [2] standards): "The Semantic Web is what we will get if we perform the same globalization process to Knowledge Representation that the Web initially did to Hypertext" [1]. Semantic Web aims to improve our relationship with the Web by just making the information contained therein "understandable" by the machine as well as by human.

Therefore, the semantic Web is related to the current Web, enhanced by semantic information. Current research on the Semantic Web is based on knowledge representation, Ontologies, Annotations and reasoning model, and also other areas such as databases.

The idea of the Semantic Web is not to make sure that computers can understand human language or operating in natural language, it is not artificial intelligence allowing the Web to think, but simply to group the information in a useful way, as a huge database, where everything is written in a structured manner.

The Semantic Web is an exchange space, which is still under construction with various promising features to name a few: it provides sufficient information on resources, adding annotations in metadata [3] form. Additionally, it describe their content in meaningful and formal ways using Ontology, to be interpreted by humans as well as machines

[1] PhpDig can be downloaded from http://www.phpdig.net

[2] W3C : word wide web consortium www.w3c.org

[3] Metadata are secondary information affixed to primary resources. They are created by the authors of the documents to be read by machine (indexing). They are more oriented towards the discovery of resources description.

978-0-7695-5044-2/13 $26.00 © 2013 IEEE
DOI 10.1109/EIDWT.2013.60

III. ONTOLOGY

A. Definition

In Philosophy, ontology refers to the science describing the different kinds of entities in the world, and the relationship of these genres between them. In the Web domain, ontology defines the terms used to describe and represent an expertise area. The ontology is represented by schemas and knowledges to describe a domain by structured ways in a readable format by computers. Ontology allows establishing the interoperability and sharing between different systems. We can imagine it as a database with a very large network of relationships between concepts.

The Ontology use can provide us with several advantages such as:

- The enhancement of the web functioning by finding pages relating to a specific concept instead of those found using ambiguous keywords.
- Sharing the common knowledge of the information between people or software agents in a specific area.
- Enable the reuse of the field knowledge on reusing its ontology for different fields.
- Facilitate the field change suppositions in case our relating knowledge has to be changed.

B. Ontology Construction

The construction of the ontology has to be elaborate from a range of well-defined knowledge by a clear operational objective, and based on objectives knowledge which the semantics can be formally and rigorously expressed.

The ontology which is constructed (built) to resolve the web retrieval issue should have the following fundamental characteristics:

- It should precisely define the terms and their meanings, the terms meanings have to be sufficiently precise so that the ontology can be used as a reference and provide a vocabulary shared by communities in different areas.
- It should be based on rigorous and formal principles in which each concept used for resources semantic markup should have a shared signification and can be reused for different applications.
- It should be multi-use and has to be generic enough in order to be reusable for different uses, different forms.

To respect those characteristics to build the Ontology that we use in our search and indexing Engine we decided to adopt a seven-steps approach proposed by Noy and McGuinness (2002) and depicted in [7].

IV. IMPLEMENTATION OF NOVEL SEARCH AND INDEXING ENGINE BASED ON ONTOLOGY

Search and Indexing Engine is a tool that permits to extract from information, principally textual, words and terms that are related and most representatives to that information and store them in an index. The same tool then traverses this index to identify the most relevant terms related to the user's query and sorting information to provide. The Search and Indexing Engine, which we use, is based on the presence or absence of a word/term in documents, without exploiting the semantic level contained in the document index.

The novelty of our Indexing and Search engine (MIRO) is the use of ontologies to exploit the semantic content of documents to better index them and reduce the silence and increase the accuracy of the research. MIRO includes the following three parts (see Figure1): Indexation part, Research and Presentation part and Ontology enrichment part.

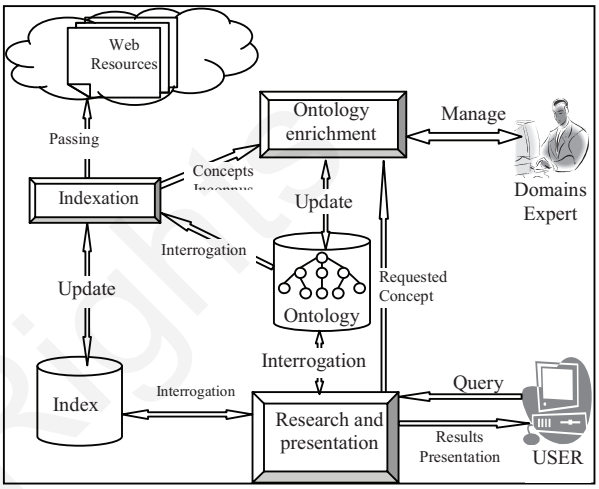

Figure 1. System functional Schema

A. Medical Ontology Implementation

Our work is a part of a global project of a high quality medical information research on the web for a urology specialist clinic, we build then our ontology for Urology domain, which has specific needs such as sharing and structuring medical information [8], especially for Urology domain.

There are several thesauruses, corpus, Ontology in a medical domain (ex: GALENS, UMLs, MENELAS), but none of them satisfies the needs of the medical research and indexing domain.

After performing our research, we have not found a specific ontology of the urology domain, it is why we have chosen to implement our own ontology structure, and this structure is then used by urology's specialists in order to fill the ontology with the domain knowledges.

Figure 2 shows a part of the constructed ontology; it represents the organs hierarchy of the Urinary system, this hierarchy uses the relation "part of". We have defined with the help of Urology experts, a fairly complete network of relationships (about fifty relationships) between concepts to cover up the semantic that worn by concepts.

For our ontology implementation we are using OWL (Ontology Web Language) [11] and we have opted for the ontology editor Protégé-2000 [9, 10], this choice is supported by several reasons:

- It is a Free and Open Source editor.
- It can, via "plug ins", import and export ontologies in different implementation languages ontology-schema RDF, OWL, DAML, OIL ... etc.
- Ontologies can be edited interactively within Protégé and accessed with a graphical user interface and Java API.
- Ontology editor for defining classes of concepts.
- Automated generation of tools for building knowledge bases that define instances of concepts.
- Knowledge-visualization systems.
- Lots of user-contributed "plug ins" and the availability of various "plug ins": JSave, Protégé Web Browser, XML Schema, Docgen, PROMT, OWL-S Editor.
- Ability to archive ontologies and knowledge bases in a variety formats.

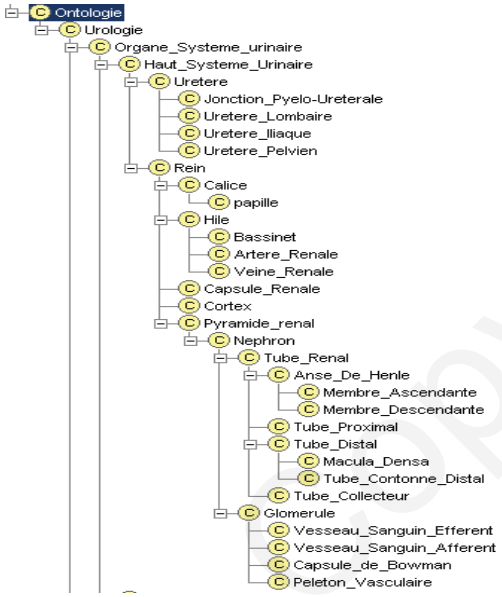

Figure 2. The Ontology built party, it represents the Organs hierarchy of the urinary system, and this hierarchy represented the "part of" relationship

The ontologies construction steps we have followed and which are described in [7] are:

Step 1: the Domain definition and the domain scope

- the covered domain by our ontology is the urology
- the ontology will be used by doctors , patient and domain researchers via a search engine
- the ontology maintenance will be ensured by the specified domain experts.

Step 2: Considering the possibility of reusing the existing ontologies

We have extracted ontologies concepts such as GALEN, MESH, related to the urology domain; we use these concepts in order to widen our ontology.

Step 3: enumerate the most important terms of Ontology

Due to the high number of terms to be treated in our ontology, we cannot mention them all in this paper, and this terms list will never be exhaustive.

As the ontology construction is an iterative process, new concepts will always be added to the ontology.

Step 4: define classes and their hierarchy

In this step, we will use the ontology model that we have developed to classify the collected terms from the previous step according to their natures (body part, symptom, disease ... etc). This ranking is the first level of our ontology. And for each class of the first level, we use the top-down approach in order to define the terms hierarchy of this class.

Step 5 and 6: define the classes' properties and their facets:

The classes' properties and their facets are defined in the designed ontology model level and each ontology concept will have the class properties of the model it belongs to.

Step 7: Creating the instances

Our ontology concepts represent terms related to the urology domain, this is why, and there is no instance for these terms. Example: the concept "Rein" has no instance.

Our system has been designed in order to use any kind of ontology, it can be edited or imported by ROTEGE2000, when the ontology administrator wants to use a new ontology, and he has to indicate the ontology file (with the extension «pprj»: protected project).

B. Semantic Indexing based on Ontology

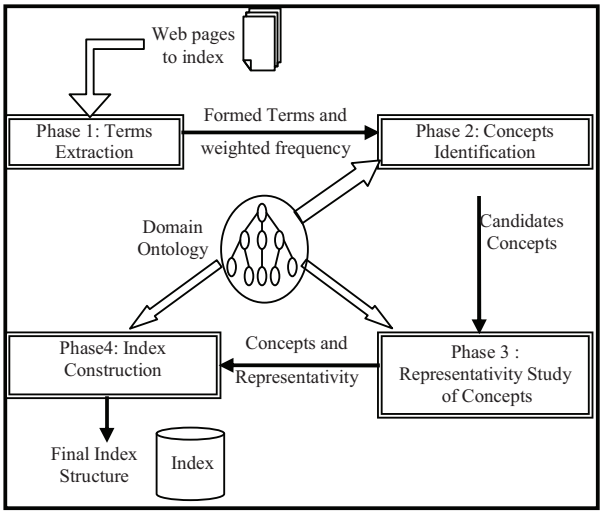

Figure 3. The Indexing general process based on Ontology.

In this section, we illustrate in this section the indexing process of a Webpage using ontologies to improve the

quality of indexing documents on the WEB. Figure.3 presents a global schema of our system.

The indexing process described in Figure 4 is a recursive process. It is recurrent for each website page we want to index.

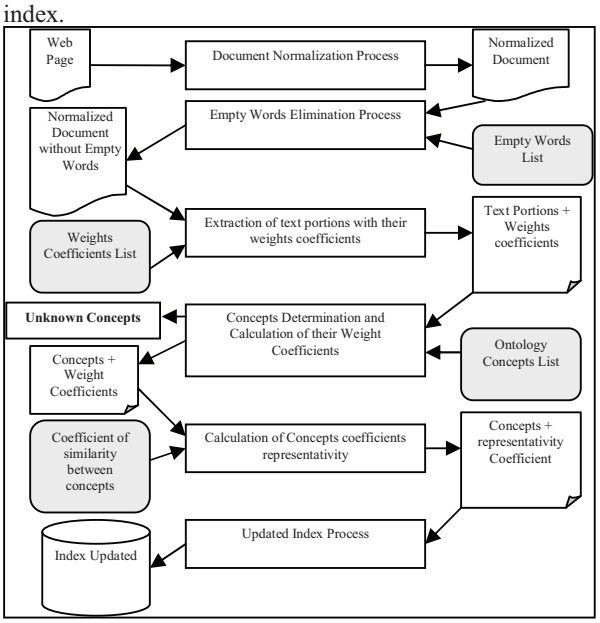

Figure 4. General scheme of Webpage Indexing Process based on Ontology.

The indexing process constitutes of several steps:

Step1: Document normalization/lemmatization and removing the stop words: consists on transforming each document term from its inflected form into its canonical form.

Step 2: Concepts Identification and calculation of their weights in the Webpage: For this we identify the ontology concepts present in the extracted text portions from the step1. For example: if one of the concept synonyms is present in the document, we consider that the concept itself is present in that document.

Step3: coefficients calculation of the concepts representativeness, for this we need:

- A table containing coefficients of similarity between each concept pair, this coefficient is calculated from the arcs number of the shortest path between these two concepts relating to ontology
- Ontology concepts and their weights calculated in step 2.

First, we have to calculate the sum of the similarity coefficients (SomSim) for each concept in the document; this sum is the same as the one of the similarity coefficients between the concept and all the other concepts of the document.

The equation to calculate the similarity coefficients for each concept is:

$$SomSim(C_j) = \sum_{k=1}^{m} sim(C_j, C_k) \quad (1)$$

SomSim (Cj): Sum of similarities between the concepts Cj and the other document concepts

Sim(C1,C2): Similarity coefficient between the concepts C1 and C2

m: number of concepts in the document.

The result calculated by (1) is normalized by dividing it by the greater value found, in order to get a result between 0 and 1. Therefore, we obtain a concept convenience coefficient (see (2)).

$$Conv(C_j) = \frac{SomSim(C_j)}{\max_{k\in[1,m]} SomSim(C_k)} \quad (2)$$

The concepts which are most related to the other concepts of the same document are enhanced by this coefficient .Otherwise, the concepts which are semantically isolated have a low coefficient, and here we can observe the document semantic processing.

Finally, we have to find the representativeness coefficient of each concept present in the document in order to identify the concepts more representative in the document among those found previously. This coefficient is calculated from the concept weight and its sum from the similarities coefficients (see (3)).

$$representativity(C_j) = \frac{\alpha * Conv(C_j) + \beta * SomSim(C_j)}{\alpha + \beta} \quad (3)$$

α and β: weight coefficient.

Step 4: Index update.

C. *Semantic Research based on Ontology*

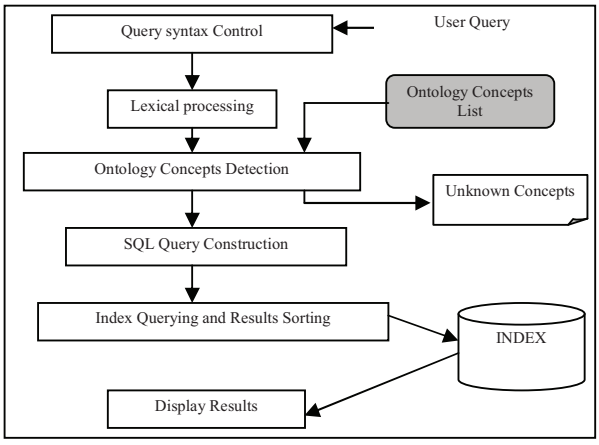

Figure 5. General scheme of Searching Process based on Ontology.

The research process represents the interface between the system and the user. Indeed, the user expresses his need using a request (query) through this process. This request follows a precise syntax defined by the research process so that it can exploit it in order to provide documents relating to the user needs.

As presented in Figure5, the research process proceeds following the steps below:

Step1: the process starts with a query syntax check, in order to make sure that the query is well formed.

Step2: the query is subject to a lexical treatment which consists of normalization (lemmatization) and stop words removal, same as in the indexing process.

Step3: This is the most important step that distinguishes our research method from the other conventional methods; it consists of searching the requested concepts in the query among the existing concepts in the ontology. This phase provides us with two different lists:

- A list that includes the recognized concepts (present in the ontology).
- A list that includes the unrecognized concepts, which we use to enrich the Ontology (see next section).

In case the user provides, for example, a concept synonym in his query, the research will be done using the concept corresponding to that synonym in the ontology. The research will be about all the concepts synonyms and not about the mentioned one. This permits to retrieve all the documents with the same meaning as the term introduced in the query, it is similar to performing a semantic research.

Finally it will be translated from the user query (using the recognized concepts) into a SQL query which permits to interrogate the index in order to get an answer for the user.

D. Ontology Enrichment Tool

Our system provides Ontologies enrichment tool from two different sources. The first source is the indexing engine which provides a list of unknown concepts encountered in documents during their indexing. The second source is the search engine that lists the concepts requested by the user, which are not included in the ontology.

V. RESULTS AND DISCUSSIONS

In order to realize our comparatives tests, we have downloaded a medical website and we have run MIRO and PhpDig on the same site and on two identical machines performance. Figure 6 and 7 illustrate our tests results.

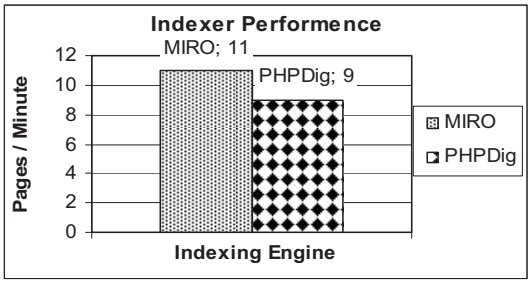

Figure 6. The performance test of our Indexing engine MIRO and PHPDig indexer, measured by the number of indexed pages per minute. The result demonstrates that MIRO is somewhat better than PHPDig, using of ontology, which is supposed slow down more the indexing process.

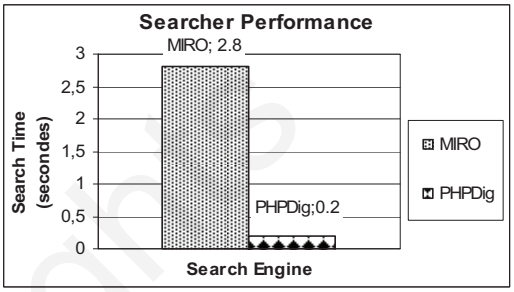

Figure 7. The performance test, of our Search engine MIRO and PHPDig searcher, measured using the time interval (per seconds) between the user query and the search engine response. The result shows that PHPDig is faster than MIRO. This is mainly due to the complexity of processing performed (Standardization, Ontology query ...etc).

TABLE I. COMPARATIVE TABLE OF THE RELEVANCE OF RESEARCH RESULTS (REL = RELEVANCE)

		MIRO	PHPDig
First link	Engine Performance	Very Rel	Very Rel
	Reel Performance	Very Rel	Medieum Rel
Second link	Engine Performance	Very Rel	Very Rel
	Reel Performance	Very Rel	Very Rel
Third link	Engine Performance	Very Rel	Medieum Rel
	Reel Performance	Very Rel	Medieum Rel
Fourth link	Engine Performance	Very Rel	Medieum Rel
	Reel Performance	Very Rel	Very Rel
Fifth link	Engine Performance	Very Rel	Medieum Rel
	Reel Performance	Medieum Rel	Very Rel

In Table 1, we have compared the top five links given y each system. MIRO shows a very high performance compared to PHPDig. Indeed, it finds the most pertinences pages for user query, instead of PHPDig which rarely provide us first with the most related pages to our query. This proves that the semantic indexing technique by ontology improves significantly the relevance of the indexing and search processes.

VI. CONCLUSION AND FUTUR WORKS

The search engines available on the web do not resolve the problem due to the noise and the silence of web pages. The best solution is to exploit the documents semantic content, by using Ontologies.

The solution we proposed is within this context; it uses the ontology in the indexing and research engine in order to enhance the research results relevance.

We have developed an Indexing and Retrieval engine (called MIRO), which has proven more efficient in pertinence respond to a user request than a classic search engine (by term), despite the fact that it is slower because of indexing and Search processes are using ontology. This problem can be solved using performance servers which are very accessible today. We plan to extend our engine to accommodate other ontologies of different domains.

REFERENCES

[1] T. Berners-Lee, J. Hendler and O. Lassila, "The semantic web", Scientific American, Vol 284(5), pp 34-43, 2001.

[2] E. Desmontils and C. Jacquin, "Indexing a Web Site with a Terminology Oriented Ontology", In The Emerging Semantic Web, I.F. Cruz, S. Decker, J. Euzenat and D. L. McGuinness Ed. IOS Press. pp 181-197. 2002. (ISBN 1-58603-255-0)

[3] F. Fürst and F. Trichet, "Integrating domain ontologies into knowledge-based systems", In Proceedings of the Eighteenth International Florida Artificial Intelligence Research Symposium Conference (FLAIRS'2005). AAAI Press. pp 826-827. 2005.

[4] C. Golbreich, S. Zhang and O. Bodenreider, "The foundational model of anatomy in OWL: Experience and perspectives". J. Web Sem. Vol, 4. pp 181-195. 2006.

[5] T. Poibeau, "Semantic annotation: Mapping text to ontologies", In the International Journal of Metadata, Semantics and Ontologies (IJMSO), Inderscience publishing. vol. 2 n°2, ISSN, 1744-2621. pp 67-78. 2007.

[6] E. Desmontils, C. Jacquin and L. Simon, "Ontology Enrichment and Indexing", Process. RR-IRIN-03.05. pp, 18. Nantes. Mai 2003.

[7] F. Natalya and L. Deborah, Mc. Guinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford University. 2000.

[8] P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet and JF, Boisvieux, "Issues in the structuring and acquisition of ontology for medical language understanding. Methods Inf Med", Vol 34 (1/2). pp15–24. 1995.

[9] M.A. Musen, R.W Fergerson, W.E Grosso, N.F. Noy, M.Y. Grubezy, and J.H. Gennari, "Component-based support for building knowledge-acquisition systems. Proc", Intelligent Information Processing (IIP 2000) Conf. Int. Federation for Processing (IFIP), World Computer Congress (WCC'2000), Beijing, China, pp 18-22. 2000.

[10] J. Gennari et al. The evolution of Protégé: An environment for knowledge-based systems development.Int. Journal of Human-Computer Interaction, 58(1),

[11] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, C. Lutz, " OWL 2 Web Ontology Language: Profiles (Second Edition)". W3C eds, 2012.