I.J. Information Technology and Computer Science, **2015, 02, 1-11**
Published Online January 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2015.02.01

Semantic Indexing of Web Documents Based on Domain Ontology

Abdeslem DENNAI
EEDIS Laboratory, Djillali Liabes University, Sidi Bel Abbes Algeria
E-mail: de_selam@yahoo.fr

Sidi Mohammed BENSLIMANE
EEDIS Laboratory, Djillali Liabes University, Sidi Bel Abbes Algeria
E-mail: Benslimane@univ-sba.dz

***Abstract*—**The first phase of reverse engineering of web-oriented applications is the extraction of concepts hidden in HTML pages including tables, lists and forms, or marked in XML documents. In this paper, we present an approach to index semantically these two sources of information (HTML page and XML document) using on the one hand, domain ontology to validate the extracted concepts and on the other hand the similarity measurement between ontology concepts with the aim of enrichment the index. This approach will be conceived in three steps (modeling, attaching and Enrichment) and thereafter, it will be realized and implemented by examples. The obtained results lead to better re-engineering of web applications and subsequently a distinguished improvement in the web structuring.

***Index Terms*—**Reverse Engineering, Ontology, Semantic Distance, Semantic Indexing, Semantic Web

I. Introduction

Extracting information from HTML pages or XML documents was considered a subject of research strongly advocated in the areas of information retrieval on the Web, web application reverse engineering, their maintenance and knowledge engineering.

Web-oriented applications have become the most important means of communication for commercial enterprises of all kinds. They provide the main engines that not only improve the brand image of the enterprise, but also act as useful resources to increase global market share of a company. However, most web-oriented applications are built in a hurry. To shorten development time, the conceptualization phase is often sacrificed, and associated documentation is neglected. In addition, during the operational phase, Web-oriented applications are modified according to the enterprise's needs. They undergo various degradations affecting both their information content and their navigational structure. The heterogeneous and dynamic components constituting a Web-oriented application, the lack of effective programming mechanisms for the production of these applications, the rapid development of these applications by processes that do not often meet traditional approaches to systems development information, make the maintenance and development of these applications complex and expensive. In practice, most conceptual schemes of information systems and databases are developed essentially from zero. However, over the last decade, several approaches have emerged, with the objective of maintenance Web oriented applications based on the reverse engineering process [1]; [2]; [3]; [4]; [5]; [6]; [7].

On the other hand, several researchers [8]; [9]; [10] have demonstrated that the concept of ontology is used to analyze knowledge in a domain by modeling the concepts relevant to one or more applications in this domain. Recently, several approaches attempt to use the ontologies as a semantic source for the derivation of conceptual schemas [11]; [12]; [13]; [14]. However, most of these approaches assume the existence of useful information for this extraction. In addition, if the domain ontology used is large enough, the derived conceptual schema may include several unnecessary concepts and relations.

The objective of this paper is to present the first three phases of web-oriented applications reverse engineering based on ontology using semantic indexing, which are extraction, validation and enrichment. The first phase allows the extraction of useful information from HTML pages including tables, lists and forms and from tag-based XML documents. This information represents the candidate elements for the identification phase. The extraction uses the domain ontology as a source for the identification of semantic concepts hidden in HTML pages or XML documents.

The rest of the paper is organized as follows: Section 2 presents the contribution of the technologies HTML and XML for a semantic web. We present the related works in section 3. In Section 4, we present our semantic indexing approach by designing it in addition to the associated detailed algorithms and in the 5th section, we give a graphic description of our software achieving and implementing the indexing approach with at the end an interpretation of the results gotten before concluding in Section 6.

II. Web Data Structuring by an Indexing

The domain of the development of the applications oriented web requires, currently, the consideration the

passage of the traditional web toward the semantic web, which is a topic of actuality research greatly landed by the web developers. The technologies HTML (HyperText Markup Language) and XML (eXtensible Markup Language) remain very important in this domain and that appear like interesting resources for everything that can constitute real numeric document reservoirs.

The use more and more the XML and of degrees less the HTML in the structuring of the web offers possibilities of combination between the information research and the data bases questioning in the web and this through their very fine ways of description of the documents and of the attachment between their different parts.

Some conception methodologies have been proposed for Web applications based on HTML. But the limits imposed by this language, notably during the process of information research, and the emergence of XML as format of data brings as a matter of course to use XML for the construction of important web sites. This use permits to exploit the enormous possibilities of representation and interoperability offered by this language. It permits to do a clean and distinct separation between the site content (data) and the presentation during the process of the site conception on the one hand and, on the other hand, to exploit the site data after its realization.

The objective of the semantic Web is to increase therefore the efficiency of the information research [15]. This while making evolve the indexing techniques based on thesauruses toward techniques that use the knowledge representation and the Artificial intelligence.

A. Unstructured web data

The fast development of the WWW (World Wide Web) and the success of the language HTML permitted the construction of thousands of web sites generating a quantity important of accessible information on Internet. At the time of the construction of most these sites, the most current approach consists in focusing a lot more on the implantation of a solution that on the development process. These web sites present a set of pages HTML statics: the content only varies when the server's administrator either modifies them or interactive and dynamic: the content depends either of the information localized on the server (connection with a data base for example), either of parameters given in a transparent way by the customer's navigator.

Some development tools permitted to bring a substantial help in the generation and the setting in fast of applications web, with the help of the ASP (Activate Server Pages of Microsoft) technologies, JSP (Java Server Pages), PHP (Personal Home Page or Hypertext Preprocessor), PL-SQL, (Oracle-Web)... These technologies permit to extract some information dynamically from various sources of data and to include them in models of pages HTML. In these applications, the inventors often privileged the aspect presentation to the detriment of the data structuring. It is at the time of the exploitation of these sites that this approach shows its limits. The calm problems are often due to the increase of the size and the complexity of the sites, if need be of an interoperability with other applications, to the necessity of modifications during the time and to the lack obvious of possibilities of the pages HTML questioning.

B. Semi structured web data

The XML norm must be seen as such like a tool permitting to define a language (one says whereas it is about a structuring language or simply of a Meta language), permitting to create documents structured with the help of tags. So, by using extensibility of XML, it is possible to represent simultaneously the content and the logical structure of document. The continuous growth in structured documents stored in companies has caused different efforts in developing retrieval systems based on document structure. These systems exploit the available structural information in documents, as marked up in XML, in order to implement a more accurate retrieval strategy and return document structural elements instead of complete documents. However, much of the information is contained in the text fields not just in tag labels [16].

This Meta language that encouraged the expression of the standards specifications and the description norms, as RDF (Resource Description Framework), DC (Dublin Core), LOM (Learning Object Metadata)..., can offer the possibility to create the documents about that can be been like an intrinsic data base. In more these documents can be in conformity with structures, based themselves on the XML language (according to two existing recommendations that are DTD and XML Schema).

C. Indexing

The overall goal of indexing is to identify the information contained in any text and represented by a set of entities called index, in order to facilitate comparison between the representation of a document and a query. Rather, the indexing process is the transfer of the information contained in the text to a different representation space treatable by a computer system [17].

The use of indexes dates back to the fifteenth century, shortly after the invention of printing. The indexes (or indexing terms) play an important role in the search for information in that they identify with what words we can find a document [18].

FLUHER.C defined the indexing as follows: «the documents are read by a librarian who deduces the main themes and translates them into a list of words called descriptors of documents. This set of words is the index of the document and represents the description of its semantic content» [19].

POM and al. 97 define indexing as an operation designed to facilitate access to the contents of documents or set of documents from a subject or a combination of subjects or any other input that is useful for research [20].

- *Indexing techniques*

Indexing is the reduction in the data volume in a document through a representation of this document by keywords. Indexing can be done in a manual way (manual indexing), automatic (automatic indexing), assisted (semi - automatic), or by annotation [21].

Manual indexing is an operation that is to inventory concepts of a document and to represent them with a documentary language; often several semantic indicators: Classification index, free descriptors, authorized terms, descriptors, or keywords of a thesaurus. [22]; [23]

Automatic indexing uses software methods to establish an ordered list (index) of all the words in the documents with the exact location of each of their occurrences, and which best match the information content of a document [24].

Current systems replace humans for a substantial part of their expertise (semi-automatic indexing), in fact, they do not replace them completely, because the term "automatic indexing" implies a total system response, this is far from the case because human intervention is still needed. One example is SINTEX ALEXDOC as software and computer-aided indexing [25].

Generally, annotating a document is attaching to one of its parts a description that corresponds to the use we wish to make later. The scholarly annotation is necessary for intellectual work on the texts and often comes in the form of comment; linking and building a network of inter texts [21].

- *Semantic Indexing*

The semantic indexing has for objective the representation of the documents and requests by the senses of the words (or the concepts) rather than by the words of indexing them even. The interest of such an approach is to raise the ambiguousness of the words and to solve the problem of disparity of the terms.

The semantic indexing is the setting of our work (it will be retailed in our proposed approach) and that rests on the indexing of the HTML page or the XML document for which we have an ontology that can be constructed from the corpus or by using different resources. The choice to deal with specialized corpora simplifies the task by limiting the vocabulary, the ambiguity and variability of syntactic forms.

III. Related Works

The appearance of XML documents after the HTML page has provoked a lot of researches on adapting information retrieval techniques to structured documents (information extraction). Taking into account the logical structure of documents affects the document representation.

Wilkinson in [26] was the first to propose an information retrieval system based on document structure. In his system, Documents are split in section and the query is compared to each section. Document relevancy depends on different aspects: the frequency of terms in document content, frequency of term in a section content and section type. He applies the TF-IDF [1] formula to section of document instead of the whole document [16].

Yosi Mass in [27] describes a method for component ranking in XML documents by creating separate indices for the most informative logical element type in the collection of documents. They have improved their approach by proposing document pivot to compensate the problem of the data outside the scope of the logical element. The document pivot scales scores of logical elements by the scores of their containing articles. Their method is based on the vector space model and TF-IDF formula [16].

Khan in [28] proposes a concept-based model using domain-dependent ontologies. In this method he uses an automatic disambiguation algorithm which prunes irrelevant concepts. Only relevant concepts are associated to documents and thus they participate in query generation [16].

Zargayouna and Salotti in [29] the computation of term weights is influenced by the context (the indexing unit) in which they appear. The computation of weight based on the TF-IDF method is applied to tags. Thus, the author proposes the TF-ITDF[2] formula, which estimates the discriminatory power of a term t for a tag b in a document d. This work uses the concept and document structure together.

Chagheri and all in [16] propose a semantic indexing model which exploits both the logical structures and the semantic contents of documents. This method is an extension of the vector model of Salton (Salton, 1968) adjusting the calculation of the TF-IDF by considering the structural element instead of whole document.

In our approach, we suggest using a semantic resource like WordNet to model the semantic of document content. This indexing allows a search based on context (for structure) and semantics (the concepts of ontology). Our main contributions compared to the above work can be summarized in the following points:

1. The use of two types of web documents (XML and HTML).
2. Knowing that the reverse engineering of the web-oriented applications passes by four phases that are: The extraction, the identification (validation), the enrichment and the conceptualization, we demonstrated, through our detailed manner of the approach description, its positive contribution in the first three phases of the reverse engineering.
3. In the semantic attachment phase and in addition to the WordNet tool, we applied the semantic distance between the concepts extracted from HTML pages or XML documents and those of the ontology.
4. In the enrichment phase, we enriched the index by other ontology concepts similar to the concepts attached of the same ontology while using the Wu and Palmer[3] measure.
5. The use of WordNet, to determine the derivative terms of each term frequently used in the Web document (HTML or XML) that was applied TreeTagger[4] and then incorporate these terms results in the index.

[1] TF-IDF: *T*erm *F*requency - *I*nverse *D*ocument *F*requency.

[2] TF-ITDF: *T*erm *F*requency - *I*nverse *T*ag and *D*ocument *F*requency.

[3] Is a measurement technique based similarity arcs.

[4] TreeTagger is a tool which makes it possible to annotate a text with information on the parts of speech (kind of words: nouns, verbs, infinitives and particles) and of information of lemmatization.

IV. Our Approach

In this section, we outline our semantic indexing approach for information extraction from unstructured and semi structured web documents while using domain ontology. The proposed approach that takes into account the structure and content of HTML pages or XML documents includes the following phases (see Fig. 1):

1. Modeling of the HTML page or the XML document,
2. Attachment of concepts using domain ontology (for validation),
3. Enrichment these concepts.

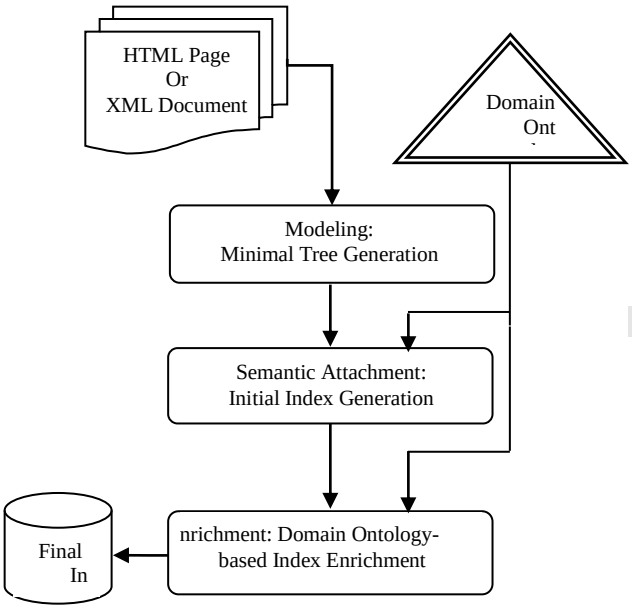

Fig. 1. General indexing approach

A. Modeling: Minimal tree generation

During this phase, first we model the HTML pages or the XML documents by using our own parser and thereafter by extracting the structure given by the tags in these two sources of information; secondary, we represent the HTML or XML structure with a labeled tree in which each element (or attribute) corresponds to a node of tree, and at the end, we generate the minimal tree structure found by eliminating the redundant ways where each semantic unit represents an information unit (single way).

The main steps of the minimal tree generation are described in Fig. 2.

Hereinafter the algorithm of the minimal tree generation:

Input: HTML Page or XML Document.
Output: Minimal Tree.

Load WEB file /*.HTML or *.XML */
To parser the document
/* creating the list chained ***LIST1*** and */
/* the labeled tree ***TREE1*** */
Pointer on beginning file
WHILE NOT END OF File DO
Fill ***LIST1*** by the TAGS
Creating the ***TREE1***
Each NODE represents a TAG
Next in the File
END WHILE
/* End of the creation of the labeled tree */
I Pointer on the ROOT of ***TREE1***
WHILE $TREE1_I \neq$ NIL-1 DO
J Pointer on the I+1 of ***TREE1***
WHILE $TREE1_J \neq$ NIL DO
/* eliminating the redundant ways */
IF $TREE1_I = TREE1_J$ THEN to delete $TREE1_J$
Next $TREE1_J$
END WHILE
Next in $TREE1_I$
END WHILE
/* Load ***TREE1*** */
Pointer on the ROOT of ***TREE1***
WHILE ***TREE1*** ≠ NIL DO
Fill the table ***TABLE1*** by nodes of the tree
Next in ***TREE1***
END WHILE

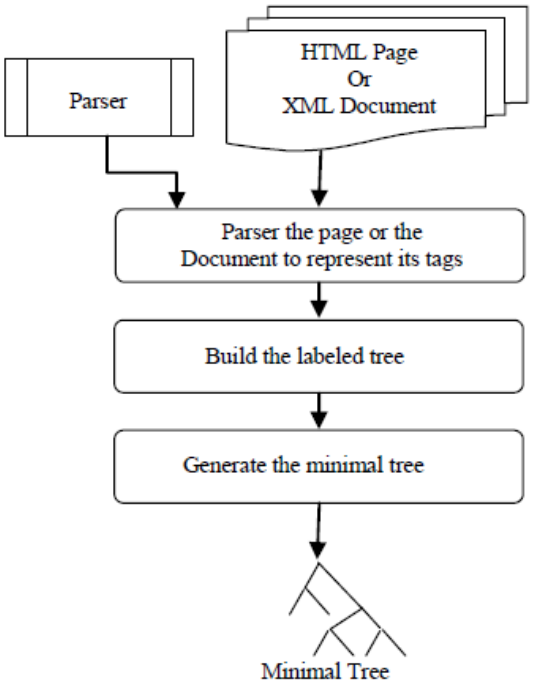

Fig. 2. Minimal tree generation.

While applying the minimal tree generation algorithm (as input: XML document, see Fig. 3), we obtain the labeled tree and the minimal tree represented respectively by Fig. 4 and Fig. 5.

```
<Sector Name = "TOURISM">
  <S-Sector> Hosting </S-Sector>
  <Residences>
      <Residence> Hotel </Residence>
      <Residence> Inn </Residence>
      <Cells>
        <Cell> Hotel Rooms </Cell>
        <Cell> Cabins </Cell>
      </Cells>
  </Residences>
  <Localities>
     <Locality> Béchar (Algeria) </Locality>
     <Locality> Oran (Algeria) </Locality>
  </Localities>
  <Cells>
    <Cell>Tourist residences </Cell>
  </Cells>
  </Sector>
```

Fig.3. XML document.

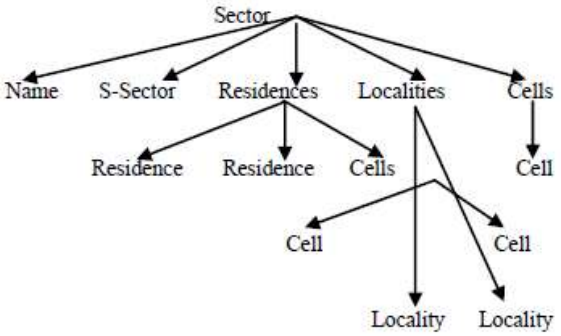

Fig.4. Labeled tree.

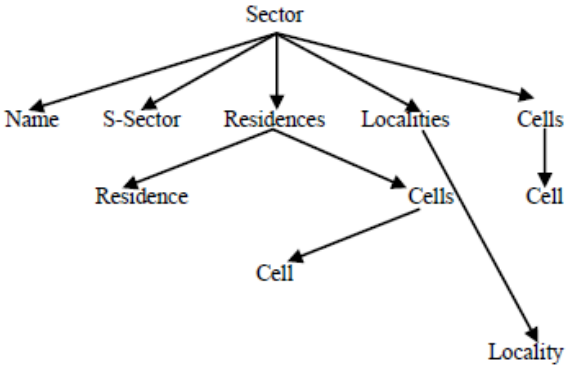

Fig.5. Minimal tree.

B. Semantic attachment for initial index generation

During this phase, an initial index is generated by attaching terms from the minimal tree with concepts of domain ontology. Node of each single path, known as information unit or semantic unit, is attached with the concept of the ontology to which it refers by calculating the semantic distance between terms and ontological concepts. The semantic attachment is achieved by semantic based similarity measure that explores the semantic meanings of the word constituents by using external resources like WordNet lexical database. While performing this attachment, semantically similar structures with different labels can be found.

Then, we continue by integrating the terms in the index. We finish this phase by the enrichment these terms by others (WordNet results) and we connect them again with the concepts of the minimal tree to integrate them into the index. The initial index generation phase is described in Fig. 6.

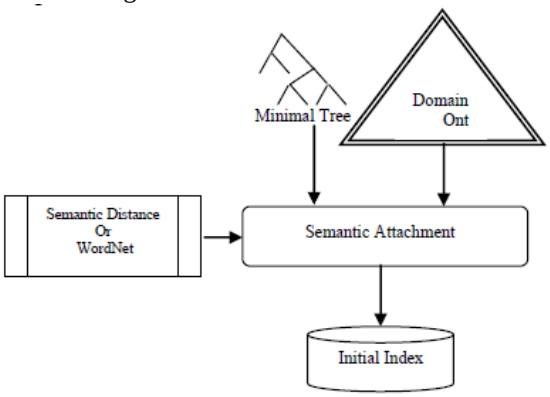

Fig.6. Initial index generation phase.

Below the algorithm of the initial index generation:

Input: Domain Ontology File + Minimal Tree.
Output: Initial Index.

```
/* Load ONTOLOGY file [*.OWL] in the same way*/
                /* domain that HTML or XML */
Pointer on beginning file
WHILE NOT END OF File DO
  Fill the table TABLE2 by the ontology concepts
  Next in file
END WHILE
FOR i=1 TO n                /* n is the size of TABLE1 */
  FOR j=1 TO m              /* m is the size of TABLE2 */
       /* creating the list chained LIST_INDEX */
    Call WordNet
    IF TABLE1 [i] ≈ TABLE2 [j] THEN
      Fill LIST_INDEX with the elements of TABLE1 [i]
          /* Using WordNet */
  END FOR
END FOR
```

Fig.7. Semantic unit terms attachment with domain ontology.

Fig. 7 show a semantic unit terms attachment with domain ontology. For this purpose we use the tourism ontology[5] that is a tutorial for the Semantic Web.

C. Domain ontology–based index enrichment

By using the tagger TreeTagger, we can produce the part of speech and lemma for each frequent term of the semantic unit result of the two weighted frequencies calculation of the terms (number of occurrences of the term in the semantic unit, number of occurrence of that term in the HTML page or XML document). This calculation allows us to select other terms results of the TreeTagger (selecting a few parts of speech: nouns, verbs, adjectives and extracting the lemmatized terms [30]; [31], removing long terms and reversing term variants -

[5] http://protege.stanford.edu/plugins/owl/owl-library/travel.owl

filtering and normalization-) and to integrate them in the index.

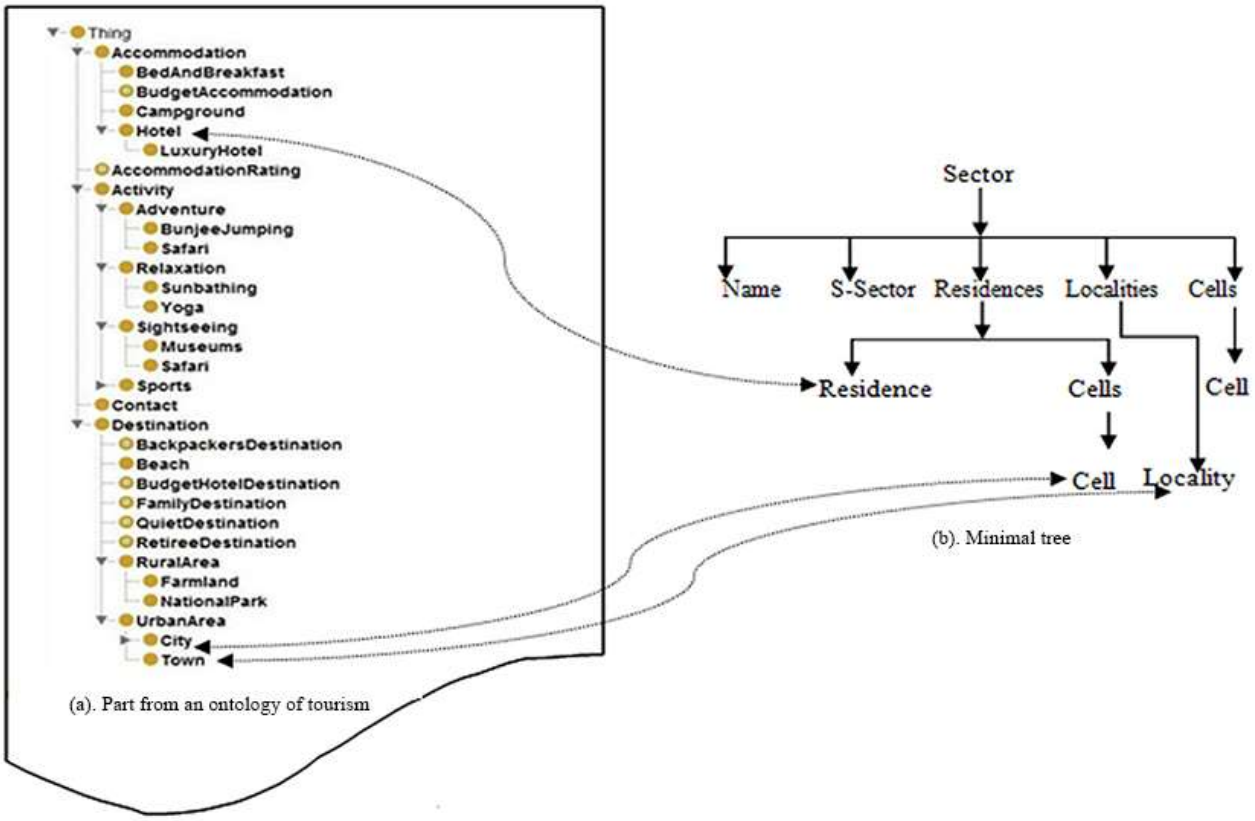

Fig. 7. Semantic unit terms attachment with domain ontology.

At the end, to calculate the similarities between the concepts relating to some terms with others co-occurring in the same semantic unit, to enrich the frequencies of words with their similarities and to integrate again the concepts of the ontology those are not attached in the index and that are semantically similar to the others concepts attached of the same ontology (While using the similarity measure of Wu and Palmer [32]).

By using WordNet, we can determine the derivative terms of each term frequently used in the Web document (HTML or XML) that was applied TreeTagger and then incorporate these terms results in the index.

The main steps of the domain ontology–based index enrichment are showed in Fig. 8.

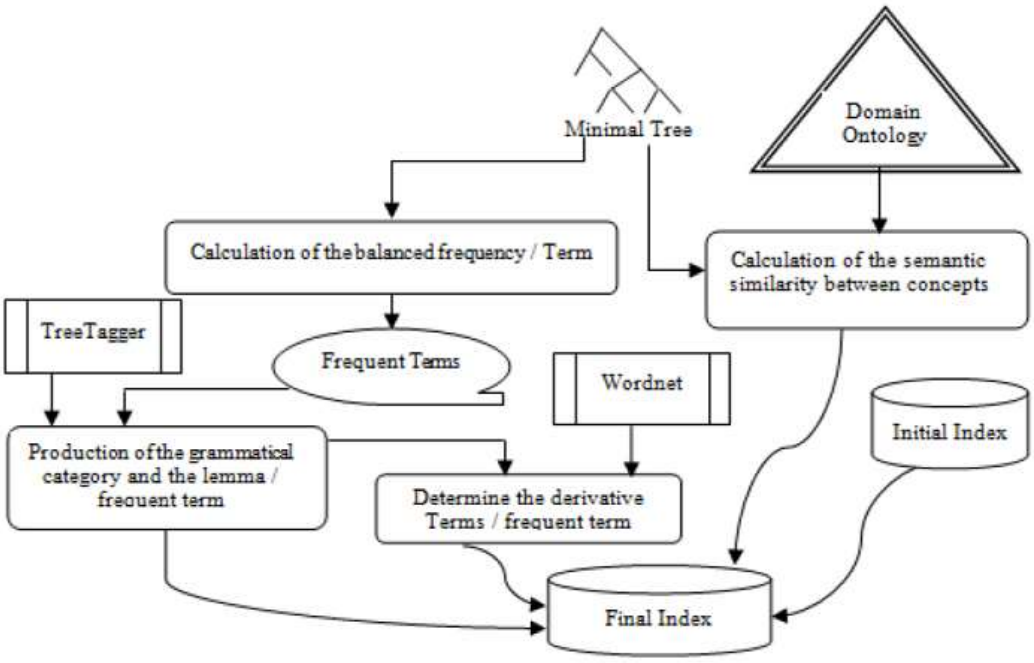

Fig.8. Domain ontology-based index enrichment.

Hereinafter the algorithm of the Final Index generation:

Input: Initial Index + Minimal Tree + Domain Ontology.
Output: Final Index.

```
FOR j=1 TO m-1   /* m is the size of TABLE2 */
  IF (SimWP (TABLE2 [j], TABLE2 [j+1]) ≈ X) AND
        (TABLE2 [j] ≈ element of LIST_INDEX) THEN
    /* SimWP() is a Calculation function of the Wu and
    */
    /* Palmer similarity measure, X → (tends toward) 1 */
    To fill LIST_INDEX with elements of TABLE2 [j+1]
  ENDIF
END FOR
WHILE  LIST_INDEX ≠ NIL DO
  Call TREETAGGER
  Creating the list chained LIST2
  /* LIST2 contained the grammatical category and the
        */
 /* lemma of each element of  LIST_INDEX  */
 Next in LIST_INDEX
END WHILE
WHILE LIST2 ≠ NIL DO
/* TF= Number of occurences for an element LIST2 in */
        /* HTML-page or XML-doc, IDF= Number of
*/
    /*occurences for an element LIST2 in semantic unit */
  IF (TF*IDF≈ 1) THEN          /* ≈1: tends toward 1 */
      /* To add to the list chained LIST_INDEX */
  FOR j=1 TO m          /* m is the size of TABLE2 */
    Call WordNet
    IF element of LIST2 ≈ TABLE2 [j] THEN
        Fill LIST_INDEX with element of LIST2
          /* Using WordNet * /
      ENDIF
  END FOR
  ENDIF
  Next in LIST2
END WHILE
```

While applying the final index generation algorithm on the minimal tree represented by Fig. 5, and using the tourism ontology, the content of the final index becomes:

{Residence, Locality, Hotel, City, Town, Cities, Hotels, Urban Area, Destination, Rural Area, …}

V. Implement and Evaluation

We provide in this section, an implementation of the proposed semantic indexing approach. For empirical evaluation, we developed an EMBARCADERO DELPHI 2010 based tool that implements all features presented above.

In the following, we will show the different screenshots that allow the description of the different phases to perform in order create a containing index of the concepts extracted from an XML document or an HTML document, enriched in the same way by other domain ontology concepts.

The Fig. 9 shows that from our application, we can:

(a) Loading the XML file for modeling the document,
(b) Generating the labeled tree,
(c) Deduct the minimal tree for the XML document,
(d) Display of the concepts extracted of the XML document after filtering.

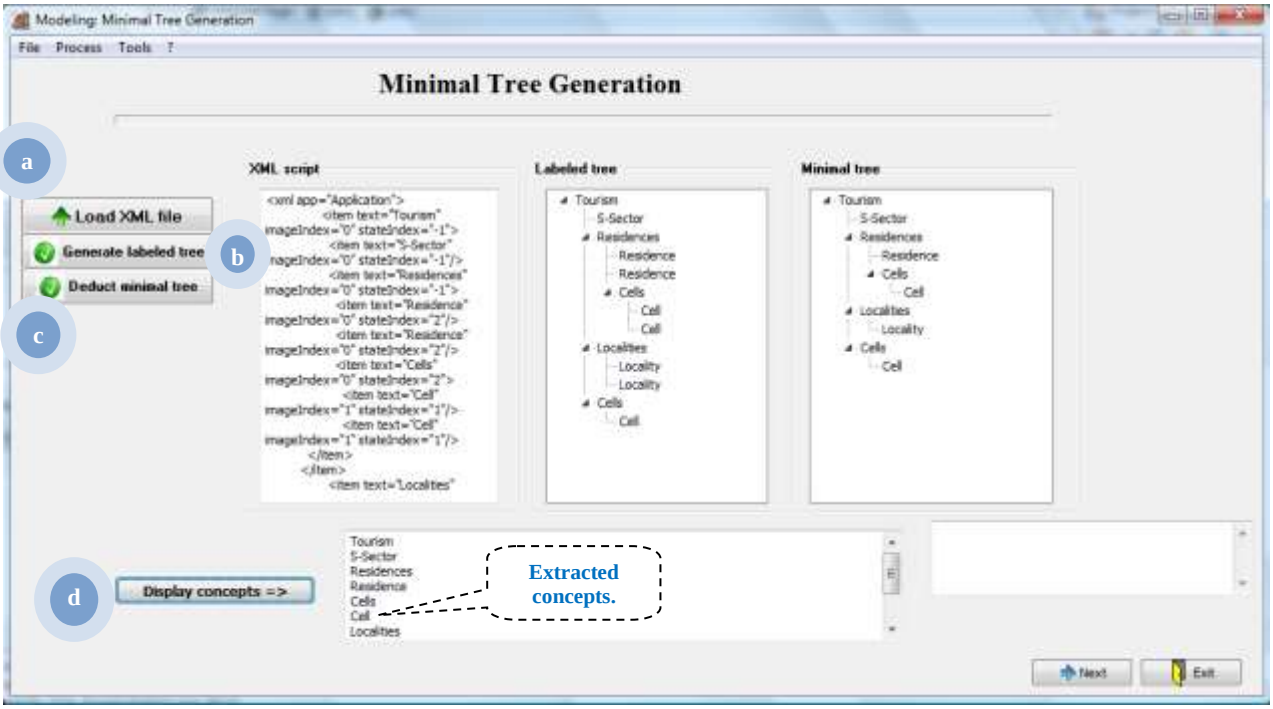

Fig.9. Extracting concepts from an XML document.

- Attachment operation between the concepts extracted of the XML document and those of the ontology.
- Creating of the initial index containing the ontology connected concepts with the possibility to save this index. (See Fig. 10, phases e then f).

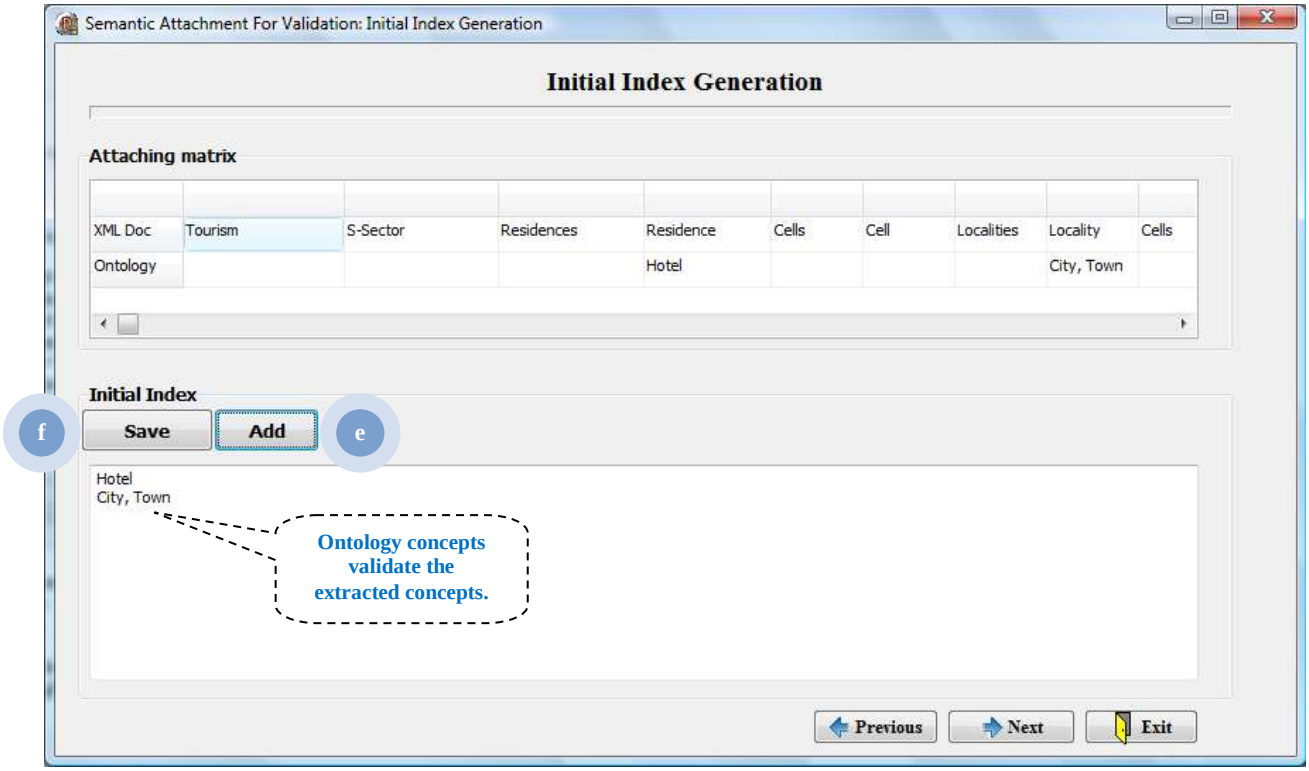

Fig. 10. Creation of the first index

- The enrichment index by other ontology concepts that are semantically near to the concepts extracted from the XML document while using the WordNet tool. (See Fig. 11, phase g).
- Update the index and save it another time. (See Fig. 11, phases h then i).
- The same process is followed in our application, but as a source of information an HTML page.
- The Fig. 12 shows the contents of the final index (Case: HTML page).

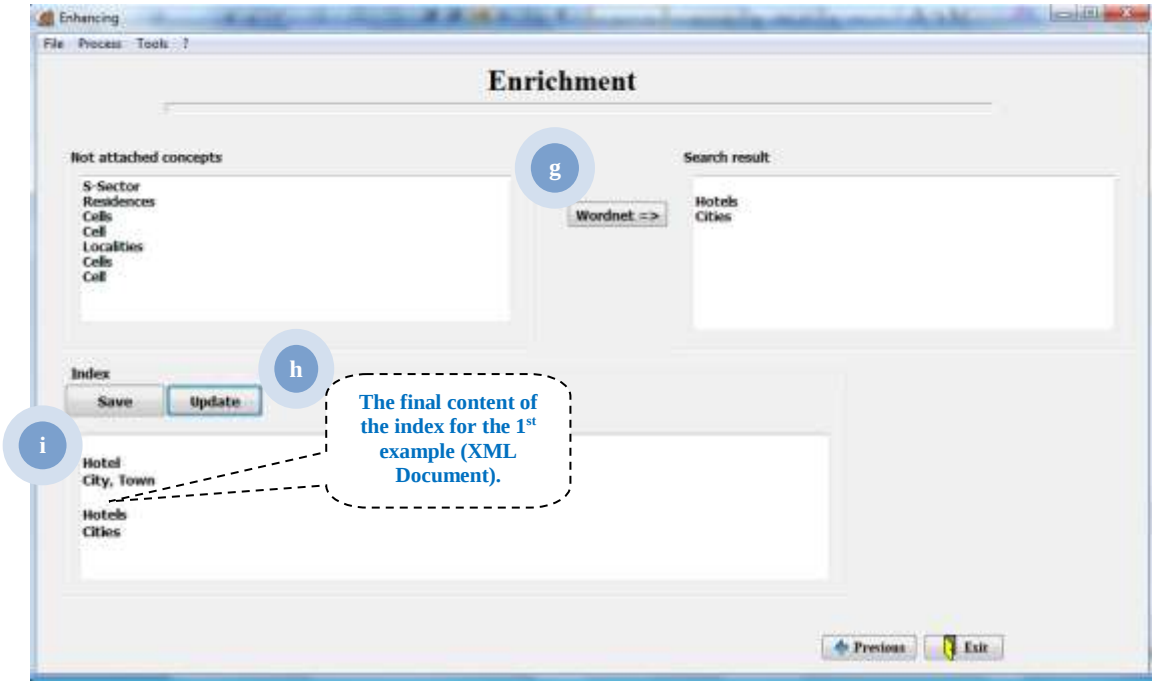

Fig. 11. Creation of the final index (XML document).

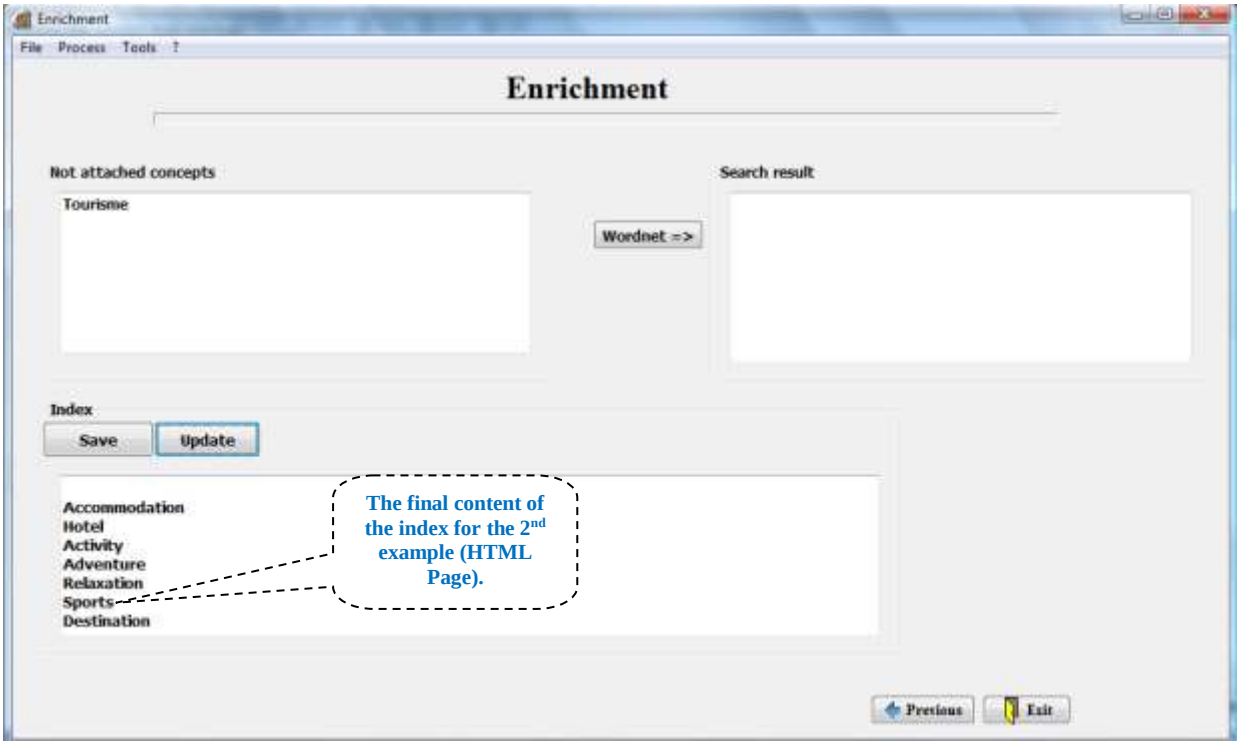

Fig. 12. Creation of the final index (HTML page).

Now, we conduct a set of experiments to illustrate the effectiveness of our approach. We perform a semantic indexing for six web documents taken from tourism domain (Three XML documents and three HTML pages). The semantic attachment and enrichment phases are performed while using the tutorial ontology for a Semantic Web of tourism[6].

Table 1. The content of the index in growth.

Size of the Index (Number of concepts)			
1st Experiment (XML files)			
Example No.	After Modeling	After Semantic Attachment	After Enrichment
1	--	**05**	**10**
2	--	**04**	**07**
3	--	**04**	**09**
2nd Experiment (HTML files)			
Example No.	After Modeling	After Semantic Attachment	After Enrichment
1	--	**07**	**11**
2	--	**06**	**08**
3	--	**04**	**07**

While reading graphs results (Fig. 13 and Fig. 14), we can deduct that the content of the index increase more and more while:

1. Executing the different phases of this approach successively,
2. Doing a better extraction of concepts from a XML document or a HTML page,
3. Using an ontology rich of concepts (in the same way domain that the XML document or the HTML page),
4. Using one of the semantic similarity measures based on the arcs between concepts of a same ontology (Enrichment phase),
5. Using the semantic distance in the same way between a concept of a XML document or a HTML page with another of an ontology domain (Attachment phase).

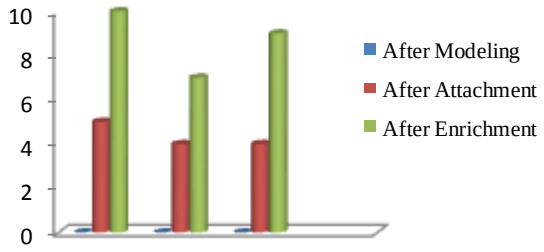

Fig. 13. First experiment results.

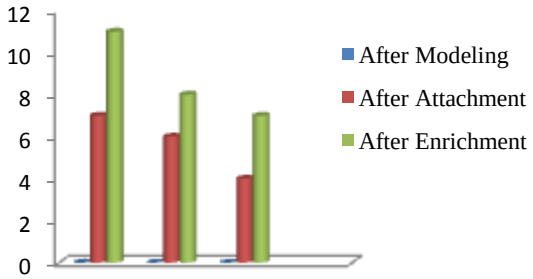

Fig. 14. Second experiment results.

[6] http://protege.cim3.net/file/pub/ontologies/travel/travel.owl

VI. Conclusion and Future Work

The indexing consists in constructing a structure of access to the documents that will facilitate the phase of research beforehand. The ontologies showed their efficiency in information research and their utility saw itself confirmed by the semantic web. The ontology permits to refine the results.

In this work, we presented a semantic indexing approach of HTML pages or XML documents in order to have a better extraction tool of the concepts from these two web information sources while using domain ontology. This phase of extraction (among others) is the beginning of a reverse engineering of web-oriented application; it permits at the end, a better reengineering of these applications. The relevance of this approach is also increases while using it in the two other phases that are attachment (identification) and the enrichment of these concepts descended of the extraction phase.

These encouraging results are stimulating a number of further researches to extend the current approach. First, we intend to update the Wu and Palmer measurement (used in the enrichment phase) of which we noticed that it gives the priority to the concepts brothers that to the concepts father-sons of a hierarchical ontology. What is to our sense, inadequate in the information research domain where, it is necessary to bring back all sons of a concept (i.e request) before its neighbors? Second, to enhance the initial index generation, we plan to propose a semantic distance calculation algorithm instead of using WordNet. Finally, future research will work towards considering the obtained result to improve the web applications.re-engineering process.

References

[1] P. Tramontana, ''Reverse engineering web applications'', in IEEE (Ed.), in Proceedings 21st International Conference on Software Maintenance (ICSM05), pp. 705–708, Budapest, Hungary, 2005.

[2] F. Ricca and P. Tonella, ''Using clustering to support the migration from static to dynamic web pages'', in Proceedings of the 11th International Workshop on Program Comprehension, pp. 207–216, Portland Oregon, USA, 2003.

[3] F. Estivenart, A. Franois, J. Henrard and J. Hainaut, ''A tool-supported method to extract data and schema from web sites'', in Proceedings of the 5th International Workshop on Web Site Evolution, pp. 3–11, Amsterdam, Netherlands, 2003.

[4] L. Paganelli and F. Paterno, ''Automatic reconstruction of the underlying interaction design of web applications'', in A. Press (Ed.), in Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering, pp. 439–445, Ishia Italy, 2002.

[5] Y. Gaeremynck, L. Bergman and A. Lau, ''More for less: Model recovery from visual interfaces for multi-device application design'', in A. Press (Ed.), in Proceedings of the International Conference on Intelligent User Interfaces, pp. 69–76, Miami Florida, USA, 2003.

[6] G. D. Lucca, A. Fasolino, F. Pace, P. Tramontana and U. D. Carlini, ''Ware: a tool for the reverse engineering of web applications'', in Proceedings of the 6th European Conference on Software Maintenance and Reengineering (CSMR2002), pp. 02-41, Budapest, Hungary, 2002.

[7] C. Bellettini, A. Marchetto and A. Trentini, ''Webuml: Reverse engineering of web applications'', in 19th ACM Symposium on Applied Computing (SAC 2004), pp. 1662–1669, Nicosia, Cyprus, 2004.

[8] P.A Gomez and D. Rojas Amaya, ''Ontological reengineering for reuse'', Fensel D. and Studer R., Eds., 11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW-99), Vol. 1621 of LNAI, pp. 26–29, Berlin, Germany, 1999 (Springer, pp. 139–156).

[9] F. Frédéric, ''L'ingénierie ontologique'', Institute for Research in Computer of Nantes France, Research Report No. 02-07, Oct. 2002.

[10] F. Gandon, ''Ontology Engineering: a survey and a return on experience'', Research Report No. 4396, INRIA, 2002.

[11] B. Peterson, W. Andersen and J. Engel, ''Knowledge bus: Generating application focused databases from large ontologies'', in Proceedings of the 5th KRDB Workshop, Seattle, WA, 1998.

[12] J. Conesa and A. Olive, ''Pruning ontologies in the development of conceptual schemas of information systems'', in ER'2004, LNCS 3288, pp. 122–135, 2004.

[13] H. El-Ghalayini, M. Odeh and R. McClatchey, ''Deriving conceptual data models from domain ontologies for bioinformatics'', in the 2nd International Conference on Information and Communication Technologies from Theory to Application ICTTA, 2006.

[14] O. Vasilecas and D. Bugaite, ''An algorithm for the automatic transformation of ontology axioms into a rule model'', in Proceedings of the 2007 International Conference on Computer Systems and Technologies (CompSysTech '07), pp. 1–6, Bulgaria, 2007.

[15] V. Jain and M. Singh, "Ontology based information retrieval in semantic web: A survey", International Journal of Information Technology and Computer Science (IJITCS), pp. 62-69, 2013.

[16] S. Chagheri, C. Roussey, S. Calabretto and C. Dumoulin, ''Semantic indexing of technical documentation'', LIRIS 2009.

[17] C. Roussey, S. Calabretto and J. M Pinon, ''Etat de l'art en indexation et recherche d'information'', Digital document, special issue : Gestion des documents et gestion des connaissances, Vol. 3, No. 3-4, pp. 121-150, Dec. 1999.

[18] J. Y Nie, ''Le domaine de la recherche d'information, survol d'une longue histoire'' in Gaussier (E.), Stefanini (M-H.), Intelligent search assistance information, Treaty Collection Science and Information Technology, pp.19-28, Lavoisier, Paris, 2003.

[19] C. Fluher, ''Le traitement du langage naturel dans la recherche d'information'', in Intelligent interface for Scientific and Technical Information ; Klingenthal : INRIA, pp. 103-130, 1992.

[20] P.D Pomart and E. Sutter, ''Indexation'', Article of the Encyclopedic Dictionary of Information and Documentation, Paris, Nathan pp. 284-287, 1997.

[21] M. Hadj Henni, ''Approche ontologique pour la modélisation sémantique, l'indexation et l'interrogation des documents Coraniques'', Computer science memory schoolmaster, School of Computer Science, Oued-Smar, Algeria, 2009.

[22] J. Maniez, ''Actualité des langages documentaires, Fondements théoriques de la recherche d'information'', ABDS Paris Ed., 2002.

[23] P. Lefevre, ''La recherche d'information du texte intégral au thésaurus'', Paris Hermès Ed., pp. 253, 2000.
[24] W. Mustafa El Hadi, ''Indexation humaine et indexation automatisée : la place du terme et des environnements'', 7th Science Meeting AUF - LTT "Words, terms and contexts", Brussels, Belgium, 2005.
[25] EM. El-Hachani, ''Indexation des documents multilingues d'actualité incluant l'arabe : équivalence interlangues et gestion des connaissances chez les indexeurs'', PhD Thesis, University of Lyon 2, France, Nov. 14, 2005.
[26] R. Wilkinson, ''Effective retrieval of structured documents''. (S.-V. New York, Ed.) pp. 311 – 317, 1994.
[27] Y. Mass, ''Component ranking and automatic query refinement for XML retrieval'', INEX 2004, pp. 134–140.
[28] L.R. Khan, ''Retrieval effectiveness of an ontology-based model for information selection'', International Journal on Very Large Data Bases (IJVLDB), vol. 13, pp. 71–85, 2004.
[29] H. Zargayouna and S. Salotti, ''Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML'' in Francophones Days of Knowledge Engineering, Lyon, France, 2004.
[30] M. Volk, B. Ripplinger and S. Vintar, ''Semantic annotation for concept-based cross-language medical information retrieval'' in International Journal of Medical Informatics, Vol. 67 pp. 1-3, Dec. 2002.
[31] M. Volk, S. Vintar and P. Buitelaar, ''Ontologies in cross-language information retrieval'', in Proceedings of 2nd Conference on Professional Knowledge Management, Lucerne, Switzerland, 2003.
[32] Z. Wu and M. Palmer, ''Verb semantics and lexical selection'', in Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, pp. 133-138, 1994.

Authors' Profiles

Abdeslem DENNAI is a PhD student in fifth year Computer Science, University of Sidi Bel Abbes, Algeria. In 1994, he received the diploma of engineering in Computer Science from the University of Sidi Bel Abbes, Algeria. In 2008, he received the diploma of teaching in Computer Science from the University of Bechar, Algeria. He is a lecturer at the University of Bechar, Algeria. His research interests are in the field of semantic web, web applications and ontology.

Sidi Mohamed Benslimane is an Associate Professor at the Computer Science Department of Sidi Bel Abbes University, Algeria. He received his PhD degree in computer science from Sidi Bel Abbes University in 2007. He also received a M.S. and a technical engineer degree in computer science in 2001 and 1994 respectively from the Computer Science Department of Sidi Bel Abbes University, Algeria. He is currently Head of Research Team 'Service Oriented Computing' at the Evolutionary Engineering and Distributed Information Systems Laboratory (EEDIS). His research interests include, semantic web, service oriented computing, ontology engineering, information and knowledge management, distributed and heterogeneous information systems and context-aware computing.

How to cite this paper: Abdeslem DENNAI, Sidi Mohammed BENSLIMANE,"Semantic Indexing of Web Documents Based on Domain Ontology", International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.2, pp.1-11, 2015. DOI: 10.5815/ijitcs.2015.02.01