Building a Semantic Index from HTML Pages or XML Documents

Abdeslem DENNAI[1], Sdi Mohammed BENSLIMANE[2]
[1, 2]EEDIS Laboratory University of SIDI BEL ABBES, ALGERIA,
[1]De_selam@yahoo.fr, [2]Benslimane@univ-sba.dz

ABSTRACT

Among the phases of reverse engineering of web-oriented applications is the extraction of concepts hidden in HTML pages or marked in XML documents. In this article, we propose an approach to index semantically these two sources of information using on the one hand, domain ontology to validate the extracted concepts and on the other hand the similarity measure 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 validated by examples. The obtained results lead to better re-engineering of web applications and subsequently a distinguished improvement in the web structuring.

Keywords

Reverse engineering, ontology, semantic distance, semantic indexing, semantic web.

1 INTRODUCTION

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 and 7].

On the other hand, several researchers [8, 9 and 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, and 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 and in the 5th section, we give an interpretation of the results gotten before concluding in Section 6.

2 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, that 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. This while making evolve the indexing techniques based on thesauruses toward techniques that use the knowledge representation and the Artificial intelligence.

2.1 HTML Pages

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.

2.2 XML Documents

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 [22].

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).

3 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 [19] 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 [22].

Yosi Mass in [20] 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 [22].

Khan in [21] 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 [22].

Zargayouna and Salotti in [15] 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.

Samaneh and all in [22] propose a semantic indexing model which exploits both the logical structures and the semantic contents of documents. This method is an extension

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

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

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.

4 OUR CONTRIBUTION

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.

[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.

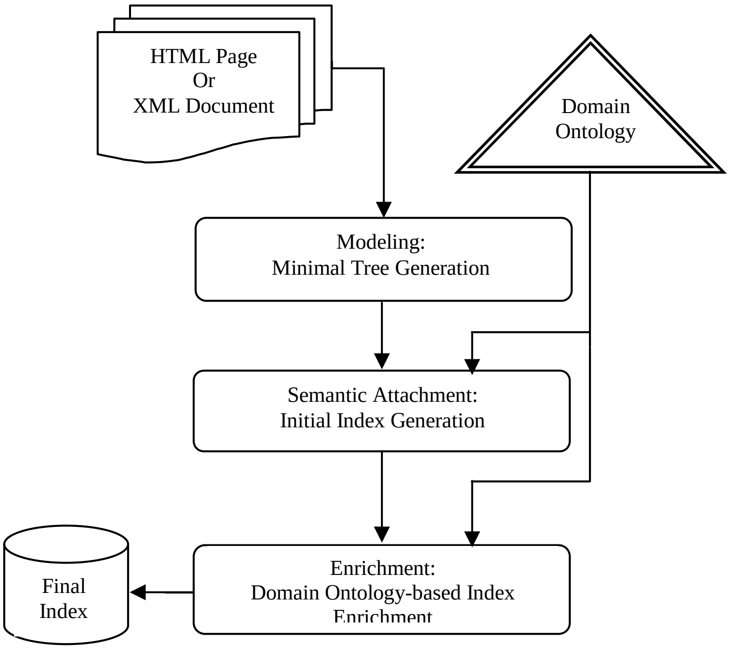


Fig.1. General Indexing Approach

4.1 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.

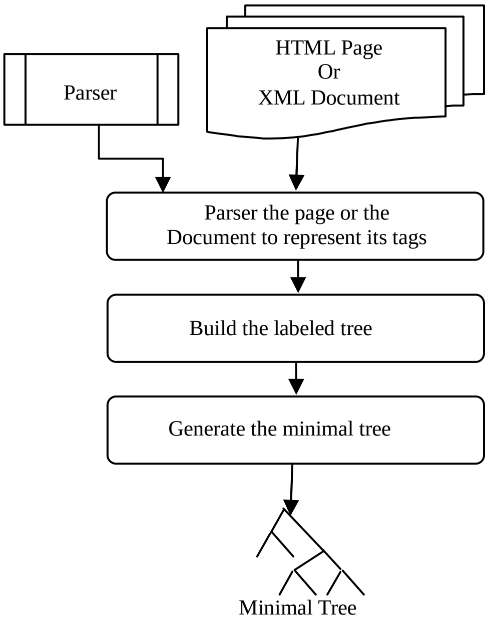
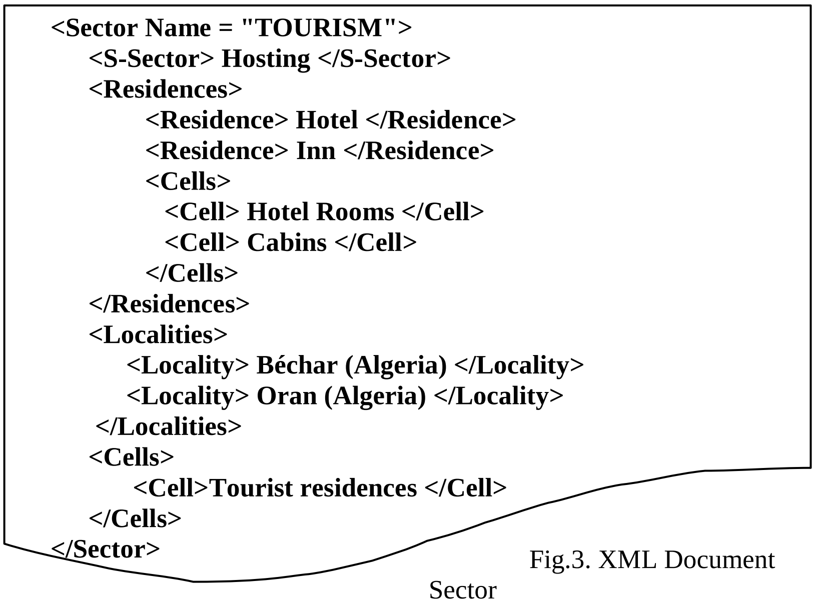


Fig.2. Minimal Tree Generation

While applying the minimal tree generation algorithm on the XML document (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

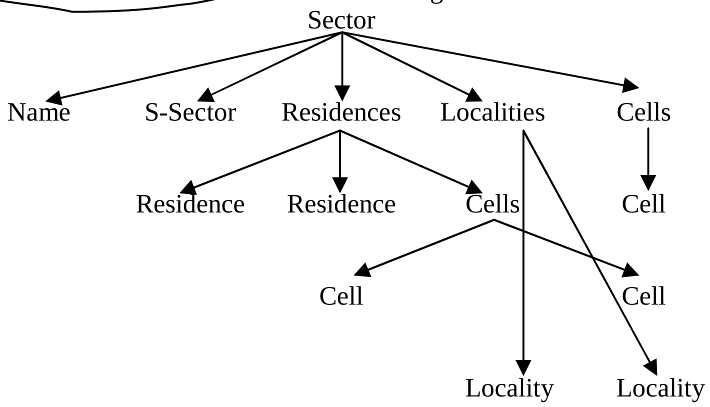


Fig.4. Labeled Tree

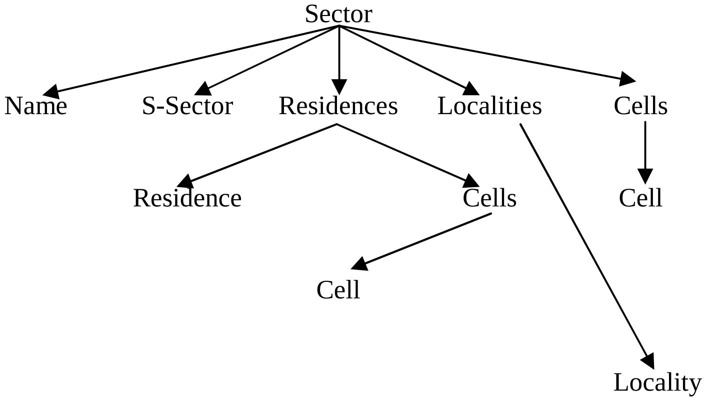


Fig.5. Minimal Tree

4.2 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 bv 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.

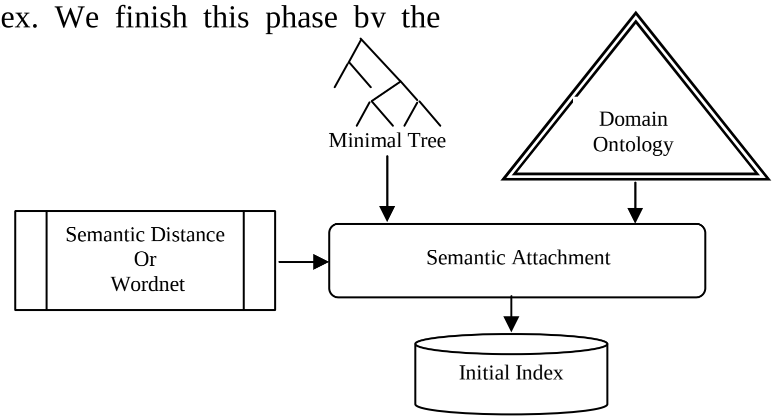
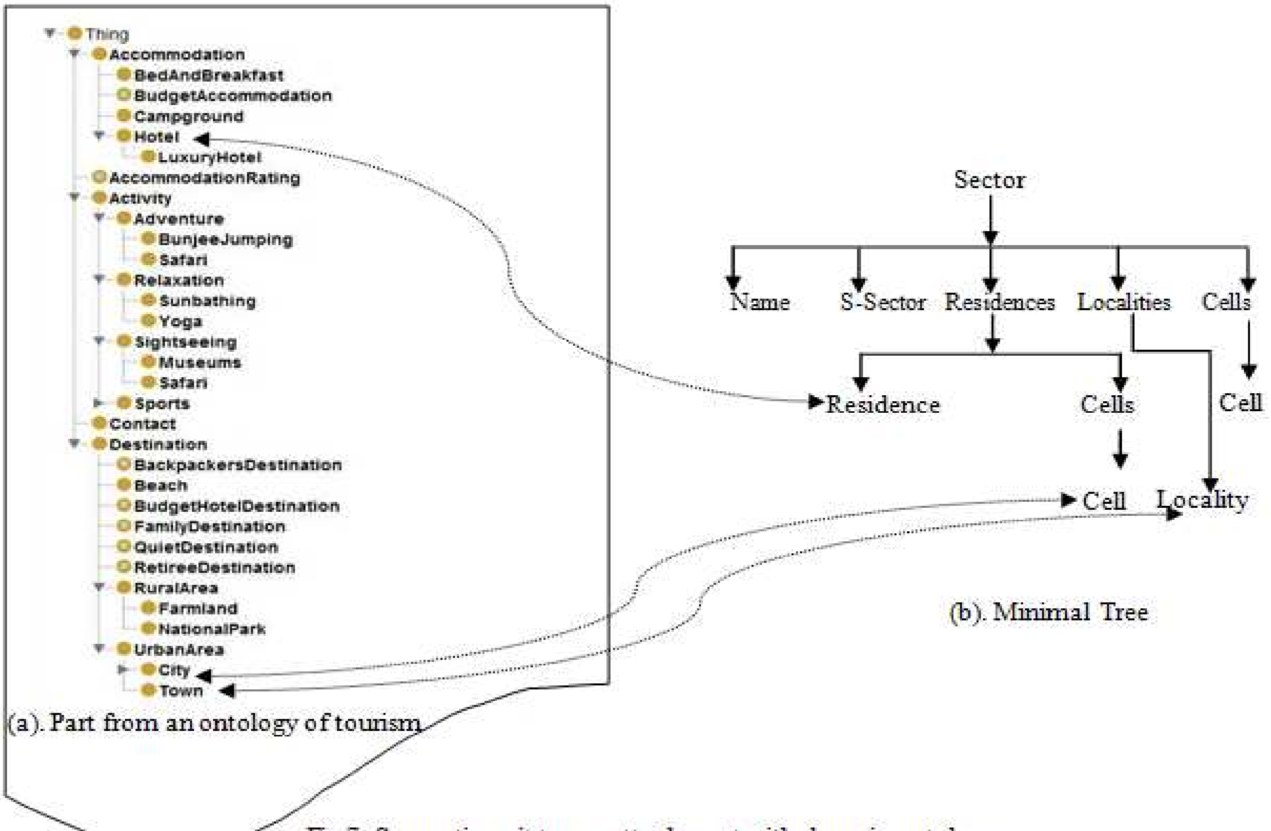


Fig.6. Initial Index Generation phase

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.



(a). Part from an ontology of tourism

(b). Minimal Tree

Fig.7. Semantic unit terms attachment with domain ontology

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

While applying the initial index generation phase on the minimal tree, the content of the initial index will be {*Residence*, *Locality*, *Hotel*, *City*, *Town*, …}

4.3 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 [25, 26], removing long terms and reversing term variants -filtering and normalization-) and to integrate them in the index. 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 measurement of Wu and Palmer [18]).

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.

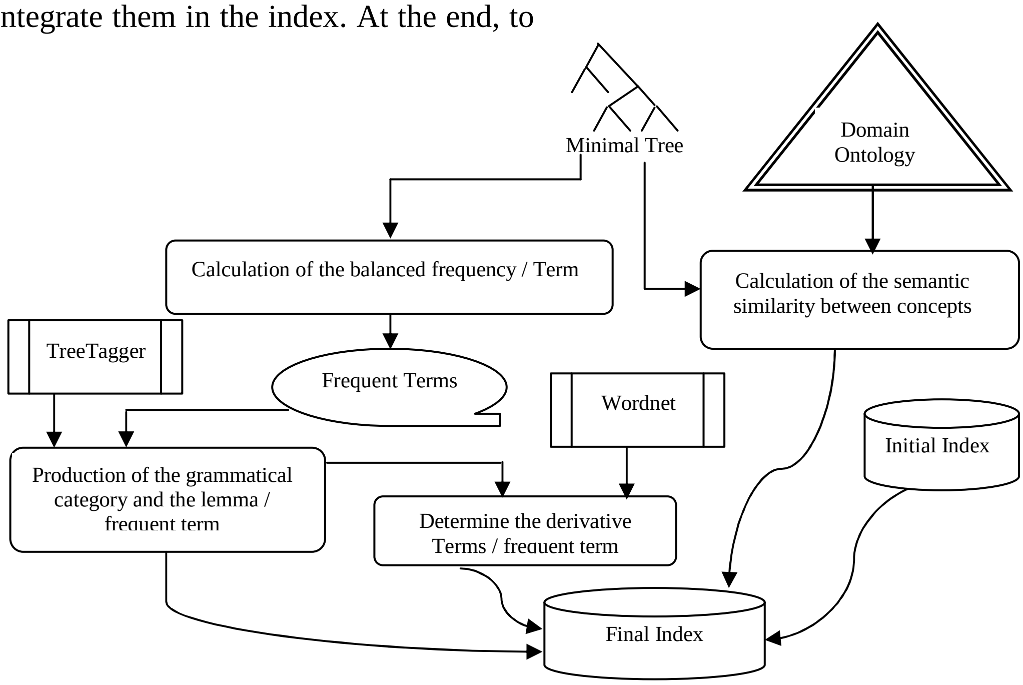

Minimal Tree
Domain Ontology
Calculation of the balanced frequency / Term
Calculation of the semantic similarity between concepts
TreeTagger
Frequent Terms
Wordnet
Initial Index
Production of the grammatical category and the lemma / frequent term
Determine the derivative Terms / frequent term
Final Index

Fig.8. Domain Ontology-Based Index Enrichment

While applying the final index generation phase on the minimal tree, and using the tourism ontology, the content of the final index becomes:
{Residence, Locality, Hotel, City, Town, Cities, Hotels, Urban Area, Destination, Rural Area, …}

5 INTERPRETATION OF THE RESULTS

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].

Table1. Size of the Index[7] (1st Experiment)

Example : XML Files	After Modeling	After Semantic Attachment	After Enrichment
1st	--	05	10
2nd	--	04	07
3th	--	04	09

Table2. Size of the Index (2nd Experiment)

Example : HTML Pages	After Modeling	After Semantic Attachment	After Enrichment
1st	--	07	11
2nd	--	06	08
3th	--	04	07

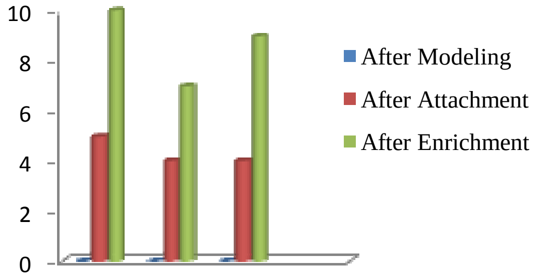


Fig.9. First experiment results

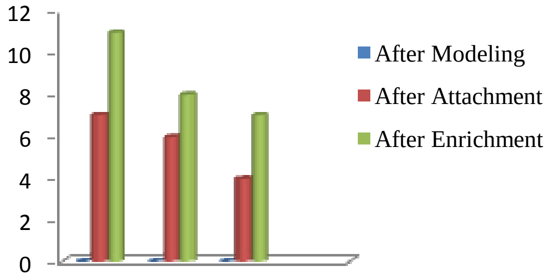


Fig.10. Second experiment results

While reading the two graphs results (Fig.16 and Fig.17), 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).

6 CONCLUSION

The indexing consists in constructing a structure of access to the documents that will facilitate the phase of research. 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 measure (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, this is inadequate in the information research domain. Second, to enhance the initial index generation, we plan to propose a semantic distance calculation algorithm instead of using Wordnet.

[6] ttp://protege.cim3.net/file/pub/ontologies/travel/travel.owl
[7] Number of concepts.

7 REFERENCES

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

[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*, Portland Oregon, USA, 2003, pp. 207–216.

[3] F. Estivenart, A. Franois, J. Henrard, 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*, Amsterdam, 2003, pp. 3–11.

[4] L. Paganelli, 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*, Ishia Italy, 2002, pp. 439–445.

[5] Y. Gaeremynck, L. Bergman, 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*, Miami Florida, USA, 2003, pp. 69–76.

[6] G. D. Lucca, A. Fasolino, F. Pace, P. Tramontana, 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)*, Budapest, 2002, pp. 02-41.

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

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

[9] F. Frédéric, ‘’L’ingénierie ontologique’’, Institut de Recherche en Informatique de Nantes, *Rapport de recherche No 02-07*, Octobre 2002.

[10] F. Gandon, ‘’Ontology Engineering: a survey and a return on experience’’, *Rapport de recherche n° 4396*, INRIA, 2002.

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

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

[13] H. El-Ghalayini, M. Odeh, 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, 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)*, Bulgaria, 2007, pp. 1–6.

[15] H. Zargayouna and S. Salotti, ‘’Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML’’ dans *Ingénierie des Connaissances*, Lyon, France, 2004.

[16] M. Volk, B. Ripplinger, S. Vintar ‘’Semantic annotation for concept-based cross-language medical information retrieval’’ in *International Journal of Medical Informatics*, Vol. 67 pp. 1-3, Déc 2002.

[17] M. Volk, S. Vintar and P. Buitelaar, ‘’Ontologies in cross-language information retrieval’’, in *Proceedings of 2nd Conference on Professional Knowledge Management*, Lucerne, 2003.

[18] 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.

[19] R. Wilkinson, ’’Effective retrieval of structured documents’’. (S.-V. New York, Ed.) pp. 311 – 317, 1994.

[20] Y. Mass, ‘’Component ranking and automatic query refinement for XML retrieval’’, *INEX* 2004, pp. 134–140.

[21] L.R. Khan, ‘’Retrieval effectiveness of an ontology-based model for information selection’’, pp. 71–85, 2004.

[22] S. Chagheri, C. Roussey, S. Calabretto, C. Dumoulin, ‘’Semantic indexing of technical documentation’’, *LIRIS* 2009.