# A PROTOTYPE FOR KNOWLEDGE EXTRACTION FROM SEMANTIC WEB BASED ON ONTOLOGICAL COMPONENTS CONSTRUCTION

Nesrine Ben Mustapha, Hajer Baazaoui-Zghal
*Riadi Laboratory., ENSI Campus Universitaire de la Manouba, 2010 Tunis, Tunisie*

Marie-Aude Aufaure
*Supelec, Computer Science Department, Plateau du Moulon, 91 192 Gif sur Yvette, France*

Keywords: Semantic web, ontology construction, knowledge extraction, OntoCoSemWeb (Ontology Construction for the Semantic Web) prototype.

Abstract: Adding a semantic dimension to web pages is a response to some problems of the present web and is known as the semantic web. Many methods and methodologies can be found in the literature. Generally, they are dedicated to particular data types like text, semi-structured data, relational data, etc. This paper presents a prototype for knowledge extraction from web pages based on ontological components construction. Our work deals with web pages. We will first study the state of the art of methodologies defined to learn ontologies from texts. Then, we will define architecture of ontological components for the Semantic web. An implementation and experimentation of the proposed architecture are presented.

## 1 INTRODUCTION

The volume of available information on the web is growing exponentially. Consequently, integration of heterogeneous data sources and information retrieval, have become more and more complex. Adding a semantic dimension to web pages is a response to this problem and is known as the semantic web (Berners-Lee, 2001). Ontologies can be seen as a fundamental part of the semantic web. They can be defined as an explicit, formal specification of a shared conceptualization (Gruber, 1993). Meanwhile, building ontology manually is a long and tedious task. We are interesting in learning ontologies from text. We present in section 2 semantic web and ontological components and our approach to build a domain ontology. In section 3 and 4, implementation and experimentation are presented. Section 5 analyses the results. At last, we conclude and give some perspectives for this work.

## 2 SEMANTIC WEB AND ONTOLOGICAL COMPONENTS

Starting from the state of the art, we propose a hybrid approach to build domain ontology; our objective is to increase the capability of this ontology to specify and extract web knowledge in order to contribute to the semantic web. Analyzing the web content is a difficult task relative to relevance, redundancies and incoherencies of web structures and information. For these reasons, proposing an approach to build automatically an ontology still remains utopian. Our approach is based on the cyclic relation between web mining, semantic web and ontology building as stated in (Berendt and al., 2002). Our proposal is based on the following statements: (1) satisfy the fact that the ontology is useful to specify and extract knowledge from the web, (2) link the semantic content within the web documents structure, and (3) combine linguistic and learning techniques taking into account the scalability and the evolution of the

Ben Mustapha N., Baazaoui-Zghal H. and Aufaure M. (2007).
A PROTOTYPE FOR KNOWLEDGE EXTRACTION FROM SEMANTIC WEB BASED ON ONTOLOGICAL COMPONENTS CONSTRUCTION.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 451-454
DOI: 10.5220/0001285304510454

ontology. Our ontology is produced using web mining techniques. We mainly focus on web content and web structure mining. Building this ontology leads us to solve two main problems. The first one is relative to the heterogeneity of web documents structure while the second one is more technical and concerns technical choices to extract concepts, relationships and axioms as well as the selection of learning sources and scalability. An architecture of ontological components is proposed to represent the domain knowledge, the web sites structure and a set of services. These ontological components are integrated into a customizable ontology building environment (Ben Mustapha and al., 2006).

## 2.1 Our Approach

Learning ontologies from web sites is more complex than texts. Indeed, web pages can contain more images, hypertext and frames than text. Learning concepts is a task that requires texts able to explicitly specify the properties of a particular domain. Starting from the state of the art, we can say that no learning method to extract concepts and relationships is better. For these reasons, we propose a customizable ontology building environment taking into consideration the criteria defined in our synthesis. In this environment, we propose a set of interdependent ontologies to build a knowledge base on a particular domain, made up of a set of web documents, their structure and associated services. We distinguish three ontologies, namely a generic ontology of web sites structures, a domain ontology and a service ontology. The generic ontology of web sites structure contains a set of concepts and relationships allowing a common structure description of HTML, XML and DTD web pages. This ontology enables users to learn axioms that specify the semantic of web documents patterns. The main objective is to ease the structure of web mining knowing that the results can help to populate the domain ontology. The domain ontology is divided into three layers according to their level of abstraction. The ontology of services is defined starting from the concept of task ontology (Gomez-Perez and al., 2003). In our web context, we speak of web services instead of tasks. This ontology specifies the domain services and will be useful to map web knowledge into a set of interdependent services. This ontology is hierarchically structured: the upper level is the root service while the leaves are elementary tasks for which a triplet "concept-relation-concept" belonging to the domain ontology is associated. These three ontological components are interdependent where the axioms included in an ontology are used to enhance another ontology component. Meanwhile, these ontologies differ from their use. The domain ontology is used to specify the domain knowledge. The service ontology specifies the common services that can be solicited by web users and can be attached to several ontologies defined on subparts of the domain. As we said previously, the axioms of the structure ontology are used to extract instances of the domain ontology.

## 2.2 Building the Domain Ontology

In this section, we focus on the domain ontology extraction. Our strategy is based on three steps. The first one is the initialization step. The second one is an incremental learning process based on linguistic and statistic techniques. The last one is a learning step based on web structure mining. Here is their definition. The initialization is based on the following steps: (1) The design and manual building of a minimal ontology related to the domain; this construction is based on concepts and relationships of Wordnet, (2) Composition of concepts and relationships learning sources which consist in: (1) Web search of documents related to our domain using the concepts defined in the minimal ontology as requests, (2) Classification of these web documents, (3) Composition of a textual corpus containing a set of phrases in which we can find at least one concept of the minimal domain ontology and (4) Composition of a corpus of HTML and XML documents indexed by their URL. Each iteration of the second stage includes two steps. The first one (Procedure A) is defined by the following tasks: (1) Enrichment of the ontology with new concepts extracted from semi-structured data found in the web pages (XML, DTD, tables), (2) Construction of a word space based on the concepts of the minimal domain ontology, (3) Lexico-syntactic patterns learning based on the method defined in (Alfonseca and Manandhar, 2002); these patterns are related to non taxonomic relationships between the concepts of the minimal ontology, (4) Lexico-syntactic patterns learning to extract synonymy, hyponymy and part-of relationships (lexical layer of the domain ontology), (5) Similarity matrix building: this matrix allows computing the similarity between pairs of concepts found in the multidimensional space word. The second step (Procedure B) consists in: (1) Updating the textual corpus and the web documents collection by searching them according to the concepts defined in the minimal ontology, (2) New concepts and non taxonomic relationships extraction by the application of lexico-syntactic patterns, (3) Attribution of a weight for each extracted relationship relative to the frequency of the relationships that apply the lexico-syntactic pattern, (4) Updating the minimal ontology. Each iteration can be validated by the domain expert. This process is incremental: procedures A and B are repeated until no integration of new data is required. The last stage consists in an enrichment of the ontology structure and an extraction of structure patterns for each relationship of the domain ontology. The implementation of this strategy is still in progress. We have realized a little case study to illustrate the first iteration (Ben Mustapha and al., 2006).

## 3 ONTOCOSEMWEB PROTOTYPE: APPLICATION TO TOURISM

An implementation and experimentation of the suggested approach were carried. The principal objective is to automate the process of ontology construction. In fact a prototype named *OntoCoSemWeb* had been developed. The main purposes of the developed prototype are: (1) to automate the construction of ontology by combining the method of (Hearst, 1998), the semantic signature and a method of text mining which consists in the construction of the space word and the ASIUM method of syntactic Frames learning (Faure and Al, 1998 ) ; (2) to proceed to the learning of the possible relationships while saving information of knowledge extraction in the metaontology ; (3) To annotate the chosen web pages by using a minimal ontology of a specific field and enrich it by learning from these same Web pages which will be then annotated by the result of the ontology learning. Thus, after the elimination of the HTML tags from web documents, using API DOM (Document Object Model) a web mining technique has been implemented, to perform the segmentation of the texts in sentences. For the construction of the minimal ontology and meta-ontology, Protege 2000 has been used. The alimentation of the meta-ontology is made by the nominal expressions of each concept, the definition of the semantic signature and Hyponyms of each concept and the research of syntactic Frames of each minimal ontology relation. Concepts are built according to existing words in the corpus and in Wordnet which is used to select the most similar words or expressions, which will be considered as the topics signatures of, as an example, the concept "hotel ". The construction of the corpus patterns was done from 10 Web sites to obtain a textual corpus of 10 groups where each group contains between 130 and 300 textual files. Each file represents one Web page of the site associated with only one group. So that the frequency of a pattern is computed according to its occurrence in a Web page and for all the pages of the Web site.

## 4 DESCRIPTION OF ONTOCOSEMWEB

The experimentation starts by choosing a concept of the minimal ontology of tourism to begin the learning step. The metaontology enrichment starts by searching synonyms, “part-of” relations and nominal expressions referring to the concept. These concepts are “part-of” the concept and will be inserted temporarily in the metaontology during step A in order to enrich the domain ontology in step B. Then, the expressions referring to it are constructed. The existing generic linguistic axioms of the metaontology are used to deduce new concepts or instances (from nominal propositions patterns or other patterns). As an example, the NP_Concept pattern, NP means « proper noun » allows us to insert the instance « Arkansas hotel » as an instance of the concept « Hotel ». Expressions referring to the concept « hotel “ are generated. The multidimensional space is built with terms existing in the corpus and in Wordnet (to select the most similar words or expressions associated to “hotel”). These terms are the semantic signature of the concept « hotel ». The concepts which are similar to the concept « hotel » (*palace* for example) are extracted. The next step consists in searching a relation between a concept and its semantic signature. The semantic signatures represent close concepts. Some of them have no taxonomic relation in the domain ontology. Then, we have to verify the existence of a taxonomic or a non taxonomic relation, in order to filter the semantic signatures list related to a given concept. With the developed tool, the user can select a concept and "Extract senses ", which allows the visualization of the possible senses of the word "Airline " and search in Wordnet. Two senses are found: the user has to choose the suitable one and "show synonym" to visualize the synonyms relating to the selected sense. The synonyms in the last stage are inserted in the metaontology. When the concept is referred by a term having at least one sense, the user cans «Enrich Nominal Proposition ». A lexico-syntactic pattern and its occurrence are associated to each nominal expression. The following step allows the enrichment of the meta-ontology with the nominal expressions of the given concept, its occurrence in the corpus and the occurrence of each nominal expression. The next step consists of the construction of the Multidimensional word space and concept similarity matrix.

## 5 RESULT ANALYSIS

The experimentation carried out on Web pages concerning tourism was presented in this paper. The main aim was to show the feasibility of the approach of ontologies construction for the semantic Web by applying techniques of knowledge extraction. The techniques of knowledge extraction are mainly text mining techniques: the extraction technique of lexical patterns, the construction of word space and the construction of matrix similarity. The originality of the suggested approach consists in ontology construction to generate knowledge. Thus, the iterative character of the approach makes it possible to obtain after the last iteration association rules such as : *If " hotel " near to " sea" is expensive*, or *80% of hotels of "Paris 2" are classified hotels 2 stars and less.* Such rules represent a knowledge being able to contribute to the decision-making.

## 6 CONCLUSIONS AND PERSPECTIVES

Learning methodologies try to give a response to the time-consuming manual ontology building task. Learning techniques can be either numeric or symbolic. They have been exploited to semi-automate some foundation tasks such as concept hierarchy building, taxonomic relationships extraction, non taxonomic relationships learning, etc. The proposed architecture is based on self-learning ontological components which define Web content semantics: the domain ontology ; Web structure semantics : the ontology structure and domain web services semantics: the web services ontology. The metaontology is reusable and can be applied to other domain building ontology processes (from corpus with a determined language). Axioms to extract concepts, relationships and instances are learned incrementally. This metaontology can be extended to other extraction techniques, where each technique can use another similarity measure. A hybrid domain ontology combining three techniques is built: lexico-syntactic patterns, syntactic frames and multidimensional word space. The proposed approach is based on extracted information weighted by its frequency in the corpus. The corpus is updated, from one iteration to another, which is useful to revise this information and the associated weightings.

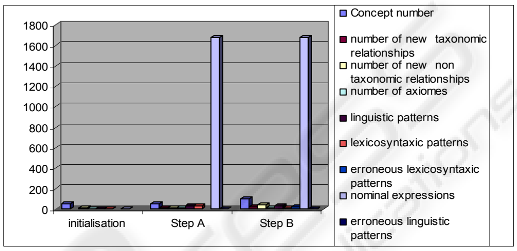

Figure 1: Results analysis.

## REFERENCES

Berners-Lee, T, Hendler, J, Lassila, O., 2001: The Semantic Web, *Scientific American.*

Berendt, B., Hotho, A., Stumme, G., 2002: Towards semantic web mining. In: *International Semantic Web Conference*, volume 2342 of Lecture Notes in Computer Science, Springer, 264–278.

Ben Mustapha, N., Aufaure, M.-A., Baazaoui-Zghal, H., 2006: Towards and Architecture of Ontological Components for the Semantic Web, *Web Information Systems Modeling Workhop (WISM)*, in conjunction with CAISE'2006, Luxembourg.

Gruber, T., 1993: Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, special issue on Formal Ontology in Conceptual Analysis and Knowledge Representation.*Eds, Guarino, N.&Poli , R.

Grüninger, M., Fox M.S., 1995: Methodology for the design and evaluation of ontologies. *IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada.

Alfonseca, E., Manandhar, S., 2002: Improving an Ontology Refinement Method with Hyponymy Patterns. *Language Resources and Evaluation (LREC)*, Las Palmas, Spain.

Gomez-Perez, A., Fernandez-Lopez, M., Corcho, O., 2003: *Ontological Engineering*. Springer.