Using Association Rules to Enrich Arabic Ontology

Asma Ksiksi
Technologies of Information Laboratory (LR-SITI)
National Engineering School of Tunis (ENIT)
Tunis, Tunisia
asma.ksiksi@gmail.com

Hamid Amiri
Technologies of Information Laboratory (LR-SITI)
National Engineering School of Tunis (ENIT)
Tunis, Tunisia
hamidlamiri@gmail.com

***Abstract*—In this article, we propose the use of a minimal generic base of associative rules between term association rules, to automatically enrich an existing domain ontology. Initially, non-redundant association rules between terms are extracted from an Arabic corpus. Then, the matching of the candidate terms is done through the matching between the concepts of the initial ontology and the premises of the association rules, with three distance measures that we define.**

***Keywords*-ontology; automatic enrichment; association rules**

I. Introduction

Ontology is a tool for representing knowledge and reasoning that serves the organization of a set of concepts in a specific field, as well as the relations between these concepts [1-3]. Ontologies are regularly subject to updates and changes. Performing these updates manually is an expensive and time-consuming task as it mobilizes experts in the field to identify and classify new vocabulary items in the ontology. To accelerate this process of evolution and adaptation and to take away any form of subjectivity, recent research has focused on the implementation of semi-automatic and automatic ontology enrichment techniques. The majority of approaches, often based on statistical or linguistic tools, focus on adding new concepts and/or relationships between them. The ontology enrichment process can be divided into two stages: the search for new concepts and relations and the placement of these concepts and relationships within the ontology [3]. The general process is depicted in Figure 1. Several works have focused on this process of enrichment of ontologies, addressing one or more of its stages:

- Extraction of representative terms in a specialized field.
- Identification of lexical relations between terms.
- Placement of new terms in an existing ontology

In these works, the term ontology takes several meanings like thesaurus, taxonomy or more generally controlled vocabulary. The work dealing with the extraction of candidate terms in the ontology enrichment process is based on statistical and syntactic methods. Statistical methods select terms according to their distribution in the corpus [1-3], as well as other measures such as mutual information, "the probability of the appearance of the word A knowing that the word B has appeared", or else measures calculating the probability of occurrences of a set of terms [4-6]. These different propositions make it possible to identify new ontology elements, but do not allow their placing in the ontology, without human intervention. Syntactic methods aim at determining the grammatical function of a word or a group of words within a sentence. They are based on the hypothesis that grammatical dependencies reflect semantic dependencies. These techniques lead to the proposition of new concepts, linked by relations that are not yet semantically identified. Regarding the identification of concepts and relationships and their placement in the ontology, the extraction of ARs is one of the major techniques proposed by the data mining community. Many other works propose the use of frequent correlations which can exist between the terms of a corpus. These approaches consist most often of extracting ARs between candidate terms, previously identified by statistical or syntactic tools [7]. At the end of the process, authors get a set of ARs, describing the existence of a relationship between two concepts [8-10].

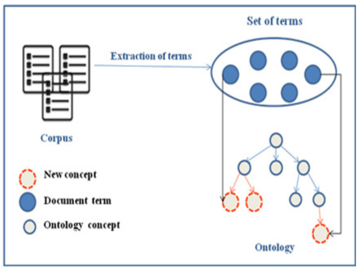


Fig. 1. General process of ontology enrichment.

In this paper, we propose a methodology for building a conceptual network formed by the combination of two types of knowledge, namely, knowledge present in the initial ontological structure specific to a domain and represented by semantic links, and knowledge derived from the minimal generic base of associative rules (ARs) between terms, essentially representing correlations that are appreciated by statistical measures.

II. Existing Techniques for Ontology Enrichment

A. Existing Approaches to Discover Candidate Concepts

We distinguish two types of methods for the discovery of candidate concepts:

- Statistical methods: they select the terms according to their distribution in the corpus [1-3], as well as other more complex measures such as mutual information, tf-idf, etc., or the use of statistical distributions of terms [4]. These different propositions make possible to identify new ontology elements, but do not make possible placing them in the ontology, without tedious human intervention [12].
- Syntactic methods: they aim to determine the grammatical function of a word or group of words within a sentence. They are based on the hypothesis that grammatical dependencies reflect semantic dependencies [13]. They define in a sentence, the verb (V) as being the relation which links the subject (S) to the complement (C). They thus have the disadvantage of identifying only the relationships labeled by the verbs. Some approaches also use syntactic patterns [12]. The extracted terms illustrate the new candidate concepts for enrichment, but also the existence of relations between them. However, these relationships are not labeled semantically. Moreover, no measure evaluating semantically new added relations is calculated.

B. Existing Approaches to the Concept Placement in Ontology

After the discovery of the candidate terms, it is essential to detect the relations between these new terms and those which link them to the initial ontology. In [2], authors propose a statistical approach based on the frequent co-occurrence of candidate terms with terms of the initial ontology. The major drawback of this work lies in the fact that they do not allow the precise addition of new concepts and relations in the ontological structure [14]. Other approaches in the literature suggest using search techniques data [10, 11, 13]. The work in [4, 15], is based on a classification method in order to bring together the candidate terms contained in the texts of the concepts present in the ontology. The principle is similar to that explained in the approaches of [1, 15], which group together terms by a clustering method according to their number of occurrences within the corpus. However, these methods do not detect the relations between the candidate terms, i.e., these new terms can therefore be added only by human intervention. In addition, several studies propose the use of frequent correlations that exist between the terms of a corpus. These approaches consist of extracting rules association [7] between candidate terms [8-10]. At the end of the search process, a set of ARs between terms is generated. Each rule expresses the existence of a relationship between two concepts of the domain. This process of enrichment remains semiautomatic because on the one hand the number of derived ARs is very important and on the other hand, a human intervention is necessary to semantically define the relations discovered and to name them.

III. Association rules

Association rule mining is a famous knowledge discovery technique for finding associations between items from a transaction database. Its definition varies according to the three main currents initiated by the following: author in [16] defines rules of statistical implication to help educationalists find relationships between acquiring basic notions in class, authors in [17] are more interested in orderly representation of concepts with informative implications, authors in [18] favored optimized extraction of ARs in large databases. Subsequently, these forms have known extensions in several directions. The binary properties are no longer required, we can now make ARs with digital properties [19, 20]. To avoid the vast increase of rule extraction time, more efficient algorithms have been proposed [21]. The semantics of the rules have been refined through many quality indices [22], which helps the user to choose the most appropriate rules for his needs. Navigation and queries by using an appropriate language have been developed [23] to facilitate the exploration of this set of rules. ARs present conditional relationships between the attributes of a database. They represent an implication of the form A→B where A and B are an itemsets. The set of items A is called antecedent and B consequent of the rule which provides information about the existing relations between A and B. It expresses how objects or items are related to each other, and how they can be grouped together. The first step of extraction in the association rules mining is finding out the frequent itemset which is called candidate (te). This transaction can be measured by two statistic measurements called support and confidence. The support (Sup(A→B)) is defined as the relative frequency of transactions in the data set D that contains the itemsets A and B.

$$\mathrm{Sup}(A \rightarrow B) = \mathrm{Sup}(A \cup B) = \frac{\left|\{t \in D : A \subseteq t \text{ et } B \subseteq t\}\right|}{|D|} \qquad (1)$$

The confidence (Conf(A→B)) of a rule measures the reliability of the inference given by rules.

$$\mathrm{Conf}(A \rightarrow B) = \frac{\mathrm{Sup}(A \cup B)}{\mathrm{Sup}(A)} \qquad (2)$$

Then, the important association rules are filtered from the candidate itemsets. A rule r is available only if Sup(A→B)>minsup and Conf(A→B)> minconf where minsup represents the threshold of support and minconf represents the threshold of confidence. These two values are specified by the user.

A. Process for Association Rules Extraction

The process of extracting association rules consists of several phases ranging from data selection and preparation to result interpretation (Figure 2). Several works have focused on this process of enrichment of ontologies, addressing one or more of its stages:

- Data selection and preparation (cleaning): In this phase, the database data used for the extraction of the association rules are selected and the transformation of these data into an extraction context occurs. This phase is necessary to be able to apply rule extraction algorithms to different kinds of data from different sources, to concentrate the search on the useful data and to minimize extraction time [24]. To have significant rules the extraction of morphological analysis of each word must follow the order described in [25] and shown in Table I

- Generation of association rules: is carried out from the frequent itemsets generated previously. In general, the generation of association rules is done directly, without access to the extraction context, and the cost of this phase in execution time is therefore low compared to the cost of extracting frequent itemsets.
- Visualization and interpretation: This phase consists in the visualization of the association rules extracted from the context and their interpretation. Thus the domain expert can judge their relevance and usefulness.

TABLE I. REPRESENTATIVE SCHEMA OF AN ARABIC WORD STRUCTURE

Enclitic	**Suffix**	**Schematic body**	**Prefix**	**Proclitic**
Base post		Radical	Near_base	

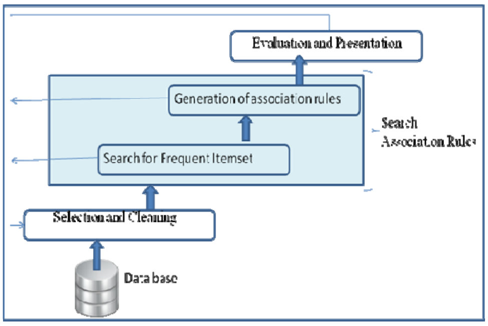

Evaluation and Presentation
Generation of association rules
Search for Frequent Itemset
Search Association Rules
Selection and Cleaning
Data base

Fig. 2. Process of Association Rules.

IV. PROPOSED APPROACH

This stage consists in bringing closer to our initial ontology, that will be noted as O, the terms which appear in the premises of the candidates rules of the base of the sequential rules. These terms are identified as candidate concepts for enrichment.

- Definition 1: An ontology is a quadruplet $O=(C_D, \leq_C, R, \leq_R)$ where C_D is the set of concepts of the domain, $\leq_C$ is the partial order defined on C_D, R is the set of relations defined on $C_D \times C_D$ and $\leq_R$ is the partial order relation defined on R. We consider that a formal extraction context is a triplet K=(D, T, R) where D represents a finite set of documents from the corpus C, T is a finite set of terms and R a binary relation. Each pair $(d,t) \in R$ means that the document $d \in D$ contains the term $t \in T$.
- Definition 2: A termset is a non-empty set of terms denoted by $(t_1, t_2 \ldots t_k)$. An associative rule R is valued by two statistical metrics, namely support and trust [18]. The support of the associative rule R: $T_i \rightarrow T_j$, denoted by Supp (R), expresses the frequency with which the two termsets T_i and T_j co-occur together in corpus C. The confidence of R, denoted by Conf(R), expresses the conditional probability for a document to contain termset T_j, knowing that it contains the termset T_i. An associative rule is valid if its confidence is greater than or equal to the minimum confidence threshold noted by minconf.

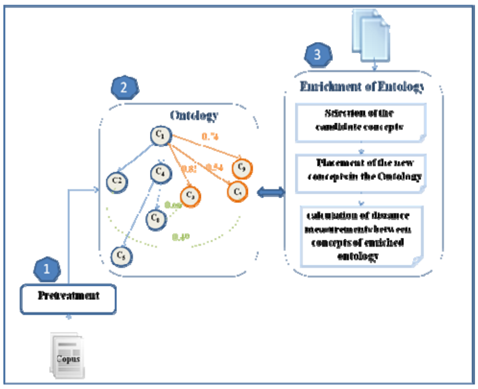

1
Pretreatment
Copus
2
Ontology
C1
C2
C4
C3
C9
C8
C5
0.74
0.81
0.54
0.40
3
Enrichment of Entology
Selection of the candidate concepts
Placement of the new concepts in the Ontology
calculation of distance measurements between concepts of enriched ontology

Fig. 3. Process of enrichment of Ontology by using Association Rules.

A. *Extraction of the Ontology and Creation of the Generic Base*

We use the GEN-MGB algorithm [26] for the extraction of the generic base of RA no redundant MGB. This base is characterized by its significant compactness, i.e., it contains a minimal core of ARs, from which all the redundant and valid rules can be deduced by means of a complete and valid axiomatic system [26, 27]. By considering the context of text extraction K, we adapt the definition of the MGB base given in [26] to the problem of Ontology enrichment. We remind that non-redundant ARs have one only term of the domain in the premise [28].

$$MGB = \mathrm{MGB} = \{\mathrm{R} : \mathrm{t} \rightarrow \mathrm{T}_k \mid \mathrm{Conf(R)} \geq \mathrm{minconf} \wedge \mathrm{T}_k = \{\mathrm{t}_1, \ldots, \mathrm{t}_k\} \subseteq \mathrm{T}\} \quad (3)$$

We use then a semi-automatic tool such as Protege 2000 [29] for the ontology O construction from CO corpus. It is validated downstream by a domain expert. The evaluation of the semantic link between O concepts are computed from the proposed similarity measure in [30] that takes into account both the depth of concepts in the hierarchy of concepts and the structure of the latter. Thus, the similarity between two concepts C1 and C2 of the ontology O is calculated as [30]:

$$\mathrm{SimWu}(C_1, C_2) = \frac{(2 \times \mathrm{depth(c)})}{(\mathrm{depth}(c_1) + \mathrm{depth}(c_2))} \quad (4)$$

where depth (c_i) corresponds to the depth level of the concept c_i and c represents the most specific concept that generalizes c_1 and c_2 in O.

B. *Adopted Approach for the Ontology Enrichment*

The enrichment process we propose is iterative and includes the following steps:

1) *Calculation of the candidate concepts for the enrichment*

We compute for each concept c_i of ontology O the set of the candidate concepts to be connected to c_i. This set includes the terms figuring in the conclusions of the valid associative rules whose premise is c_i as well as those of the redundant rules [31].

According to the example shown in Figure 4, the candidate concepts for enrichment related to the concept c_1 are $\{c_{10}, c_{12}, c_5, c_{15}\}$.

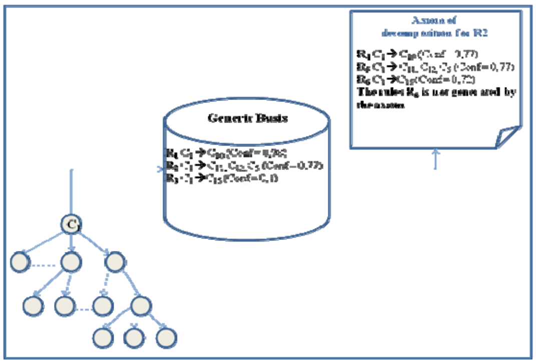

Axioms of
decomposition for R2
R_4 $C_1 \rightarrow C_{10}$(Conf 0,77)
R_5 $C_1 \rightarrow C_{11}, C_{12}, C_5$ (Conf=0,77)
R_6 $C_1 \rightarrow C_{15}$(Conf=0,72)
The rules R_6 is not generated by the axiom
Generic Basis
R_1 $C_1 \rightarrow C_{10}$(Conf=0,98)
R_2 $C_1 \rightarrow C_{11}, C_{12}, C_5$ (Conf=0,77)
R_3 $C_1 \rightarrow C_{15}$(Conf=0,1)
C_1

Fig. 4. Example of calculating candidate concepts.

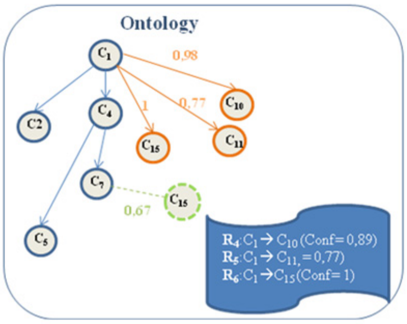
2) *Placement of the new concepts*

This step consists in placing the candidate concepts while preserving the coherence of the concepts and pre-established relations in the initial ontology. This makes possible not to add relational redundancies in the case of a concept being candidate to be related to several concepts of ontology O [32]. Figure 5 shows the addition of the new concepts c_{10} and c_{11} and the displacement of c_{15} because $\mathrm{Conf}(c_1 \Rightarrow c_{15}) > \mathrm{Conf}(c_7 \Rightarrow c_{15})$.

3) *Calculating the neighborhood of ci and distance* measurements

We define the notion of neighborhood of a concept of the ontology O as:

- Definition 3: The neighborhood of a concept represents the set of corners connected to it in the ontology, by one or more valid association rules [32]. The relations between c_i and its neighbors, are evaluated on the basis of a statistical metric that we call measure of distance between c_i and its neighborhood, and denoted by $\mathrm{Dist}_{O\ MGB}$. It is computed according to the measure of confidence of associative intervening during the ontology enrichment and the measures of similarities calculated between the concepts of the initial ontological structure [33].



R_4:$C_1 \rightarrow C_{10}$(Conf=0,89)
R_5:$C_1 \rightarrow C_{11,}$=0,77)
R_6:$C_1 \rightarrow C_{15}$(Conf=1)

Fig. 5. Example of placement of candidate concepts.

The measure of distance that we define is calculated according to three possible cases [34]:

- Case 1: If the two concepts c_i and c_j come from the base C then $\mathrm{Dist}_{O\ C}(c_i, c_j) = \mathrm{Conf}(R: c_i \Rightarrow c_j)$.
- Case 2: If the two concepts c_i and c_j belong initially to ontology O then $\mathrm{Dist}_{O\ C}(c_i, c_j) = \mathrm{SimWu}(c_i, c_j)$. The similarity between the two concepts c_1 and c_2 of ontology O is calculated as in (3).
- Case 3: If c_i is a concept added to the ontology O and it is related to the concept c_k of the initial ontology O in a way that $\mathrm{Dist}_{O\ MGB}(c_k, c_i) = \mathrm{Conf}(R{:}c_k \Rightarrow c_i) = \beta$ then any concept c_x of the ontology O in relation to c_k such that $\mathrm{SimWu}(c_k, c_x) = \alpha$, is also in relation with c_i. In this case, the distance measure is mixed, i.e., $\mathrm{Dist}_{O\ MGB}(c_i, c_x) = \alpha \times \beta$.

The three cases are illustrated in Figure 3. Thanks to this enrichment technique we are able to add new concepts and relationships.

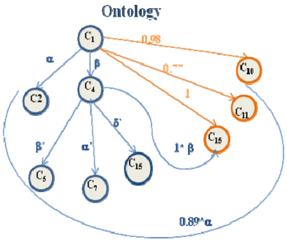


Fig. 6. Example of a figure caption.

V. CONCLUSION

Various ontology enrichment techniques have been proposed in the literature. Their limitations come from the fact that they do not allow the entire enrichment process without the intervention of the domain expert. In this article, we presented an automatic ontology enrichment process with a generic base of associative rules. The originality of our approach is that it exploits the maximum of concepts for enrichment without resorting to a priori knowledge. Its advantage is that it allows the learning of the distance represented by any relation of the enriched ontology.

REFERENCES

[1] E. Agirre, O. Ansa, E. Hovy , D. Martinez, "Enriching very large ontologies using the WWW", ECAI 2000 Workshop on Ontology Learning, Berlin, Germany, August 2000

[2] A. Faatz, R. Steinmetz , "Ontology enrichment with texts from the WWW", 2nd Semantic Web Mining Workshop at ECMLI/PKDD, WS'02, Helsinki, Finland, pp. 20-33, 2002

[3] V. Parekh, J. Gwo, T. Finin, "Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies", Proceedings of the International Conference of Information and Knowledge Engineering, Las Vegas, USA, June 21, 2004

[4] K. Neshatian, M. R. Hejazi, "Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies", Proceedings of the 2nd Workshop on Information Technology and its Disciplines, pp. 43-48, 2004

[5] P. Velardi, M. Missikoff, R. Basili, "Identification of relevant terms to support the construction of Domain Ontologies", ACL-EACL Workshop on Human Language Technologies, Toulouse, France, July 2001

[6] A. Xu, S.-K. Park, S. D'Mello, E. Kim, Q. Wang, C. Pikielny, “Novel genes expressed in subsets of chemosensory sensilla on the front legs of male Drosophila melanogaster”, Cell and Tissue Research, Vol. 307, No. 3, pp. 381-392, 2002

[7] R. Srikant, R. Agrawal, “Mining generalized association rules”, Future Generation Computer Systems, Vol. 13, No. 23, pp. 161-180, 1997

[8] R. Bendaoud, “Construction et enrichissement d’une ontologie à partir d’un corpus de textes”, Actes des Rencontres des Jeunes Chercheurs en Recherche d’Information (RJCRI’06), Lyon, France, pp. 353-358, March, 2006 (in French)

[9] A. Maedche, S. Staab, “Mining ontologies from text”, Lecture Notes in Computer Science, Vol. 1937, pp. 189-202, Springer-Verlag, 2000

[10] G. Stumme, A. Hotho, B. Berendt, “Semantic web mining : State of the art and future directions”, Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 4, No. 2, pp. 124-143, 2006

[11] L. Jorio, L. Abrouk, C. Fiot, D. Hérin, M. Teisseire, “Enrichissement d’ontologie basé sur les motifs séquentiels”, Actes de la Plateforme AFIA 2007, Atelier Ontologies et gestion de l’hétérogénéité sémantique, 2007(in French)

[12] A. Maedche, V. Pekar, S. Staab, “Ontology Learning Part One - On Discovering Taxonomic Relations from the Web”, in: Web Intelligence, pp. 301-319, Springer Verlag, 2002

[13] N. Hernandez, J. Mothe, C. Chrisment, D. Egret, “Modeling context through domain ontologies”, Information Retrieval, Vol. 10, No. 2, pp. 143-172, 2007

[14] P. Cimiano, A. Hotho, G. Stumme, J. Tane, “Conceptual Knowledge Processing with Formal Concept Analysis and Ontologies”, Lecture Notes in Computer Science, Vol. 2961, pp. 189-207, Springer-Verlag, 2004

[15] E. Han, G. Karypis, “Centroid based document classification : Analysis and experimental results”, Lecture Notes in Computer Science, Vol. 1910, pp. 424-431 Springer-Verlag, 2000

[16] R. Gras, Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse d'Etat, Universit e de Rennes, 1979, (in French)

[17] J. L. Guigues, V. Duquenne, “Familles minimales d'implications informatives résultant d'un tableau de données binaires”, Mathématiques et Sciences Humaines, Vol. 95, pp. 5-18, 1986

[18] R. Agrawal, T. Imielinski, A. Swami, “Mining Association Rules between sets of items in large Databases”, Proceedings of ACMSIGMOD Conference,Washington, USA, pp. 207-216,May 25-28, 1993

[19] S. Guillaume, Traitement des données volumineuses, mesures et algorithmes d'extraction de RA et règles ordinales, PhD Thesis, Nantes, 2000

[20] M. Cadot, “RA et codage flou des données”, 11èmes Rencontres de la Société Francophone de Classification (SFC'04), Bordeaux, France, pp. 130-133, 2004, (in French)

[21] N. Pasquier, “Data Mining : Algorithmes d'Extraction et de Réduction des RA dans les Bases de Données”, PhD Thesis, Université Blaise Pascal-Clermont-Ferrand II, 2000 (in French)

[22] F. Guillet. “Mesure de qualité des connaissances en ECD”, Cours donné lors des journées de la conférence EGC 2004, Clermont-ferrand, January 2004, (in French)

[23] M. Botta, J. F. Boulicaut, C. Masson, R. Meo, “A Comparison between Query Languages for the Extraction of Association Rules”, Lecture Notes in Computer Science, Vol. 2454, pp. 1-10, Springer-Verlag, 2002

[24] M. Jarrar, “Building a Formal Arabic Ontology”, Experts Meeting on Arabic Ontologies and Semantic Networks, Alecso, Arab League: Tunis, pp. 26-28, July 26-28, 2011

[25] F. Z. Belkredim, F. Meziane, “DEAR-ONTO: A Derivational Arabic Ontology Based on Verbs”, International Journal of Computer Processing of Languages, Vol. 21, No. 03, pp. 279-291, 2008

[26] C. C. Latiri, L. B. Ghezaïel, L. B. Ahmed, T. Tunsisie “Fast-MGB: Nouvelle base générique minimale de règles associatives”, EGC’2006, Lille, France, pp. 217-222, January, 2006, (in French)

[27] C. L. Cherif, W. Bellagha, S. Ben Yahia, G. Guesmi, “VIE-MGB : A Visual Interactive Exploration of Minimal Generic Basis of Association Rules”, 3rd International Conference on Concept Lattices and their Applications (CLA’05), Olomouc, Czech Republic, pp. 179-196, September, 2005

[28] C. Fankam, OntoDB2: un système flexible et efficient de Base de Données à Base Ontologique pour le Web sémantique et les données techniques, PhD Thesis, ISAE-ENSMA Ecole Nationale Supérieure de Mécanique et d’Aérotechique-Poitiers, 2009, (in French)

[29] N. F. Noy, R. W. Ferguson, M. A. Musen, “The Knowledge Model of Protégé-2000 : Combining Interoperability and Flexibility”, Lecture Notes in Computer Science, Vol. 1937, pp. 17-32, Springer-Verlag, 2000

[30] Z. Wu, M. Palmer, “Verb semantics and lexical selection”, 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA, pp. 133-138, June 27-30, 1994

[31] T. R. Gruber, “The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases”, Proceedings of the Second International Conference, Cambridge, pp. 601-602, Morgan Kaufmann, 1991

[32] D. J. Abadi, A. Marcus, S. R. Madden, K. Hollenbach, “Scalable Semantic Web DataManagement Using Vertical Partitioning”, 33rd International Conference on Very Large Data Bases, Vienna, Austria, pp. 411-422, September 23-27, 2007

[33] P. Gamallo, M. Gonzalez, A. Agustini, G. Lopes, V. S. de Lima, “Mapping Syntactic Dependencies onto Semantic Relations”, Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT’2002), Lyon, France, pp. 15-22, 2002

[34] F. Cerbah, ”Learning highly structured semantic repositories from relational databases”, Lecture Notes in Computer Science, Vol. 5021, pp. 777-781, Springer-Verlag, 2008