Combining Semantic Search and Ontology Learning for Incremental Web Ontology Engineering

Nesrine Ben Mustapha[1], Hajer Baazaoui Zghal[1], Marie-Aude Aufaure[2], and Henda Ben Ghezala[1]

[1] Laboratory RIADI-GDL, National School of Computer Sciences, University of Manouba, 2010 la Manouba, Tunisia
{nesrine.benmustapha, hajer.baazaouizghal, henda.benghezala}@riadi.rnu.tn
[2] Mas Laboratory, SAP BusinessObjects Academic Chair, Ecole Centrale Paris, Grande Voie des Vignes,
92295 CHATENAY-MALABRY Cedex, France
Marie-Aude.Aufaure@ecp.fr

Abstract. In the context of the semantic Web, ontologies improve the exploitation of Web resources by adding a consensual piece of knowledge. The need for using domain ontology for information retrieval (IR) has been explored by some approaches to better answer users queries. However, ontology in IR system requires a regular updating, especially the addition of new concepts and relationships. Besides, IR systems are generally based on few domain ontology that cannot be expanded. This paper proposes a framework that describe semantic Web search for ontology learning. In a previous work, we have proposed an incremental approach for ontology learning using an ontological representation called Metaontology In this paper, we describe how the processes of semantic search and ontology learning from texts can collaborate to facilitate Web ontology engineering using case based reasoning.

Keywords: Semantic Web, ontology, ontology engineering, ontology learning, semantic search, case base reasoning

1 Introduction

Over the past few years, with the continued rapid growth in Web information volume, information access and knowledge management has become challenging. Thus, adding a semantic dimension to the Web, by the deployment of ontologies, contributes to solve many problems (Information retrieval, knowledge sharing, communication between Web agent, etc.). In the context of the semantic Web [1], ontologies improve the exploitation of Web resources by adding a consensual knowledge. The need for using domain ontology for information retrieval (IR) has been explored by some approaches to better answer users queries; however, ontologies in an IR system require a regular updating, especially with the addition of new concepts and relationships. In addition, semantic IR systems are generally based on a few domain ontologies that cannot be expanded. Manual ontology building is a

long and tedious task. In the last decade, several approaches of ontology learning have appeared and proposed a partial automation of knowledge acquisition from structured, semi-structured or unstructured data sources i.e. (database, knowledge base, texts,). In fact, ontology learning (OL) is defined as an approach of ontology building from knowledge sources using a set of machine learning techniques and knowledge acquisition methods. Keeping in mind that a unique data source cannot cover all concepts of a target domain of knowledge and that the Web is a rich textual source, the Web can be considered as a learning corpus from which domain ontologies are extracted to be used in semantic search systems. Our main objective is to make the semantic search engine more flexible and autonomous to construct domain ontologies from relevant documents in an incremental manner, by combining ontology learning from text and semantic search technology.

This paper is organized as follows: Section 2 presents the related work to ontology learning approaches from the Web. In section 3, we also describe our motivation to propose an incremental Ontology learning approach in semantic Web search systems for Web ontology engineering. The proposed framework is presented with an illustration scenario. Finally, we conclude and give some perspectives for this research work.

2 Related work

In the past decade, with the enormous growth of Web information, the Web has become an important source of information for knowledge acquisition, due to its size and heterogeneity. This has led to the development of two categories of OL approaches: OL from online Web ontologies [2, 3, 4, 5] and OL from textual content of the Web [6, 7, 8, 9, 10, 11].

Firstly, the idea of online ontology building from the Web is not a new one. Several approaches were proposed to use ontology search engines or ontology meta-search engines to build ontologies by aggregating many searched domain ontologies. In [5], a new approach consists in searching online ontologies for representations of certain concepts, ranks the retrieved ontologies according to some criteria, then extract the relevant parts of the top ranked ontologies, and merge those parts to acquire the richest domain representation as possible. Such approaches could easily lead to construct domain ontologies but some issues to be considered, still remain such as:

- The reliability of existing Web ontology;
- The availability of Web ontologies in terms of numbers and domain variety;
- The quality of output ontology which depends on the quality of input ontologies;
- The complexity of the use of Ontology searching, ontology ranking, ontology mapping, ontology merging, and ontology segmentation methods.

Secondly, OL approaches from Web content consist generally of enriching a small ontology, called "minimal ontology" or "granular ontology", with new concepts and new relationships using text mining techniques. Five categories of text mining techniques are mainly distinguished: linguistic techniques and lexico-syntactic patterns [12], clustering techniques and/or classification techniques [6, 13, 14],

statistical techniques [15, 16, 17, 18], association rule based techniques [Maedche], and hybrid techniques [19, 20]. The most well-known approaches exploit the textual Web content to enrich concepts using Wordnet. In these approaches, a priori domain knowledge is required. For this reason, they are dependent of the ontology domain and the collect of Web documents related to this domain needs user intervention.

On one hand, any process of ontology learning from text depends on the relevance of the textual corpus in addition to applied machine learning techniques. On the other hand, the main purpose of a semantic search is to provide users with the most relevant Web documents according to their query, and with the use of specific domain ontology. Having established this fact, a semantic search can be a useful way to perform ontology learning from Web content. Indeed, other approaches proposed an incremental approach of ontology learning from Web using Web search engine. In [21], one can find a study of several types of available Web search engine and how they can be used to assist the learning process (searching Web resources and compute IR measures).

In this context, an approach presented in [11] proposed to use an ontology-based search engine [22] to collect textual sentences from which new concepts and new relations are discovered. In [20], distinguishable and incremental process was proposed based on three phases: an initialization phase, an incremental phase of domain ontology learning, and lastly, a phase of results analysis. Indeed, the initialization phase includes the preparation and the pretreatment of the data sources which are composed of a minimal ontology, a metaontology, the linguistic ontology "Wordnet" and a set of Web documents relating to the target domain. The second phase is characterized by its incremental and iterative aspect. Each iteration has two successive steps: the first one is the creation of a metaontology [23] and the second is the application of the axioms related to ontology element learning. The first step consists of applying the techniques specified by the metaontology to instanciate metaconcepts and metarelations. These techniques are applied according to the process described in [11]. The second step includes the discovery of new concepts, new relations, and new axioms related to a domain.This approach leads to the implementation of the OntoCoSemWeb prototype and has been used to build ontology of tourism [24]. An online IR based on this ontology [22] collects and classifies the results selected by users. These results are used it as the input of "OntoCosemWeb prototype" Integrating an ontology learning task into the semantic search process and defining how the two processes work together will allow more domain ontologies to be built from selected documents and will also ameliorate the semantic search.

3. Incremental Ontology learning in semantic Web search systems

In this section, the main objectives of this work, the proposed approach and related framework are described.

3.1 Objectives and motivation

According to [25], the problem in contextual semantic search systems resides in building a new domain ontology which has not been already defined before. Knowing that the Web is an enormous information source and a dynamic one, integrating the ontology learning process in the search process is challenging. Besides, to fulfill this target, many objectives are fixed such as:

- Modularity and reuse of learned ontologies;
- Scalability and evolution of ontology building;
- Facility of learning axioms on ontology modules by linking the search request to search results;
- Personalization of the constructed ontology.

In fact, to have networked ontologies in a multi-contextual search engine is a key requirement to cover user needs. However, when many domain ontologies are used by a semantic search system, taking consideration of the modularity aspect facilitates the management task of modular ontologies. In many cases, a search query can be translated to an ontology module (a sub-part of ontology), which can be reused by other users to express a similar query or to enrich it with new concepts, instances, or relations. As a result, the contextual search system will become multi-contextual and more adaptable to users queries. The searcher will also participate also in ontology building by selecting more relevant documents which will be the input of ontology learning process to enrich the initial submitted query, the details of which are presented in the next section.

3.2 A multi-layer ontology Warehouse

A multi-layer ontology warehouse is proposed to be built for semantic search systems (cf. figure 1). The first layer represents the topic ontology, which is an ontological classification of topics, domain and contexts, regardless of the used language. Each topic (*T*) can be the sub-domain of a domain (D), depending on the position of the topic in the hierarchy. The second layer represents a set of networked domain ontology schema. Each domain ontology (Od) is a networked modules (M). A Module (M) is seen as a dimension in the domain ontology which consists of a main concept (C) with its common properties (relations with others concept i). Proprieties of a concept are defining as the more frequent relations that characterize this concept and that are used in query interfaces and relevant Web documents. A given Module could be present in many ontologies and in relation to other modules. For example, the module having as main concept conference could be in many domain ontologies (computer science, physics, mathematical, etc.), as we can find conferences related to many domains. A concept is the following tuple *(id, {(ti, language, context)}* $_{i=1..n}$, state, credibility Degree) where:

- *Id*: a concept identifier associated with a sense regardless of the terminological labels and the language referencing it;
- {(*ti, language, context*)}$_i$: a set of triple (*t*, *language, context*) where *t* is a nominal phrase referencing the concept in a target language and used in a specific context which can be the topic that represents the concept role in a specific domain;

– *State*: the state of the discovered concept. A discovered concept from text could be "new candidate", "validated", "rejected", "average candidate";
– *Credibility degree*: a degree of the concepts accuracy according to its module.

The third layer is made up with instantiated and personalized modular ontologies which are associated to each user and which represents the most used ontology fragments in the search activity of the user besides of their used terminology. The fourth layer includes the indexed resources with ontology modules.

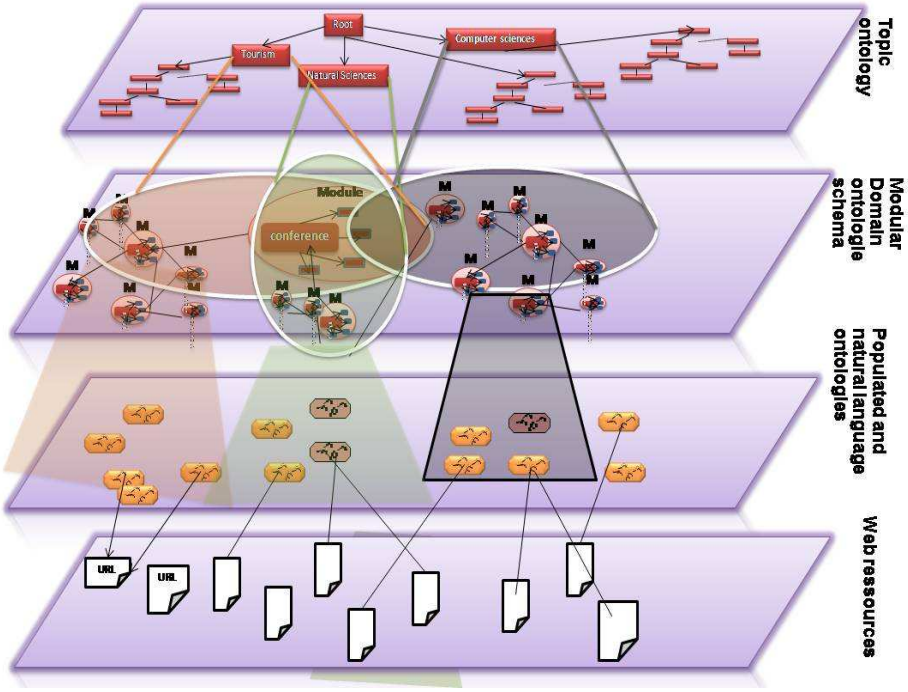


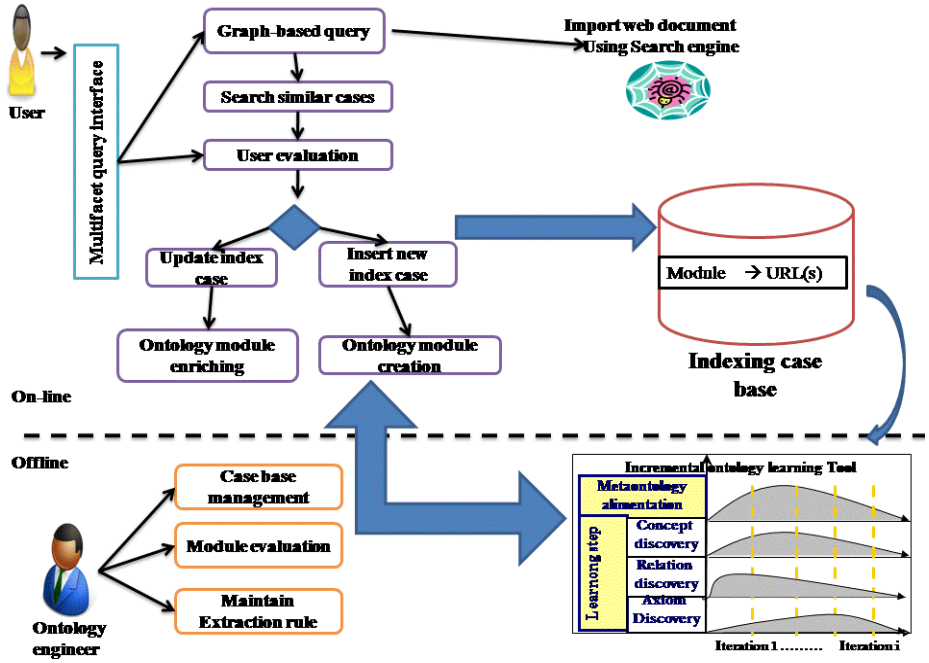
Fig. 1. Multi-layer ontology warehouse for semantic search

3.2 Approach based on combining semantic search and ontology learning for Web modular ontology engineering

The study of the ontology learning process and semantic search process allows for the conclusion that collaboration between the two processes could be useful, to have both "incremental ontology building" and "performed search."

For example, the user selects an existing topic from the Topic ontology, if it is a new one, the user can create the topic and place it in the appropriate position in the ontology and then formulate a search goal. According to [26], we distinguish between many search goals at which there are three categories: navigational goal, informational goal and resource search. By using the type of search goal selected by the user, the search purpose is better understood. Two possible scenarios are

distinguished in the proposed approach: creating a new ontology without background knowledge (an initial search process with any domain ontologies according to user request) and enriching an existent one by a new ontology module (search process based on existent domain ontology according to user request).



–

Fig.2. Combined process.

Creating new domain ontology without background knowledge. This process is described by the following steps:

- In the absence of domain ontology, there is not a similar case to learn new query reformulation. Using a traditional search engine, the user selects relevant Web documents.
- Then, a new case defined as the couple (*Prob*, *Sol*), is added to the case base. A case is We distinguish two types of cases: a primitive case and a validated case. In this scenario, we deal with primitive case (initially added) where:
  - *Prob* is (*T*, *D*, *Q*, *MC*, *St*) where *T*: the type of goal search, *D*: Topic of search, *Q*: formal query as a graph, *MC*: main concept, t*S*: state of the case (primitive or valid).
  - *Sol* is the solution formed by a set of Web documents addresses.

An iteration of process of ontology learning can start in off-line mode to create a new granular ontology for the selected domain and enrich it by extracting new ontology elements from Web documents (the solution of the target case); We have adapted our approach for ontology learning proposed in [13] which is based on

metaontology that use a search engine and formalized rules to extract new elements. The case is refined to be a valid one and the formal graph-based will be substituted by an ontology module (see Figure 2). A new rule associated with the case is added according to the type of search goal. For example, if the goal is navigational, a possible rule could be the following: "if *instanciated_module* then ", but in the case of closed and direct informational goal, a rule could be : "If *instanciated_module_1* then *instanciated_module_2*".

Enriching existent domain ontology by a new ontology module. Given a user query based on a central concept C, a list of domain ontology modules is listed from the case base.
If the user request is similar to another one existent in the case base **Then** the user chooses it to reformulate his request by adding other terms to restrict his search. Many steps are started simultaneously such as:

- Searching documents indexed by this module and classified by topic.
- Searching additional Web documents with a search engine and the relevance feedback method based on similarity to selected documents
- User evaluation of results
- The application of Ontology learning process to the most relevant documents for the enrichment of the target module of ontology.
- The update of the case with the enriched module and new Web documents.

3.4 Illustration example for creating a first ontology module:

Suppose that a user wants to know the URL of the workshop WISM 2009. The type of search goal is a navigational search. The user selects his domain of search ("*Computer science-Workshops*"), from the topic ontology. The main concept of the query is "workshop"

Table 1. Application of syntactic patterns and verb-based patterns.

Extraction of nominal phrases and applying lexico-syntactic patterns		
Noun phrases	**Lexico-syntactic pattern**	**New ontology elements discovered**
Workshop Program	Concept_noun	New concept: program and possessive relation between this phrase and "program"
Workshop dates	Concept_noun	New concept: date and possessive relation between this phrase and "date"
Extraction of sentences and applying verb based pattern		
Workshop proceedings will be published by the official CAiSE workshop proceedings.		New concept: proceeding, participants New relation: is_published, submit, invited
Workshop participants are invited to submit a paper related to one (or more) of the workshop topics.		

Suppose also that modular domain ontology is related to computer science in the ontology Warehouse. However, the concept "workshop" does not exist. So, the

formulation of this first request will be a new core ontology module to be enriched in computer science ontology. The main concept of this request is "workshop" (figure 3). Since the case base is empty, a query is submitted to a search engine and the user selects the Web document corresponding to the WISM workshop. This document will be the input of the ontology learning phase of the process. Then, the application of learning techniques (syntactic patterns and verb based pattern) specified by the above mentioned meta-ontology . The results of these techniques are illustrated by table 1, when new concepts and relations are discovered.

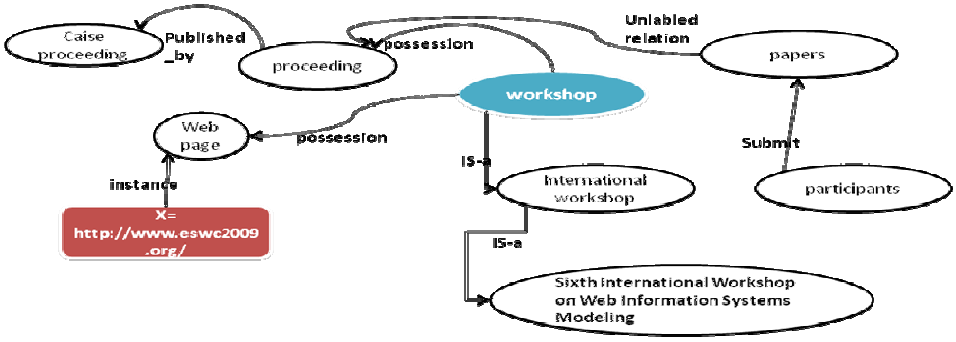


Fig. 3. Enriched ontology module.

4. Case study and first Experimental Observation

In this section, two ontologies will be compared. The first one is an ontology resulting from our previous approach "OntoCoSemWeb" This approach is based on the metaontology which is based on extraction of all textual elements from Web documents which are imported by a search engine. The second is modular ontology resulting from the approach described below, using a modified version of OntoCosemWeb.

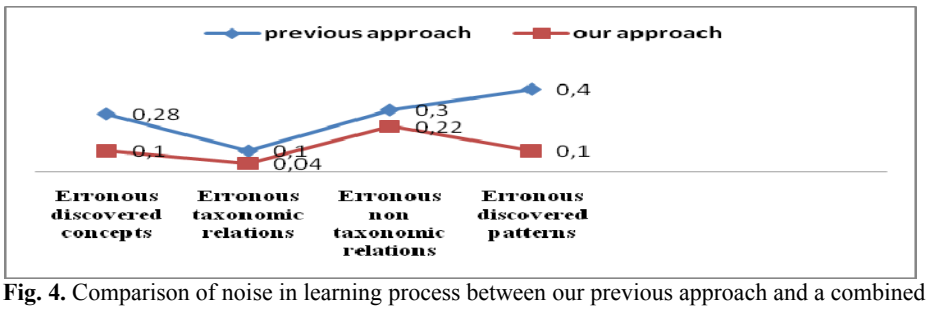


Fig. 4. Comparison of noise in learning process between our previous approach and a combined one.

The number of errors in discovered concepts and learned patterns has been compared. Noise in learning results was incredibly decreased by the first iteration. So,

the combination of the two processes can produce a more relevant Web document from which only an ontology fragment (module) will be enriched. This has also an effect on processing time.

4. Conclusion and future Works

This paper focuses on the possible combination of semantic search approaches and ontology learning methods to facilitate the integration of personalized and evolutionary ontology building in semantic search systems. The proposed framework is based on two main ideas: query transformation to "granual ontology" to be enriched and combining semantic search and ontology learning from texts can collaborate to facilitate Web ontology engineering and semantic indexing of Web documents using case based reasoning.. The main contribution of this work is to facilitate the Web semantic engineering using semantic search and ontology learning from Web documents and to link users' requests to ontology modules constructed by using case base reasoning. The next step is to implement the proposed approach using Google API and to compare ontology learning results and semantic search results to prove that the proposed multilayer approach is valuable for indexing documents and enriching ontology.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, Scientific American pp. 34-43. (2001)
2. Deitel, A., Faron, C., Dieng, R. : Learning Ontologies from RDF Annotations. In: Workshop in Ontology Learning, Seventeenth International Joint Conference on Artificial Intelligence, Seattle, CEUR-WS (2001)
3. Stojanovic, N., Stojanovic, L., Volz, R.: A Reverse Engineering Approach for Migrating Data-intensive Web Sites to the Semantic Web. In: 17th World Computer Congress, pp. 141-154. Kluwer Academic Publishers (2002)
4. Tijerino, Y., Embley, D., Lonsdale, D., Nagy G.: Ontology Generation from Tables. In: 4th International Conference on Web Information Systems Engineering, pp. 242-252. IEEE Computer Society (2003)
5. Allani, H.: Position paper: Ontology Construction from Online Ontologies. In: 15th International Conference on World Wide Web, pp. 491-495. ACM (2006)
6. Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching very Large Ontologies using the WWW, 1st Ontology Learning Workshop, 14th European Conference on Artificial Intelligence (2000)
7. Faatz, A., Steinmetz, R.: Ontology Enrichment with Texts from the WWW, In: First International Workshop on Semantic Web Mining, European Conference on Machine Learning, Helsinki, Finland, pp. 20-34. (2002)
8. Missikoff, M., Navigli, R., Velardi, P.: Integrated approach to Web Ontology Learning and Engineering, Journal of IEEE Computer, 35(11):60-63. (2002)
9. Navigli, R., Velardi, P.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites, In. Journal of Computational Linguistics, MIT press, 30(2):151-179 (2004)

10. Sanchez, D. and Moreno, A.: Creating Ontologies from Web Documents. Recent Advances in Artificial Intelligence Research and Development, 113:11-18. IOS Press (2004)
11. Baazaoui Zghal, H., Aufaure, M.-A., Ben Mustapha, N.: A Model-Driven Approach of Ontological Components for On-line Semantic Web Information Retrieval, Journal on Web Engineering, Special Issue on Engineering the Semantic Web, Rinton Press, 6(4):309-336 (2007)
12. Hearst, M.A.: Automated Discovery of WordNet Relations, Wordnet An Electronic Lexical Database, Word{N}et: An Electronic Lexical Database, pp. 132-152. MIT Press (1998)
13. Faure, D., Nedellec, C.: A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition, In: workshop on Adapting Lexical and Corpus Resources to sublanguages and Applications, 1st International Conference on Language Resources and Evaluation , pp. 1-8. Granada, Spain (1998)
14. Alfonseca, E., Manandhar, S.: An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In: International Conference on General WordNet, India (2002)
15. Lin, C-Y., Hovy, E.H.: The Automated Acquisition of Topic Signatures for Text Summarization. In: 8th International Conference on Computational Linguistics Conference, Strasbourg, France pp.495-501. Morgan Kaufmann (2000)
16. Velardi, P., Fabriani, P. , Missikoff, M.: Using Text Processing Techniques to Automatically enrich a Domain Ontology. In: International Conference on Formal Ontologies and Information Systems, pp. 270-284. ACM Press (2002)
17. Sugiura, N., Masaki, K., Naoki, F., Noriaki, I., Takahira, Y.: A Domain Ontology Engineering Tool with General Ontologies and Text Corpus. In: 2nd International Workshop on Evaluation of Ontology-based Tools, 2nd International Semantic Web Conference, pp.71-82. (2003)
18. Resnik, P.: Using Information Content to Evaluate Similarity in Taxonomy. In: 14th joint conference in Artificial Intelligence, pp. 448-452. Morgan Kaufmann (1995)
19. Maedche, A., Staab, S.: Mining Ontologies from Text. In: 12th International Conference on Knowledge Engineering and Knowledge Management, Springer Lecture Notes in Artificial Intelligence pp. 169-189 (2000)
20. Baazaoui Zghal, H., Aufaure, M.-A., Ben Mustapha, N.: Extraction of Ontologies from Web Pages: Conceptual Modeling and Tourism, Journal of internet Technologies, 8(4):411-421 (2007)
21. Sánchez, D.: Domain Ontology Learning from the Web, Thesis in Artificial Intelligence (2007)
22. Aufaure, M.-A., Soussi, R., Baazaoui Zghal, H., Ben Ghezala, H.: SIRO: On- Line Semantic Information Retrieval using Ontologies, In: 2nd International Conference on Digital Information Management, pp. 321- 326. IEEE Computer Society Press, Lyon, France (2007)
23. Ben Mustapha, N., Soussi, R., Baazaoui zghal, H., Aufaure, M.-A. : A Metaontology for Domain Ontology Enriching in an Information Retrieval System. In : second Francophone Day of ontologies, pp. 63-72. ACM Press, Lyon (2008)
24. Ben Mustapha, N., Baazaoui Zghal, H., Aufaure, M.-A.: A Prototype for knowledge Extraction from Semantic Web based on Ontological Components Construction, In: 3rd International Conference on Web Information Systems and Technologies, pp. 451-454. Barcelona, Spain (2007)
25. Esmaili, K. S., Abolhassani, H.: A Categorization Scheme for Semantic Web Search Engines. In: 4th International Conference on Computer Systems and Applications, pp. 171-178. IEEE Computer Society, UAE (2006)
26. Rose, D., Levinson, D.: Understanding User Goals in Web Search. In: 13th International Conference on World Wide Web, pp.13-19. USA, ACM Press (2004)