DYNAMic Ontology for Information Retrieval

Damien Dudognon, Gilles Hubert, Joel Marco, Josiane Mothe, Bachelin Ralalason
IRIT, Université de Toulouse
UMR 5505 CNRS
31062, Toulouse
+33 5 61 55 64 44
prenom.nom@irit.fr

Jérôme Thomas, Axel Reymonet
ACTIA Automotive
25, chemin de Pouvourville
B.P. 74215 - 31432 TOULOUSE cedex 04 (FRANCE)
+33 5 61 17 68 32
prenom.nom@actia.fr

Hervé Maurel, Mohamed Mbarki
Artal Technologies
Rue Pierre-Gilles de Gennes
Ensemble "La Rue" - Bat. 9 - BP 38138
31681 LABEGE CEDEX
herve.maurel@artal.fr
mohamed,mbarki@artal,fr

Philippe Laublet
STIH (LaLIC) Université de Paris-Sorbonne
28 rue Serpente 75006 Paris
+33 (0)1 53 10 58 40
Philippe.Laublet@paris-sorbonne.fr

Valentine Roux,
Préhistoire et Technologie
CNRS, Paris 10
valentine.roux@mae.u-paris10.fr

ABSTRACT
In this paper, we describe the DYNAMO project which defines a semantic information retrieval system.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval

Keywords
Semantic information retrieval, ontology.

1. OBJECTIFS DE DYNAMO
DYNAMO (DYNAMic Ontology for information retrieval, 2008-2011) is a project that aims at designing and developing an information retrieval system that considers both semantic indexing and retrieval and the dynamics of the documents [1] and of the underlying indexing language. The main goal of the project is to conceive a methodological approach and a set of software tools which take into account the building and the maintenance of ontological resources starting from documents. These ontological resources are then used for semantic indexing and thus making information retrieval more semantic as well. We define two modules, one for building and updating ontologies, the other module is dedicated to document annotation and indexing and to information matching. These two modules work in a cooperative way.

DYNAMO project considers three components of the information retrieval system: an ontology, a document set and annotations associated with these documents based on the use of the ontology. It focuses on the updating of these components (cf Figure 1). We propose to maintain ontology and annotations considering the document collection and its analysis. One of the originality of DYNAMO lies in the fact that there is a strong correlation between the maintenance of ontology and of the search module. We take into account, on the one hand, the evolution of the document collection and its implication on the ontological resources and, on the other hand, the dynamics of the document annotation according to the evolutions in the ontology.

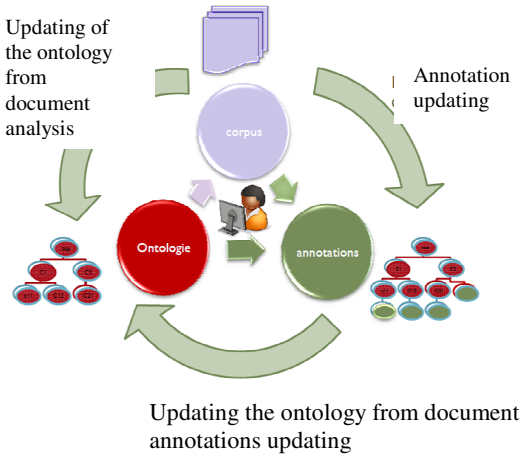

Figure 1 : Dynamics in Dynamo

Although the prototype considers the various aspects presented in figure 1, this presentation is focused on the annotation and semantic search parts.

2. ANNOTATION USING CONCEPT GRAPHS

Information is separate according to two levels: information resulting from the OR (Ontological Resource) on the one hand, and from the annotations on the other hand (cf figure 2):

- Concepts (*Smoke* and *Motorization* in figure 2), terms (T37 and T51 corresponding to the labels "*engine*" and "*to smoke*") as well as the existing denotation relations (*dénote* in figure 2) belong to the OR model and are completely independent of the corpus to which the process of annotation applies;
- Concepts instances (*f1* and m1), occurrences of terms (*occ3* and *occ15*) as well as the existing relations of designation (*désigne* in figure 2) cannot be dissociated from annotated documents. Indeed, the term occurrences contain some information like the relative position (*doc_offset*) of the term in the current document (*doc_id*). Concerning the concept instances, Dynamo considers them as anonymous and thus with local range: we do not aim at finding known specific objects; rather we aim at knowing if an object (i.e an instance) of a certain type (i.e a concept) either occurs or not in a document.

Considering the semantic process of annotation, the process consists first in finding the occurrence of certain terms of the OR in documents and building the corresponding occurrences. Then the system associates these occurrences with the corresponding concept instances. The concept instances are associated according to the semantic relations that exist in the OR. Thus, an annotation corresponds to a graph of instances.

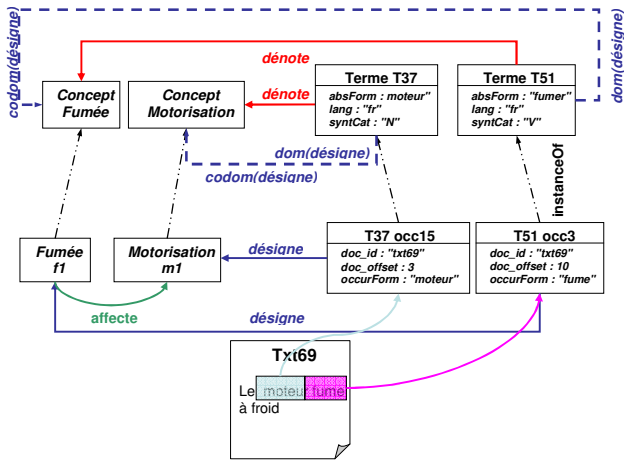

Figure 2 : OR and annotations in Dynamo.

3. SEMANTIC SIMILARITY

The similarity of a query and a document is based on the similarity between the associated annotation graphs. In turn, the similarity between annotation graphs rests on the similarity between the concepts associated with these graphs. Inspired by the principle of genealogy, the similarity between two concepts is comparable with the proximity of two family members: the more two members of the family have ancestors in common, the closer they are. The distance of a family member starting from a common ancestor influences his distance compared to the other family members. The similarity between two concepts is thus a question of relationship between the number of the common ancestors and the genealogy of these concepts.

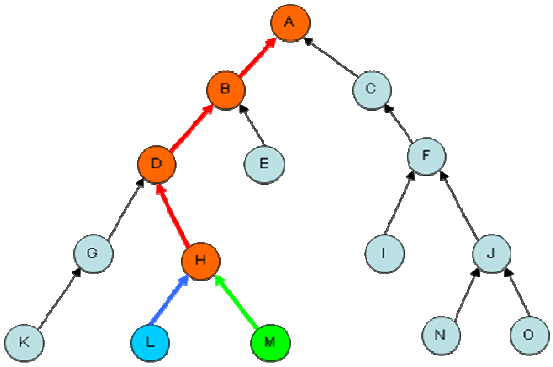

Figure 3 : Concept hierarchy : example.

The ascendant genealogy « Gen » of M is:

$$Gen(M) = \{A, B, D, H, M\}$$

The common ancestors of L and M are:

$$Ancestors(L, M) = Gen(L) \cap Gen(M)$$
$$= \{A, B, D, H, L\} \cap \{A, B, D, H, M\} = \{A, B, D, H\}$$

The similarity of the concept L (respectively M) with regard to the common ancestors with the concept M (respectively L) is express by the following formula:

$$\frac{Card(Ancestors(L, M))}{Card(Gen(M))}$$

The genealogic conceptual similarity *ProxiGenea* of L and M is defined as:

$$ProxiGenea(L, M) = \frac{Card(Ancestors(L, M))}{Card(Gen(M))} * \frac{Card(Ancestors(M, L))}{Card(Gen(L))}$$

$$ProxiGenea(L, M) = \frac{Card(Ancestors(L, M))^2}{Card(Gen(M)) * Card(Gen(L))}$$

This principle is similar to the one used in other semantic similarity such as the ones defined by Wu and Palmer [2] or Lin [3] Our function is a variant.

We defined the similarity between a document and a query as the weighted mean of the conceptual similarities of the annotations involved in the query:

$$Similarity(A_{req}, A_{doc}) = \frac{\sum_{i=0}^{taille(A_{req})} Coef[i] * ProxiGenea(A_{req}[i], A_{doc}[i])}{\sum_{i=0}^{taille(A_{req})} Coef[i]}$$

4. PROTOTYPE

Figure 4 presents the DYNAMO prototype. The user submits his query in natural language ("Démarrage difficile à froid " in figure 3 ["*Difficult to start the engine when it is cold]*") in the top-left part of the window. The query is automatically translated into a semantic graph of annotation and is represented in the top-right part of the window. The list of the documents the system retrieves is presented in the bottom-left part of the screen. The first document has been selected by the user and its content is presented in the bottom-right part of the screen.

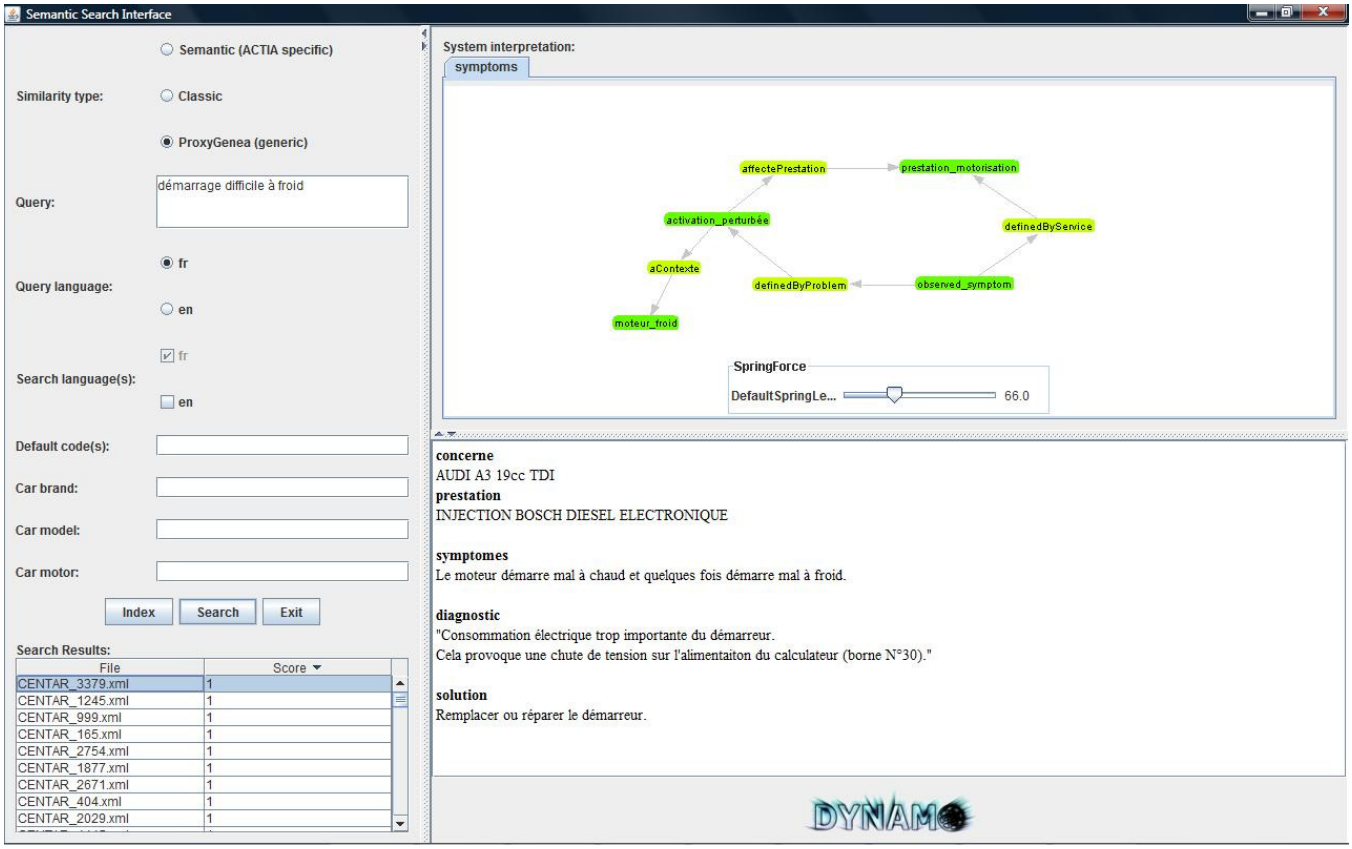

Figure 4 : Dynamo – Search part

5. IMPACT

Currently, three partners are associated with the project to test the application. The experimental frameworks are: the assistance with automobile diagnosis (ACTIA), the assistance with data-processing projects follow-up (ARTAL) and the capitalization and the re-use of experiments (Préhistoire et technologies).

6. ACKNOWLEDGMENTS

We would like to thank the ANR pour its financial support to the project, as well as the other participants to the project: research teams IC3 and SMAC from IRIT lab.

7. REFERENCES

[1] Hubert G., Mothe J., Ralalason B., Ramamonjisoa A.. Modèle d'indexation dynamique à base d'ontologies, Conférence francophone en Recherche d'Information et Applications (CORIA 2009), 169-184, 2009.

[2] Wu Z., Palmer M., Verb semantics and lexical selection, Proceedings of the 32nd annual meeting of the Association for Computational Linguistics, 133-138, 1994.

[3] Lin D. An information-theoretic definition of similarity. 15th International Conference On Machine Learning, 1998.