SEMANTIC-BASED COMPOSITION OF MODULAR ONTOLOGIES APPLIED TO WEB QUERY REFORMULATION

Manel Elloumi-Chaabene[1], Nesrine Ben Mustapha[1], Hajer Baazaoui-Zghal[1],
Antonio Moreno[2] and David Sánchez[2]
[1] *Riadi-GDL Laboratory, ENSI Campus Universitaire de la Manouba, 2010, Tunis, Tunisie*
[2] *ITAKA Research Group, Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili*
Av. Països Catalans, 26, 43007, Tarragona, Spain

Keywords: Ontology composition, Modular ontology, Module clustering, Semantic measures.

Abstract: When a Web query is submitted and relevant documents are not found, users are faced with the difficult task of reformulating the query. It may be argued that ontologies can be useful to automate the query reformulation process, taking advantage of the domain knowledge. This paper proposes a novel way of building a modular ontology that may be suitable for this task, composed of interrelated modules focused on specific topics. We propose to use well-known Web-based semantic relatedness measures to improve the content and structure of the modular ontology. Some experiments on query reformulation based on the obtained ontology modules show satisfactory results.

1 INTRODUCTION

With the growing amount of available Web resources, users are faced with the task of finding the precise information suited for their purpose. Users frequently modify a previous search query to have better results. These modifications are called *query reformulations* or query refinements.

Ontologies have proven to be useful to interpret queries and documents if they are adapted to the particular needs of the search process. Big ontologies usually do not cover a specific field of search. With smaller and more compact ontologies, it can be easier for the user to find the appropriate terms to make or refine a search, restricted to a specific field in which terms are semantically linked. Therefore, *ontology modularization* processes can have a significant contribution in this field. Existing methods on this field are based only on taxonomic relationships between terms and do not consider the semantics of concepts. Thus, their results have a poor taxonomic structure and lack non-taxonomic relationships. This work proposes the composition of modular ontologies based on semantic relatedness measures, and shows how these modules can be used to improve the reformulation of Web queries. The main contributions of this work are the definition of a Web-based semantic modular ontology building algorithm and the use of modular ontologies for query reformulation.

The next section presents a brief state of the art on compositional approaches.

2 COMPOSITION OF ONTOLOGY MODULES

In this paper we take the expression *ontology module* to refer to a fragment of a domain ontology that may be reused in different tasks. It represents a set of concepts which are strongly interrelated. A *modular ontology* includes a set of independent modules, which are linked through *inter-module connectors* (Stuckenschmidt et al., 2009). An inter-module connector is defined by a relationship between two concepts belonging to two modules within the ontology.

Ontology composition approaches can be classified into three categories. The first one includes techniques based on an algebra (Jarrar, 2005) (Mitra et al., 2004), in which a composition operator has been defined to put modules together.

Elloumi-Chaabene M., Ben Mustapha N., Baazaoui-Zghal H., Moreno A. and Sánchez D..
SEMANTIC-BASED COMPOSITION OF MODULAR ONTOLOGIES APPLIED TO WEB QUERY REFORMULATION.
DOI: 10.5220/0003536403050308
In *Proceedings of the 6th International Conference on Software and Database Technologies* (ICSOFT-2011), pages 305-308
ISBN: 978-989-8425-76-8

The second one is based on the object-oriented paradigm (Henriksson et al., 2007). Thus, role modelling for ontologies helps to remedy the deficiencies of class-based ontologies without losing their advantages. The last category is based on description logics (Stuckenschmidt et al., 2009).

This state of art has identified the following limitations. First, it shows a lack of appropriate approaches for obtaining modular ontologies usable for Web semantic search. Second, the composed modular ontologies are not well structured. Finally, ontology modules have a big size which make it hard to the user to choose the appropriate search terms. To remedy these problems, we propose a novel method to compose ontology modules using well known techniques like clustering and Web-based co-occurrence statistics. The final aim is to unravel the implicit taxonomic and non-taxonomic relationships between the concepts in the ontology modules. Besides, the method intends to improve the modular ontology structure and to reduce the size of the modules, in order to improve the process of query reformulation.

3 A NEW METHOD FOR THE COMPOSITION OF MODULAR ONTOLOGIES

The final objective of the work is to improve the query reformulation process using a modular ontology. Our hypothesis is that it may be more beneficial to the user to specify some strongly related concepts and relations. Thus, we propose to analyze, improve and re-structure these modules to obtain a coherent modular ontology. This section aims to describe the proposed compositional approach based on clustering and co-occurrence similarity measures. We start in the following subsection with a general description of the approach, and then we turn to a detailed explanation of each step.

3.1 General Description of the Proposed Approach

Figure 1 depicts the main phases of the new method to compose ontology modules. The first step deals with the reorganization of the ontology modules. The second step concerns the classification of these concepts into modules based on their semantic similarity, computed from Web-scale co-occurrence statistics. The third and last step ensures a proper structure of the ontological modules.

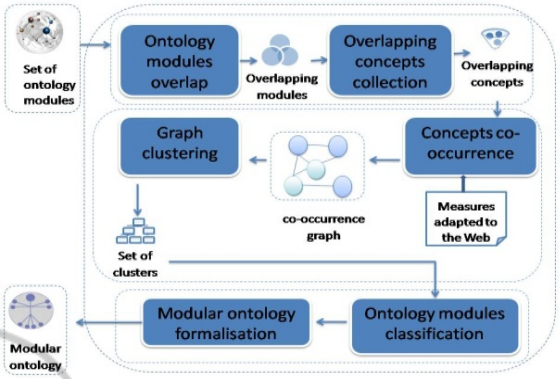

Figure 1: Global presentation of the approach.

3.2 Ontology Modules Reorganization

The reorganization of the ontology modules is performed in two sub-steps. The first one aims at determining the overlap between the different modules. The second one consists in reducing the size of the modules by collecting the overlapping concepts for each pair of modules. We propose to store a set of overlapping concepts *for each pair of modules*, and not just a single set of overlapping concepts.

3.3 Concept Classification using Web-based Semantic Measures

The classification of concepts allows a meaningful redistribution of the concepts within the modules. In fact, we propose in this section to form new modules with new semantic relations based on the co-occurrence between concepts.

3.3.1 Web-based Semantic Similarity between Concepts

This sub-section aims to uncover semantic relations between concepts using the Web.

The co-occurrence between concepts may be determined from the hits returned by Web search engines. To calculate the semantic similarity between terms based on co-occurrence, the well-known PMI (Pointwise Mutual Information) measure was applied. This measure is computed with the following formula (1):

$$PMI\_IR(a,b) = \log_2 \left(\frac{\frac{hits(a\ AND\ b)}{total\_webs}}{\frac{hits(a)}{total\_webs} \times \frac{hits(b)}{total\_webs}} \right) \quad (1)$$

This measure is then represented in a co-occurrence graph. Let $G = (V, E)$ be a co-occurrence graph (Newman, 2005), where $V = \{v_1,\ldots, v_N\}$ is a set of vertices and $E \subseteq V \times V$ is a set of edges. Each concept is represented by a node. The edge between nodes v_1 and v_2 is included in the graph if the degree of relatedness between those two concepts, as measure by the PMI estimation, is above a certain threshold, which basically determines the intended density of the graph. The co-occurrence graph constitutes the input of the clustering algorithm which will be detailed in the next sub-section.

3.3.2 Clustering based on the Co-occurrence Graph

Due to the number of clusters is unknown, it is necessary to employ an unsupervised clustering algorithm. In this case, the Newman clustering algorithm was chosen. Newman and Girvan (Newman, 2005) impose weights on the edges based on structural properties of the graph G. The weight used by Newman and Girvan is the between's of an edge $\{v_1, v_2\}$, which is the number of shortest paths connecting any pair of vertices that pass through the edge. The cost of a path is calculated as the sum of the PMI values of the edges of the path. As a computational detail, it may be noted that there may exist multiple paths of the same length between a given pair of vertices.

This step has ensured the organization of semantically related concepts within the same module, but lacks a proper structuring of the whole set of modules. In the next section, we describe the modules classification process using Web-based semantic relationships.

3.4 Ontology Modules Structuring

This section describes the last phase, in which the final modular ontology is properly structured and formally represented. This phase basically aims to find relationships between concepts in different modules. Besides, it has the objective to obtain a hierarchy of modules that may be formalized using P-DL.

3.4.1 Ontology Modules Classification

The clusters (modules) obtained with the Newman algorithm in the previous section, are classified with the vector model to create the structure of the final modular ontology. This process consists of two phases: term weighting and similarity computation.

Thus, the space is of dimension K, being K the number of common concepts between the overlapping modules. In addition, we define a *vector concept* $VC_i = (w_i(c_1),\ldots, w_i(c_k))$ that represents the common concepts between the modules. $w_i(c_k)$ is the weight of concept c_k in the Web. This weight is determined by the inverse occurrence of the concept c_k in the Web.

$$w_i(c_k) = \frac{1}{hits\ (c_k)} \tag{2}$$

Each module is represented by a *vector module* $VM_j = (w_{ij}(c_1),\ldots,w_{ij}(c_k))$, where $w_{ij}(c_k)$ is the average of the PMIs of the concept c_k with the concepts of the module M_j. $w_{ij}(c_k)$is described by the formula (5), where N is the number of concepts in the module M_j.

$$w_{ij}(c_k) = \frac{\sum_{l=1}^{l=N} PMI\ (c_k, c_l)}{N} \tag{3}$$

Then, it is calculated the cosine between the vector modules and the vector concept to classify the modules.

$$Sim\ (M_j, VC_i) = \frac{\sum_{l=1}^{k} w_{ij}(c_l)\ w_i(c_l)}{\sqrt{\sum_{l=1}^{k} w_{ij}(c_l)^2 \bullet \sum_{l=1}^{k} w_i(c_l)^2}} \tag{4}$$

Next, we propose to formalize these modules into a modular ontology, as detailed in the following section.

3.4.2 Modular Ontology Formalisation

We detail in this subsection the last step of module composition, which consists on the formalisation of the modular ontology with P-DL (Bao et al., 2007). P-DL, a description logic based on packages, uses the importation of relations in order to compose a local model.

4 EVALUATION

In order to evaluate the approach presented in this paper, the impact of the use of ontology modules during query reformulation is tested. We have computed the precision of results retrieved by means of query reformulation.

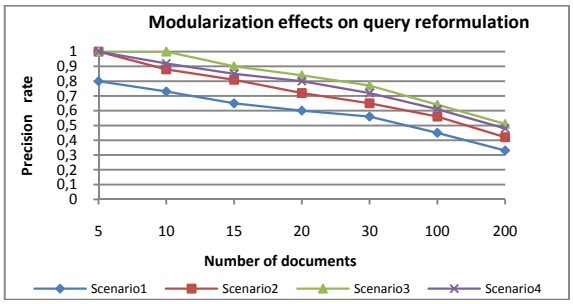

Figure 2: Modularization effects on query reformulation.

In the *first scenario*, documents are retrieved via queries proposed from the INEX (INEX, 2010) topics without performing any query reformulation. The *second scenario* consists on reformulating the queries by means of the input ontologies provided by INEX. The *third scenario* consists on reformulating queries using concepts of the same module than the queried concept. The *fourth scenario* consists on reformulating queries using concepts related to the queried concept in other modules (these relationships are obtained in the third step of the proposed approach). In each scenario, the amount of implicit and explicit knowledge considered during the query reformulation increases.

5 CONCLUSIONS

The challenge addressed in this paper is to propose an approach to improve query reformulation based on ontology modularization. Indeed, the use of ontological modules created by means of case-based reasoning has improved the relevance of results (Elloumi et al., 2010). Given that the approaches of modularization of ontologies are based mainly on the structure of the ontology, the resulting modules only rely on explicitly modelled relationships. In order to consider as much explicit and implicit knowledge between modules, we propose to estimate concept relatedness from term co-occurrence in the Web. As a result, better precision is achieved by query reformulation tasks relying on these better structured modules.

ACKNOWLEDGEMENTS

This work has been supported by the Spanish-Tunisian AECID project A/030058/10, “A framework for the integration of Ontology Learning and Semantic Search”.

REFERENCES

Bao J., Slutzki G., and Honavar V., «A Semantic Importing Approach to Knowledge Reuse from Multiple Ontologies». *In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, July 22-26, 2007, Vancouver, British Columbia, Canada, AAAI Press, 2007, p. 1304–1309, 2007.

Elloumi-Chaabene M., Ben Mustapha N., Baazaoui-Zghal H., Moreno A. and Sánchez D. «Evolutive content-based search system- Semantic Search System based on Case-based-Reasoning and Ontology Enrichment». *In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2010, p. 24-34.

Henriksson J., Assmann U., Johannes J., and Zschaler S., «Reuseware - Adding modularity to your language of choice». *In Proceedings of Technology of Object-Oriented Languages and Systems Europe 2007*, Zurich, Switzerland (24th–27th June 2007), 2007.

Jarrar M., Towards Methodological Principles for Ontology Engineering. PhD thesis, Vrije Universiteit Brussel, 2005.

Mitra P. and Wiederhold G., «An Ontology-Composition Algebra. *International Handbooks on Information Systems*. Springer-Verlag, handbook on ontologies edition», pages 93-117, 2004.

Newman, M. E. J. A measure of Betweenness Centrality based on Random Walks. *In Social Networks*, 27, pp. 39–54, 2005.

Stuckenschmidt H. et al. (Eds.): Modular Ontologies, LNCS 5445, pp. 321–347, 2009.

INEX, Overview of the INEX 2010 Ad Hoc Track, http://staff.science.uva.nl/~kamps/publications/2010/arvo:over10.pdf, 2010.