Artificial Intelligence Master (URV, UPC, UB)
Master Thesis Proposal

Automatic refinement of ontologies through the analysis of the results of an ontology-based information retrieval system

Directors: Dr Antonio Moreno, Dr David Sánchez
Research Group: ITAKA-Intelligent Technologies for Advanced Knowledge Acquisition (http://deim.urv.cat/~itaka) - Universitat Rovira i Virgili

Given the tremendous growth rate of the World Wide Web, it is increasingly necessary to develop appropriate tools that can assist users looking for information. The techniques used by search engines are based on statistical and syntactical analysis, and do not consider the semantics contained in the user's request and the available documents. Adding semantics to the data, as suggested in the Semantic Web field, will improve the process of information retrieval. Ontologies [1] are an explicit and formal specification of a shared conceptualization, and have contributed to the emergence of semantic search engines. Ontologies can be used to reformulate and refine queries, annotate information, index documents, etc. However, one of the problems of semantic search engines is the availability of ontologies to formalize a specific domain. The manual construction of an ontology is usually a long and expensive process. One possible solution is the analysis of texts for the automatic construction of ontologies, despite the difficulties inherent in the exploration of textual content. The texts available on Web pages provide a wealth of information, although not necessarily relevant or structured. You have to select and treat a large volume of information to obtain a complete and correct ontology.

An interesting idea to be explored is to benefit from the quality of the results of an ontology-based information retrieval system, considering it as a filtered corpus with relevant domain documents that can be analysed in order to improve or refine the same domain ontology that has been used to guide the process of information retrieval. The bottomline is that the system could automatically improve the ontology every time that the user makes a search and, at the same system, the improvement of the ontology would provoke an increase in the quality of the results delievered in the next search. In that way, a “virtuous circle” would be constructed, in which the ontology guides the search and the search results improve the ontology.

This work is framed in a starting collaboration between the ITAKA research group of URV (which has experience on the ontology learning area, [2]) and the RIADI-GDL Laboratory of the National School of Computer Sciences at the University of Manouba in Tunisia (that has worked on semantic information retrieval, [3]).

The main objectives of the Master Thesis are the following:

- To develop a comprehensive state of the art on the use of ontology learning methods on the results of information retrieval systems, or more generally on the combination of ontology learning and information retrieval.
- To design new ontology learning methods (or think how to use existing ontology learning methods) that allow to improve/refine a domain ontology through the analysis of the results of an ontology-based information retrieval system and other parameters which may be available (e.g. the user query, the user rating of the retrieved results, etc).
- To implement the methods considered in the previous step in a prototype, which could be joint at a later stage to an independently developed semantic information retrieval system.
- To validate the resulting system with tests in different domains, starting with ontologies of different sizes.

References

[1] Fensel, D. Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, 2003.

[2] Sánchez, D. Domain ontology learning from the Web. PhD thesis, Univ. Politècnica de Catalunya, 2007.

[3] 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, Vol.6, No.4, pp. 309-336, 2007.