Domain Ontology Extractor from Wikipedia's Categories Structure in Portuguese Language

Clarissa Castellã Xavier[1], Vera Lúcia Strube de Lima[1]

[1] Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) – Porto Alegre – RS – Brasil
{clarissa.xavier, vera.strube}@pucrs.br

Abstract. In this abstract we describe a web based domain ontology extractor in Portuguese language, developed according to the semi-automatic method for extracting domain ontologies from Wikipedia's category structure developed by the authors.

Keywords: ontologies extraction, Wikipedia.

1 Description

Wikipedia has demonstrated to be a very interesting source for ontologies extraction, due to the large amount of organized content in it, freely available and covering a wide range of issues. Therefore, we have developed a web based prototype that performs the extraction of a Tourism ontology in Portuguese language from Wikipedia's category system.

The prototype was built in order to validate a semi-automatic method of domain ontologies extraction from Wikipedia's categories structure developed by the authors in [1]. It was implemented in PHP[1], same programming language of Mediawiki system. It access Portuguese Wikipedia's database in MySql, obtained from the Wikipedia's Download Website[2] and generate a Tourism ontology described in OWL language, recommended by W3C since January 2004 as the standard of the Semantic Web project.

To validate the structure generated by the prototype we have developed a web based program in PHP that calculates the following metrics: Precision, Recall and F-Measure. This ontology evaluator receives as inputs two OWL files: one containing the ontology that will be evaluated and other with the golden map. The output are the metrics, and the main similarities and differences between the two ontologies.

To evaluate the Tourism ontology generated by the prototype, a reference model (Golden Map) was drawn manually from the structure of the Tourism category in the Portuguese Wikipedia, revised and refined by a linguist. The obtained results are

[1] www.php.net
[2] http://download.wikimedia.org/backup -index.html

promising, demonstrating the feasibility of extracting domain ontologies in Portuguese from the categories of Wikipedia, through the proposed method.

1.1 Architecture

The prototype architecture, illustrated in Figure 1, follows the four steps of the method being validated. In the first stage (E1) is performed the taxonomic structure selection from the Tourism category. Then in the second stage (E2), located-in relations, new classes and instances are obtained from the taxonomy generated in the previous step. In the third stage (E3) is performed the standardization of classes and instances names, and finally the fourth stage (E4) generates the OWL file.

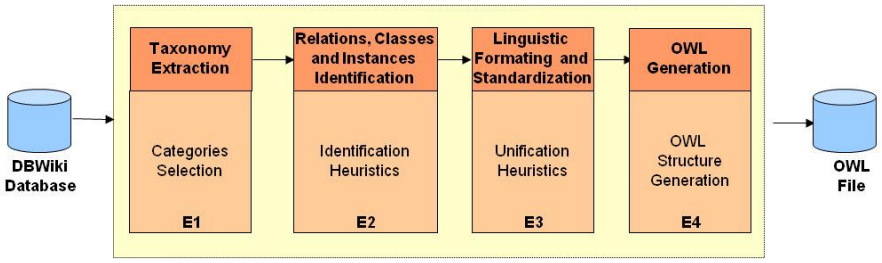

Fig. 1. Software Architecture.

1.2 Demonstration

To demonstrate the prototype and the evaluator, we need to perform its execution in a web browser. For this purpose, the system must be installed on a machine containing Apache, PHP 5, MySQL with MediaWiki's database populated with Portuguese Wikipedia data.

After that, we suggest to open the OWL file containing the ontological structure generated using a simple text editor, as Windows Notepad or Linux Kate, and using a graphic ontology editor, as Protegé.

All the software and hardware needed can be provided by the authors.

References

1. Xavier, C.C.: Um método semi-automático para extração de estruturas ontológicas de domínio a partir das categorias da Wikipédia em língua portuguesa. Master Thesis, PPGCC-PUCRS (2010).