Chapter IV

Ontology Learning from a Domain Web Corpus*

Roberto Navigli
University of Rome “La Sapienza”, Italy

ABSTRACT

Domain ontologies are widely recognized as a key element for the so-called semantic Web, an improved, “semantic aware” version of the World Wide Web. Ontologies define concepts and interrelationships in order to provide a shared vision of a given application domain. Despite the significant amount of work in the field, ontologies are still scarcely used in Web-based applications. One of the main problems is the difficulty in identifying and defining relevant concepts within the domain. In this chapter, we provide an approach to the problem, defining a method and a tool, OntoLearn, aimed at the extraction of knowledge from Websites, and more generally from documents shared among the members of virtual organizations, to support the construction of a domain ontology. Exploiting the idea that a corpus of documents produced by a community is the most representative (although implicit) repository of concepts, the method extracts a terminology, provides a semantic interpretation of relevant terms and populates the domain ontology in an automatic manner. Finally, further manual corrections are required from domain experts in order to achieve a rich and usable knowledge resource.

INTRODUCTION

The next generation of the World Wide Web, the so-called semantic Web, aims at improving the “semantic awareness” of computers connected via the Internet. The semantic Web requires that information be given a well-defined meaning through a machine-processable representation of the world, often referred to as an ontology.
The goal of a domain ontology is to reduce (or eliminate) the conceptual and terminological confusion among the members of a virtual community of users (for example, tourist operators, commercial enterprises, computer scientists) that need to share electronic documents and information of various kinds. This is achieved by identifying and properly defining a set of relevant concepts that characterize a given application domain. An ontology is therefore a shared understanding of some domain of interest (Uschold, 1996). In other words, an ontology is an explicit, agreed specification about a shared conceptualization. The construction of such a shared understanding, that is, a unifying conceptual framework, fosters communication and cooperation among people, interoperability among systems, system engineering benefits (reusability, reliability, and specification), and so forth.

Ontologies may have different degrees of formality but they necessarily include a vocabulary of terms with their meaning (definitions) and their relationships. Thus the construction of an ontology requires a thorough domain analysis that is accomplished by:

- Carefully identifying the vocabulary that is used to describe the relevant concepts within the domain;
- Coding complete and rigorous definitions about the terms (concepts) in the vocabulary;
- Characterizing the conceptual relations among those terms.

The definition of the basic kinds and structures of concepts that are applicable in every possible domain usually requires the joint work of specialists from several fields, like philosophical ontologists and Artificial Intelligence researchers. The issue of identifying these very few “basic” principles, referred to as the top ontology (TO), is not a purely philosophical one, since there is a clear practical need of a model that has as much generality as possible, to ensure reusability across different domains (Smith et al., 2001).

Domain modelers and knowledge engineers are involved in the task of identifying the key domain conceptualizations, and describing them according to the organizational backbones established by the top ontology. The result of this effort is referred to as the upper domain ontology (UDO), which usually includes a few hundred application-domain concepts.

While many ontology projects eventually succeed in the definition of an upper domain ontology, very few projects can overcome the actual barrier of populating the third level that we call the specific domain ontology (hereafter, SDO), at the price of inconsistencies and limitations. On the other hand, general-purpose resources like Wordnet (Fellbaum, 1995), Cyc (Lenat, 1993) and EDR (Yokoi, 1993) — while dealing with thousands of concepts — do not encode much of the domain knowledge needed by specialized applications like information retrieval, document management, (semantic) Web services and so on.

Figure 1 reports the three levels of generality of a domain ontology.

It turns out that, although domain ontologies are recognized as crucial resources for the semantic Web (Berners-Lee, 1999), in practice, full-fledged resources are not available, and when available they are not used outside specific research or community environments.

We identify three features needed to build usable ontologies: coverage, consensus and accessibility.
Blending Virtual Campuses: Managing Differences Through Web 2.0 Experiences in Transnational Cooperation Projects
www.igi-global.com/chapter/blending-virtual-campuses-managing-differences/37709?camid=4v1a