Chapter 5

Construction of Domain Ontologies: Sourcing the World Wide Web

Jongwoo Kim
University of Massachusetts Boston, USA

Veda C. Storey
Georgia State University, USA

ABSTRACT

As the World Wide Web evolves into the Semantic Web, domain ontologies, which represent the concepts of an application domain and their associated relationships, have become increasingly important as surrogates for capturing and representing the semantics of real world applications. Much ontology development remains manual and is both difficult and time-consuming. This research presents a methodology for semi-automatically generating domain ontologies from extracted information on the World Wide Web. The methodology is implemented in a prototype that integrates existing ontology and web organization tools. The prototype is used to develop ontologies for different application domains, and an empirical analysis carried out to demonstrate the feasibility of the research.

INTRODUCTION

The World Wide Web is a massively distributed reservoir of information, but the information does not have well-defined machine-understandable meaning attached to it, prohibiting automated manipulation and reasoning about such information (Ram & Zhao, 2007). The next generation of the World Wide Web, the Semantic Web, is intended to enable more intelligent use of data and information for effective electronic interoperability and collaboration (Horrocks, 2008). A successful Semantic Web, however, depends upon the ability to manage, integrate, and analyze data and is driven by the role of semantics for automated approaches to exploiting Web resources (Berners-Lee, Hendler, & Lassila, 2001). Ontologies, which are at the heart of the Semantic Web, define the concepts and relationships that make global interoperability possible, facilitate sharing and integration (Hor-
Construction of Domain Ontologies

rocks, 2008; Leukel & Sugumaran, 2009; Tun & Tojo, 2008) and serve as surrogates for semantics. Ontologies are also useful for digital libraries and personalized information management (Katifori, Halatsis, Lepouras, Vassilakis, & Giannopoulou, 2007; Meng & Chatwin, 2010). Although their need is well-documented, ontology development is often performed manually and is challenging and time-consuming (Ding & Foo, 2002; Farquhar, Fikes, & Rice, 1997). One of the major reasons for this difficulty is finding relevant knowledge sources to use to create ontologies.

The World Wide Web is a great resource of information for almost all imaginable domains. If this information could be properly extracted and organized, it should be possible to effectively use it to create domain ontologies, especially if a process to do so could be automated to some extent (Sánchez & Moreno, 2008). The objectives of this research, therefore, are to:

- Develop a methodology for semi-automatically generating domain ontologies by extracting and organizing terms and relationships among those terms using the World Wide Web as a source;
- Establish the feasibility of the ontology creation methodology by creating a prototype; and
- Assess the performance of the methodology through an empirical analysis.

The contribution of the research is to develop a way to semi-automatically create domain ontologies by using the World Wide Web as a source and integrating web tools. Libraries could be used for the Semantic Web and other applications (e.g., heterogeneous databases, conceptual modeling, and web queries (Horrocks, 2008; Ram & Zhao, 2007)).

The section below examines related research on domain ontologies and their role in the Semantic Web. A six-step ontology creation methodology is presented in the section following. Later, this chapter details the implementation of the methodology in a prototype, WebtoOnto. It then evaluates the methodology using an empirical study. A summary and concluding remarks are found in the last section.

RELATED RESEARCH

Ontologies

An ontology is a way of describing one’s world and can be used as a surrogate for semantics (Dahlgren, 1995). An ontology represents a set of concepts and the relationships among them for a specific domain. Ontologies have been developed in both Artificial Intelligence and knowledge management research to facilitate knowledge use and reuse with the main idea being to develop an understandable, complete, and sharable system of categories, labels, and relationships that represent the real world in an objective manner (Cristani & Cuel, 2005; Horrocks, 2008; March & Allen, 2007). They are useful because they formalize a shared view of a domain (Grootjen & van der Weide, 2005). An example of an ontology for carpel tunnel syndrome (resulting from repetitive stress) created by our proposed methodology is shown in Figure 1.

There are a number of challenges to developing ontologies. Ontologies are specific to each domain and are time-consuming to create (Herman, 2007; Maedche & Staab, 2000). Large-scale ontologies such as Cyc require a collaborative, community effort from knowledgeable people. Applications can be developed with small, domain specific ontologies (Herman, 2007), the creation of which is the focus of this research.

Organizations may use existing documents for domain ontology creation (Kietz, Maedche, & Volz, 2000; Alexander Maedche & Staab, 2000; Sugiura, Kurematsu, Fukuta, Izumi, & Yamaguchi, 2003). However, when they start a new business or expand an existing one, they may not have