Extending Metadata Definitions by Automatically Extracting and Organizing Glossary Definitions

Eduard Hovy[1], Andrew Philpot[1], Judith Klavans[2], Ulrich Germann[1], Peter T. Davis[2]

[1] Digital Government Research Center
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292

[2] Digital Government Research Center
Department of Computer Science
Columbia University
535 West 114th Street, MC 1103
New York, NY 10027

Abstract

Metadata descriptions of database contents are required to build and use systems that access and deliver data in response to user requests. When numerous heterogeneous databases are brought together in a single system, their various metadata formalizations must be homogenized and integrated in order to support the access planning and delivery system. This integration is a tedious process that requires human expertise and attention. In this paper we describe a method of speeding up the formalization and integration of new metadata. The method takes advantage of the fact that databases are often described in web pages containing natural language glossaries that define pertinent aspects of the data. Given a root URL, our method identifies likely glossaries, extracts and formalizes aspects of relevant concepts defined in them, and automatically integrates the new formalized metadata concepts into a large model of the domain and associated conceptualizations. This demo will show the end-to-end performance of this system.

The demonstration will show the concept acquisition and placement process. Using the AskCal interface (Philpot et al, 2002), we will ask the viewer to interact with the system to retrieve some data. We will then introduce a new topic, one for whom there is no concept yet in the ontology. We will browse the ontology to verify this. We will identify candidate glossary web pages or sites using prior work (Klavans et al, 2002) and/or a web-accessible term finder. In a separate window, we will then display a glossary-containing file from www.eia.gov or similar, in which the concept is defined in text. Using the system described in the accompanying paper, we will enter the root URL and activate the glossary analysis and concept alignment procedures. Upon conclusion, the system will announce the acquisition and placement of as many concepts as it has found. We will then re-display the ontology in the browser. Newly acquired concepts will be displayed in a different color. The user will be free to click on the concepts in order to examine the formalized contents, as well as hyperclick back to the glossary source page for verification. This demo will thus illustrate not only our latest work, but the major part of the EDC system that we have been building over the past 4 years.

Klavans, J.L., P.T. Davis, and S. Popper. 2002. Building Large Ontologies using Web-Crawling and Glossary Analysis Techniques. *Proceedings of the NSF's National Conference on Digital Government dg.o 2002.* Los Angeles, CA.

Philpot, A., J.L. Ambite, and E. Hovy. 2002. DGRC AskCal: Natural Language Question Answering for Energy Time Series. *Proceedings of the NSF's National Conference on Digital Government dg.o 2002.* Los Angeles, CA.