Chapter IX

Ontology Extraction and Conceptual Modeling for Web Information

Hyoil Han
The University of Texas at Arlington, USA

Ramez Elmasri
The University of Texas at Arlington, USA

ABSTRACT

A lot of work has been done in the area of extracting data content from the Web, but less attention has been given to extracting the conceptual schemas or ontologies of underlying Web pages. The goal of the WebOntEx (Web ontology extraction) project is to make progress toward semiautomatically extracting Web ontologies by analyzing a set of Web pages that are in the same application domain. The ontology is considered a complete schema of the domain concepts. Our ontology metaconcepts are based on the extended entity-relationship (EER) model. The concepts are classified into entity types, relationships, attributes, and superclass/subclass hierarchies. WebOntEx attempts to extract ontology concepts by analyzing the use of HTML tags and by utilizing Part-of-Speech tagging. WebOntEx applies heuristic rules and machine learning techniques, in particular, inductive logic programming (ILP).
INTRODUCTION

The amount of data on the Web is growing tremendously, and management of this huge amount of data is not easy, particularly because most Web data lack explicit structure. People try to search and manage data on the Web, and many researchers are working on the integration of data from heterogeneous data sources. Most of this work assumes that the schema of Web data is given in advance. However, most data on the Web does not have a predefined structure, and even if it is well structured, the structure is not known in advance. Users usually do not know the source of data on the Web and the structure of the data can be a proprietary schema. Knowing the schema is the first required step to manage, integrate, analyze, or query data on the Web. Our work is to try to capture semiautomatically the structure and ontology of Web data, to model them using the extended entity-relationship (EER) model (Elmasri & Navathe, 2000) and to build an ontology database for Web applications that is stable, evolvable, and scalable.

The ontology (or conceptual structure) of a particular application domain changes over time. Currently, ontologies are designed by domain experts. This standard ontology building process is slow and inflexible. On the other hand, the goal of our WebOntEx (Web ontology extraction) system is to semiautomatically extract ontologies from Web pages and store them in a relational database so that an ordinary user can use WebOntEx to create an ontology, save it, and utilize it easily in an application domain.

The extracted ontologies can be used in various important applications, such as understanding Web information content, querying Web metadata, more intelligent Web searching, and automatic conversion of HTML Web pages to other formats, such as XML (Bray, Paoli, Sperberg-McQueen, & Maler, 2000). This chapter describes the WebOntEx (Web ontology extraction) project, its architecture, and each component module.

WEB DATA EXTRACTION VERSUS SCHEMA (ONTOLOGY) EXTRACTION

A lot of work has been done in the area of extracting data content from the Web, but less attention has been given to extracting the types or conceptual schema of underlying Web data, which is the goal of WebOntEx.

Some research related to extracting data from the Web is accomplished using wrappers, which basically provide interfaces for extracting information from Web pages (Ashish & Knoblock, 1997; Adelberg, 1998; Constantopoulos, Christophides, & Plexousakis, 2001). When extracting data from Web pages,