Capturing Semantics in XML Documents

Tok Wang Ling
Department of Computer Science
National University of Singapore
3 Science Drive 2
Singapore 117543
lingtw@comp.nus.edu.sg

Abstract

Traditional semantic data models, such as the Entity Relationship (ER) data model, are used to represent real world semantics that are crucial for the effective management of structured data. The semantics that can be expressed in the ER data model include the representation of entity types together with their identifiers and attributes, n-ary relationship types together with their participating entity types and attributes, and functional dependencies among the participating entity types of relationship types and their attributes, etc.

Today, semistructured data has become more prevalent on the Web, and XML has become the de facto standard for semi-structured data. A DTD and an XML Schema of an XML document only reflect the hierarchical structure of the semistructured data stored in the XML document. The hierarchical structures of XML documents are captured by the relationships between an element and its attributes, and between an element and its subelements. Element-attribute relationships do not have clear semantics, and the relationships between elements and their subelements are binary. The semantics of n-ary relationships with n > 2 cannot be represented or captured correctly and precisely in DTD and XML Schema. Many of the crucial semantics captured by the ER model for structured data are not captured by either DTD or XML Schema. We present the problems encountered in order to correctly and efficiently store, query, and transform (view) XML documents without knowing these important semantics. We solve these problems by using a semantic-rich data model called the ***O***bject, ***R***elationship, ***A***ttribute data model for ***S***emi***S***tructured Data (ORA-SS). We briefly describe how to mine such important semantics from given XML documents.