Discovering Semantics from Data-Centric XML

Luochen Li¹, Thuy Ngoc Le¹, Huayu Wu², Tok Wang Ling¹, and Stéphane Bressan¹

¹ School of Computing, National University of Singapore
{luchen,ltngoc,lingtw,step}@comp.nus.edu.sg
² Institute for Infocomm Research, Singapore
huwu@i2r.a-star.edu.sg

Abstract. In database applications, the availability of a conceptual schema and semantics constitute invaluable leverage for improving the effectiveness, and sometimes the efficiency, of many tasks including query processing, keyword search and schema/data integration. The Object-Relationship-Attribute model for Semi-Structured data (ORA-SS) model is a conceptual model intended to capture the semantics of object classes, object identifiers, relationship types, etc., underlying XML schemas and data. We refer to the set of these semantic concepts as the ORA-semantics. In this work, we present a novel approach to automatically discover the ORA-semantics from data-centric XML. We also empirically and comparatively evaluate the effectiveness of the approach.

1 Introduction

To improve the conceptual quality, we need to discover the intended semantics in the logical XML schemas and data. This requires finding such semantic information as object classes, relationship types, object identifiers (OIDs), etc., as present in conceptual models for semi-structured data such as Object-Relationship-Attribute for Semi-Structured data (ORA-SS) [6]. We refer to this semantics as the ORA-semantics. Once discovered, the ORA-semantics is useful not only for users to understand the data and schemas but also for improving both the effectiveness and efficiency of processing. Let us use the XML document in Fig. 1 to illustrate how the availability of such semantics help applications.

XML Query Processing
To process an XPath query, e.g. //Student[Matric# =‘HT001’]/Name, most approaches match the query pattern to the data to find all occurrences. However, if we have the semantics that Matric# is the OID of student, after getting an answer, we can stop searching the rest of data.

XML Keyword Search
The use of semantics in current keyword search approaches [7] is still on object level. For a query {CS5201, CS5208} to find common information of two courses, only by knowing there is a relationship type between object classes Student and Course, one can infer the meaningful answer should be all students taking these two courses. Otherwise, the root node will be returned by most LCA-based XML keyword search approaches [12].
Schema/Data Integration

Most existing approaches integrate elements based on their structural and linguistic similarities. *Grade* is an attribute of the relationship type between *Course* and *Student*. Without this semantics, when we integrate this schema with another, in which *Student* has an object attribute *Grade* which means the year of his study in school, we may wrongly integrate these two different attributes with the same attribute name *Grade* and the same parent node *Student*, because of their high structural and linguistic similarities.

![Fig. 1. An XML data tree](image)

However, most practical applications are semantics-less, as most existing XML schema languages, e.g., DTD and XSD, cannot fully represent the semantics such as object class, relationship type, OID, etc. Despite the existence of semantically rich XML models, e.g., ORA-SS, they still requires manual provision of semantics from the initial design or model transformation. We believe only if the automatic semantics discovery technique is developed to a satisfactory level, the achievements in semantics-based query optimization, keyword search, schema/data integration, etc., will be widely adopted by different applications.

In this paper we present a novel approach to automatically discover the ORA-semantics from data-centric XML schemas and data. Different from the existing approaches that only focus on object identification, we consider a comprehensive set of ORA-semantics, including OID, relationship type as well as the distinction between object attribute and relationship attribute.

2 Preliminary

We refer the tree structure derived from XML schemas as *XML schema trees*. For ease of description, all following concepts are defined on XML schema trees.

In XML schema tree, **object class** is an internal node representing a real world entity or concept. An object class has a set of **object attributes** to describe its properties. Each object class has an **object identifier (OID)** to uniquely identify its instance. Several object classes may be connected through a **relationship type** which may or may not explicitly appear in the XML schema tree. We call them **explicit relationship type** and **implicit relationship**