Chapter 3
A Semantic Similarity Analysis for Data Mappings between Heterogeneous XML Schemas

Jaewook Kim
University of Maryland Baltimore County, USA

Yun Peng
University of Maryland Baltimore County, USA

ABSTRACT

One of the most critical steps to integrating heterogeneous e-business applications using different XML schemas is schema mapping, which is known to be costly and error-prone. Past research on schema mapping has not made full use of semantic information imbedded in the hierarchical structure of the XML schema. This chapter investigates the existing schema mapping approaches and proposes an innovative semantic similarity analysis approach to facilitate XML schema mapping, merging and reuse. Several key innovations are introduced to better utilize available semantic information. These innovations include: (1) a layered structure analysis of XML schemas, (2) layer-specific semantic similarity measures, and (3) an efficient semantic similarity analysis using parallel and distributed computing technologies. Experimental results using two different schemas from a real world application demonstrate that the proposed approach is valuable for addressing difficulties in XML schema mapping.

INTRODUCTION

The electronic business (e-Business) requires interoperability between different e-Business systems for the seamless exchange of information either within or across enterprises. One of the most critical steps to achieving a successful integration and interoperability between heterogeneous e-Business systems is schema mapping, which is known to be costly and error-prone. Schema mapping is, roughly speaking, to identify how information can be shared between heterogeneous
A Semantic Similarity Analysis for Data Mappings between Heterogeneous XML Schemas

schemas and how they can be mapped, merged, or reused for integration and interoperability of their e-Business systems.

A schema typically defines the syntax of business documents (or instances), but it also contain semantic information, i.e., the meaning of elements in business documents. Sometimes, the schema refers to other semantic sources such as ontology, dictionary, or documentation for additional semantic information. The typical way for schema mapping is to identify semantically identical or similar elements between the two schemas. Many approaches have been proposed, but the challenge is still daunting because of the complexity of schemas and immaturity of technologies in semantic representation, measuring, and reasoning.

The goal of this chapter is to investigate the existing schema mapping approaches and to propose an innovative semantic similarity analysis approach to facilitate XML schema mapping, merging and reuse. Several key innovations are introduced to better utilize available semantic information. These innovations includes: (1) a layered structure analysis of XML schemas, (2) layer-specific semantic similarity measures, and (3) an efficient semantic similarity analysis using parallel and distributed computing technologies. Experimental results using two different schemas from a real world application demonstrate that the proposed approach is valuable for addressing difficulties in XML schema mapping.

BACKGROUND

The Challenges for Data Mappings between Heterogeneous XML Schemas

Over the past decades, the eXtensible Markup Language (XML) has emerged as one of the primary languages to help information systems in sharing structured data. Especially, XML schemas have been widely used in the e-Business for enterprises to exchange the business documents with their partners in a supply chain. The popularity of the XML and XML schema leads to an exponential growth of Business-to-Business (B2B) transactions. This success, however, leads to several problems: (1) individual enterprises often create their own XML schemas with information most relevant to their own needs; (2) different enterprise groups define different but similar XML schemas; and (3) the enterprises often extend or redefine the existing standard XML schema for their own needs. To successfully integrate heterogeneous e-Business systems, therefore, it is now critical to integrate their respective different XML schemas. This is what is called schema mapping.

The schema mapping is the process of identifying if and how two schemas are semantically related (Miller et al., 1994; Rahm & Bernstein, 2001; Shvaiko & Euzenat, 2005). It is one of the most important steps to integrate heterogeneous e-Business systems; however, it is typically largely performed manually by human engineers who are at best supported by some graphical interface tools. This manual mapping process is known to be very labor-intensive, costly, and error-prone (Gal, 2006; Rahm & Bernstein, 2001). As the e-Business systems grow to handle more complex databases and applications, their schemas become larger and more complicated. This further increases the search space to be examined as well as the number of correspondences to be identified. As a result, it is critical to automate the schema mapping task as much as possible to reduce the costs of labor-intensive data integration work and to reduce the mapping errors.

The XML schema mapping can be classified into two types depending on the types of the e-Business standard schemas: component schema and document schema. The component schema only contains reusable and extensible components (types or elements) as global type definition (e.g., OAG Common Core Component schema), while the document schema contains a global root ele-
Related Content

B2C Failures: Toward an Innovation Theory Framework
www.igi-global.com/chapter/b2c-failures-toward-innovation-theory/9300?camid=4v1a

Successful Implementation of Emerging Communication Technologies in a Mobile-Intense Organization: A Case Study of Sydney Airport
www.igi-global.com/chapter/successful-implementation-emerging-communication-technologies/19514?camid=4v1a

Drivers and Inhibitors of Mobile-Payment Adoption by Smartphone Users
www.igi-global.com/article/drivers-inhibitors-mobile-payment-adoption/68175?camid=4v1a

AGATHE: An Agent- and Ontology-Based System for Gathering Information about Restricted Web Domains
www.igi-global.com/article/agathe-agent-ontology-based-system/3927?camid=4v1a