Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 65 (2008) 266–303
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

An XML Schema integration and query mechanism system ☆

Sanjay Madria [a], Kalpdrum Passi [b,*], Sourav Bhowmick [c]

[a] Department of Computer Science, University of Missouri–Rolla, Rolla MO 65401, USA
[b] Department of Mathematics and Computer Science, Laurentian University, Sudbury, ON, Canada P3E2C6
[c] School of Computer Engineering, Nanyang Technological University, Singapore

Available online 21 September 2007

Abstract

The availability of large amounts of heterogeneous distributed web data necessitates the integration of XML data from multiple XML sources for many reasons. For example, currently, there are many e-commerce companies, which offer similar products but use different XML Schemas with possibly different ontologies. When any two such companies merge, or make an effort to service customers in cooperation, there is a need for an integrated schema and query mechanism for the interoperability of applications. In applications like comparison-shopping, there is a need for an illusionary centralized homogeneous information system. In this paper, we propose XML Schema integration and querying methodology. We define an object-oriented data model called XSDM (XML Schema Data Model) and present a graphical representation of XML Schema for the purpose of schema integration. We use a three-layered architecture for XML Schema integration. The three layers included are namely *pre-integration*, *comparison*, and *integration*. The three layers can conceptually be regarded as three phases of the integration process. During *pre-integration*, the schemas present in XML Schema notation are read and converted into the XSDM notation. During the *comparison* phase of integration, correspondences as well as conflicts between elements are identified. During the *integration* phase, conflict resolution, restructuring and merging of the initial schemas takes place to obtain the global schema. We define integration policies for integrating element definitions as well as their datatypes and attributes. An integrated global schema forms the basis for querying a set of local XML documents. We discuss various strategies for rewriting the global query over the global schema into the sub-queries over local schemas. Their respective local schemas validate the sub-queries over the local XML documents. This requires the identification and use of mapping rules and relationships between the local schemas.

Keywords: XML; Schema; Integration; Query

1. Introduction

The integration of heterogeneous data sources has become a central problem of modern computing [8]. Data integration involves data from a variety of applications, repositories, and legacy systems. With the

☆ Partially supported by NSERC grant 232038.
* Corresponding author.
E-mail addresses: madrias@umr.edu (S. Madria), kpassi@cs.laurentian.ca (K. Passi), assourav@ntu.edu.sg (S. Bhowmick).

0169-023X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2007.09.008

advent of improvements to Internet technologies, there has been a greater demand on the integration of data from diverse sources, especially in e-business where companies need to connect their online systems with those of their suppliers. Besides e-commerce, integration of data from different sources is required when two or more organizations merge. In recent times web sites are linked to various sources of data which necessitates integration of data sources.

The availability of large amounts of XML data necessitates the integration from multiple XML sources for many reasons. Each organization or application creates its own document structure according to specific requirements. These documents/data may need to be integrated or restructured in order to efficiently share the data with other applications. Interoperability between applications can be achieved through an integration system, which automates the access to heterogeneous schema sources and provide a uniform access to the underlying schemas.

In e-commerce applications, XML documents can be used to publish any data ranging from product catalogs and airline schedules to stock reports and bank statements. XML forms can be used to place orders, make reservations and schedule shipments. XML eliminates the need for custom interfaces with every customer and supplier applications, allowing buyers to compare products across many vendors and catalog formats, and sellers to publish their catalog information to reach many potential buyers.

XML enables online businesses to build on one another's published content and services to create innovative virtual companies, markets, and trading communities. With a global view of the Internet-wide shopping directories, a query system can locate all merchants carrying a specific product or service, and then query each local schema in parallel to locate the best deals. The query system can use the integrated schema and can sort the offers according to criteria set by the buyers–the cheapest flight, the roomiest aircraft, or some weighted combination. Other examples where integrated view is useful are given below.

- When companies merge or endeavor to service customers jointly, their local schemas need to be merged to provide an integrated view of the data present with the companies and to enable conflict-free information sharing and retrieval.
- Applications like comparison-shopping that have to retrieve the data from different heterogeneous data sources and compare the prices and specifications of various items have a need for the integrated view of all the sources and a query mechanism.
- Any application that needs to interact with data from two or more XML sources need to have an integrated view of the schemas of their local schema and a mechanism to retrieve the data from different sources.

There are two popular styles for integrating heterogeneous sources, data integration and schema integration. During data integration the physical data from the heterogeneous sources is combined. However, during schema integration the data are not touched but rather the schemas of the sources are combined. In either case, the goal is to provide a uniform interface to a multitude of data sources. To mask the heterogeneity, a mediator presents a unified context to users. One of the key advantages of integration is that it frees the user from having to locate and interact with every source, which is related to their query.

For seamless information access, the mediation systems have to cope with different data representations and search capabilities [9]. A mediator presents a unified context for uniform information access, and consequently must translate original user queries from the unified context to a target source for native execution [7]. This translation problem has become more critical now that the Internet and Intranets have made available a wide variety of disparate sources, such as multimedia databases, web sources, legacy systems, and information retrieval (IR) systems. Integrating a number of heterogeneous sources [29] is difficult in part because each source has its own set of vocabulary and semantics, which can be used when formulating queries. Hence, the query processor needs to be able to efficiently collect related data from multiple sources, minimize the access to redundant sources, and respond flexibly when some sources are unavailable. To ensure semantic interoperability, information must be appropriately mapped from its source context to its target context where it will be used [8]. For this reason, mapping rules and algorithms must be created to ensure a query is rewritten properly.

The currently available integration systems for semistructured data [1,15,8,16,17,31] use the approach where they integrate the data by using mediated schemas to reformulate queries on the disparate data sources