Semantic Query Planning Mechanism on XML Based Web Information Systems

Jinguang Gu\textsuperscript{1,2}, Bo Hu\textsuperscript{1}, and Yi Zhou\textsuperscript{1}

\textsuperscript{1} College of Computer Science and Technology, 
Wuhan University of Science and Technology, Wuhan 430081, China
\textsuperscript{2} College of Computer Science and Engineering, 
Southeast University, Nanjing 210096, China

sam@seu.edu.cn, simon@wust.edu.cn

Abstract. To enable accessing web information at semantic level, this paper develops a semantic query planning mechanism on XML based web information system with complex ontology mapping technology. It discusses the patterns of complex ontology mappings at first, and then the ontology-based query planning in Mediator-Wraper based environment with GAV style querying request. The extension of XML query algebra and XML query planning algorithm are discussed in detail.

1 Introduction

We witness a rapid increase in the number of web information sources that are available online. The World-Wide Web (WWW), in particular, is a popular medium for interacting with such sources\textsuperscript{[1]}. This sources include both free and commercial databases, flat files services, web services or web based applications, HTML files and XML files. Due to its ability to express semi-structured information, XML based information System (X-WIS), which is an extension to Database Information System, plays a key role in web information community, and XML is rapidly becoming a language of choice to express, store and query information on the web, other kinds of web information such as HTML-based web information can be transferred to XML based information with annotation technologies. Users can query information with XML languages, XPath based languages such as XQuery, XUpdate are suitable for retrieving information in distributed integration systems. Problems that might arise due to heterogeneity of the data are already well known within the distributed database systems community: structural heterogeneity and semantic heterogeneity. Structural heterogeneity means that different information systems store their data in different structures. Semantic heterogeneity considers the content of an information item and its intended meaning\textsuperscript{[2]}. How to accessing distributed information with a consistent semantic environment and how to make the XML query mechanism with semantic enabled are the main problems that should be discussed in distributed X-WISs.

The use of ontologies for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity. Ontologies can be used to describe the semantics of the X-WIS sources and to
make the content explicit. With respect to the data sources, they can be used for the identification and association of semantically corresponding information concepts.

This paper focuses on how to use ontology technology to enable semantic level querying on XML based WISs. It uses ontology mapping technology to get a consistent semantic environment, and extends the XML based querying technologies to enable semantic querying on X-WISs. The remainder of this paper is structured as follows. Section 2 gives the general discussion about ontology enabled XML based information systems (WISs). Section 3 defines the patterns of ontology mapping. Section 4 discusses ontology enabled query planning on XML based semi-structured web information systems, such as ontology enhanced XML query algebra and XML query planning algorithm. Section 5 summarizes the whole paper.

2 The XML Based Information Systems

Because XML has been the standard language to represent web information and semantic web resources, using XML to represent web based or semi-structured information systems is a good choice. In the distributed environment, every local site contains the local ontology based structured or semi-structured information source, this information source may be a relational database, native XML database, web site, XML based application or other autonomous system.

2.1 The Formal Definition of XML Based Web Information Systems

From the point of view of web based or semi-structured information processing, all local information sites can be expressed as collections of XML instances. An XML instance can be described as a structure

\[ I_d := (V_d, E_d, \delta_d, T_d, O_d, t_d, oid_d, root_d) \]

consisting of:

i. \( I_d \) is a rooted and directed tree, and \( V_d \) is the node set of the tree, \( E_d \) is the edge set of the tree and \( E_d \subseteq V_d \times V_d \), \( \delta_d \) is the mapping function between nodes which identify the direction of the edges, \( \text{root}_d \) is the root of the tree;

ii. Every node has a unique identifier which belongs to \( O_d \), which means for every node \( e \) exists \( \forall(e) \{ e \in V_d \rightarrow \exists(o) \{ o = \text{oid}_d(e) \land o \in O_d \} \} \);

iii. There exist a mapping function \( t_d(e, \text{string}) \) which maps the attribute of the node a type \( \tau \in T_d \), and the function has the following rules:

- if \( \text{string} = \text{tag} \), then maps it to the type of tag;
- if \( \text{string} = \text{content} \), then maps it to the type of content.

iv. Every type \( \tau \in T_d \) has a domain denoted as \( \text{dom}(\tau) \);

v. \( \delta_d \) is the identifier of edge, with which the parent nodes can access the child nodes. On the other hand, the child nodes can access parent nodes by the identifier \( \delta_d^{-1} \).