Web Information Integration
Based on Compressed XML

Hongzhi Wang, Jianzhong Li, Zhenying He, and Jizhou Luo

Department of Computer Science and Technology
Harbin Institute of Technology
wangzh@hit.edu.cn
lijz@banner.hl.cninfo.net
hzy_hit@sina.com
luojizhou@x263.net

Abstract. Nowadays, information integration to web data sources and XML becomes a favorite information exchange format. New application motivates the problems that massive information is often transmitted in network and must be processed in limited buffer in mediator. To process query on massive data from web data source effectively, we present a method of XML compression based on edit distance for information transmission in information integration. By compressing XML, this method can reduce both the transmission time and buffer space. Two different strategies of XML compression for transmission and process in mediator are designed. Optimization of the combination of these strategies is discussed. We also propose the query execution algorithms on compressed XML data in buffer of mediator. We focus on main operators of data from wrapper in mediator, namely sort, union, join and aggregation. Implementation of these operators on compressed data using two different methods is described in this paper.

1 Introduction

Information integration system aims to integrate autonomy distributed and heterogeneous data sources to form a single data source. There are mainly three information integration methods, federal databases, data warehouse and mediation [1]. Mediation [2] becomes the most favorite information integration form, because of its capability of representing both structured and semi-structured data with semantics tags and XML [3] becomes the most important format of information exchange in information integration system [4].

Nowadays, Internet, which holds massive information, becomes the largest data source (or can be treated as lots of data sources). Information integration is not limited in only traditional known data source but can also be extended to data sources of web. The flexibility of XML makes it adaptive to be used as the format of integration of web data sources. The information integration of web with XML as information exchange format brings new problems:
- Massive information in web makes the information exchange process slow. How to speed up this process of the information exchange in information integration system?
- The web is a dynamic data source with both static web pages and hidden data source with only a query interface [5]. The traversal of the whole web needs long time. How to answer user's query timely?
- The environment of Internet is not stable and the efficiency of various data sources is not the same. How to schedule the execution of the query in various wrappers?

To solve the second problem, an XML warehouse is used as a semantics cache for store history information from web. The query to related information could be answered quickly directly from the XML warehouse. The architecture of the information integration is shown in fig.1.

Since the information obtained from web may be massive, XML warehouse need to be compressed. [6] discusses the compression method based on edit distance and the query process techniques on compressed data. Such XML compression techniques could be used to compress exchange information in XML format, thus solves the first problem presented above.

To solve the third problem, query buffer is used in mediator to store the result returned temporarily. In order to make plenary use of the query buffer, the data transferred into the query buffer should be kept compressed and decompressed only when query process needs so.