Chapter VII

Effective Processing of XML-Extended OLAP Queries Based on a Physical Algebra

Xuepeng Yin, Aalborg University, Denmark
Torben Bach Pedersen, Aalborg University, Denmark

Abstract

In today’s OLAP systems, physically integrating fast-changing data, for example, stock quotes, into a cube is complex and time-consuming. This data is likely to be available in XML format on the World Wide Web (WWW); thus, instead of physical integration, making XML data logically federated with OLAP systems is desirable. In this chapter, we extend previous work on the logical federation of OLAP and XML data sources by presenting simplified query semantics, a physical query algebra and a robust OLAP-XML query engine as well as the query evaluation techniques. Performance experiments with a prototypical implementation suggest that the performance for OLAP-XML federations is comparable to queries on physically integrated data.
Introduction

Online analytical processing (OLAP) technology enables data warehouses to be used effectively for online analysis, providing rapid responses to iterative complex analytical queries. Usually, an OLAP system contains a large amount of data, but dynamic data, for example, stock prices, is not handled well in current OLAP systems. To an OLAP system, a well-designed dimensional hierarchy and a large quantity of pre-aggregated data are the keys. However, trying to maintain these two factors, when integrating fast-changing data physically into a cube, is complex and time-consuming, or even impossible. However, the advent of XML makes it very possible that this data is available in XML format on the WWW. Thus, making XML data accessible to OLAP systems is greatly needed.

Our overall solution is to logically federate the OLAP and XML data sources. This approach decorates the OLAP cube with virtual dimensions, allowing selections and aggregations to be performed over the decorated cube. In this chapter, we describe the foundation of a robust federation query engine with the focus on query evaluation, which includes the query semantics, a physical algebra, and query evaluation techniques. First, a query semantics that simplifies earlier definitions (Pedersen, Riis, & Pedersen, 2002) is proposed. Here, redundant and repeated logical operators are removed and a concise and compact logical query plan can be generated, after a federation query is analyzed. Second, a physical query algebra, unlike the previous logical algebra, is able to model the real execution tasks of a federation query. Here, all concrete data retrieval and manipulation operations in the federation are integrated. This means that we obtain a much more precise foundation for performing query optimization and cost estimation. Third, the detailed description of the query evaluation introduces how the modeled execution tasks of a query plan are performed, including the concrete evaluation algorithms and techniques for each physical operator and the general algorithm that organizes and integrates the execution of the operators in a whole plan. In addition, algebra-based query optimization techniques, including the architecture of the optimizer, cost estimation of physical operators, and plans, also are presented. Experiments with the query engine suggest that the query performance of the federation approach is comparable to physical integration.

Related Work

There has been a great deal of previous work on data integration, for instance, on relational data (Chen, Wu, Wang, & Mao, 2006; Hellerstein, Stonebraker, & Caccia, 1999; IBM, n.d.; Oracle, 2005), semi-structured data (Chawathe, Garcia-Molina,
Detecting Users’ Dietary Preferences and Their Evolutions via Chinese Social Media
www.igi-global.com/article/detecting-users-dietary-preferences-and-their-evolutions-via-chinese-social-media/218928?camid=4v1a