IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB13496

This paper appears in the book, Information Quality Management: Theory and Applications edited by Latif Al-Hakim © 2007, Idea Group Inc.

Chapter II

Quality-Extended Query Processing for Mediation Systems

Laure Berti-Équille, IRISA, France

Abstract

For noncollaborative distributed data sources, quality-driven query processing is difficult to achieve because the sources generally do not export data quality indicators. This chapter deals with the extension and adaptation of query processing for taking into account constraints on quality of distributed data. This chapter presents a novel framework for adaptive query processing on quality-extended query declarations. It proposes an expressive query language extension combining SQL and QML, the quality of service modeling language proposed by Frølund and Koistinen (1998) for defining in a flexible way dimensions, and metrics on data, sources, and services quality. The originality of the approach is to include the negotiation of quality contracts between the distributed data sources competing for answering the query. The principle is to find dynamically the best trade-off between the local query cost and the result quality. The author is convinced that quality of data (QoD) and quality of service (QoS) can be advantageously conciliated for tackling the problems of quality-aware query processing in distributed environments and, more generally, open innovative research perspectives for quality-aware adaptive query processing.

Introduction

For classical mediator/wrapper architectures, the access to distributed information sources is carried out in a declarative way. The mediator processes the queries of the users at the global level and optimizes the query plans according to the wrappers that reach respectively their underlying data sources. In this type of distributed environment, the sources are usually noncollaborative and do not export information describing the local costs of query processing, neither indicators of their quality of service (e.g., resource accessibility, reliability, security, etc.) nor information describing the quality of their content[1] (e.g., data accuracy, freshness, completeness, etc.). However this information is useful and has to be dynamically computed and periodically updated. The lack of scalability and flexibility face to the growing number and the changing structure of the distributed data sources harm seriously the efficiency and effectiveness of execution of the queries and the quality of the results. Although there are several approaches that deal with the assessment and management of quality metadata, the dual problem of fixing the query cost and optimizing the result quality, or fixing the result quality and optimizing the query cost, still remains. Querying simultaneously several data sources with different degrees of quality and trust in a dynamic and distributed environment raises several interesting problems and open issues:

- **Selecting dynamically appropriate sources:** Different information sources may answer a global query with different response times, query costs, and various levels of result quality. How do we define strategies for selecting adaptively the most appropriate sources for answering the whole or some parts of the global query with the appropriate quality at a given time? What are the criteria to select dynamically the sources with "the best relative quality"?
- **Defining semantically and qualitatively correct distributed query plans:** The result of a global query is built according to the particular order for execution of subquery plans. This must combine in a coherent way both information and quality metainformation from the various sources; but data quality levels are often unknown, heterogeneous from one source to another (intersource data quality heterogeneity), and locally nonuniform (intrasource data quality nonuniformity). In this context, the aim is to control and merge data quality indicators in a consistent and meaningful way for both correctly integrating data and quality metadata.
- **Making trade-offs between the cost of the query and the perceived quality of the result (including both the quality of service and the quality of content):** Because we may accept a query result of lower quality (if it is cheaper or has a shorter response time than if the query cost is higher), it is necessary to adapt query costs to users' quality requirements (and tolerance thresholds). The objective is to measure and optimally reduce the query cost and bargain query situations where the system searches for solutions that "squeeze out" more gains (in terms of result quality) than the classical query without specified quality constraints.
- **Developing concrete cost models to evaluate whether the expected benefits from the improved (or quality-extended) query plan compensate for the cost of collecting and evaluating feedback from the environment during execution time:** The difficulty here is to adapt existing query processing techniques to environments where resource availability, allocation, quality, and cost are not, by definition, decidable at compile time.

Related Content

An Algebraic Approach to Data Quality Metrics for Entity Resolution over Large Datasets

John Talburt, Richard Wang, Kimberly Hess and Emily Kuo (2007). *Information Quality Management: Theory and Applications (pp. 1-22).*
www.igi-global.com/chapter/algebraic-approach-data-quality-metrics/23022?camid=4v1a