J. Albrecht, W. Hämmer, W. Lehner, L. Schlesinger

Using Semantics for Query Derivability in Data Warehouse Applications

Erstveröffentlichung in / First published in:

Flexible Query Answering Systems. Warsaw, 25.-28.10.2000. Springer, S. 3-4. ISBN 978-3-7908-1834-5.

DOI: http://dx.doi.org/10.1007/978-3-7908-1834-5_1

Final edited form was published in "Flexible Query Answering Systems. Warsaw 2001", S. 3-4, ISBN 978-3-7908-1834-5
https://doi.org/10.1007/978-3-7908-1834-5_1

Using Semantics for Query Derivability in Data Warehouse Applications

J. Albrecht, W. Hümmer, W. Lehner, L. Schlesinger
Department of Database Systems, University Erlangen-Nuremberg
Martensstr. 3, 91058 Erlangen, Germany
{albrecht, huemmer, lehner, lutz}@immd6.informatik.uni-erlangen.de

Abstract. Materialized summary tables and cached query results are frequently used for the optimization of aggregate queries in a data warehouse. Query rewriting techniques are incorporated into database systems to use those materialized views and thus avoid accessing the possibly huge raw data. A rewriting is only possible if the query is derivable from these views. Several approaches can be found in the literature to check the derivability and find query rewritings. However, most algorithms either find rewritings only in very restricted cases or in complex cases which rarely occur in data warehouse environments. The specific application scenario of a data warehouse with its multidimensional perspective allows the consideration of much more semantic information, e.g. structural dependencies within the dimension hierarchies and different characteristics of measures. The motivation of this article is to use this information to present simple conditions for derivability in a large number of relevant cases which go beyond previous approaches.

1 Introduction

Data warehousing has nowadays become a common technology. The goal of a data warehouse is to provide analysts and managers with strategic information about the key figures of the underlying business. Since microdata are of no interest at this level, almost all queries on data warehouses involve aggregates.

A common optimization technique in data warehousing is the use of materialized summary tables. Because in general the fact tables storing the summary values of interest are very large, it is especially in an OLAP environment infeasible to query the fact tables directly. Instead, queries should be answered by materialized aggregate views if possible. The question of derivability in the presence of redundancy is as old as the theory of relations [4.]. In order to rewrite queries three questions have to be investigated:

1) Under which circumstances is an aggregate query derivable from one or more materialized views?
2) How must the query be rewritten in order to make use of the materialized views?
3) If there are several possibilities to use materialized views, which is least expensive?

Some of the large relational database vendors like Oracle [14.] and IBM [10.] provide mechanisms to transparently rewrite certain types of queries so that appropriate

Final edited form was published in "Flexible Query Answering Systems. Warsaw 2001", S. 3-4, ISBN 978-3-7908-1834-5
https://doi.org/10.1007/978-3-7908-1834-5_1

materialized views are used instead. However, in general the problem is NP-hard ([18.]) and in some cases unsatisfiable. Therefore, many algorithms for query rewriting especially for aggregate queries are of exponential complexity (see related work in section 3).

The main focus of this article is on the first two questions. In contrast to both, commercial products which can utilize materialized views only in a very limited number of cases and very complex and expensive approaches in literature, we want to identify simple cases for the derivability of aggregate queries with high practical use in data warehousing and statistical databases. A very interesting special case which has not been considered before is the derivation of composite measures, like the turnover which can be computed from a sales quantity and the respective price.

The article is organized as follows: In section 2 an example is given to motivate our case. Section 3 gives information about related work and the shortcomings of previous approaches. A formalism to describe the class of queries which is to be investigated is defined in section 4. Sufficient conditions for the derivability of such queries are discussed in section 5. Section 6 closes with a short summary.

2 Motivation

We will motivate our case with an illustrative example. Consider the conceptual schema of the data warehouse of some retail store chain as depicted in figure 1. By adopting a multidimensional terminlogy, there are three dimensions, Product, Location and Time with several category attributes. The arrows define functional dependencies in database terms. Each path from a terminal attribute, e.g. Article, to the Top category in the respective dimension defines a classification hierarchy (figure 1). Parallel paths result in parallel hierarchies. For the OLAP user these paths describe the basic navigation constructs for drill-down and roll-up operations. In the sample database there are four measures. Only the quantity sold (Qty) is given per article, day and shop. The other measures are gathered at a higher granularity. For example the retail chain has the policy that the prices are the same in all of its shops and they change only quarterly.

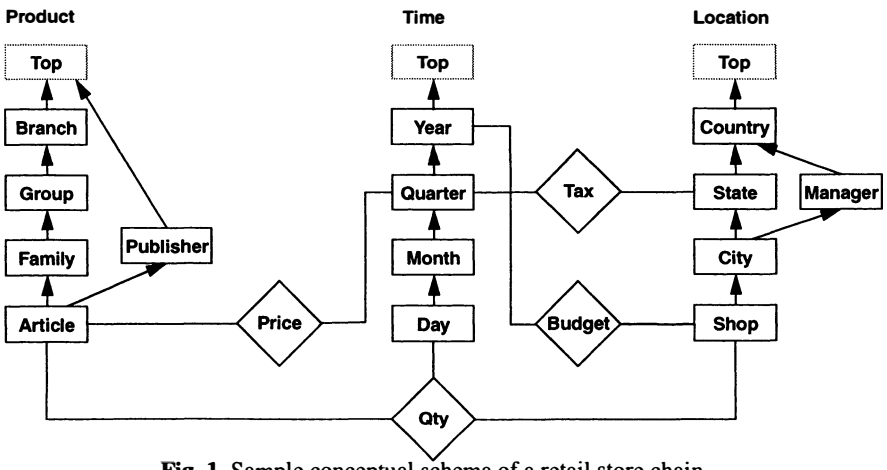

Fig. 1. Sample conceptual schema of a retail store chain