Data & Knowledge Engineering 89 (2014) 1–20

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Editorial

Detecting summarizability in OLAP

Tapio Niemi [a,*], Marko Niinimäki [a,b], Peter Thanisch [c], Jyrki Nummenmaa [c]

[a] Helsinki Institute of Physics, Technology Programme, CERN, CH-1211 Geneva 23, Switzerland
[b] HEPIA, University of Applied Sciences of Western Switzerland, Rue de la Prairie 4, CH-1202 Geneva, Switzerland
[c] School of Information Sciences, University of Tampere, FIN-33014 Tampere, Finland

ARTICLE INFO

Article history:
Received 17 February 2012
Received in revised form 13 November 2013
Accepted 13 November 2013
Available online 27 December 2013

Keywords:
Business intelligence
Database design
Modelling and management
Summarizability
Additivity
OLAP

ABSTRACT

The industry trend towards self-service business intelligence is impeded by the absence, in commercially-available information systems, of automated identification of potential issues with summarization operations. Research on statistical databases and on data warehouses have both produced widely-accepted categorisations of measure attributes, the former based on general summarizability properties and the latter based on measures' additivity properties. We demonstrate that neither of these categorisations is an appropriate basis for precise identification of measure types since they are incomplete, ambiguous and insufficiently refined.
Using a new categorisation of dimension types and multidimensional structures, we derive a measure categorisation which is a synthesis and a refinement of the two aforementioned categorisations. We give formal definitions for our summarizability types, based on the relational model of data, and then construct rules for correct summarization by using these definitions. We also give a method to detect whether a given MDX OLAP query conforms to those rules.

1. Introduction

The current trend for Business Intelligence (BI) to become increasingly pervasive [39] has led to an increasing reliance on self-service BI. Spahn et al. [36] note that business users face considerable challenges when trying to mine data in a self-service manner. One of these challenges is finding what data should be summarised and how. Here, we aim at answering this question by studying the OLAP cube, the multidimensional structure used in practical BI. More specifically, we study the measures (like units sold, distance travelled) that appear in the cube.

In the past, there have been two main sets of ideas about categorising measures:

1. The Lenz and Shoshani [16] categorisation of flow, stock and value-per-unit has been adopted by several groups of researchers.
2. The Kimball and Ross [13] categorisation of additive, semi-additive and non-additive has also been adopted by various groups of researchers.

In our paper, we examine how these two categorisations relate to each other and we develop a new categorisation which is a synthesis of the two earlier categorisations.

When a user's query extracts information from a multidimensional database ("cube", for short) the output is typically in summary form. The user's query specifies:

(a) The aggregation operation (e.g. a summation, a mean or a count),
(b) The measure which is to be aggregated, and
(c) Either an aggregation level or a member for each dimension.

* Corresponding author. Tel.: +41 22 767 6179; fax: +41 22 767 3600.
E-mail addresses: tapio.niemi@cern.ch (T. Niemi), man@cern.ch (M. Niinimäki), peter.thanisch@cs.uta.fi (P. Thanisch), jyrki@cs.uta.fi (J. Nummenmaa).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.11.001

The query can only produce meaningful results if:

1. The aggregation operation is appropriate for the measure and
2. The measure is appropriate for the aggregation levels in the cube's dimensions.

When both of the above conditions are true, we say that the selected data is "summarizable" with respect to the aggregation operation. The first condition is determined by the inherent statistical properties of the measure. For example, if our data comprises wind directions taken at time intervals and we want to determine the prevailing wind during some period of time, then it is not appropriate to compute the sum or the arithmetic mean, but it is appropriate to compute the mode. By contrast, the second condition is determined by the semantics of the selected data. For example, while it is appropriate to sum the inventory levels of a particular product at a particular time over all of the stores, it is not appropriate to sum inventory levels for the product at one particular store over a number of points in time.

The problem of detecting summarizability has attracted the attention of researchers for many years. An indication as to why detecting summarizability is such a hard research problem can be gleaned from our use of the word "appropriate" in both of the above conditions. In particular, for the second of the above pair of conditions, what is "appropriate" is context-sensitive to a heterogeneous collection of contextual elements. As is well-known, such elements include the semantics of the dimensional data and the kind of measure involved. However, as we demonstrate in the present paper, whether or not a particular measure is appropriate can also depend on the kind of cube. Indeed, one of the contributions of the present paper is to provide categorisations of measures and cubes and to demonstrate that it is the combination of the measure and the cube category which determines summarizability.

In this paper we propose a method that can be used to detect when a query intended to summarise data violates one or both of the above conditions. We focus on additivity, i.e. cases in which the sum operation is used. In order for our method to work, it is necessary for the cube design to represent the units of measurement and the statistical scale of each of the cube's measures. We present a model of multidimensional databases which allows us to provide decision criteria for summarizability that are provably correct. Our model places restrictions on cube designs, for example we only allow simple hierarchical relationships in dimension and we do not model many-to-many relationships. Furthermore, our model also places an extra responsibility on the cube designer, since he/she must include the units of measurement for each of the measures as separate attributes in the schema for the cube. However, the benefit of our model is that we can provide provably-correct conditions for determining additivity.

Our model is sufficiently detailed to permit the automatic detection of additivity. Our approach requires that we model the types of measures in a more formal manner than has been done in earlier research. To demonstrate our method, we show how the correctness of summarizability can be checked when using the MDX [38] query language for expressing OLAP queries.

The paper is organised as follows. We illustrate the problems of existing work on summarizability in Section 2. In Section 3, we give a brief review of related work and then present a running example in Section 4. We present a formal model of OLAP based on relational calculus in Section 5. The core of the paper is presented in Sections 6 and 7, where we introduce the types of measures (tally, semi-tally, reckoning, and snapshot), and the definition of additivity. In Section 8, we describe how the designer can determine the summarizability types of measure attributes and apply this information in OLAP design. We also show how additivity rules are applied to OLAP queries, expressed in the MDX language. In Section 9, we give a brief evaluation of the method. Conclusions and future work are given in Section 10.

2. Motivation

According to Lenz and Shoshani [16], correct summarizability requires the following three conditions to be true:

1. disjointness of attribute groups,
2. completeness of grouping, and
3. the combination of types of the attribute, the dimension, and the aggregation function must be consistent.

The first rule means that an attribute value may roll-up to only one group on the higher level in the hierarchy, while the completeness rule means that each value must roll-up to some group. These two rules are quite straightforward to check but some real-world cases do not necessarily fit into them. The classical examples are Russia belonging to both Europe and Asia, and Washington D.C. being without a state.

The third rule is not so straightforward, since semantics plays an important role in it. Lenz and Shoshani's suggested solution is to divide dimensions into temporal and non-temporal ones and summary attributes ("measures") into three groups:

- Flow, "cumulative effect over a period", the unit/the period of time, such as Euro/month, e.g. monthly sales.
- Stock, "state at specific point at time", a simple unit such as Euro, meter, kilogramme, e.g. inventory.
- Value-per-unit, x/y units such as price/item, e.g. item price or exchange rate.

The flow type refers to a period of time and it is recorded at the end of the period, while the stock type is recorded at a particular point of time. The difference between the stock and the value-per-unit type is the unit. For example, the unit of the product price is e.g. Euro/product, while the unit of the daily sales, i.e. the total sales of all products sold on a day, is Euro or the number of products. Horner and Song [9] give a more detailed description of this topic and provide a classification of aggregation functions based on their applicability to summarization. The flow type describes the change during a period, the stock type the value recorded at a specific point in time, and the value-per-unit type a fraction or rate that can be used to convert units.