Regular Sparsity in OLAP System

Kalinka Kaloyanova¹, Ina Naydenova²

Abstract One of the primary challenges of storing multidimensional data is the degree of sparsity that is often encountered. Because the extremely sparse cubes are frequent phenomenon, OLAP engines offer different methods of increasing the performance of sparse cubes, but all of these methods do not take account of the sparsity nature and did not divide the sparsity into any types. Our experience leads us to a following division of the empty areas in the multidimensional cubes: (a) areas that are empty because of the semantics of the business (the semantics enforces lack of value) and (b) areas that are empty by a chance. To formally distinguish these types of sparsity, we introduce a new object (“regular sparsity map”) which provides business analysts with the ability to define rules and place data constraints over the multidimensional cube. In this paper we present our regular sparsity map editor and discuss how it can be used for the purpose of data errors detection and selection of relevant dimension elements.

Introduction

Prior to the start of the Information Age in the late 20th century, businesses had to collect data from non-automated sources. Businesses then lacked the computing resources necessary to properly analyze the data, and as a result, companies often made business decisions primarily on the basis of intuition [1]. The modern technologies of computers and networks have made data collection and organization much easier. However, the captured data needs to be converted into information and knowledge to become useful. This puts a new class of problems related to the data integrity and correctness, results reliability, performance of the analytical query and others ([2],[3],[4]).

During the 1990s, a new type of data model – the multidimensional data model – has emerged that has taken over from the relational model when the objective is to analyze data, rather than to perform on-line transactions. Multidimensional models lie at the core of On Line Analytical Processing (OLAP) systems. Such systems provide fast answers for queries that aggregate large amounts of detail data to find overall trends, and they present the results in a multidimensional fashion, which renders a multidimensional data organization ideal for OLAP [5]. In a multidimensional data models, there is a set of numeric measures (facts) that are the objects of analysis. Each of the numeric measures depends on a set of dimensions,
which provide the context for the measure. For example, the dimensions associated with a sale amount can be the store, product, and the date when the sale was made. The dimensions together are assumed to uniquely determine the measure. Thus, the multidimensional data views a measure as a value in the multidimensional space of dimensions. Often, dimensions are hierarchical; time of sale may be organized as a day-month-quarter-year hierarchy, product as a product-category-industry hierarchy [6]. In OLAP cube, cross product of dimensional members forms the intersections for measure data. But in reality most of the intersections will not have data. This leads to sparsity. The multidimensional cross relationships which exist in all OLAP applications and the fact that the input data is usually very sparse are the main reason most of the OLAP applications to suffer from data explosion consequences [7]. Because the extremely sparse cubes are frequent phenomenon, OLAP engines offer different methods of increasing the performance and reducing the size of the cubes. But all of these methods do not take account of the sparsity nature and did not divide the sparsity into any types.

In Naydenova [8] we introduce a new classification of multidimensional cube sparsity phenomena, define an object named “regular sparsity map” (RSM) and investigate the RSM applicability. The RSM saves information about empty domains of multidimensional cubes and provides analysts with the ability to define business rules and place data constraints over the multidimensional model. The map can be used at many stages of the business intelligence system life cycle (storage and performance consideration, user-interface improvements), but its primary function is to support the process of discovering inaccurate and inconsistent information.

We developed an editor for RSM creation and implement an algorithm that performs set operations between RSM and arbitrary multidimensional domains in the map space. We are going to briefly describe our implementation approach and discuss the data error and relevant dimension elements detection process. RSM can also be used for data storage and query optimization and also allows report enhancements. But these applications are not still implemented and they are out of the scope of this paper.

**The Regular Sparsity Map**

To explain what a regular sparsity map is, first of all we will introduce some definitions.

**Multidimensional Data Model Definition**

To define a regular sparsity map object we assume a simplified conceptual cube model that treats data in the form of n-dimensional cubes. The hierarchies between the various levels of aggregation in dimensions are of no interest to us.