

A Model of Regular Sparsity Map Representation

Ina Naydenova

Institute for Parallel Processing Bulgarian Academy of Science, Sofia, Bulgaria

e-mail: naydenova@gmail.com

Zlatinka Covacheva

Higher College of Technology, Muscat, Oman & Higher College of Telecommunications and Post, Sofia, Bulgaria,

e-mail: zkovacheva@hotmail.com

Kalinka Kaloyanova

“St. Kliment Ohridski” University of Sofia Faculty of Mathematics and Informatics, Sofia, Bulgaria,

e-mail: kkaloyanova@fmi.uni-sofia.bg

Abstract

Sparse data causes the data explosion problem in precomputation process of multidimensional cubes and decreases the performance of OLAP methodology. The regular sparsity map is an innovative object that saves information about specific empty domains of the OLAP hyper-cubes and enables business analysts to define rules and place data constraints over the multidimensional cube. The preserved information can be used for several purposes – data validation, storage consideration, user-interface improvements. In this paper we present an approach for a regular sparsity map representation. It allows implementation of set operations between a regular sparsity map and multidimensional domains and requires reasonable storage and computational resources for map implementation.

Sparsity refers to a natural phenomenon evident in all multidimensional data to some degree: not all of the cells in the logical cube will ever contain data. Cube sparsity has many impacts on the storage size, calculations, loading and query performance. The current methods for overcoming of data explosion work mainly on physical level and do not take into account the nature of sparsity.

In this paper we present a model of map representation that allows practical implementation of set operations between a map object and rectangular domains over a multidimensional space. To perform more efficiently set operations with rectangular domains, we propose an algorithm that works with dimension subsets instead of dimension elements. In our solution the input rectangular domain is split into a set of rectangular sub-domains, each of which is entirely inside or outside the map. To facilitate the process of the input domain splitting, we represent the regular sparsity map as a union of nonintersecting rectangular domains. The tests of the presented approach over real data cubes show significant reduction of the cube storage size.

Представяне на модел на карта на регулярна рехавост

Ина Найденова

Институт за паралелна обработка, Българска Академия на науките, София, България

e-mail: naydenova@gmail.com

Златинка Ковачева

Висш колеж по технологии, Маскат, Оман & Висш колеж по телекомуникации и пощи, София,

България,

e-mail: zkovacheva@hotmail.com

Калинка Калоянова

Софийски Университет „Св. Климент Охридски“, Факултет по математика и информатика, София,

България

e-mail: kkaloyanova@fmi.uni-sofia.bg

Резюме

Наличието на рехави данни (sparse data) в многомерните кубове от данни причинява експлозивно разрастване на куба в процеса на предварително калкулиране на обобщените му нива и намалява предимствата на OLAP методологията. Картата на регулярна рехавост е иновативен обект към модела, който запазва информация за конкретни празни домейни от OLAP хиперкубовете, давайки възможност на бизнес анализаторите да налагат правила и ограничения върху данните от многомерния куб. Тази информация може да се използва за няколко цели - валидиране на коректността на данните, оптимизиране на съхранението на кубовете, подобрения в потребителския интерфейс за работа с OLAP обекти. В тази статия представяме подход за реализация на новопредложения обект, наречен карта на регулярна рехавост. Той позволява осъществяването на операции за работа с множествата, описани в картата на регулярна рехавост и произволни домейни от многомерния куб и изисква разумни ресурси за реализацията на картата.

Рехавостта е естествен феномен, до известна степен във всички многомерни данни: не всички клетки в логическия куб съдържат данни. Разредеността на куба оказва голямо въздействие върху размера на паметта, скоростта на обработката, зареждането и изпълнението на заявките. Настоящите методи за преодоляване на експлозията на данни действат главно на физическо ниво и не отчитат естеството на рехавостта.

В тази статия представяме модел на карта на регулярна рехавост, който позволява практическо изпълнение на операции с множества между области от картата и правоъгълни домейни от многомерното пространство на куба, върху който е дефинирана тя. За да изпълняваме по-ефикасно множествени операции с правоъгълни домейни, предлагаме алгоритъм, който работи с подмножества на измеренията, вместо с конкретни точки в пространството (елементи на куба). В нашето решение входният правоъгълен домейн е разделен на множество от правоъгълни поддомейни, всеки от които е изцяло вътре или извън картата. За да улесним процеса на разделяне на входния домейн, ние представяме картата на регулярна рехавост като обединение на непресичащи се правоъгълни домейни. Тестовите на описания подход върху реални кубове с данни показват значително намаляване на размера на заеманата от куба памет.