Chapter VII

Bitmap Indices for Data Warehouses

Kurt Stockinger
Lawrence Berkeley National Laboratory, University of California, USA

Kesheng Wu
Lawrence Berkeley National Laboratory, University of California, USA

Abstract

In this chapter we discuss various bitmap index technologies for efficient query processing in data warehousing applications. We review the existing literature and organize the technology into three categories, namely bitmap encoding, compression, and binning. We introduce an efficient bitmap compression algorithm and examine the space and time complexity of the compressed bitmap index on large datasets from real applications. According to the conventional wisdom, bitmap indices are only efficient for low-cardinality attributes. However, we show that the compressed bitmap indices are also efficient for high-cardinality attributes. Timing results demonstrate that the bitmap indices significantly outperform the projection index, which is often considered to be the most efficient access method for multidimensional queries. Finally, we review the bitmap index technology currently supported by commonly used commercial database systems and discuss open issues for future research and development.
Introduction

Querying large datasets to locate some selected records is a common task in data warehousing applications. However, answering these queries efficiently is often difficult due to the complex nature of both the data and the queries. The most straightforward way of evaluating a query is to sequentially scan all data records to determine whether each record satisfies the specified conditions. A typical query condition is as follows: “Count the number of cars sold by producer P in the time interval T”. This search procedure could usually be accelerated by indices, such as variations of B-Trees or kd-Trees (Comer, 1979; Gaede & Günter, 1998). Generally, as the number of attributes in a dataset increases, the number of possible indexing combinations increases as well. To answer multidimensional queries efficiently, one faces a difficult choice. One possibility is to construct a separate index for each combination of attributes, which requires an impractical amount of space. Another possibility is to choose one of the multidimensional indices, which is only efficient for some of the queries. In the literature, this dilemma is often referred to as the curse of dimensionality (Berchtold, Boehm, & Kriegel, 1998; Keim & Hinneburg, 1999).

In this chapter we discuss an indexing technology that holds a great promise in breaking the curse of dimensionality for data warehousing applications, namely the bitmap index. A very noticeable character of a bitmap index is that its primary solution to a query is a bitmap. One way to break the curse of dimensionality is to build a bitmap index for each attribute of the dataset. To resolve a query involving conditions on multiple attributes, we first resolve the conditions on each attribute using the corresponding bitmap index, and obtain a solution for each condition as a bitmap. We then obtain the answer to the overall query by combining these bitmaps. Because the operations on bitmaps are well supported by computer hardware, the bitmaps can be combined easily and efficiently. Overall, we expect the total query response time to scale linearly in the number of attributes involved in the query, rather than exponentially in the number of dimensions (attributes) of the dataset, thus breaking the curse of dimensionality.

These statements omitted many technical details that we will elaborate in this chapter. In the next section we give a broad overview of the bitmap index and its relative strengths and weaknesses to other common indexing methods. We then describe the basic bitmap index and define the terms used in the discussions. We devote a large portion of this chapter to review the three orthogonal sets of strategies to improve the basic bitmap index. After reviewing these strategies, we give a more in-depth discussion on how the word-aligned-hybrid (WAH) bitmap compression technique reduces the bitmap index sizes. We will also present some timing results to demonstrate the effectiveness of the WAH compressed bitmap indices for two different application datasets. Our performance evaluation is deliberately based on datasets
Related Content

Inexact Field Learning Approach for Data Mining
www.igi-global.com/chapter/inexact-field-learning-approach-data/10669?camid=4v1a

Modeling Quantiles
www.igi-global.com/chapter/modeling-quantiles/10993?camid=4v1a

Neural Networks for Prediction and Classification
Kate A. Smith (2005). Encyclopedia of Data Warehousing and Mining (pp. 865-869).
www.igi-global.com/chapter/neural-networks-prediction-classification/10718?camid=4v1a
A Java Technology Based Distributed Software Architecture for Web Usage Mining
www.igi-global.com/chapter/java-technology-based-distributed-software/7667?camid=4v1a