Chapter 8.9
Vertical Fragmentation in Databases Using Data-Mining Technique

Narasimhaiah Gorla
American University of Sharjah, UAE

Pang Wing Yan Betty
Hong Kong Polytechnic University, Hong Kong

ABSTRACT

A new approach to vertical fragmentation in relational databases is proposed using association rules, a data-mining technique. Vertical fragmentation can enhance the performance of database systems by reducing the number of disk accesses needed by transactions. By adapting Apriori algorithm, a design methodology for vertical partitioning is proposed. The heuristic methodology is tested using two real-life databases for various minimum support levels and minimum confidence levels. In the smaller database, the partitioning solution obtained matched the optimal solution using exhaustive enumeration. The application of our method on the larger database resulted in the partitioning solution that has an improvement of 41.05% over unpartitioned solution and took less than a second to produce the solution. We provide future research directions on extending the procedure to distributed and object-oriented database designs.

INTRODUCTION

Vertical fragmentation (or partitioning) is a physical database design technique that is aimed at improving the access performance of user transactions. In vertical partitioning, a relation is split into a set of smaller physical files, each with a subset of the attributes of the original relation. The rationale is that database transactions normally require access only to subset of the attributes. Thus, if we can split the relation into sub files that closely match the requirements of user transactions, the access time for transactions reduces significantly.

The fragmentation problem is computationally complex. Consider a relational schema with N relations, with A_i attributes for relation i. A relation with A attributes can be partitioned in B(A) different ways (Hammer & Niamir, 1979), where B(A) is the A^{th} Bell number (for A=30, $B(A) = 10^{15}$). Using exhaustive enumeration, the number of possible fragmentations for the N-relation schema is approximately $B(A_1)B(A_2)$ ... $B(A_N)$. Yu, Chen, Heiss, and Lee (1992) find out that the number of attributes for base tables and views in a typical relational environment are 18 and 41, respectively. Even if we consider a small schema of 10 relations with 15 attributes per relation, the number of possible fragments is approximately $(10^9)^{10} = 10^{90}$. Since the problem is intractable, solving large problems requires the use of heuristic techniques.

The objective of this research is to provide a general approach for vertically fragmenting a relation. Since the problem is computationally intractable, we use a heuristic procedure to solve the problem using association rules. Our approach is based on Apriori algorithm developed by Agarwal and Srikanth (1994). We believe that "association rules" provide a natural way to represent the linkage between attributes as implied by the database transactions, thus providing a convenient way of solving the problem. Though several authors have studied vertical partitioning problem in databases, there is no study that employed association rules approach. The objective of the research is to develop a methodology for attribute partitioning with the least database operating cost, given the characteristics of relations and database transactions. The application of our algorithm using standard database workload (Yu et al., 1992) on large database resulted in an improvement of 41% over unpartitioned solution. Our association rules-based algorithm took only a few second to produce the solution, since it is relatively less complex compared to other approaches. Furthermore, the application of our methodology on small problems yielded optimal solutions as obtained by exhaustive enumeration.

The organization of the article is as follows. Section 2 provides related research in database partitioning. Section 3 provides background on association rules. Section 4 has the methodology for vertical partitioning using association rules. Section 5 contains experiments using the proposed method employing two real life data sets for various support and confidence levels. Section 6 has discussion on effectiveness of the proposed method. Section 7 deals with discussion of results and section 8 contains future research directions.

RELATED WORK IN VERTICAL PARTITIONING

Because of the criticality of the database performance, several researchers have contributed enormously to vertical partitioning problem for over two decades. Database partitioning has been applied in centralized relational databases (Ceri, Navathe, & Wiederhold, 1983; Cornell & Yu, 1990; Hoffer & Severance, 1976; Ng, Gorla, Law, & Chan, 2003; Song & Gorla, 2000), distributed databases (Baiao et al, 2004; Cheng, Lee, & Wong, 2002; March & Rho, 1995; Ozsu & Valduriez, 1996;), Data Warehouse Design (Ezeife, 2001; Furtado, Lima, Pacitti, Valduriez, & Mattoso, 2005; Labio, Quass, & Adelberg, 1997), and Object-Oriented Database design (Fung, Karlapalem, & Li, 2002; Gorla, 2001).

Hoffer and Severance (1976) consider the vertical partitioning problem by applying bond energy algorithm on similarity of attributes, which are based on access patterns of transactions. Their work was extended by Navathe, Ceri, Wiederhold, and Dou (1984) by presenting vertical partitioning algorithms for three contexts: a database stored on devices of a single type; in different memory levels; and a distributed database. They used af-