

Access Path Selection in a Relational Database Management System

Author: P. Griffiths Selinger et al.

Reviewer: Gae-won You

Summary:

This paper introduces the mechanism on how to optimize query processing. Especially, it focuses on query optimizer in System R which deals with how to find an optimal access path out of numerous ones. It defines a cost model through statistical analysis and strives to find the access path minimizing the cost with the various statistics. First, it defines a cost model on single relation by using selectivity estimation under the assumption that every data distribution is uniform and independent. It also extends the model up to not only 2-way but also n-way join relation. Especially, n-way join is combined as the sequence of the optimized 2-way joins. Meanwhile, to prevent to increase the search space explosively as n increases, it adopts a kind of dynamic programming technique along with another heuristic. In addition, it discusses the way of scheduling more complex queries such as nested queries.

Comments:

This access path selection is an essential problem to efficiently compute the query results as the scale of data size increases more explosively and the form of query becomes more complicated. I believe this paper is the evolutionary work leading database area to a novel field. Moreover, it is very clear and intuitive to gradually extend the optimization mechanism from single relation to 2-way join and from 2-way join to n-way join.

However, I believe this paper will be able to be strengthened as follows:

First, in the selectivity estimation, it assumes that the data distribution is even and every column is independent each other, but I think this assumption is very strong. For instance, practically, many different distributions can exist, e.g., Gaussian, Zipfian, etc. In addition, the correlation between two columns included in a predicate makes the query result more selective and the cost can be underestimated. Thus, the relaxation of such assumptions will make the estimation more accurate and reduce the cost of the query processing.

Second, it suggests the strategies for the reduction of search space and decreases it at most up to 2^n . However, if n is large, it is considerably expensive still. It seems to need another interesting approximation approach.

Third, it needs to experimentally validate the efficiency and effectiveness of the approaches as mentioned as future work in this paper.