Algorithms for Index-Assisted Selectivity Estimation

Paul M. Aoki*
Deparment of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720-1776 USA

Abstract

The standard mechanisms for query selectivity estimation used in relational database systems rely on properties specific to the attribute types. The query optimizer in an extensible database system will, in general, be unable to exploit these mechanisms for user-defined types, forcing the database extender to invent new estimation mechanisms. In this work, we discuss extensions to the generalized search tree, or GiST, that simplify the creation of user-defined selectivity estimation methods. An experimental comparison of such methods with multidimensional estimators from the literature has demonstrated very competitive results.

1. Motivation and General Approach

There has been considerable research in the development of query selectivity estimation techniques, but relatively little in the context of frameworks that can be applied to arbitrary user-defined types. The general frameworks that have been proposed tend to be conceptually simple APIs that impose significant implementation complexity on the database extender. In this research, we have examined several techniques for using the generalized search tree, or GiST, to implement a selectivity estimation framework for extensible database management systems.

From an engineering viewpoint, the main benefit of our index-based approach is that it applies a solution to a relatively well-understood problem (search) to a relatively poorly-understood problem (estimation). This will help database extenders, who are typically domain knowledge experts in areas such as computer vision, to produce estimators of reasonable quality without becoming experts in database-specific domains (statistics, query processing cost models, *etc.*). The intuitive appeal of this approach is supported by an empirical trend observed by extensible database vendors: third-party extenders are far more likely to try to integrate search structures than they are to produce non-trivial selectivity estimators.

*Research supported by NSF under grant IRI-9400773 and by NASA under grants FD-NAG5-6587 and FD-NAGW-5198.

2. Proposed Solutions

We have studied a variety of algorithms based on two main techniques, *prioritized traversal* and *random sampling*, as well as various methods for combining them. Prioritized traversal exploits the clustering inherent in an index structure. A GiST constitutes a recursive partitioning of a data set down to an arbitrary resolution; as has been previously observed, this can be used like an external memory histogram, and we propose prioritized traversal as a new heuristic for doing so. The random sampling algorithms are based on the pseudo-ranking technique for tree sampling combined with statistical estimators of varying complexity.

Our main design contribution lies in recognizing the effects of (1) varying index effectiveness and (2) a cost-limited environment. Our algorithms make "best effort" use of an explicit, limited I/O budget to produce interval estimates. For example, the best strategy depends on how effectively the index answers the search predicate (a good index is often a good "histogram") and the I/O budget (sampling produces very bad estimates unless enough samples can be obtained). Since the first factor is not known in advance, we must consider adaptive combination algorithms as well as fixed algorithms.

The algorithms, experimental results, and an extensive discussion of both background and related work may be found in [1]. In particular, results from an experimental comparison between our estimation algorithms and several multidimensional estimators (*i.e.*, those based on the uniformity assumption, Hausdorff fractal dimension, correlation fractal dimension and density) have been promising.

References

[1] P. M. Aoki. Algorithms for Index-Assisted Selectivity Estimation. Technical Report UCB//CSD-98-1021, University of California, Berkeley, CA, Oct. 1998.