Dynamic Histograms: Capturing Evolving Data Sets

Donko Donjerkovic
University of Wisconsin–Madison
donko@cs.wisc.edu

Yannis Ioannidis
University of Athens
yannis@di.uoa.gr

Raghu Ramakrishnan
University of Wisconsin–Madison
raghu@cs.wisc.edu

Abstract

Conventional histograms are ‘static’ since they cannot be updated but only recalculated. In this paper, we introduce a ‘dynamic’ version of V-optimal histograms, which is constructed and maintained incrementally.

1. Introduction

The best plan to execute a database query depends on data distributions of the relations involved in a query. Typically, these distributions are captured by histograms and therefore the optimality of a query plan is dependent on the quality of relevant histograms.

Histograms are considered static structures, since they cannot be updated when the underlying relation is updated. Therefore, frequent updates may easily make a histogram obsolete and mislead the query optimizer. The traditional solution to this problem is to periodically recalculate histograms. Unfortunately, this leaves the database administrator with a difficult choice: (1) choose a long update period and risk having outdated histograms, or (2) choose a short update period and risk overloading the system.

As a solution to this problem, we propose a dynamic histogram that can be updated and maintained incrementally with low overhead. Dynamic histograms are an important contribution towards the widespread deployment of self administering database systems.

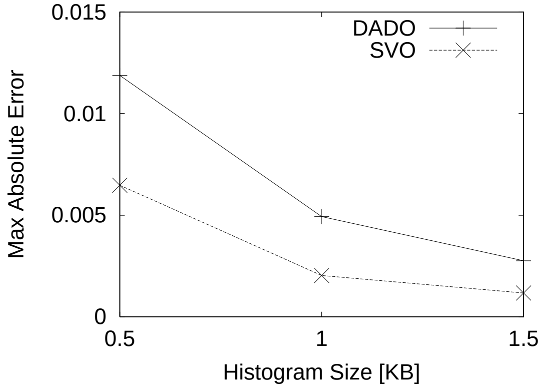
2. Dynamic v-optimal (DVO) histogram

The objective of V-Optimal histograms is to minimize the sum of frequency variations within each bucket. The exact variation in frequency within each bucket of a DVO histogram cannot be measured due to the limited space given to the histogram. For the sole purpose of approximating this variation, we maintain two subbuckets in every bucket of a DVO histogram. To capture distribution changes, a DVO histogram may merge and split buckets. The best bucket to split is the one with the largest internal variance. Similarly, the best pair to merge is the one with the smallest combined variance.

The objective of overall frequency variance minimization can be dynamically achieved as follows: (1) Upon every insertion, increment the appropriate subbucket counter. (2) Check if the insertion bucket has become the best one to split or to merge. (3) Split and merge the best candidates only if the overall variance would be reduced. Step (2) can be executed efficiently by storing the variance of the best candidates.

3. Performance result

We show the maximum absolute error (as a fraction of the relation size) in selectivity estimates for different histogram sizes. Algorithms presented are: (1) Dynamic Absolute Deviation Optimal (DADO), which is a variation of a DVO algorithm, (2) Static V-optimal (SVO) algorithm, recomputed after the final update. We note that the two algorithms have comparable precision.



4. Conclusion

Dynamic histograms approximate evolving data sets without serious quality degradation, compared to the periodical reconstruction technique. More details are available in the full paper [1].

References

[1] D. Donjerkovic, Y. Ioannidis, and R. Ramakrishnan. Dynamic histograms: Capturing evolving data sets. *UW-Madison Technical Report CS-TR-99-1396*, March 1999.