A Single Pass Computing Engine for Interactive Analysis of VLDBs

Ted Mihalisin
Mihalisin Associates, Inc.
P.O. Box 3183, Maple Glen, PA 19002
tmihal@bellatlantic.net

Abstract

A single pass computing and visualization engine will be demonstrated that allows one to interactively analyze VLDBs that contain tens of millions of multivariate records. The engine allows one to compute millions of different quantities from a single pass over the records. Each computation can be performed for the entire multivariate domain and for all sub-domains that can be obtained by constraining one or more discrete variables to span any subset of their values and one or more continuous variables to span any subset of their predefined bins. Each new computation takes less than one second irrespective of the number of records. Our patented visualization method (U.S. Patent No. 5,228,119) allows one to quickly detect highly conditional multidimensional correlations, clusters and associations.

The analysis of very large multivariate databases is a topic of growing importance for business, research and government endeavors. Analysts of all types are struggling to find effective means of extracting information from huge data warehouses. The methods being utilized range from neural nets to OLAP (On Line Analytical Processing) to statistics and data visualization. Many of these methods cannot be applied to the entirety of data due to their computational complexity which would lead to horrendously slow performance when applied to tens of millions of records. Some methods require multiple scans of the data records which also results in very poor performance. Finally, some methods utilize precomputed results (to enhance performance) in an attempt to anticipate analyses that will be desired at a future date. Unfortunately these methods require enormous amounts of storage space.

We have developed a method of analyzing VLDBs where one can perform a wide range of analyses using all of the data. The method involves a single pass over the records and provides extremely fast performance. Moreover, the method does not require large amounts of storage since only one data view needs to be stored. Other data views can be computed in about one second. Two phases of computing are performed. The first phase is an unattended batch phase and it has an execution time that is proportional to the number of records. The single pass over the records occurs during the batch phase. During this phase certain "core" statistics are computed for a set of dependent variables (response variables or measures) for each cell in a multivariate (multidimensional) space of independent variables (explanatory variables or dimensions). That is, the statistics are computed for all of the records in each such cell. The second phase is an interactive phase in which the analyst gets sub-second responses, irrespective of the number of records (data points) in the dataset, as he or she requests new computations based on the results of previous computations which are quickly comprehended using our patented method for visualizing multidimensional data. Analysts can literally perform point and click interactive exploratory data analysis on tens of millions of records. Each new computation would, generally speaking, require hours of computing time if one used conventional methods instead of the computing engine described here. The "trick" of course is that our method of computing does not require additional passes over the tens of millions of records for each new calculation. Nor does it require associated slow processes such as SQL queries and hard disk accesses.

Proceedings of the 24th VLDB Conference
New York, USA, 1998