

Extending and Optimizing SQL for Data Mining

Carlos Ordonez
Teradata, NCR

Talk Outline

The talk summarizes my research on using the SQL language to enable data mining inside a relational database system. The focus will be on query optimization, a classical problem in databases. Most research issues discussed will be based on the Teradata parallel relational DBMS, which as of today remains one of the most scalable data warehousing system capable of storing very large tables with billions of rows. I will start explaining ideas to extend SQL aggregations to compute percentages and to build tabular (denormalized) data sets. Tabular data sets are the basic input for machine learning or statistical algorithms and generally require significant manual work, with long SQL queries during the data pre-processing stage. On the other hand, percentages are ubiquitous in any type of statistical or data mining analysis. Then I will discuss alternatives to make machine learning and statistical algorithms work inside the database system. Alternatives go from internally manipulating matrices to expressing mathematical equations as SQL queries. I will briefly discuss SQL implementations of the K-means algorithm and the EM algorithm. I will conclude my talk discussing future research directions. My first research direction is on the integration of multivariate linear models with a database system combining SQL and User-Defined Functions. The second research direction is on the improvement of heart disease prediction helped by association rules, clustering and Bayesian classifiers.

Speaker Bio

Carlos Ordonez received a degree in Applied Mathematics and an M.S. degree in Computer Science, both from the UNAM University, Mexico, in 1992 and 1996 respectively. Carlos got a Ph.D. degree in Computer Science from the Georgia Institute of Technology, USA, in 2000. Carlos currently works for Teradata conducting research on databases and machine learning. He has published more than 20 scientific articles in international journals and conferences.