

Caching Statistics in Spatial Data Structures: Fast Real-time Massive Science Data Analysis

Andrew W. Moore (Department of Computer Science, Carnegie Mellon University)
`awm@cs.cmu.edu`

Abstract

This talk is about the algorithmic challenges involved in allowing biologists and astrophysicists to continue using the modeling and inference tools they've been happily applying to megabytes of data, when they start drawing in terabytes of data.

We'll discuss new algorithms and data structures that fall into the class of "cached sufficient statistics". These are summary data structures that live between the statistical algorithm and the database, intercepting the kinds of operations that have the potential to eat up valuable time if they were answered by direct reading of the dataset. Some structures may be familiar (kd-trees and R-trees, for example) while some are new (All-dimensions trees, and the Anchors Hierarchy for high dimensions), but for all structures we introduce new search algorithms operating on the cached structures that have interesting properties which call for further development.

I will give some computer demonstrations showing various classes of accelerations broadly covering kernel density speedups (1000 fold), k -means, mixture and hierarchical clustering speedups (1000-10000 fold), anomaly detection (100-fold) and 2-, 3-, 4- and 5-point correlation function computation (100-fold up to about a trillion-fold). If time permits we will also discuss (i) "racing" methods to accelerate expensive model selection operations by early termination of models that have less than delta probability of being more than epsilon better than the best model and (ii) preliminary results of new optimization approach to non-linear regression of image morphology parameters for tens of millions of galaxy images.

In collaboration with: Brigham Anderson, Alex Gray, Dan Pelleg, Mary Soon Lee, Jeff Schneider, Bob Nichol, Andy Connolly (U Pitt), Alex Szalay (JHU), Larry Wasserman, Weng-Keen Wong. Related papers and software download information: www.autonlab.org.