May 2013

MADALGO seminar by Zhewei Wei, Aarhus University

Range Summary Queries

Abstract:

Database queries can be broadly classified into two categories: reporting queries and aggregation queries. The former retrieves a collection of records from the database that match the query's conditions, while the latter returns an aggregate, such as count, sum, average, or max (min), of a particular attribute of these records. Aggregation queries are especially useful in business intelligence and data analysis applications where users are interested not in the actual records, but some statistics of them. They can also be executed much more efficiently than reporting queries, by embedding properly precomputed aggregates into an index.

However, reporting and aggregation queries provide only two extremes for exploring the data. Data analysts often need more insight into the data distribution than what those simple aggregates provide, and yet certainly do not want the sheer volume of data returned by reporting queries. In this paper, we design indexing techniques that allow for extracting a statistical summary of all the records in the query. The summaries we support include frequent items, quantiles, various sketches, and wavelets, all of which are of central importance in massive data analysis. Our indexes require linear space and extract a summary with the optimal or near-optimal query cost.

Joint work with: Ke Yi