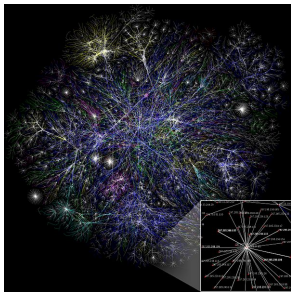


Mergeable Quantile Summaries

Dan Stubbs

Big Data

- ▶ In the modern world, with the size of available data sets increasing so quickly, RAM-model algorithms become increasingly unsuitable to answer many important questions, mostly related to the internet.



- ▶ Generally the problem of oversized data is solved by either viewing the data one element at a time (streaming), or by splitting the data into multiple segments (MUD).

Summary Data Structures

- ▶ A standard procedure in any situation with big data is to maintain some summary as the bulk of the data gets processed and then discarded.
- ▶ In short, a summary data structure is any data structure that maintains information sufficient to answer some questions about an underlying data set in much less space than it would take to maintain the entire data set.

ε -Approximate Quantile Summary

- ▶ An ε AQS is a summary that can be given arbitrary quantile queries, and gives an answer whose rank is within ε times the size of the data set of the desired quantile.
- ▶ This can be accomplished by maintaining only $O(\varepsilon^{-1})$ different values, with each serving as the answer for an $O(\varepsilon)$ -wide band of quantile queries.

Mergeability of Summaries

- ▶ We say a summary is *mergeable* if two summaries S_1 and S_2 for disjoint data sets D_1 and D_2 with approximation bounds ε_1 and ε_2 can be used to construct a new summary S that $\max(\varepsilon_1, \varepsilon_2)$ approximates $D_1 \cup D_2$.
- ▶ Many summaries are in effect multiplication of the input's frequency vector by some random matrix. By associativity of matrix multiplication, summing two summary vectors constructed this way provides a summary of the sum of their frequency vectors, i.e. the union of their underlying data sets.

Benefits of Standardization

- ▶ Since mergeable summaries are a young field, there are multiple different systems of notation across the literature and empirical tests are thin on the ground.
- ▶ Synthesis and simulation of these algorithms may expose underlying properties – one algorithm is always better than another, or particularly powerful under certain assumptions – that are not yet apparent in the theoretical work.

Example Mergeable Summary: Count-Min Sketch

- ▶ The Count-Min Visualizer shows one linear sketch summary in action, it supports quantiles, as well as a number of other statistics of the summarized data set.