Compact Data Representations and their Applications

Moses Charikar
Princeton University

Several algorithmic techniques have been devised recently to deal with large volumes of data. At the heart of many of these techniques are ingenious schemes to represent data compactly. This talk will present some constructions of such compact representation schemes (also referred to as sketches) for estimating distances between sets, vectors, and distributions on an underlying metric (where distance is measured by the Earth Mover Distance (EMD)). The construction of these compact representation schemes is motivated by techniques used in approximation algorithms to round solutions of LP and SDP relaxations for optimization problems. There are interesting connections between such sketching methods and low distortion embeddings of metric spaces into ℓ_1. Such compact representation directly lead to efficient approximate nearest neighbor search algorithms. We will also see how such schemes lead to efficient, one-pass algorithms for processing large volumes of data (streaming algorithms).

Finally, I will talk about some recent work applying these ideas to designing compact data structures for content-based image retrieval systems. The main challenge here is to achieve high-quality similarity searches while using very compact meta-data. We adapt the ideas from the sketch constructions for EMD, as well as other ideas from embeddings of normed spaces to produce compact sketches for images. I will discuss results from a prototype implementation on a database with 10,000 images. Our results show that our method can achieve more effective similarity searches than previous approaches with meta-data significantly smaller than previous systems.

The work on content-based image retrieval is joint work with Qin Lv and Kai Li at Princeton.