Towards overcoming the reranking bottleneck

Paul Erdős Memorial Lecture

Piotr Indyk[*]

Reranking is a popular approach to information retrieval. It proceeds in two stages. In the first stage, a "quick-and-dirty" data structure retrieves a shortlist of r points closest to the query, where the length of the shortlist r is larger than the desired output k. In the second stage, the shortlist is post-processed to identify $k \ll r$ points that satisfy the desired objective. For example, the postprocessing could identify the k most "diverse" points in the shortlist or use a "slower-but-accurate" distance metric to identify the best answers. Despite its popularity, it has various drawbacks; notably the quality of the output is limited by the accuracy of the first stage.

In this talk, I will discuss an alternative to reranking, which fuses the two stages into a single search procedure. The new approach crucially uses recent developments in graph-based algorithms for high-dimensional similarity search, as well the tools developed to analyze such algorithms.

---

[*]CSAIL, MIT, `indyk@mit.edu`