A randomized approximate nearest neighbors algorithm

Andrei Osipov, Yale University

We present a randomized algorithm for the approximate nearest neighbor problem in d-dimensional Euclidean space. Given N points $\{\boldsymbol{x}_j\}$ in $\mathbb{R}^d$, the algorithm attempts to find k nearest neighbors for each of $\boldsymbol{x}_j$, where k is a user-specified integer parameter. The algorithm is iterative, and its CPU time requirements are proportional to $T \cdot N \cdot (d \cdot (\log d) + k \cdot (d + \log k) \cdot (\log N)) + N \cdot k^2 \cdot (d + \log k)$, with T the number of iterations performed. The memory requirements of the procedure are of the order $N \cdot (d + k)$.

A byproduct of the scheme is a data structure, permitting a rapid search for the k nearest neighbors among $\{\boldsymbol{x}_j\}$ for an arbitrary point $\boldsymbol{x} \in \mathbb{R}^d$. The cost of each such query is proportional to $T \cdot (d \cdot (\log d) + \log(N/k) \cdot k \cdot (d + \log k))$, and the memory requirements for the requisite data structure are of the order $N \cdot (d + k) + T \cdot (d + N)$.

The algorithm utilizes random rotations and a basic divide-and-conquer scheme, followed by a local graph search. We analyze the scheme's behavior for certain types of distributions of $\{\boldsymbol{x}_j\}$, and illustrate its performance via several numerical examples.

Joint work with Peter W. Jones, Andrei Osipov and Vladimir Rokhlin