Abstract

We show a new approach to the approximate near neighbor problem, which improves upon the classic Locality Sensitive Hashing (LSH) scheme. Our new algorithms obtain query time (roughly) quadratically better than the optimal LSH algorithms of [Indyk-Motwani’98] for the Hamming space, and [Andoni-Indyk’06] for the Euclidean space. For example, for the Hamming space, our algorithm has query time $n^{\cdot r}$ and space $n^{\cdot \{1+r\}}$, where $r=1/(2c-1)+o(1)$ for $c$-approximation. Our algorithms bypass the lower bounds for LSH from [ODonnell-Wu-Zhou’11].

The new approach is based on hashing that itself depends on the given pointset. In particular, one of the main components is a procedure to decompose an arbitrary pointset into several subsets that are, in a certain sense, pseudo-random. Our data-dependent hashing scheme is optimal.

Based on a few joint papers with Piotr Indyk, Huy Nguyen, and Ilya Razenshteyn.