Large-scale similarity joins with guarantees*

Rasmus Pagh

IT University of Copenhagen, Denmark

Abstract. The ability to handle noisy or imprecise data is becoming increasingly important in computing. In the information retrieval community the notion of similarity join has been studied extensively, yet existing solutions have offered weak performance guarantees. Either they are based on deterministic filtering techniques that often, but not always, succeed in reducing computational costs, or they are based on randomized techniques that have improved guarantees on computational cost but come with a probability of not returning the correct result.
The aim of this talk is to give an overview of randomized techniques for high-dimensional similarity search, and then proceed to discuss two recent advances. First, we consider ways of improving the locality of data access by using a recursive approach. This provably lowers the I/O cost of large-scale similarity joins. Second, we consider new methods for eliminating the probability of error inherent in classical locality-sensitive hashing methods for similarity join in Hamming space, while almost matching their theoretical performance.

* The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 614331.