Locality Sensitive Hashing Schemes, Similarities, and Distortion (Invited Talk)

Flavio Chierichetti(✉)

Dipartimento di Informatica, Sapienza University of Rome, Rome, Italy
flavio@di.uniromal.it

Abstract. Locality sensitive hashing (LSH) is a key algorithmic tool that lies at the heart of many information retrieval and machine learning systems [1, 2, 8]. LSH schemes are used to sketch large objects (e.g., Web pages, fields of flowers, or – more generally – sets and vectors) into fingerprints of few bits each; the fingerprints are then used to quickly, and approximately, reconstruct some similarity relation between the objects.

A LSH scheme for a similarity (or, analogously, for a distance) can significantly improve the computational cost of many algorithmic primitives (e.g., nearest neighbor search, and clustering). For this reason, in the last two decades, researchers have tried to understand which similarities admit efficient LSH schemes: such schemes were obtained for many similarities [1–3, 7–9], while the non-existence of LSH schemes was proved for a number of other similarities [3].

In our talk, we will introduce the class of LSH-preserving transformations [4] (functions that, when applied to a similarity that admits a LSH scheme, return a similarity that also admits such a scheme). We will give a characterization of this class of functions: they are precisely the probability generating functions, up to scaling. We will then show how this characterization can be used to construct LSH schemes for a number of well-known similarities.

We will then discuss a notion of similarity distortion [6], in order to deal with similarities which are known to not admit LSH schemes — this notion aims to determine the minimum distortions that these similarities have to be subject of, before starting to admit a LSH scheme. We will introduce a number of general theoretical tools that can be used to determine the optimal distortions of some important classes of similarities.

Finally, we will consider the computational problem of checking whether a similarity admits a LSH scheme [5], showing that, unfortunately, this problem is computationally hard in a very strong sense.

Supported in part by the ERC Starting Grant DMAP 680153 and by a Google Focused Research Award.

B. Catania et al. (Eds.): SOFSEM 2019, LNCS 11376, pp. 531–532, 2019.
https://doi.org/10.1007/978-3-030-10801-4

References

1. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: FOCS, pp. 459–468 (2006)
2. Broder, A.Z.: On the resemblance and containment of documents. In: Proceedings of the SEQUENCES, pp. 21–29 (1997)
3. Charikar, M.: Similarity estimation techniques from rounding algorithms. In: Proceedings of the STOC, pp. 380–388 (2002)
4. Chierichetti, F., Kumar, R. LSH-preserving functions and their applications. J. ACM **62**(5), 33:1–33:25 (2015)
5. Chierichetti, F., Kumar, R., Mahdian, M.: The complexity of LSH feasibility. Theor. Comput. Sci. **530**, 89–101 (2014)
6. Chierichetti, F., Kumar, R., Panconesi, A., Terolli, E.: The distortion of locality sensitive hashing. In: ITCS (2017)
7. Christiani, T., Pagh, R.: Set similarity search beyond minhash. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, pp. 1094–1107, New York, NY, USA. ACM (2017)
8. Gionis, A., Indyk, P., Motwani, R., et al.: Similarity search in high dimensions via hashing. In: VLDB, pp. 518–529 (1999)
9. Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: STOC, pp. 604–613 (1998)

Author Index