

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

Flavio Chierichetti^(✉)

Dipartimento di Informatica, Sapienza University of Rome, Rome, Italy
flavio@di.uniroma1.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.

© Springer Nature Switzerland AG 2019

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

- Aßmann, Uwe 1
- Banerjee, Indranil 54
Baste, Julien 67
Bielikova, Maria 435
Bodlaender, Hans L. 81
Bonchi, Francesco 21
Bottesch, Ralph 504
- Carosi, Raffaello 94
Chen, Taolue 206
Chierichetti, Flavio 531
Chromý, Miloš 108
Chudá, Daniela 298
Cicalese, Ferdinando 122
- D'Emidio, Mattia 136
Das, Shantanu 150, 164
Demeyer, Serge 419
Di Luna, Giuseppe A. 150
Di Stefano, Gabriele 136
Donselaar, Nils 179
- Fernau, Henning 192
Fioravanti, Simone 94
- Gao, Chong 206
Gasieniec, Leszek A. 150
Giachoudis, Nikos 164
Gözüpek, Didem 67
Grzelak, Dominik 1
Gualà, Luciano 94
Gupta, Manoj 221
Gurski, Frank 234
- Hemaspaandra, Lane A. 247
- Interian, Ruben 324
- Kalinowski, Marcos 324
Kapoutsis, Christos A. 28
Kaufmann, Michael 260
Komusiewicz, Christian 272
- Kučera, Petr 108
Kumar, Hitesh 221
Kuppusamy, Lakshmanan 192
Kutrib, Martin 285
- Labaj, Martin 298
Lipták, Zsuzsanna 122
Luccio, Flaminia L. 164
- Makarov, Vladislav 310
Markou, Euripides 164
Mendoza, Isela 324
Mey, Johannes 1
Misra, Neeldhara 221, 341
Molontay, Roland 354
Monaco, Gianpiero 94
Mousavi, Mohammad Reza 490
Murta, Leonado Gresta Paulino 324
- Nagy, Benedek 406
Nahimovs, Nikolajs 368
Narváez, David E. 247
Navarra, Alfredo 136
- Okhotin, Alexander 310
Ono, Hiroataka 379
Osula, Dorota 392
Otto, Friedrich 406
- Parsai, Ali 419
Pikuliak, Matus 435
Pukhkaiev, Dmytro 1
Püschel, Georg 1
- Raman, Indhumathi 192
Rástočný, Karol 298
Raszyk, Martin 447
Rehs, Carolin 234
Richards, Dana 54
Rossi, Massimiliano 122
- Schöne, René 1
Shalom, Mordechai 67
Shinkar, Igor 54

- Simko, Marian 435
Skorski, Maciej 461
Sommer, Frank 272
Sonar, Chinmay 341
Souza, Uéverton 324
Staron, Mirosław 39
- Theobald, Martin 50
Thilikos, Dimitrios M. 67
Torenvliet, Leen 504
- van der Wegen, Marieke 81
van Rooij, Johan M. M. 473
- van Rooij, Sebastiaan B. 473
van der Zanden, Tom C. 81
Varga, Kitti 354
Varshosaz, Mahsa 490
- Wendlandt, Matthias 285
Werner, Christopher 1
Witteveen, Jouke 504
Wu, Zhilin 206
- Yamakami, Tomoyuki 519
Yamanaka, Hisato 379