The distortion of locality sensitive hashing

Flavio Chierichetti*
Sapienza University of Rome

Ravi Kumar
Google, Mountain View

Alessandro Panconesi*
Sapienza University of Rome

Erisa Terolli*†
Sapienza University of Rome

ABSTRACT

Given a pairwise similarity notion between objects, locality sensitive hashing (LSH) aims to construct a hash function family over the universe of objects such that the probability two objects hash to the same value is their similarity. LSH is a powerful algorithmic tool for large-scale applications and much work has been done to understand LSHable similarities, i.e., similarities that admit an LSH. In this work we focus on similarities that are provably non-LSHable and propose a notion of distortion to capture the approximation of such a similarity by a similarity that is LSHable. We consider several well-known non-LSHable similarities and show tight upper and lower bounds on their distortion.We also experimentally show that our upper bounds translate to effective algorithms in practice.

This work is on the topic of similarities and locality sensitive hashing (LSH). LSH is a powerful algorithmic paradigm for computing similarities between data objects in an efficient way. Informally, an LSH scheme for a similarity is a probability distribution over a family of hash functions such that the probability the hash values of two objects agree is precisely the similarity between them. In many applications, computing similar objects (i.e., finding nearest neighbors) can be computationally very demanding and LSH offers an elegant, cost-effective, and practical alternative.

Intuitively, large objects can be represented compactly and yet accurately from the point of view of similarity, thanks to LSH. Thus, the similarity between two objects can be quickly estimated by picking a few random hash functions from the family and estimating the fraction of times the hash functions agree on the two objects. This paradigm has been very successful in a variety of applications dealing with large volumes of data, from near-duplicate estimation in text corpora to nearest-neighbor search in a multitude of domains. The success and importance of LSH has been recognized by the theory community. Researchers have constantly looked for LSH schemes for more and more similarities. Thus a natural question arises: which similarities admit an LSH scheme?

Charikar [1] introduced a couple of necessary criteria for a similarity to admit an LSH, which can be used to to rule out the existence of LSH schemes for various similarities, for instance, the Sørensen–Dice and Sokal–Sneath similarities [2]. This leads to a very natural question that is addressed in the work: 'If a similarity does not admit an LSH scheme, then how well can it be approximated by another similarity that admits an LSH?'

The conceptual contribution of the work is the novel notion of locality sensitive distortion. Deriving inspiration from distortion in metric space embeddings, the work develops a robust notion of distortion for LSH. A similarity has a certain distortion factor if there is another similarity defined on the same universe that admits an LSH and such that the two similarities are always (multiplicatively) within the distortion factor.

Once this definition is in place, the work launches a systematic investigation of the notion of distortion for LSH schemes and proves optimal distortion bounds for several well-known and widely used similarities such as cosine, Simpson [2], Braun–Blanquet [2], Sørensen–Dice [2] and several others. As concrete examples, the work shows that the distortion of cosine similarity is $\Theta\left(\sqrt{n}\right)$ and that of Braun–Blanquet and Sørensen–Dice similarities is two.

Technically speaking, the lower bounds are the most challenging. They are obtained by two new combinatorial tools introduced in the work, namely, the center method and the k-sets method. The center method, which is easier to apply than the k-sets method, is mostly based on a geometric averaging argument. The k-sets method is based on certain extremal properties of intersecting families of sets and uses some deep results in extremal combinatorics. The center method is applicable to many instances of similarity but the k-sets method is unavoidable in the following sense. Braun–Blanquet similarity not passes the Charikar tests, but also the test provided by the center method. However, the more powerful k-sets method can instead be used to show a distortion bound of two. Other similarities to which the k-sets method applies are Sørensen–Dice and the infinite family of Sørensen similarities. Interestingly, in nearly all cases, the work also exhibit matching distortion upper bounds by explicitly constructing an LSH. This is what makes the work quite interesting and complete.

The main motivation behind the work is to extend the range of applicability of LSH as far as possible and the concept of distortion should be understood in these terms. For instance, even if a similarity is shown not to admit an LSH scheme it might be possible to approximate it efficiently by means of LSH schemes of other similarities that are close to it. The results show that some cases, such as cosine, are a forlorn hope (since the distortion is not a constant), but in other instances, such as Sørensen–Dice and Braun–Blanquet, our bounds give reasons to be optimistic. As a first "proof of concept" of the notion of distortion the work performs a series of experiments with real-world text corpora. The results are encouraging, for they show that the distortion of real data sets is smaller than the worst case.

REFERENCES

[1] Moses S. Charikar. 2002. Similarity Estimation Techniques from Rounding Algorithms. In *Proceedings of STOC'02*. 380–388.
[2] Michel Marie Deza and Elena Deza. 2009. *Encyclopedia of distances*. Vol. 94. Springer.

*Supported in part by a Google Focused Research Award, by the Sapienza Grant C26M15ALKP, by the SIR Grant RBSI14Q743, and by the ERC Starting Grant DMAP 680153
†Corresponding author: terolli@di.uniroma1.it