Fast and Compact Hamming Distance Index

Simon Gog[1] and Rossano Venturini[2]

[1] Institute of Theoretical Informatics, Karlsruhe Institute of Technology
[2] Department of Computer Science, University of Pisa and Istella Srl

Abstract. This paper proposes new solutions for the approximate dictionary queries problem. These solutions combine the use of succinct data structures with an efficient representation of the keys to significantly reduce the space usage of the state-of-the-art solutions without introducing any time penalty. Finally, by exploiting triangle inequality, we can also significantly speed up the query time of the existing solutions.

Indexing large collections of objects for similarity search is a core task of many applications in databases, pattern recognition, and information retrieval (IR) (e.g., near-duplicate detection in a collection of web pages or content-based retrieval in a collection of images). In a general framework, each object is modeled with a vector of features and the similarity of two objects is measured by means of a similarity function, e.g., the Jaccard and the cosine similarity, on their vectors. Sketching techniques are used to reduce the dimensionality of these objects by producing succinct sketches of the vectors, such that the similarity of objects can be estimated by computing the Hamming distance of their sketches [1]. The *Approximate Dictionary Queries* problem is formulated as follows. We are given a dictionary set of binary keys, a k-query Q asks for identifying all the keys in D which are at Hamming distance at most k from Q. While there exist several theoretical solutions which provide guarantees on their space and time complexities (see e.g., [2] and references therein), the most efficient solutions in practice [6, 5] are based on the multi-index approach proposed by Greene et al. [4].

In this extended abstract we shortly summarizes the work in [3]. This paper improves the performance of any multi-index based solution by introducing the use of succinct data structures together with a suitable representation of the keys, SIMD-based optimizations, and the use of the triangle inequality. An empirical evaluation shows that the time improvement of the proposed solutions is by a factor up to 3.4 while the space usage reduction ranges between 22% and 40%.

References

1. M. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. STOC*, pages 380–388, 2002.
2. R. Cole, L.-A. Gottlieb, and M. Lewenstein. Dictionary matching and indexing with errors and don't cares. In *Proc. STOC*, pages 91–100, 2004.
3. S. Gog and R. Venturini. Fast and compact Hamming distance index. In *SIGIR*, 2016.
4. D. H. Greene, M. Parnas, and F. F. Yao. Multi-index hashing for information retrieval. In *Proc. FOCS*, pages 722–731, 1994.
5. A. X. Liu, K. Shen, and E. Torng. Large scale hamming distance query processing. In *Proc. ICDE*, pages 553–564, 2011.
6. G. S. Manku, A. Jain, and A. D. Sarma. Detecting near-duplicates for web crawling. In *Proc. WWW*, pages 141–150, 2007.