technische universität
dortmund

sisap

An Alternating Optimization Scheme for Binary Sketches for Cosine Similarity Search

Erik Thordsen and Erich Schubert

TU Dortmund, Informatik 8 Data Mining

{firstname.lastname}@tu-dortmund.de

SISAP 2023

Binary Sketching & Indexing

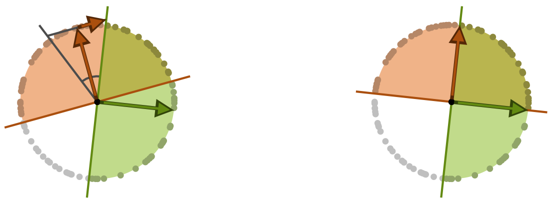
- Binary sketching defines a map $H : X \to \{0,1\}^B$
- "Quality" of sketches is induced by downstream applications
- For indexing, quality of sketches often
  $Q(H) \approx \mathrm{Corr}_{x,y\in X}\left(d_{Hamming}(H(x),H(y)), d_X(x,y)\right)$
- Approximate k-nn search by, e.g.:
  - Scan $H(X)$ for k'-nn with $d_{Hamming}$ (cheap; many)
  - Refine k'-nn with d_X to k-nn (expensive; few)

The (Euclidean) Spherical Case

- Most natural separation by dot product
  ⇒ B hyperplanes – one per bit
  ⇒ Tessellation of the d-sphere
- "Optimal" tessel. should have homogeneous sample counts, surface density integrals, and shapes
- "Balance" (∝ "entropy") of bits can be maximized without affineness

Alternating Optimization (HIOB)

- Idea: Improve initial hyperplanes by rotation
  - Homogeneous sample counts induced by pairwise independent bits of hyperplanes
  - Hope for surface area and shape to "work out" (by adding noise to X)
- Rotation by additive tangential vector (see Figure 1)
  ⇒ Aggregation of multiple updates if desired
- Scale rotation angle to help with convergence
- Work on varying subsamples to speed process up
- Observation: With "good" initialization, always only updating "worst offenders" works best

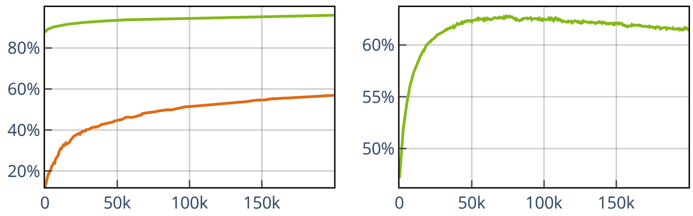


(a) Before update **(b)** After update

Figure 1: Example of an update with exactly two planes.

Evaluation

- Our approach improves bit "balance" (see Figure 2)
- "Indexing quality" of sketches improves aswell but can fall off (see Figure 2)
- Bruteforce search on optimized sketches can outperform some indices (see challenge)
- HNSW is still much faster, but builds much slower



(a) Bit "balance" (min & mean) **(b)** 10@50-recall

Figure 2: Balance and recall over iterations of HIOB

Erik Thordsen

Corresponding Author
PhD Student at TU Dortmund
Other topics:

- Intrinsic Dimensionality
- High-dimensional data

* Source code at https://github.com/eth42/hiob/