Permutation Search Methods are Efficient, Yet Faster Search is Possible

Bilegsaikhan Naidan, Leonid Boytsov, and Eric Nyberg

NTNU
Norwegian University of Science and Technology

Carnegie Mellon University
School of Computer Science

Problem

- **Nearest neighbor (NN)** search is a **fundamental** operation in computer vision, machine learning, and natural language processing;
- The problem is **hard** especially when the search space is **non-metric**.

Sample Applications

- Query by image content
- Classification
- Entity detection
- Spell-checking

State of the Art

- VP-trees
- kNN-graphs
- Multiprobe LSH (MPLSH)

Research Subject (Contestants)

Permutation search methods:

- Neighborhood APProximation Index (NAPP)
- Brute-force filtering of permutations

Basics of Permutation Methods

Filter-and-refine using **pivot-based projection** to the Euclidean space (L_2):

- Select **pivots** randomly;
- **Rank** pivots by their distances to data points;
- **Filter** by comparing **pivot rankings**;
- **Refine** by comparing remaining points to the query.

Research Questions

- How well do permutation methods fare against state of the art?
- How good are permutation-based projections?

Selected Distance Functions

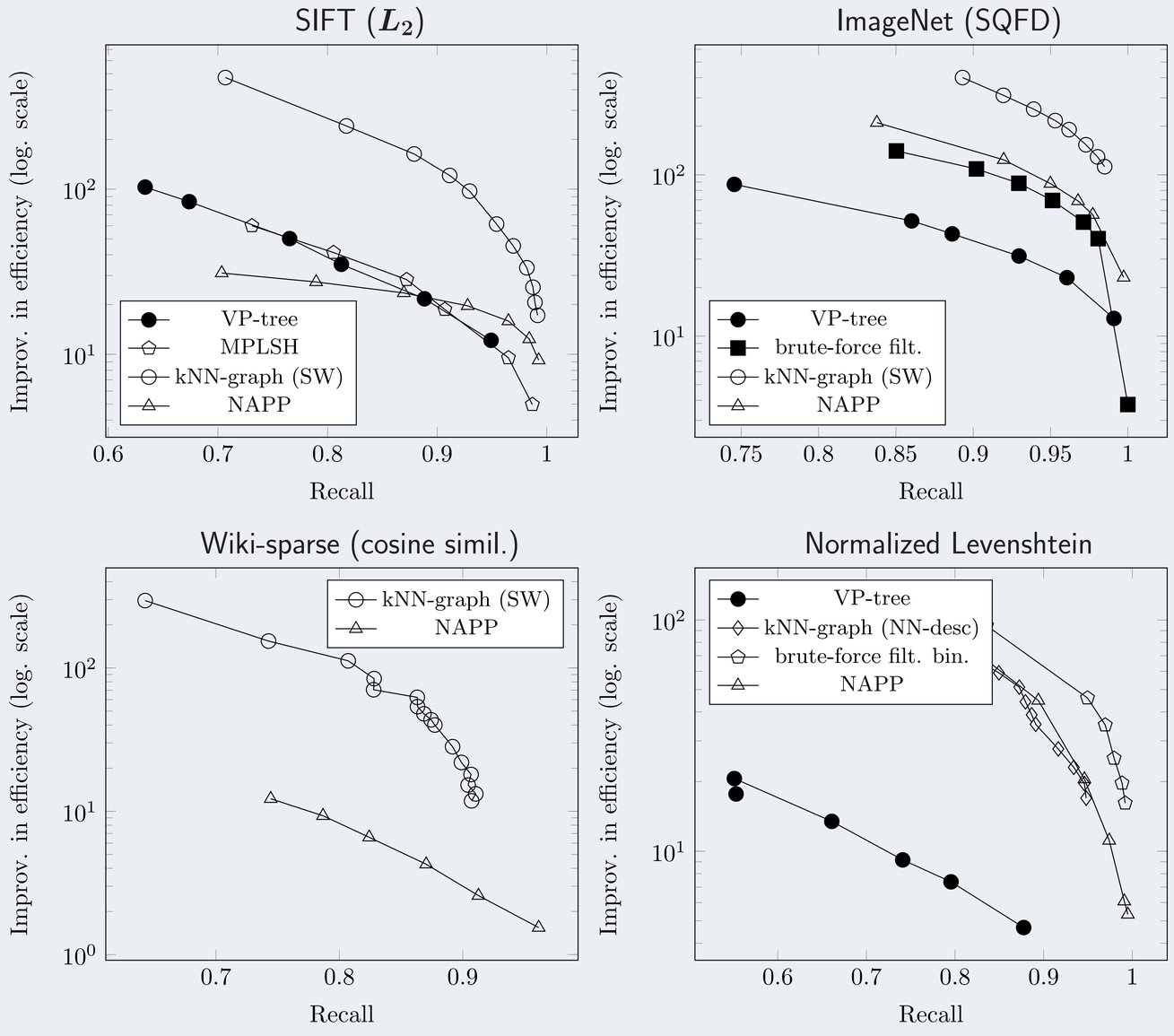
- Euclidean (L_2)
- Cosine similarity
- Jensen-Shannon divergence (JS-div.)
- Normalized Levenshtein
- Signature-Quadratic Form Distance (SQFD)

Selected Data Sets

Name	Distance function	Number of points	Brute-force (sec.)	Dimens.
SIFT	L_2	$5 \cdot 10^6$	0.3	128
ImageNet	SQFD	$1 \cdot 10^6$	4.1	N/A
Wiki-sparse	cosine sim.	$4 \cdot 10^6$	1.9	10^5
Wiki-128	JS-Div.	$2 \cdot 10^6$	4	128
DNA	norm. Leven.	$1 \cdot 10^6$	3.5	N/A

Efficiency Evaluation

Improvement in efficiency over brute-force search vs. accuracy. Higher and to the right is **better**:

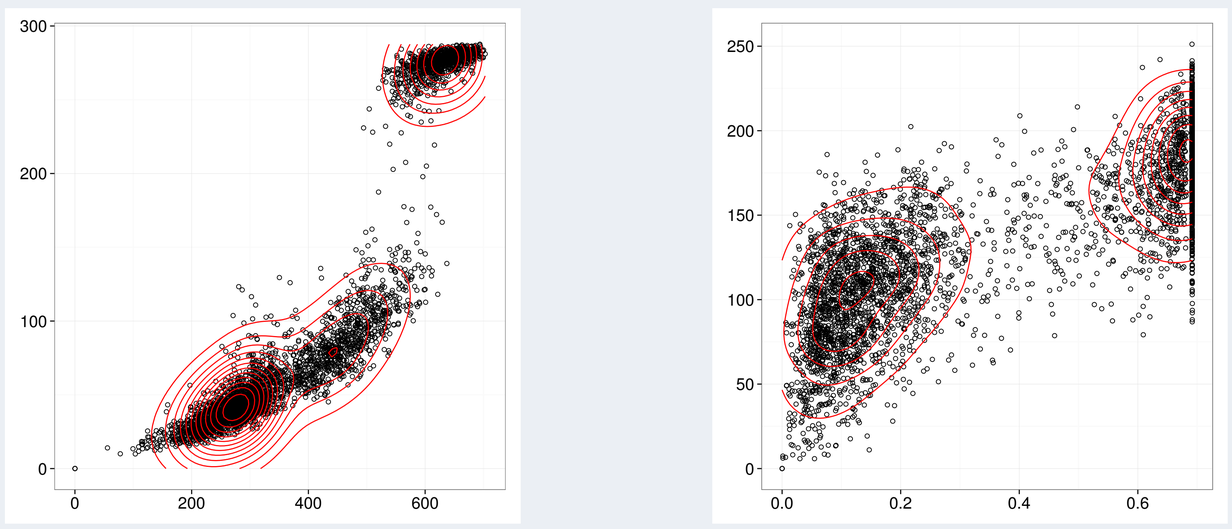


Indexing Time (min)

	VP-tree	NAPP	MPLSH	Brute-force filt.	kNN graph
SIFT	0.4	5	18.4		**52.2**
ImageNet	4.4	33		32.3	**127.6**
Wiki-sparse		7.9			**231.2**
Wiki-128	1.2	**36.6**			**36.1**
DNA	0.9	15.9		15.6	**88**

Projection Quality Evaluation

Distance in the original space vs. distance in the projected space. The closer to a monotonic mapping, the **better**:



Good projection (original distance: L_2) **So-so** projection (original distance: JS-div.)

Results and Conclusions

- Permutation methods beat state-of-the-art methods for some data sets:
  - NAPP beats MPLSH & VP-tree for SIFT, as well as VP-tree for Wiki-128;
  - Brute force filtering beats all methods including kNN graph for DNA;
  - Yet, kNN graph is the best for SIFT, Wiki-128, and Wiki-sparse.
- The quality of permutation-based projection varies;
- Permutations method are most useful when the distance function is expensive (e.g., SQFD, Levenshtein, or data is on disk) and/or indexing cost of kNN graphs is unacceptable.

http://www.idi.ntnu.no, http://scs.cmu.edu/

bileg@idi.ntnu.no,{srchvrs,ehn}@cs.cmu.edu