Scalable Nearest Neighbor Search for Optimal Transport

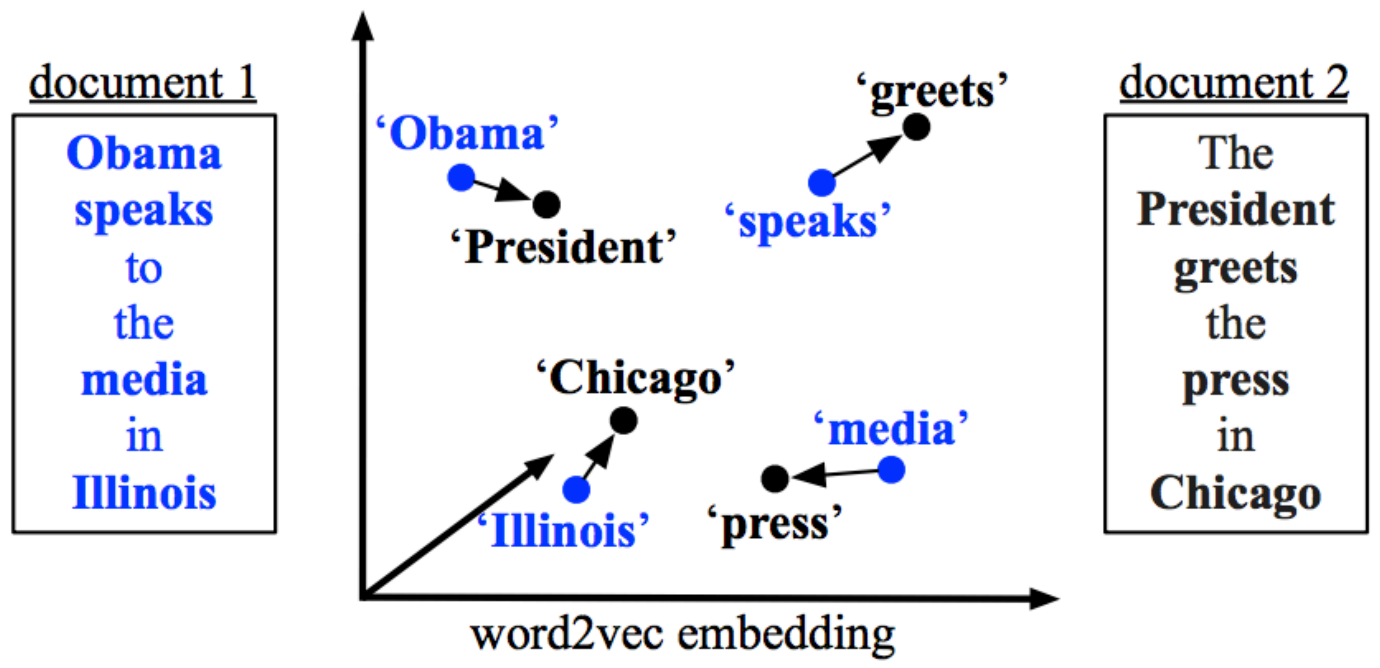
Arturs Backurs TTIC | **Yihe Dong** Microsoft | **Piotr Indyk** MIT | **Ilya Razenshteyn** MSR Redmond | **Tal Wagner** MIT

Nearest Neighbor Search for OT in High-Dimensional Spaces

Setting: Sparse distributions supported on a high-dimensional Euclidean space $\mathbb{R}^d$.

Goal: Given a collection of distributions $\boldsymbol{\mu_1, \dots, \mu_n}$, and a query distribution $\boldsymbol{\nu}$, find the **nearest neighbor** of $\boldsymbol{\nu}$ in $\boldsymbol{\mu_1, \dots, \mu_n}$.

Example: Word-Mover Distance between text documents [Kusner et al. 2015]



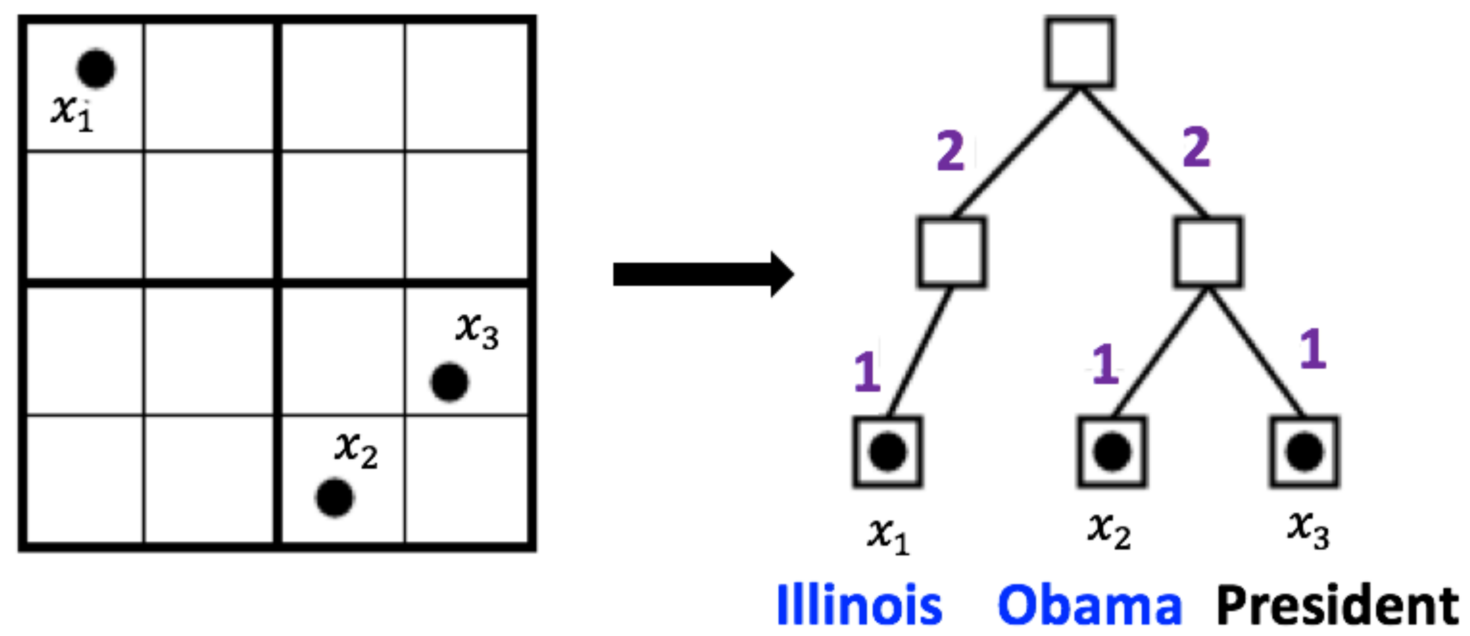
References:
- Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. **Earth mover distance over high-dimensional spaces**. SODA 2008.
- Arturs Backurs and Piotr Indyk. **Better embeddings for planar Earth-Mover Distance over sparse sets**. SoCG 2014.
- Moses Charikar. **Similarity estimation techniques from rounding algorithms.** STOC 2002.
- Marco Cuturi. **Sinkhorn distances: Lightspeed computation of optimal transport**. NeurIPS 2013.
- Piotr Indyk and Nitin Thaper. **Fast image retrieval via embeddings**. International workshop on statistical and computational theories of vision, 2003.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. **From word embeddings to document distances.** ICML 2015.
- Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. **Tree-sliced approximation of wasserstein distances**. NeurIPS 2019.

Tree-based Methods for Fast OT

Classical method: Quadtree

[Charikar 2002, Indyk & Thaper 2003, Le et al. 2019]

1. Embed support in tree of nested hypercubes.
2. Solve OT on the tree metric (linear time).



Our method: Flowtree

Solve for the optimal flow on the tree, but compute its cost w.r.t. the **original distance**.

Taxonomy of fast approximate OT methods:

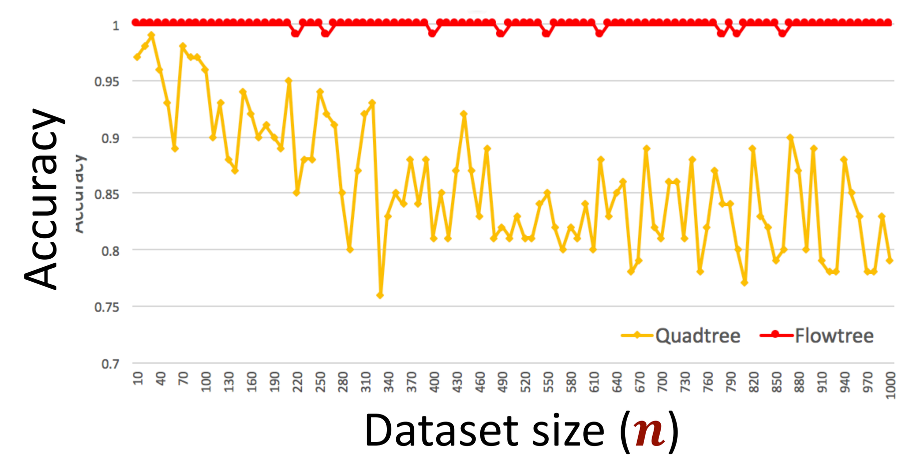
- Coarse linear time: **Mean** [Kusner et al. 2015], **Overlap/TF-IDF**, **Quadtree**
- Fine quadratic time : **R-WMD** [Kusner et al. 2015], **Sinkhorn iterations** [Cuturi 2013]
- "Slower" linear time: **Flowtree**
  Nearly as accurate and much faster than quadratic time methods.

Results

Flowtree, unlike Quadtree, does not degrade in NN-accuracy as the datasets size n grows.

Random input model

--Flowtree --Quadtree



Worst-case analysis:

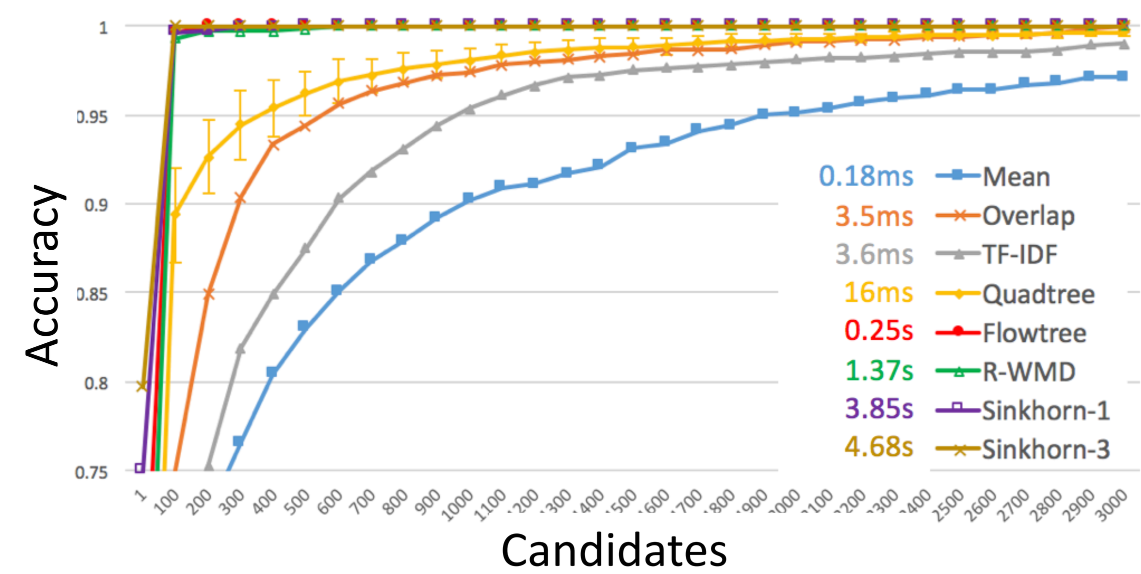
Based on [Andoni et al. 2008, Backurs & Indyk 2014]

Theorem: **Flowtree** finds an $O(\min\{\log^2 s, \log s \cdot \log(d\Phi)\})$-approximate nearest neighbor, where s is the max. support size, d is the dimension, Φ is the aspect ratio.
Note: This is independent of the dataset size n.

In comparison, **Quadtree** finds an $O(\log(sn) \cdot \log(d\Phi))$-approximate nearest neighbor, and the dependence on $\log n$ is necessary.

20newsgroups dataset:

All methods:



High-accuracy methods: