Similarity searching, or how to find your neighbors efficiently

Robert Krauthgamer
Weizmann Institute of Science

CS Research Day for Prospective Students
May 1, 2009
Background

- Geometric spaces and techniques are useful in tackling computational problems
  - Arise in diverse application areas, e.g. data analysis, machine learning, networking, combinatorial optimization

- Definition: A metric space is a set of points $M$ endowed with distance function $d_M(\cdot,\cdot)$
  - Points can model various data objects e.g. documents, images, biological sequences (or network nodes)
  - Distances can model (dis)similarity between objects (or latency)

- Common examples: Euclidean, $L_p$-norms, Hyperbolic, Hamming distance, Edit distance, Earthmover distance

- Arises also in Linear and Semidefinite Programming (LP,SDP) relaxations
Similarity Searching [Basic Problem]

- Nearest Neighbor Search (NNS):
  - **Preprocess**: a dataset $S$ of $n$ points
  - **Query**: given point $q$, quickly find closest $a \in S$, i.e. $\arg\min_{a \in S} d_M(a, q)$

- Naive solution:
  - No preprocessing, query time $O(n)$

- Ultimate goal (holy grail):
  - Preprocessing $O(n)$, and query time $O(\log n)$

- Key problem, many applications, extensively studied:
  - Difficult in high dimension (curse of dimensionality)
  - Algorithms for $(1 + \varepsilon)$-approximation in $L_1$ and in $L_2$:
    - Query time $O(\log n)$ and preprocessing $n^{O(1/\varepsilon^2)}$
    - Query time $O(n^{1/(1 + \varepsilon)})$ and preprocessing $O(n^{1+1/(1 + \varepsilon)})$ [Indyk-Motwani'98, Kushilevitz-Ostrovsky-Rabani'98]
NNS in General Metrics

- **Black-box model:** Access to pairwise distances (only).
- **Suppose M is a uniform metric**
  - i.e. \( d(x,y) = 1 \) for all \( x,y \in M \),
- **Depicts difficulty of NNS for high-dimensional data**
  - Such data sets exist in \( \mathbb{R}^{\log n} \),
- **Is this the only obstacle to efficient NNS?**
  - What about data that “looks” low-dimensional?
- **For some queries, time \( \geq \Omega(n) \), even for approximate NNS.**
A Metric Notion of Dimension

- Definition: Ball $B(x,r) = \{ \text{all points within distance } r>0 \text{ from } x \in M \}$.

- The dimension of $M$, denoted $\text{dim}(M)$, is the minimum $k$ such that every ball can be covered by $2^k$ balls of half the radius.
  - Defined by [Gupta-K.-Lee’03], inspired by [Assouad’83, Clarkson’97].
  - Call a metric doubling if $\text{dim}(M) = O(1)$
  - Captures every norm on $\mathbb{R}^k$

- Robust to:
  - taking subsets,
  - union of sets,
  - small distortion in distances, etc.
  - Unlike previous suggestions based on cardinality $|B(x,r)|$ [Plaxton-Richa-Rajaraman’97, Karger-Ruhl’02, Faloutsos-Kamel’94, K.-Lee’03, …]

Here $2^k \leq 7$. 

Similarity searching, or how to find your neighbors efficiently
Theorem [K.-Lee’04a]: There is a simple \((1+\varepsilon)\)-NNS scheme
- Query time: \((1/\varepsilon)^{O(\text{dim}(S))} \cdot \log \Phi\) \([\Phi = d_{\text{max}}/d_{\text{min}} \text{ is spread}]\)
- Preprocessing: \(n \cdot 2^{O(\text{dim}(S))}\).
- Insertion/deletion time: \(2^{O(\text{dim}(S))} \cdot \log \Phi \cdot \log \log \Phi\).

Outperforms previous schemes [Plaxton-Richa-Rajaraman’98, Clarkson’99, Karger-Ruhl’02]
- Simpler, wider applicability, deterministic, no apriori info
- Nearly matches the Euclidean low-dim. case [Arya et al.’94]
- Explains empirical successes—it’s just easy…

Subsequent enhancements
- Optimal storage \(O(n)\) [Beygelzimer-Kakade-Langford’06]
  - Also implemented and obtained very good empirical results
- Bounds independent of \(\Phi\) [K.-Lee’04b, Mendel-HarPeled’05, Gottlieb-Cole’06]
- Improved for Euclidean metrics [Indyk-Naor’06, Dasgupta-Fruend’08]
**Nets**

- **Motivation:** Approximate the metric at one scale $r > 0$.
  - Provide a spatial “sample” of the metric space
  - E.g., grids in $\mathbb{R}^2$.

- **General approach:**
  - Choose representatives $Y$ iteratively

- **Definition:** $Y \subseteq S$ is called an $r$-net if both:
  1. For all $y_1, y_2 \in Y$, $d(y_1, y_2) \geq r$ [packing]
  2. For all $x \in M \setminus Y$, $d(x, Y) < r$ [covering]
Navigating Nets

NNS scheme (vanilla version):

- **Preprocessing:**
  - Compute a $2^i$-net $Y_i$ for all $i \in \mathbb{Z}$.
  - Add "local links".

- **Query algorithm:**
  - Iteratively go to finer nets
  - Find net-point $y_i \in Y_i$ closest to query

From a $2^i$-net point to *nearby* $2^{i-1}$-net points
\[ d(q, y_i) \leq \text{OPT} + 2^i \Rightarrow d(y_i, y_{i-1}) \leq 2\text{OPT} + 2^i + 2^{i-1} \]
Thus: # "local" links $\leq 2^{O(\text{dim}(S))}$. 

Similarity searching, or how to find your neighbors efficiently
**Embeddings** [Basic Technique]

- **An embedding** of $M$ into $l_1$ is a mapping $f: M \rightarrow \mathbb{R}^m$
  - We say $f$ has distortion $K \geq 1$ if
    $$d_M(x,y) \leq \|f(x)-f(y)\|_1 \leq K \cdot d_M(x,y) \quad \forall x,y \in M$$

- **Example:**
  - **Tree metric:**
    - Embedding into $\mathbb{R}^2$:
    - **distortion** = $1$
    - **distortion** $\geq 4/3$

- **Another example:**
  - discrete cube $\{0,1\}^r$ with $d(x,y)=\|x-y\|_2$
  - under identity map: $\|x-y\|_2 \leq \|x-y\|_1 \leq \sqrt{r} \cdot \|x-y\|_2$
  - distortion = $\sqrt{r}$? Nah…

- **Very powerful concept, many applications (including NNS)**
A Few Embeddings Theorems

Every n-point **metric** embeds
- into $l_2$ (thus into $l_1$) with distortion $O(\log n)$ [Bourgain’86]
- Tight on expanders

If $M$ is **doubling**, $\sqrt{d_M}$ embeds into $l_2$ (thus into $l_1$) with distortion $O(1)$ [Assouad’83]

Every n-point **tree metric** embeds
- into $l_1$ isometrically
- into $l_2$ with distortion $O(\log \log n)^{1/2}$ [Matousek’99]

Every doubling **tree metric** embeds into $l_2$ with distortion $O(1)$ [Gupta-K.-Lee’03]
Some Open Problems [Embeddings]

- **Dimension reduction in** $l_2$:
  - **Conjecture:** If $M$ is Euclidean (a subset of $l_2$) and doubling, then it embeds with distortion $O(1)$ into low-dimensional $l_2$ (or $l_1$)
  - Known: dimension $O(\log n)$ where $n=$# points (not doubling dimension) [Johnson-Lindenstrauss’84]
  - Known: Can embed metric $\sqrt{d_M}$ [Assouad’83]

- **Planar graphs**:
  - **Conjecture:** Every planar metric embeds into $l_1$ with distortion $O(1)$
Edit Distance [Specific Metric]

Edit Distance (ED) between two strings \( x, y \in \Sigma^d \):
- Minimum number of character insertions / deletions / substitutions to transform one string into the other
- Extensively used, many applications, variants

Computational problems:
1. Computing ED for two input strings
   - Currently, best runtime is quadratic \( O(d^2/\log^2 d) \) [Masek-Paterson’80]
2. Quick approximation (near-linear time)
   - Currently, best approximation is \( 2^{O(\sqrt{d})} \) [Andoni-Onak’08]
   - Smoothed model: \( O(1/\varepsilon) \) approximation in time \( d^{1+\varepsilon} \) [Andoni-K.’07]
3. Estimate ED in restricted computational model
   - Sublinear time, data-stream model, or limited communication model
4. NNS under ED (polynomial preprocessing, sublinear query)
   - Currently, best bounds are obtained via embedding into \( L_1 \)

Examples:
- \( ED(\text{and}, \text{an}) = 1 \)
- \( ED(0101010, 1010101) = 2 \)
Ulam’s metric

- **Definitions:**
  - A string $s \in \Sigma^d$ is called a *permutation* if it consists of distinct symbols
  - *Ulam's metric* = Edit distance on the set of permutations [Ulam’72]
  - For simplicity, suppose $\Sigma = \{1, \ldots, d\}$ and “ignore” substitutions

\[
\begin{align*}
X &= 123456789 \\
Y &= 234657891
\end{align*}
\]

- **Motivations:**
  - A permutation can model ranking, deck of cards, sequence of genes, …
  - A special case of edit distance, useful to develop techniques
Embedding of permutations

**Theorem [Charikar-K.’06]:** Edit distance on permutations (aka Ulam’s metric) embeds into $l_1$ with distortion $O(\log d)$.

**Proof.** Define $f : \Sigma^d \to \mathbb{R}^{|\Sigma|^2}$ by

$$f_{a,b}(P) = \frac{1}{P^{-1}[a] - P^{-1}[b]}$$

**Intuition:**
- $\text{sign}(f_{a,b}(P))$ indicates whether “a appears before b” in P
- Thus, $|f_{a,b}(P) - f_{a,b}(Q)|$ “measures” if $\{a,b\}$ is an inversion in P vs. Q

**Claim 1:** $||f(P) - f(Q)||_1 \leq O(\log d) \text{ ED}(P,Q)$
- Assume wlog $\text{ ED}(P,Q)=2$, i.e. Q obtained from P by moving one symbol ‘s’
- General case then follows by triangle inequality on $P=P_0,P_1,\ldots,P_t=Q$
  namely $||f(P) - f(Q)||_1 \leq \sum_{j=1}^{t} ||f(P_j) - f(P_{j-1})||_1$
- Total contribution
  - From coordinates where $s \in \{a,b\}$: $\leq 2 \sum_{k} (1/k) \leq O(\log d)$
  - From other coordinates: $\leq \sum_{k} k(1/k - 1/(k+1)) \leq O(\log d)$
Embedding of permutations

**Theorem [Charikar-K.’06]:** Edit distance on permutations (aka Ulam’s metric) embeds into $l_1$ with distortion $O(\log d)$.

**Proof.** Define $f : \Sigma^d \rightarrow R^{|\Sigma|^2}$ by $f_{a,b}(P) = \frac{1}{P^{-1}[a] - P^{-1}[b]}$

Claim 1: $||f(P)-f(Q)||_1 \leq O(\log d) \text{ ED}(P,Q)$

Claim 2: $||f(P)-f(Q)||_1 \geq \frac{1}{4} \text{ ED}(P,Q)$ [alt. proof by Gopalan-Jayram-K.]

- Assume wlog that $P=$identity
- Edit $Q$ into $P=$identity using quicksort:
  - Choose a random pivot,
  - Delete all characters inverted wrt to pivot
  - Repeat recursively on left and right portions

- Now argue $\text{ ED}(P,Q) \leq 2 \mathbb{E}[\#\text{quicksort deletions}] \leq 4 ||f(P)-f(Q)||_1$

For every inversion $(a,b)$ in $Q$: $\text{ Pr}[a \text{ deleted “by” pivot } b] \leq \frac{1}{|Q^{-1}[a] - Q^{-1}[b]|+1} \leq 2 |f_{a,b}(P) - f_{a,b}(Q)|$

Surviving subsequence is increasing, thus $\text{ ED}(P,Q) \leq 2 \#\text{deletions}$

Similarity searching, or how to find your neighbors efficiently
Open Problems [Sublinear Algorithms]

- Estimate in sublinear time the distance between input permutation $P$ and the identity [testing distance to monotonicity]
  - If distance $= \delta \cdot d$, use only $\tilde{O}(1/\delta)$ queries to $P$
  - An $O(\log d)$-approximation follows from the embedding
  - Actually, a factor 2-approximation is known

- Lower bound for approximation $= 1 + \varepsilon$?
  - Question: deciding whether distance is $< d/100$ or $> d/99$ requires $d^{\Omega(1)}$ queries to $P$?

- Similar but for block operations [transposition distance]:
  - Approximation 3 is known (exercise)
  - Is there a lower bound for $1 + \varepsilon$ approximation?
  - Remark: distance is not known to be computable in polynomial time
Research Objectives

Two intertwined goals:
1. Understand the complexity of metric spaces from a mathematical and computational perspective
   - E.g., identify geometric properties that reveal a simple underlying structure
2. Develop algorithmic techniques that exploit such characteristics
   - E.g., design tools that dissect a metric into smaller pieces, or faithfully convert them into simpler metrics

Concrete directions:
- Find new NNS algorithms for different metrics (in particular for high-dimension)
- Classify and characterize metric spaces via embeddings