Sublinear Time Nearest Neighbor Search over Generalized Weighted Space

Yifan Lei **Qiang Huang** Mohan S. Kankanhalli Anthony K. H. Tung

School of Computing, National University of Singapore

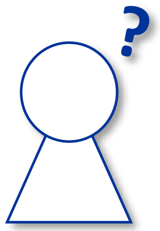
Source code: https://github.com/1flei/aws_alsh

Applications

- Nearest Neighbor Search (NNS) is widely used

- Example: booking hotel for ICML 2019
  - Considering the conditions to the convention centre, i.e., price, distance, and rating
  - Query q: a hotel that the user booked before and felt excellent
  - Weight vector w: different users have *different preference* to the hotel conditions, which lead to different choices of hotels

	Price	Distance	Rating
Hotel q	300	7	10



$w = (0.001, 1, 1)$ → Hotel 2
$w = (0, 1, 3)$ → Hotel 1
$w = (0.001, -1, 1)$ → Hotel 3
$w = (-0.001, -1, -1)$ → Hotel 4

	Price	Distance	Rating
Hotel 1	400	8	10
Hotel 2	350	6	8
Hotel 3	250	9	8
Hotel 4	200	6	6

Problem Definition

- Given
  - A dataset $\mathcal{D}$ of n data objects in $\mathbb{R}^d$
  - A query $q \in \mathbb{R}^d$ with a weight vector $w \in \mathbb{R}^d$
  - Measure: the *Generalized Weighted Square Euclidean Distance (GWSED)* d_w

$$d_w(o,q) = \sum_{i=1}^{d} w_i (o_i - q_i)^2$$

- Nearest Neighbor Search (NNS) over d_w
  - To find $o^* \in \mathcal{D}$ s.t. $o^* = \arg\min_{o \in \mathcal{D}} d_w(o,q)$

- This problem is *very fundamental*
  - Furthest Neighbor Search (FNS) and MIPS can be reduced to NNS over d_w,
  - i.e., $w_i = -1, \forall i \implies \arg\min_{o \in \mathcal{D}} d_w(o,q) = \arg\max_{o \in \mathcal{D}} \|o - q\|$

Background and Motivations

- Locality-Sensitive Hashing (LSH)
  - *Sublinear time* for Near Neighbor Search
  - Insight: construct a hash function h s.t. $Pr[h(o) = h(q)]$ *is monotonic in* $Dist(o, q)$
  - Hidden condition: $Dist(o, q)$ *must be a metric*

- *LSH schemes cannot solve NNS over d_w directly* (d_w is no longer a metric if $w_i < 0$)

- There is **NO** sublinear method for this problem

- Motivations
  - Similar to d_w, inner product (i.e., $o^T q$) is also *not* a metric
  - However, Shrivastava & Li (2014) introduced a *sublinear time* method based *Asymmetric LSH* which constructs $P(o)$ and $Q(q)$ for data objects $o \in \mathcal{D}$ and each query q, respectively.

Spherical Asymmetric Transformation

- Negative result:
  - *There is no Asymmetric LSH family over* $\mathbb{R}^d$ *for NNS over* d_w (Lemma 1 and Theorem 2)

- Spherical Asymmetric Transformation (SphAT): $\mathbb{R}^d \rightarrow \mathbb{R}^{2d}$

$$P(o) = [COS(o); SIN(o)]$$

$$Q(q, w) = [w \otimes COS(q); w \otimes SIN(q)]$$

  - where $w \otimes COS(q) = (w_1 \cos q_1, w_2 \cos q_2, \ldots, w_d \cos q_d)$

- Properties of SphAT:
  - $d_w(o, q) \sim$ Euclidean distance (or Angular distance) between $P(o)$ and $Q(q, w)$
  - SphAT is *weight-oblivious* (because $P(\cdot)$ is independent of w) $\Longrightarrow$ *build index before q and w*

Two Proposed Methods

- SL-ALSH = SphAT + E2LSH
  - SphAT: $\arg\min_{o\in\mathcal{D}} d_w(o,q) \Rightarrow \arg\min_{o\in\mathcal{D}} \|P(o)-Q(q,w)\|$
  - Apply E2LSH on $P(o)$ and $Q(q,w)$ for NNS over Euclidean distance

- S2-ALSH = SphAT + SimHash
  - SphAT: $\arg\min_{o\in\mathcal{D}} d_w(o,q) \Rightarrow \arg\max_{o\in\mathcal{D}} \frac{P(o)^T Q(q,w)}{\|P(o)\|\|Q(q,w)\|}$
  - Apply SimHash on $P(o)$ and $Q(q,w)$ for NNS over Angular distance

- Main Results
  - $Pr[h(P(o)) = h(Q(q,w))]$ *is monotonic in* $d_w(o,q)$ (Lemmas 3 and 4)
  - SL-ALSH and S2-ALSH solve the problem of NNS over d_w with *sublinear time* (Theorems 3 and 4)

Datasets and Settings

- Datasets
  - Mnist ($n = 60{,}000$ and $d = 784$)
  - Sift ($n = 1{,}000{,}000$ and $d = 128$)
  - Movielens ($n = 52{,}889$ and $d = 150$)

- Five types of weight vector w

Types	Illustrations
Identical	All “1”
Binary	Uniformly distributed in $\{0,1\}^d$
Normal	d-dimensional normal distribution $\mathcal{N}(0, I)$
Uniform	Uniformly distributed in $[0,1]^d$
Negative	All “-1”

Bucketing Experiments

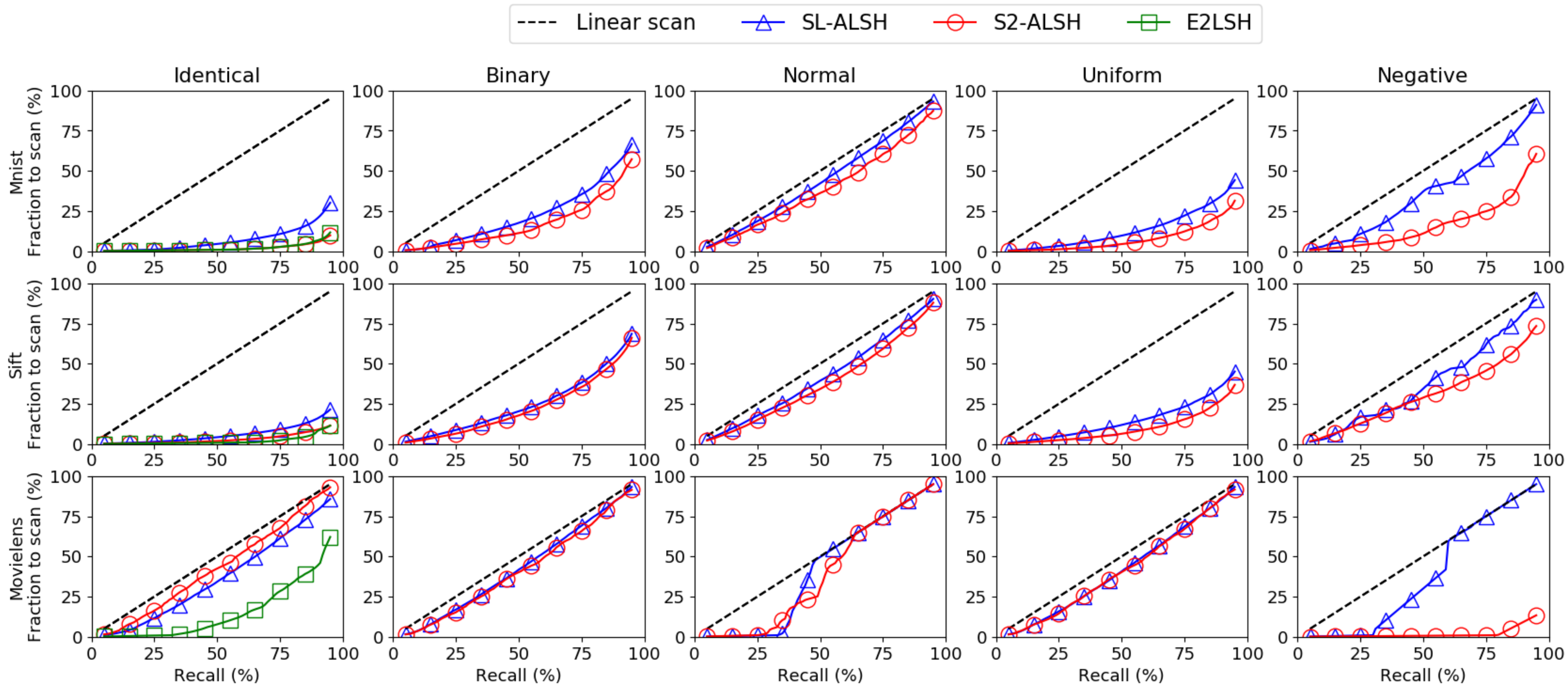


Figure: The best fraction of dataset to scan to achieve certain level of recalls (**lower is better**).

Conclusions

- Demonstrate that there is *no Asymmetric LSH family over* $\mathbb{R}^d$ for the problem of NNS over d_w

- Introduce a novel *SphAT* from $\mathbb{R}^d$ to $\mathbb{R}^{2d}$
  - SphAT is *weight-oblivious*
  - $Pr[h(P(o)) = h(Q(q,w))]$ *is monotonic in* $d_w(o,q)$

- Propose the first two *sublinear time* methods SL-ALSH and S2-ALSH for NNS over d_w

- Extensive experiments verify that SL-ALSH and S2-ALSH answer the NNS queries in sublinear time and support various types of weight vectors.

Poster Session

[Poster #82: Tue Jun 11th 06:30—09:00 PM @Pacific Ballroom]

Thank you for your attention!