Low Distortion Embeddings for Edit Distance

Rafail Ostrovsky (UCLA)
Yuval Rabani (Technion)

edit (or Levenshtein) distance

Let x,y be two character strings.

$ed(x,y)$ = minimum # edit operations needed to convert x into y.

edit operations: insert, delete (substitute)

We will restrict our attention to $\{0,1\}^d$

applications: text processing, genomics, www, image matching, ...

edit distance computation

- dynamic programming (1965 ?) $O(d^2)$
- Masek & Paterson (1980) $O(d^2/\log d)$
- BEKMRRS (2003) d^ϵ vs. d, sublinear time
- BJKK (2004) $d^{3/7}$ approx. in $\tilde{O}(d)$ time
- BES (2006) $d^{1/3+\epsilon}$ approx. in $\tilde{O}(d)$ time

- sketching: BJKK (2004) k vs. $(kd)^{2/3}$
- communication complexity
- NNS: Indyk, BJKK (2004) d^ϵ approx.

- block ed: CPSV, MS (2000), CM (2002)

low distortion embedding

Map $(\{0,1\}^d, \mathrm{ed})$ to a normed space which we know more about.

Natural candidate: ℓ_1 ($\approx$ Hamming distance)

φ will denote the mapping.

The distortion = $\|\varphi\|_{Lip} \cdot \|\varphi^{-1}\|_{Lip}$ =

$$\max_{x,y} \frac{\|\varphi(x) - \varphi(y)\|_1}{\mathrm{ed}(x,y)} \cdot \max_{x,y} \frac{\mathrm{ed}(x,y)}{\|\varphi(x) - \varphi(y)\|_1}$$

our results

- $2^{O(\sqrt{\log d \, \log\log d})}$ distortion;
- efficiently computable: embedding a point takes poly(d) time;
- implies same guarantee for sketching, communication complexity, nearest neighbor search.

the embedding

Partition the string into blocks of length b:

0010111100101010000010001000001111111000001010101011101

In each block:

Take "shingles" shifted by 0,1,2,...,s-1:

0 0 1 0 1 1 1 1 0 0 1 0 1

embedding (cont.)

We get a (multi-) set of strings:

0010111100 0101111001 1011110010 0111100101

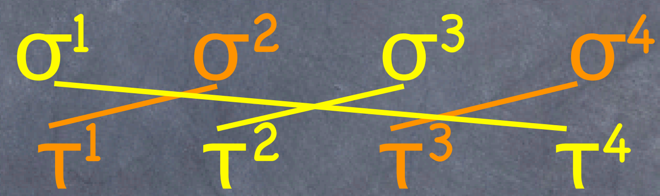
Recursively embed each shingle into the Hamming cube: S = $\{\sigma^1, \sigma^2, \sigma^3, \sigma^4\}$

Define a metric on s-sets of strings:

$$\mathrm{dist}(S,T) = \frac{1}{s} \cdot \min_{\text{matchings } \mu} \left\{ \sum_{\sigma \in S} \min\{s, c \cdot H(\sigma, \mu(\sigma))\} \right\}$$

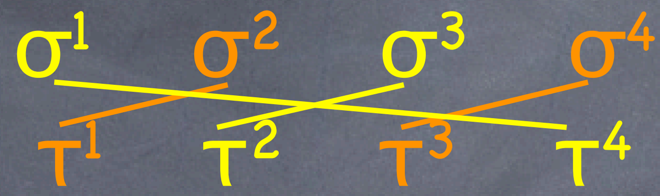
embedding (cont.)

$$\mathrm{dist}(S,T) = \frac{1}{s} \cdot \min_{\text{matchings } \mu} \left\{ \sum_{\sigma \in S} \min\{s, c \cdot H(\sigma, \mu(\sigma))\} \right\}$$



Use $c = 2 \ln(2s)$

embedding (cont.)



Embed dist into ℓ_1 (ψ is the embedding)

We don't know how to get low distortion.

<u>Guarantee:</u>

1. Always $||\psi(S)-\psi(T)||_1 \le c \cdot \mathrm{dist}(S,T)$
   recall: $c = 2 \ln(2s)$

2. If $\forall \sigma,\tau\ \mathcal{H}(\sigma,\tau) \ge s$, then $||\psi(S)-\psi(T)||_1 \ge s/2$

constructing ψ

S contains s strings of length n

I is a sample of $(1/s) \cdot n \cdot \ln(2s)$ positions
z is a $(1/s) \cdot n \cdot \ln(2s)$ bit string

Coord. I,z = #σ-s with $\sigma_I = z$.

Scaling: divide by #coordinates.

analysis of ψ's construction

Simple probabilistic analysis:

Let $J = \{j: \sigma_j \neq \tau_j\}$, so $\mathcal{H}(\sigma,\tau) = |J|$.

I is a u.a.r. sample of $(1/s) \cdot n \cdot \ln(2s)$ positions (with repetition).

$\Pr[I \cap J = \emptyset] \approx \exp(-(1/s) \cdot \mathcal{H}(\sigma,\tau) \cdot \ln(2s))$

choice of parameters

The block size $b = d / 2^{\sqrt{\log d \, \log\log d}}$

Use several values for s:

$s = (\log d)^j$, $\forall j$ s.t. $s \leq b$. Tot: $\frac{\log d}{\log\log d}$ values.

Each block and each s-value generates a set of coordinates (using ψ).

analysis

Crucial observation:

1. If #edit operations k in block $\leq s$, then $\leq$ $ed(x,y)$ shingles σ have $ed(\sigma,\mu(\sigma)) > k$:

000111111111000

011111010000001

2. If $\exists \sigma,\tau$ with $ed(\sigma,\tau) \leq s$, then the two x,y blocks align with cost $\leq 2s + ed(\sigma,\tau)$.

upper bound

Cost of "bad" shingles: $(1/s) \cdot ed(x,y) \cdot s$

"good" shingles: $(1/s) \cdot s \cdot O(||\varphi_{\leq b}||_{Lip} \cdot k \cdot \ln(s))$

Summing over blocks, s gives:

$$||\varphi_d||_{Lip} \leq \#blocks \cdot \#s + \#s \cdot \ln(d) \cdot ||\varphi_b||_{Lip}$$

lower bound

In each block i, let s_i =
max s s.t. $\forall \sigma,\tau\ ed(\sigma,\tau) \geq \|\varphi_b^{-1}\|_{Lip} \cdot s$

1. $ed(x,y) \leq \Sigma_i\ (\|\varphi_b^{-1}\|_{Lip} + 2) \cdot s_i \cdot \log(d)$

2. $\|\varphi(x)-\varphi(y)\|_1 \geq \Sigma_i\ s_i/2$

$$\|\varphi_d^{-1}\|_{Lip} \leq \log(d) \cdot \|\varphi_b^{-1}\|_{Lip} + \log(d)$$

analysis (cont.)

1. $\|\varphi_d\|_{Lip} \leq \log^2(d) \cdot \|\varphi_b\|_{Lip} + \#blocks \cdot \#s$

2. $\|\varphi_d^{-1}\|_{Lip} \leq \log(d) \cdot \|\varphi_b^{-1}\|_{Lip} + \log(d)$

We need to balance #blocks against the depth of the recurrence.

analysis (cont.)

1. $\|\varphi_d\|_{Lip} \le \log^2(d) \cdot \|\varphi_b\|_{Lip} + \#blocks \cdot \#s$

2. $\|\varphi_d^{-1}\|_{Lip} \le \log(d) \cdot \|\varphi_b^{-1}\|_{Lip} + \log(d)$

We will use #blocks = $2^{\sqrt{\log d \, \log\log d}}$

Both recurrences solve to $2^{O(\sqrt{\log d \, \log\log d})}$

The recurrence depth is $O\left(\sqrt{\frac{\log d}{\log\log d}}\right)$

concluding remarks

- For efficient implementation, sample the coordinates of ψ.
- Failure prob. δ, dim = $O(d \cdot \log(d/\delta))$.
- To embed entire cube, dim = $O(d^2)$.

Lower bounds:

- ADGIR (2003) 3/2
- Khot & Naor (2005) $\Omega(\sqrt{\log d})$
- Krauthgamer & R. (2006) $\Omega(\log d)$
- CK? (2006) $d^{\Omega(1)}$ into Hilbert space

The End


Peet's At Home