Counting Distance Permutations

Matthew Skala
David R. Cheriton School of Computer Science
University of Waterloo

April 11, 2008
SISAP'08

Outline

- Definitions
- Distance permutations
- Tree metric results
- Vector L_p metric results
- Experiments
- Open problems

Metric spaces

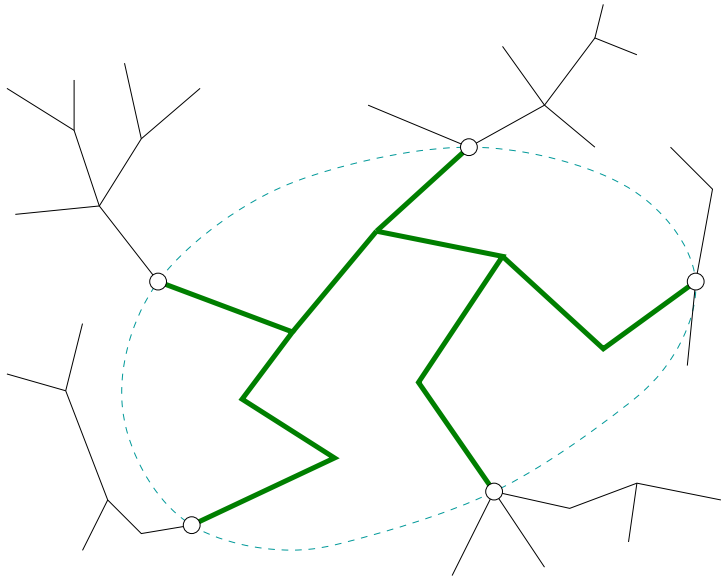
Metric spaces are a general, but rigorous, way of describing any situation where there are things and distances between them.

A metric space is a tuple $\langle S, d \rangle$ of a set S and a function $d : S \times S \to \mathbb{R}$ satisfying $d(x, y) \geq 0$; $d(x, y) = 0$ iff $x = y$; $d(x, y) = d(y, x)$; and the Triangle Inequality $d(x, z) \geq d(x, y) + d(y, z)$.

Many kinds of data can be represented by points in metric spaces; and many kinds of queries can be described in terms of metric spaces.

Tree metrics

Points are vertices of a tree, distances are the (possibly weighted) lengths of the unique paths through the tree. *Prefix distance* is one important example: edit a string at one end only. Then the tree is the trie of legal strings.



Vectors with L_p distance

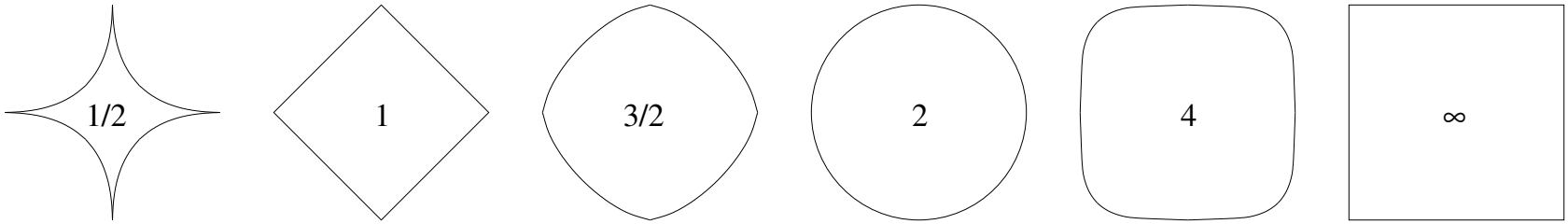
The L_p metrics generalize the Pythagorean Theorem to other exponents:

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

for real $p \geq 1$ (why not $p < 1$?) or

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_{i=1}^{n} |x_i - y_i|$$

for $p = \infty$.



Similarity search

Here's a book: 

Here's a library containing $\sim 2 \times 10^6$ books: 

How can I find other books like *Heart of Darkness*?

nearest-neighbour query

$$\arg\min_{x \in} \; d\left(x, \right)$$

range query

$$\left\{ x \,\middle|\, d\left(x, \right) < r \right\}$$

Label all the books

Call number prefixes map differences among books into a tree metric:

- "P": language and literature
- "PR": English literature
- "PR60. . . ": 20th Century English authors
- "PR6005": before 1950, surname starts with C. . .
- "PR6005.O": surname starts with CO. . .
- "PR6005.O4": Joseph Conrad
- "PR6005.O4H4": *Heart of Darkness*

Distance Permutations [Chávez, Figueroa, and Navarro, 2005]

I have a fixed list of 10 books everybody knows.

I evaluate the distance from *Heart of Darkness* to each of the books on the list.

If I wrote down that 10-vector to use with the triangle inequality, I'd be doing LAESA. [Micó and Vidal, 1994]

But instead, I'll just sort it and write down the resulting permutation. That will be my label I put on the book.

Similar books ought to have similar labels.

Labelling *Heart of Darkness*

r_i	i	Title
1	6	*Life of Pi*
2	7	*Ragtime*
3	0	*Historia universal de la infamia*
4	2	*The Snarkout Boys and the Avocado of Death*
5	5	*El ingenioso hidalgo don Quijote de la Mancha*
6	1	*2001, a Space Odyssey*
7	8	*Teatro herético*
8	3	*The Art of Computer Programming*
9	9	*Good to Great*
10	4	*LaTeX: a Document Preparation System*

The label for *Heart of Darkness* is "6702518394". How many distinct labels are possible?

Tree metrics: quadratic in number of sites

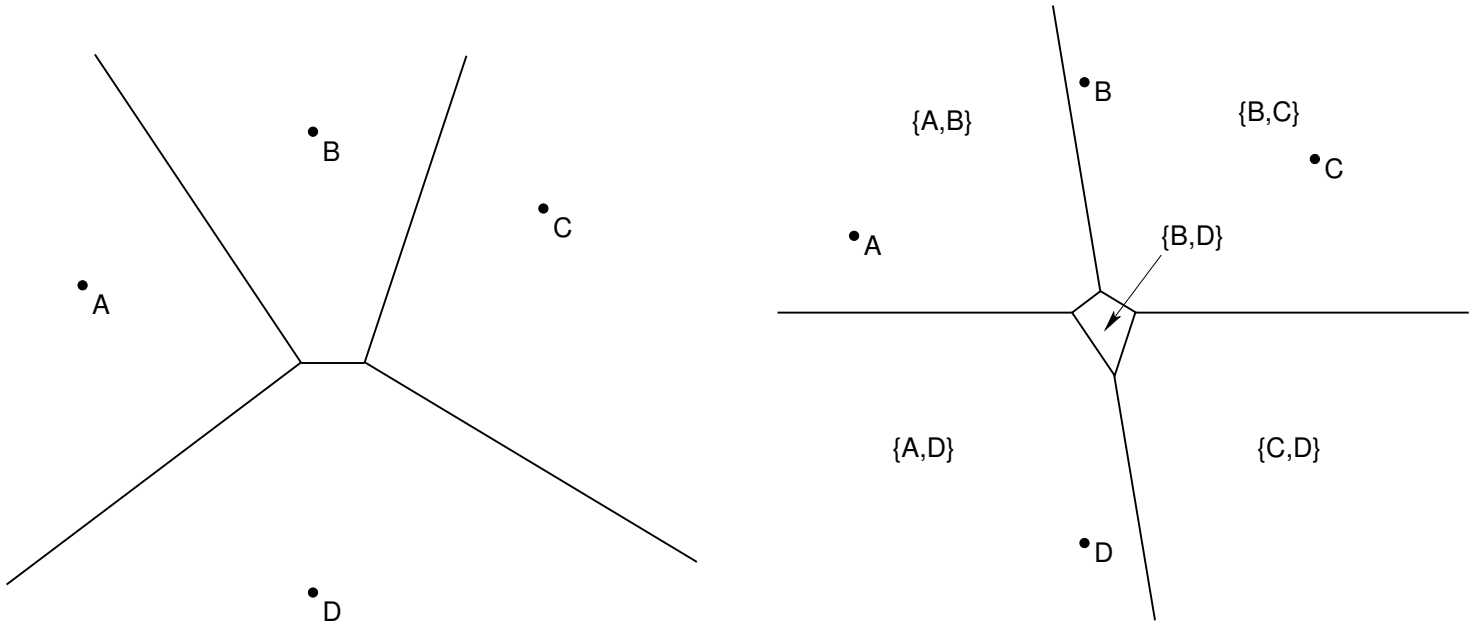
Proof: each of the $\binom{k}{2}$ comparisons between sites can cut the tree on at most one edge, so they all cut it into at most $\binom{k}{2} + 1$ components corresponding to distance permutations.

Consequence: storage space for entire distance permutation is about the same as storage space for two site indices.

Tree metrics behave like the one-dimensional real line.

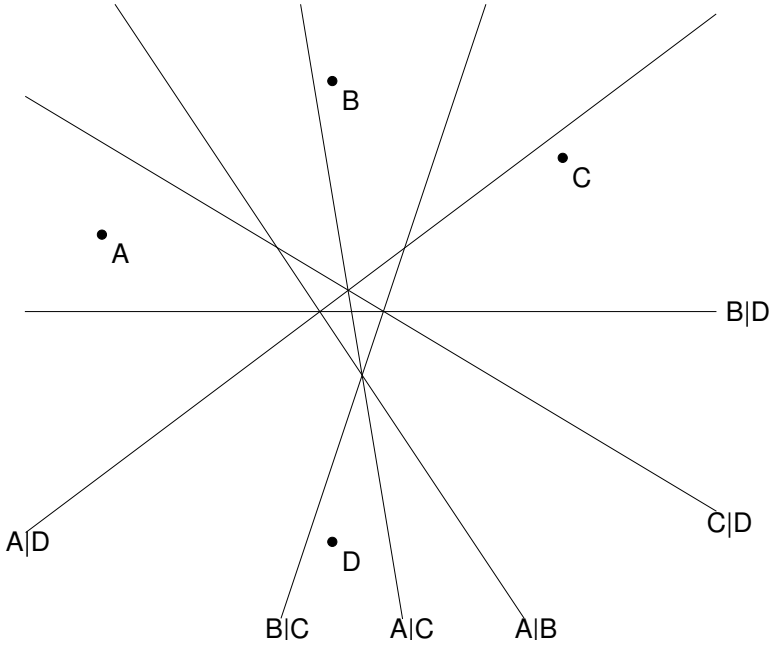
Generalized Voronoi diagrams

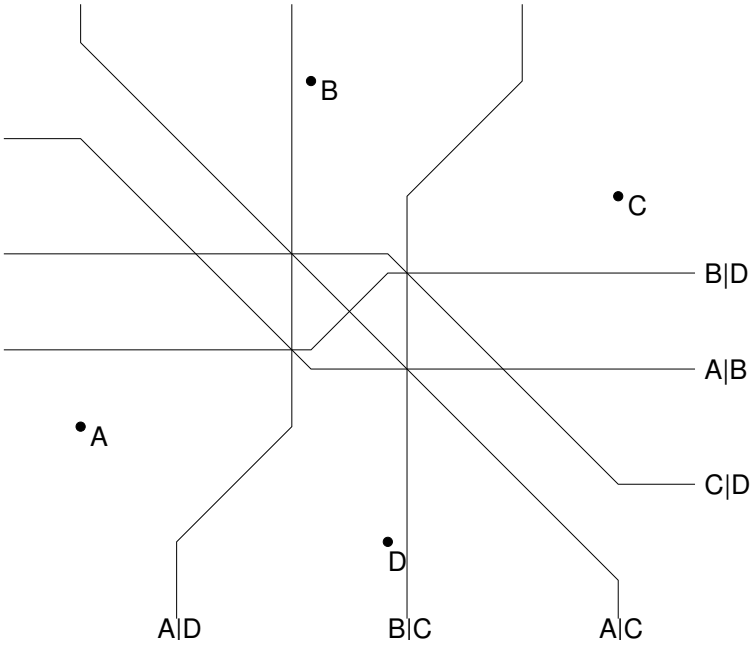
Divide space according to the k nearest neighbours:



Voronoi diagrams generalized further

Divide space according to the entire distance permutation:





The cake-cutting problem [Price, 1946]

Let $S_d(m)$ be the number of pieces formed by m cuts in d-dimensional Euclidean space

Obviously $S_0(m) = S_d(0) = 1$.

The m-th cut is a $(d-1)$-dimensional space itself, and is cut up by its intersections with the previous $m-1$ cuts, into $S_{d-1}(m-1)$ pieces.

Each piece of the new cut, cuts off a new piece of d-space.

$S_d(m) = S_d(m-1) + S_{d-1}(m-1) = \Theta(m^d)$

$S_d(m) = W(d, m)$, "Whitney numbers," Sloan's A0004070

Already we have a bound on L_2 distance permutations

With k sites, the bisector system contains $\binom{k}{2} = \Theta(k^2)$ hyperplanes.

In d-dimensional space, by cake-cutting there can't be any more than $S_d\left(\binom{k}{2}\right) = \Theta(k^{2d})$ distance permutations.

Chávez, Figueroa, and Navarro were pleased to get from $nk \log n$ bits for LAESA to $nk \log k$ bits, by going to an approximate search.

But now we can store their index in $O(nd \log k)$ bits with no further sacrifice.

Taking into account the transitivity of equality we can also get an exact count for L_2 space.

Other L_p metrics

For other metrics the bisectors are badly behaved. They can intersect multiple times, or not at all.

Oriented matroids may help with the combinatorics of this problem, but the existing results do not immediately solve it.

However: L_1, L_2, and L_∞ are the ones people care about, and in those cases bisectors are piecewise linear.

By counting all the linear pieces, we get

$$N_{d,1}(k) = O(2^{2d^2}k^{2d})$$
$$N_{d,2}(k) = O(k^{2d})$$
$$N_{d,\infty}(k) = O(2^{2d}d^{2d}k^{2d})\,.$$

All three of these are $O\left(k^{2d}\right)$ for constant d.

Experimental results for sample databases

			k:							
database	n	ρ	3	4	5	6	7	8	9	10
Dutch	229328	7.159	6	24	119	577	2693	11566	34954	74954
English	69069	8.492	6	24	120	645	2211	7140	16212	28271
French	138257	10.510	6	24	118	475	2163	8118	19785	35903
German	75086	7.383	6	24	119	517	1639	4839	10154	19489
Italian	116879	10.436	6	24	120	653	3103	10872	27843	45754
Norwegian	85637	5.503	6	24	118	632	2530	7594	15147	25872
Spanish	86061	8.722	6	24	118	598	2048	5428	13357	23157
listeria	20660	0.894	4	11	19	29	49	85	206	510
long	1265	2.603	5	10	22	47	51	98	114	163
short	25276	808.739	6	24	111	508	2104	6993	13792	20223
colors	112544	2.745	6	18	44	96	200	365	796	1563
nasa	40150	5.186	6	24	115	530	1820	3792	7577	13243

Experimental results for random vectors

			mean perms			max perms		
	d	ρ	$k=4$	8	12	$k=4$	8	12
L_1	1	1.00	7.00	29.00	67.00	7	29	67
	4	4.00	23.95	4705.35	82253.85	24	5663	94537
	7	7.00	24.00	20811.65	569807.35	24	27824	653015
	10	10.00	24.00	30715.55	884013.40	24	35698	917237
L_2	1	1.00	7.00	28.95	67.00	7	29	67
	4	4.88	22.70	4214.20	67179.40	24	5079	75850
	7	9.09	23.95	17349.30	502957.40	24	23944	613857
	10	13.35	24.00	25562.25	815217.05	24	33097	905490
L_∞	1	1.00	7.00	29.00	67.00	7	29	67
	4	5.05	23.50	3664.60	54838.10	24	4912	70354
	7	9.80	23.70	14384.65	357331.00	24	23983	466484
	10	14.90	24.00	22415.00	648613.15	24	34281	770769

Euclidean count as a limit

Recall that in the two-dimensional figures, the L_1 example had exactly 18 distance permutations, known to be the maximum (and nearly always achieved) for the equivalent L_2 case.

In practice it is hard to find even that many, and note that the means in our vector experiment were generally much smaller than the Euclidean limit.

So is the Euclidean count an upper bound on all the L_p counts? In other words, does $N_{d,p}(k) = N_{d,2}(k)$?

Answer: no. One counterexample in the paper, and we've found others.

Open problems

Exact counts or tighter bounds for L_1 and L_∞

Any counts or bounds for other spaces? I have some for strings.

What do distance permutation counts reveal about dimensionality of spaces?

What are the consequences for indexing?