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:

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

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

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

for $p = \infty$. 

![Diagram showing different shapes for different values of p]

- $p = 1/2$
- $p = 1$
- $p = 3/2$
- $p = 2$
- $p = 4$
- $p = \infty$
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

$$\arg \min_{x \in \text{library}} d(x, \text{Heart of Darkness})$$

range query

$$\{ x \mid d(x, \text{Heart of Darkness}) < r \}$$
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*

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

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.

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

$$S_d(m) = W(d, m), \text{ “Whitney numbers,” Sloan’s A0004070}$$
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 $L_\infty$ 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 $O(k^{2d})$ for constant $d$. 
Experimental results for sample databases

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>(d)</th>
<th>(\rho)</th>
<th>(k = 4)</th>
<th>(8)</th>
<th>(12)</th>
<th>(k = 4)</th>
<th>(8)</th>
<th>(12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_1)</td>
<td>1</td>
<td>1.00</td>
<td>7.00</td>
<td>29.00</td>
<td>67.00</td>
<td>7</td>
<td>29</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4.00</td>
<td>23.95</td>
<td>4705.35</td>
<td>82253.85</td>
<td>24</td>
<td>5663</td>
<td>94537</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>7.00</td>
<td>24.00</td>
<td>20811.65</td>
<td>569807.35</td>
<td>24</td>
<td>27824</td>
<td>653015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10.00</td>
<td>24.00</td>
<td>30715.55</td>
<td>884013.40</td>
<td>24</td>
<td>35698</td>
<td>917237</td>
<td></td>
</tr>
<tr>
<td>(L_2)</td>
<td>1</td>
<td>1.00</td>
<td>7.00</td>
<td>28.95</td>
<td>67.00</td>
<td>7</td>
<td>29</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4.88</td>
<td>22.70</td>
<td>4214.20</td>
<td>67179.40</td>
<td>24</td>
<td>5079</td>
<td>75850</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>9.09</td>
<td>23.95</td>
<td>17349.30</td>
<td>502957.40</td>
<td>24</td>
<td>23944</td>
<td>613857</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>13.35</td>
<td>24.00</td>
<td>25562.25</td>
<td>815217.05</td>
<td>24</td>
<td>33097</td>
<td>905490</td>
<td></td>
</tr>
<tr>
<td>(L_\infty)</td>
<td>1</td>
<td>1.00</td>
<td>7.00</td>
<td>29.00</td>
<td>67.00</td>
<td>7</td>
<td>29</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5.05</td>
<td>23.50</td>
<td>3664.60</td>
<td>54838.10</td>
<td>24</td>
<td>4912</td>
<td>70354</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>9.80</td>
<td>23.70</td>
<td>14384.65</td>
<td>357331.00</td>
<td>24</td>
<td>23983</td>
<td>466484</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>14.90</td>
<td>24.00</td>
<td>22415.00</td>
<td>648613.15</td>
<td>24</td>
<td>34281</td>
<td>770769</td>
<td></td>
</tr>
</tbody>
</table>
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 $L_\infty$

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?