Rank-Approximate Nearest Neighbor Search
Retaining Meaning and Speed in High Dimensions

Problem: For a query q, find a point p in S such that
\[ d(p, q) = \min_{r \in S} d(r, q). \]

Existing exact methods:
- Data structures: kd-tree (Freidman, et al., '77), ball tree (Omohundro, '89), metric tree (Preparata, et al., '85)
- Algorithms: Dual-tree algorithm (Gray, et al., '00)

Issues:
- Pairwise distances get concentrated in a small interval in high dimensional data (Hammersley, '50)
- True even in datasets in practice
- Tree-based methods cannot partition the data in the useful manner
- Runtimes end up being no better than naive method
- Approximation necessary to make the problem scalable!!

Issues:
- PAC model for NN (Ciaccia, et al., '00)
- Random projection based: Locality Sensitive Hashing (LSH) (Indyk, et al., '98)
- Tree-based: Spill trees (Liu, et al., '05)

Solution proposed:
- In NN only ordering of the distances matter, not the actual values (Beyer, et al., '99, Alon, et al., '08)
- INTRODUCE APPROXIMATION IN ORDERING/ RANKS OF THE NN

Distance-approximate methods:
Find p' in S such that
\[ d(p', q) \leq (1 + \epsilon) \min_{r \in S} d(r, q) \]
- PAC model for NN (Ciaccia, et al., '00)
- Random projection based: Locality Sensitive Hashing (LSH) (Indyk, et al., '98)
- Tree-based: Spill trees (Liu, et al., '05)

Rank-approximation with Trees:
- Use stratified sampling on the tree
- Can use dual-tree algorithm for multiple queries

Points to take home:
- DIRECT CONTROL OF RANK ERROR VS. SPEED AVAILABLE TO THE USER
- AUTOMATIC – NO TWEAKING OF PARAMETERS REQUIRED, UNLIKE LSH
- RELIABLE – ROBUST BEHAVIOR OVER A LARGE RANGE OF ERROR TOLERANCE
- AVERAGE ERRORS COMPAREABLE IN BOTH FORMS OF APPROXIMATION
- MAXIMUM ERROR MUCH BETTER – STABILITY IN THE QUALITY OF RESULT PROVIDED BY USE OF A TREE

Comparison with LSH:
- For datasets LayoutHistogram (dim=30) & MNIST (dim=784)
- Comparing average rank errors (lower rank error implies better distance approximation) for same query retrieval times
- Comparing the maximum rank errors (the worst case accuracy)

Comparison with Exact Search:
- Comparing speedup of both methods over brute force
- Dual-tree NN used for exact search
- Allowed rank-error relative to the size of the dataset (\( \tau = \epsilon N, \epsilon = [0.001\% - 10\%] \))

Notes:
- Significant speedups for low values of error
- Our solution is faster than LSH with better overall accuracy

RANK-APPROXIMATE NN FORMULATION: With the order statistics D for distances of q to points in S, find p' such that
\[ d(p', q) \leq D_{1(1+\tau)} \]
- Find rank (1+\tau) NN instead of rank 1 NN
- Use PAC model: Find rank-approximate NN with high probability as following:
\[ P(d(p', q) = d(q) \leq D_{1(1+\tau)}) \geq \alpha. \]
- Sampling based method:
  - Use probability bounds on the sample order statistics in comparison to the order statistics of the whole set
  \[ P_d(1) \leq D_{1(1+\tau)} = \sum_{i=0}^{n-1} \left( \frac{N - \tau + i - 1}{n} \right) \left( \frac{N - \tau + i}{n} \right) \]

Contact:
- P.Ram@gatech.edu
- agray@cc.gatech.edu
- www.fast-lab.org