Rank Cover Trees
for Nearest Neighbor Search

Michael E. HOULE
NII

Michael NETT
RWTH Aachen University, GERMANY

Why

Text, images, market data, biological data, scientific data, and other forms of information are currently being gathered in large data repositories at a rate that greatly outstrips our ability to analyze and to interpret. Together with this explosion of information, the demand for effective methods for searching, clustering, categorizing, summarizing and matching within data sets continues to grow. For such applications, solutions based on similarity search are among the most effective proposed in statistics, pattern recognition, and machine learning. The design and analysis of effective similarity search structures has consequently been the subject of intensive research for many decades.

What

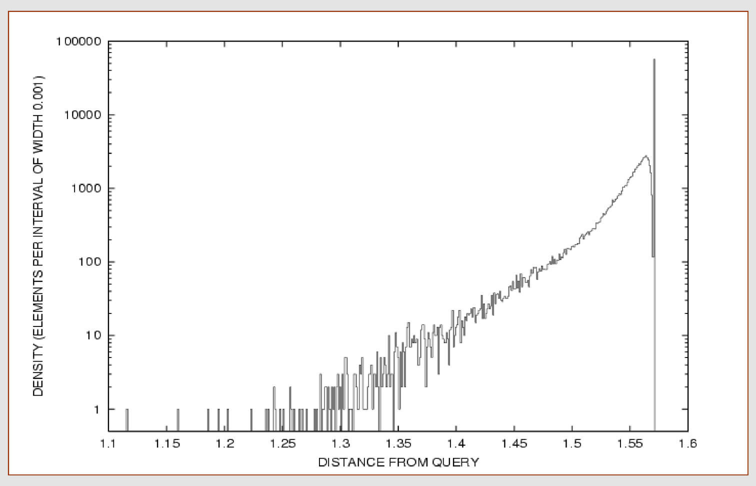
We propose the *Rank Cover Tree* (RCT), a probabilistic data structure for similarity search in general metric spaces. The RCT reinterprets the design and analysis of the Cover Tree in terms of neighbor ranks as measured from the query point, rather than explicit distances. The ranked-based analysis results in a significantly smaller dependence on the intrinsic dimensionality over practical data set sizes. This allows the RCT to find approximations of very high qualities, orders of magnitudes faster than structures for exact similarity search. Moreover, the RCT is highly competitive with the SASH heuristic in terms of its speed-up accuracy trade-off.

Similarity Search

Curse of Dimensionality

Observations

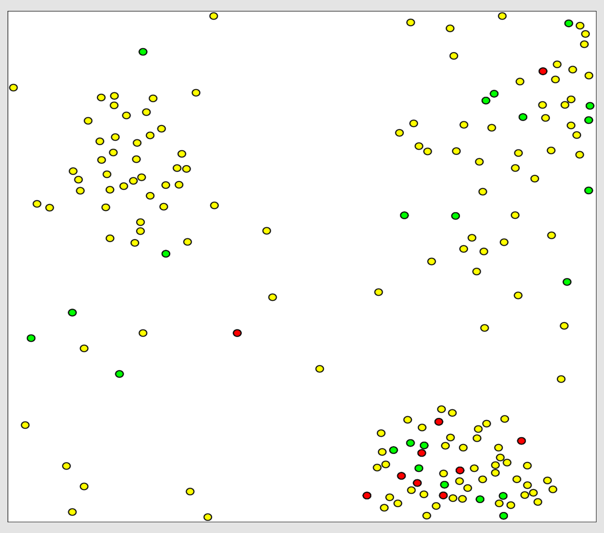
- Sequential scans outperform classical search structures.
- Similarity values concentrate around their mean.
- Similar and dissimilar points are distinguishable.
- Spatial intuitions from 3D spaces are invalid.



→ Use approximate similarity search!

Idea of Sampling

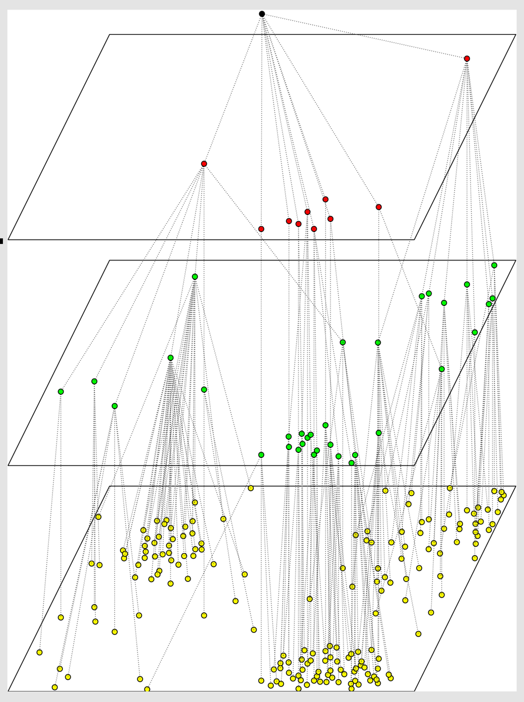
- We try to find items similar to a query q with respect to some data set T.
- Suppose for a subset S we already know an item x that is similar to q.
- An item $y \in T \setminus S$ that is similar to x is likely to be similar to q, and the probability of this can be bounded!



Rank Cover Trees

Construction

- For each item $x \in T$, introduce x into levels $0, \ldots, \lambda(x)$. For a tree of height h, $\lambda(x)$ follows a geometric distribution with $p = \frac{1}{\sqrt[h]{n}}$.
- Build a partial RCT on the highest level by connecting items in that level to an artificial root.
- Connect the next level by using approximate nearest neighbors found in the partial RCT.
- Well-formed with high probability.



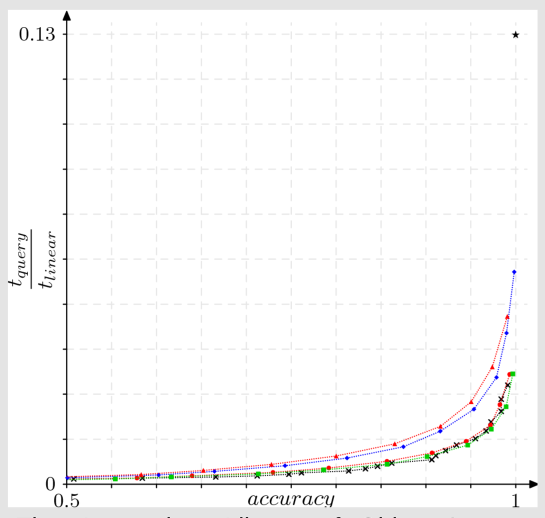
Search

Find the k items most similar to a query q.

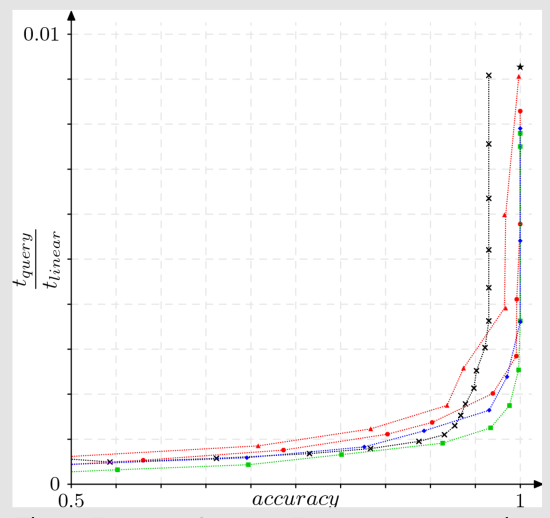
- For each level i, maintain a set of ancestors C_i covering the query result.
- Start with C_h containing the artificial root.
- C_i is constructed from the set C_{i+1} by keeping the k_i children of all elements in C_i, which are most similar to the query q.
- The set C_0 contains the query result.
- $k_i = \omega \cdot max\{k \cdot n^{-1/h}, 1\}$, where ω is parameter allowing to trade-off between accuracy and query time.
- If $\omega \geq O(\delta^{\log_\phi(\sqrt{5}h)} \cdot h + max\{h, en^{1/h}\})$, then the approximation is free of error with very high probability.
- The expansion rate δ measures intrinsic dimensionality.

Performance on real data sets

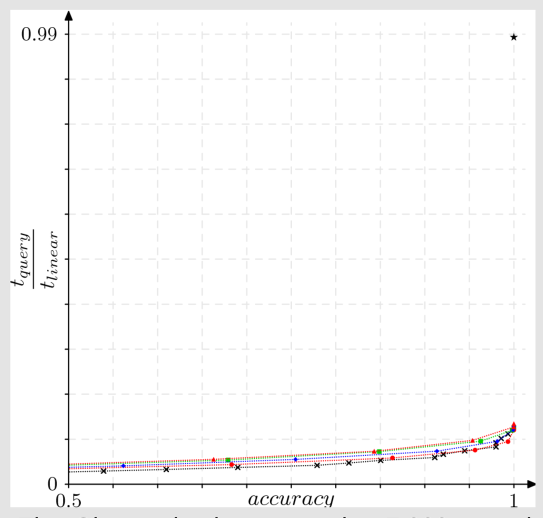
The following figures display the trade-off between approximation quality and query time. Query times are measured as factors of the time consumed by a sequential scan. The query time of the Cover Tree structure for exact similarity searches is provided as a reference point (★). The plots include the RCT for heights 3, 4, 5 and 8 (●, ■, ◆ and ▲) and the SASH heuristic (X).



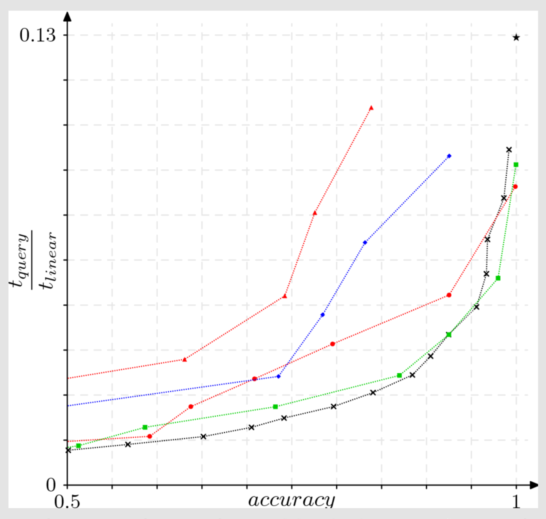
The Amsterdam Library of Object Images (ALOI) contains **108,750** photos of 1,000 objects that are taken with different angle and illumination. Items represent **641** dimensional histogram vectors which are compared using the Euclidean distance.



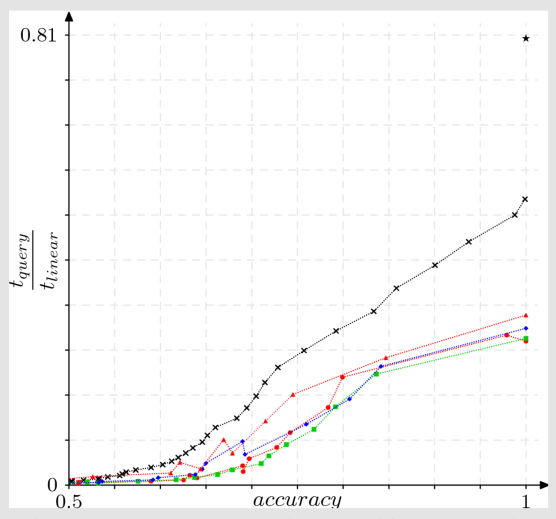
The Forest Cover Type set contains topological information on **581,012** forest-cells of 900 square meters each. The **54** attributes include elevation, slope, soil- and cover types. Normalized vectors are compared using the Euclidean distance.



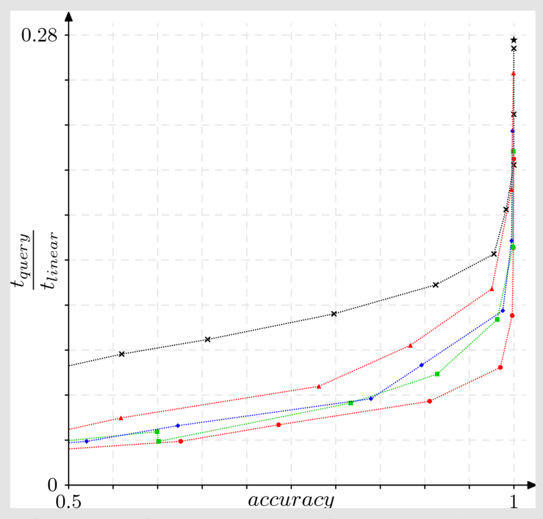
The Gisette database contains **7,000** records of the handwritten digits *4* and *9* from the NIPS 2003 feature selection challenge. Gisette contains **5,000** different features. Similarity of feature vectors is measured in terms of their Euclidean distance.



Poker Hand contains **1,025,010** sets of 5 cards each, drawn from a regular poker deck. The **10** features contain rank and suit of each card. Dissimilarity between categorical vectors is measured using the Hamming distance.



The Reuters news-wire archive contains **554,651** text documents preprocessed with Porter stemming. The **320,648** items are encoded as sparse TF-IDF vectors. Similarities are expressed in terms of the cosine similarity.



The Spambase set contains **4,601** feature vectors describing e-mail. The **57** attributes measure the frequency of certain keywords in the text. We compare items using their Euclidean distance.

NII 連絡先： **Michael E. HOULE (フール　マイケル) ／ 国立情報学研究所　客員教授**
TEL : 03-4212-2538　　FAX : 03-3556-1916　　Email : meh@nii.ac.jp