Rank Cover Trees for Nearest Neighbor Search

Michael E. HOULE [1] **Michael NETT** [1,2]

[1] National Institute of Informatics, Japan
[2] The University of Tokyo, Japan

Summary

Virtually all known distance-based similarity search indexes make use of some form of numerical constraints (triangle inequality, additive distance bounds, ...) on similarity values for pruning and selection. The use of such numerical constraints, however, often leads to large variations in the numbers of objects examined in the execution of a query, making it difficult to control the execution costs. We introduce a probabilistic data structure for similarity search, the *Rank Cover Tree* (RCT), that entirely avoids the use of numerical constraints. All internal selections are made according to the ranks of the objects with respect to the query, allowing much tighter control on the overall execution costs. A rank-based probabilistic analysis shows that with very high probability, the RCT returns a correct query result in time that depends competitively on a measure of the intrinsic dimensionality of the data set.

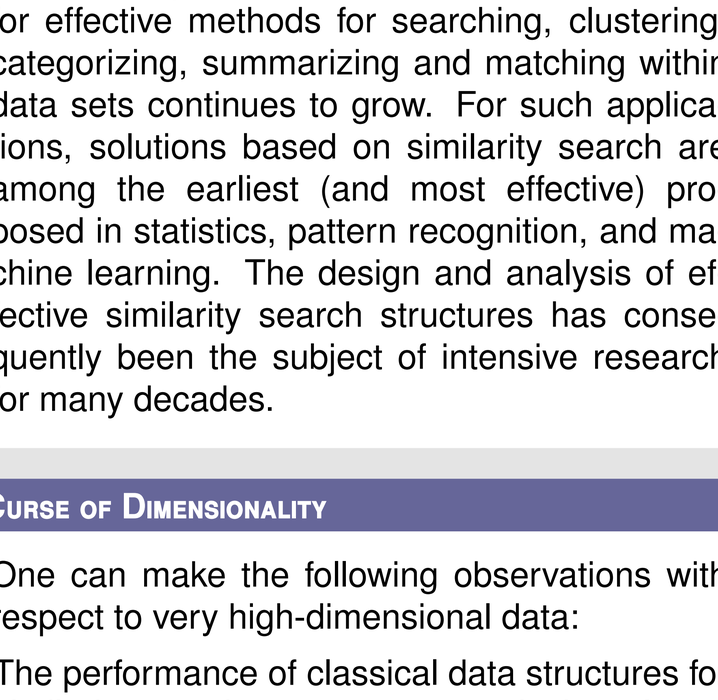
Motivation

Text, images, market data, biological data, scientific data, and many other forms of information are currently being accumulated 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 earliest (and 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.

Curse of Dimensionality

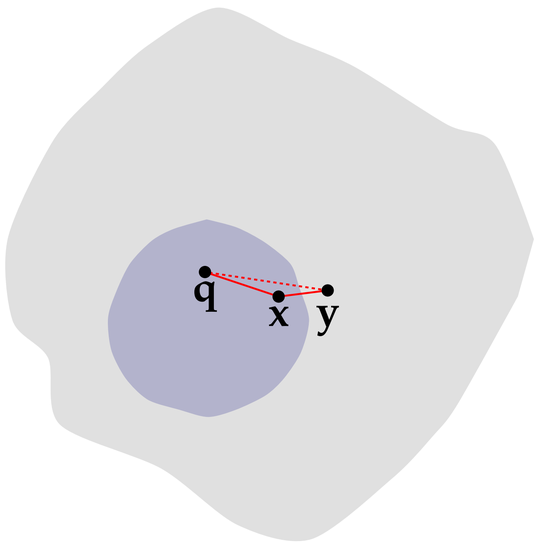
One can make the following observations with respect to very high-dimensional data:

- The performance of classical data structures for similarity search converges towards that of a sequential scan.
- Point-to-point distances become indistinguishable as they concentrate heavily around their mean value.
- Individual search paths within a similarity search data structure can no longer be effectively excluded from consideration.



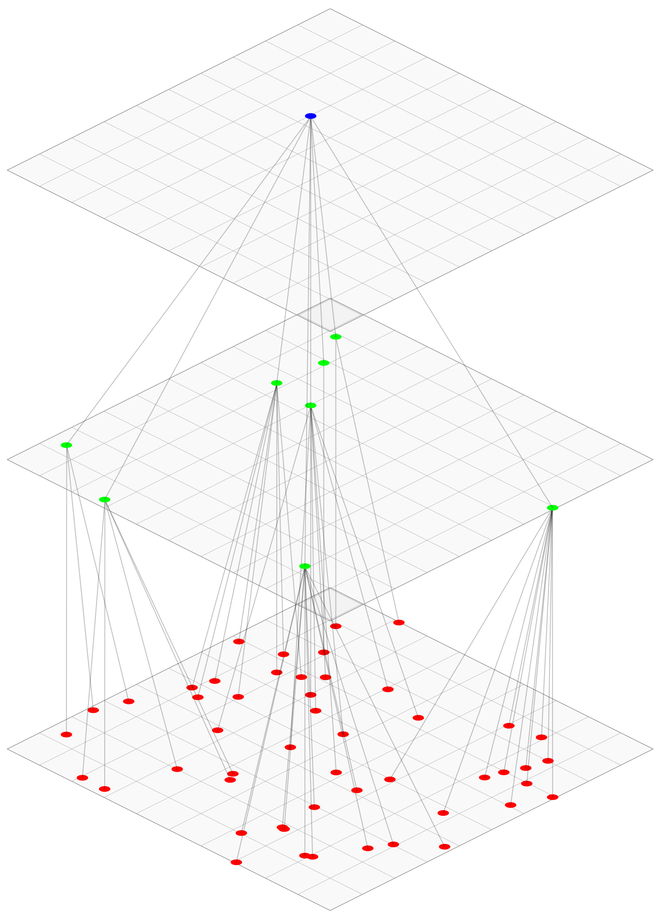
Idea of Sampling

- We try to find items similar to a query object $\mathbf{q}$ with respect to some data set $X \subseteq \Omega$.
- Suppose we found a similar point $\mathbf{x}$ with respect to a (small) subset $X' \subseteq X$, for example, by means of a sequential scan.
- We are likely to observe transitivity: an item $\mathbf{y} \in X \setminus X'$ which is similar to the item $\mathbf{x}$ is also similar to $\mathbf{q}$.
- The probability of observing this kind of transitivity can be bounded!



Construction

- For each item $\mathbf{x} \in X$, introduce $\mathbf{x}$ into levels $0, \ldots, \lambda_{\mathbf{x}}$. For a tree of height h, $\lambda_{\mathbf{x}}$ follows a geometric distribution with $p = |X|^{-1/h}$.
- 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.



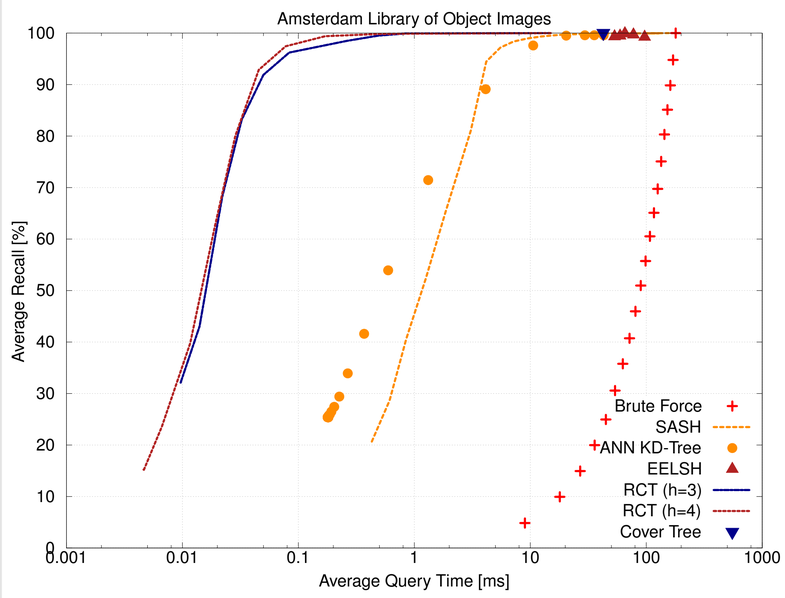
Additional Material

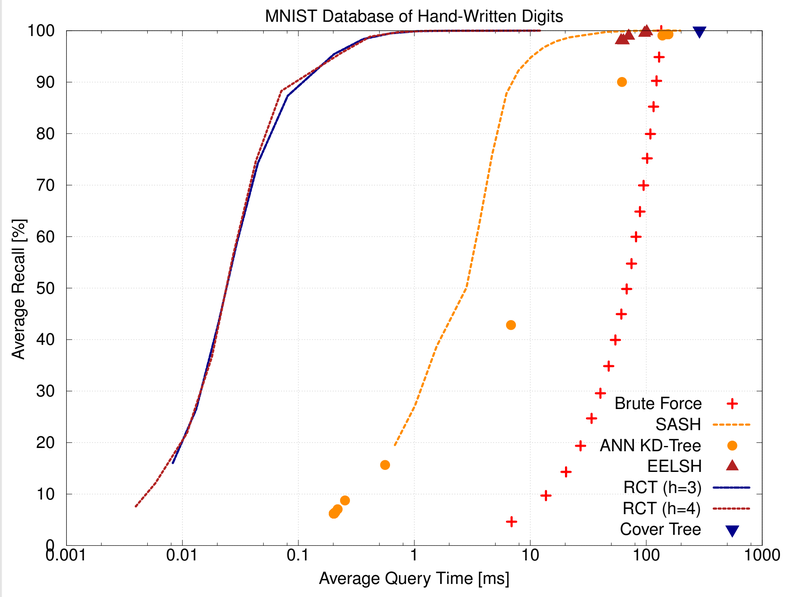
- Technical Report
- Poster
- Implementation
- Documentation

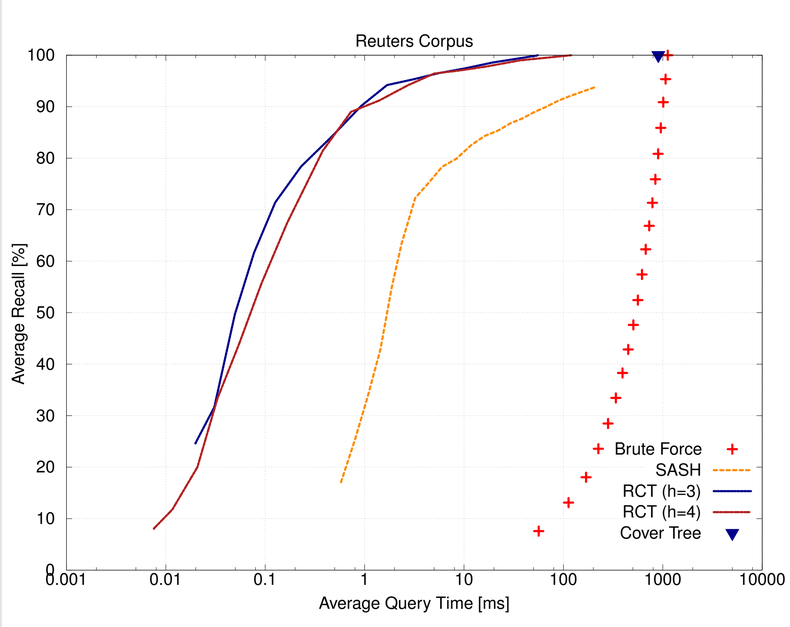
K-Nearest Neighbor Search

- Maintain level-wise sets C_i covering the query results. 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+1}, which are most similar to the query $\mathbf{q}$.
- The set C_0 contains the query result.
- We choose $k_i = \omega \cdot \max\{k/\sqrt[h]{|X|}, 1\}$, where ω is a parameter allowing to trade-off between accuracy and query time.
- Our analysis shows: if ω is chosen greater than $\delta^{\log_\phi(\sqrt{5}h)}h + \max\{h, e\sqrt[h]{|X|}\}$, then the approximation is free of error with very high probability.
- The expansion rate δ measures intrinsic dimensionality.

Recall Rates







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
☎: 03-4212-2538 FAX: 03-3556-1916 ✉: meh@nii.ac.jp