From Cover Trees to Net-Trees

Mahmoodreza Jahanseir* Donald R. Sheehy†

1 Introduction

Cover trees are a popular data structure for (approximate) nearest neighbor search on metric spaces of low intrinsic dimension [1]. They are superficially similar to the net-trees of Har-Peled & Mendel [2] as both structures may be interpreted as generalizations of compressed quadtrees beyond the Euclidean setting. Cover trees are the simplest of these data structures, requiring linear space, independent of the ambient or intrinsic dimension of the data set. An efficient implementation of cover trees is available. On the other hand, net-trees have better theoretical guarantees for preprocessing and query times. However, to achieve subquadratic preprocessing time, it uses complex techniques which are not efficient in practice.

We show how a slight modification to the definition of a cover tree allows it to satisfy the stronger conditions of a net-tree. Leveraging this structural result, we give a simple algorithm to turn a given cover tree into a net-tree in linear time.

2 Background

The input is a set of n points P in a metric space. The *closed metric ball* centered at p with radius r is denoted $\mathrm{B}(p,r) := \{q \in P \mid \mathbf{d}(p,q) \leq r\}$. The *doubling constant* ρ of P is the minimum $\rho \in \mathbb{R}$ such that every ball $\mathrm{B}(p,r)$ can be covered by ρ balls of radius $r/2$. We assume ρ is constant.

Cover trees and net-trees are both examples of hierarchical trees. In such trees, the input points are leaves and each point p can be associated with many internal nodes. Each node is uniquely identified by its associated point and an integer called its *level*. The node in level ℓ associated with a point p is denoted p^ℓ. If $q^{\ell'}$ is the parent of p^ℓ, then $\ell' > \ell$ and if $\ell' > \ell + 1$ then $p = q$. Moreover, if p^ℓ is a non-leaf node, then it has a child p^m, where $m < \ell$. Let L_ℓ be the points associated with nodes in level at least ℓ. Let P_{p^ℓ} denote leaves of the subtree rooted at p^ℓ. The levels of the tree represent the metric space at different scales.

The constant τ, called the *scale factor* of the tree determines the change in scale between levels as will be seen in the following definitions.

Cover Trees. A cover tree is defined by the following properties. (*Packing*) For all distinct $p, q \in L_\ell$, $\mathbf{d}(p,q) > \tau^\ell$. (*Covering*) If $q^{\ell'}$ is the parent of p^ℓ then $\mathbf{d}(p,q) \leq c\tau^{\ell'}$. We call c the *covering constant*, and in a usual cover tree $c = 1$.

Net-trees. For each node p^ℓ in a net-tree, the following invariants hold. (*Packing*) $\mathrm{B}(p, \frac{\tau-5}{2(\tau-1)} \cdot \tau^\ell) \bigcap P \subset P_{p^\ell}$. (*Covering*) $P_{p^\ell} \subset \mathrm{B}(p, \frac{2\tau}{\tau-1} \cdot \tau^\ell)$.[1]

The main difference in the definitions lies in the packing conditions. The net-tree requires the packing to be consistent with the hierarchical structure of the tree, a property not necessarily satisfied by the cover trees. Also, Har-Peled and Mendel [2] set $\tau = 11$, whereas optimized cover tree code sets $\tau = 1.3$.

A net-tree can be augmented to maintain a list of nearby nodes with no additional cost. For each node p^ℓ, $\mathrm{Rel}(p^\ell)$ is defined as follows.

$$\mathrm{Rel}(p^\ell) = \{q^i \in T \text{ with parent } r^j \mid i \leq \ell < j, \text{ and } \mathbf{d}(p,q) \leq 13 \cdot \tau^\ell\}$$

We add a new, easy to implement condition on cover trees. We require that children of a node p^ℓ are closest to p than to any other point in L_ℓ.

3 From cover trees to net-trees

Theorem 1. *A cover tree T with scale factor $\tau > 3$ is a net-tree, such that for each node p^ℓ, $\mathrm{B}(p, \frac{(\tau-3)\tau^\ell}{2(\tau-1)}) \cap P \subset P_{p^\ell}$ and $\mathrm{B}(p, \frac{\tau^{\ell+1}}{\tau-1}) \supset P_{p^\ell}$.*

Proof. Fix a node $p^\ell \in T$. The covering property and the triangle inequality imply that $\mathbf{d}(p, P_{p^\ell}) \leq \sum_{i=0}^{\infty} \tau^{\ell-i} = \frac{\tau^{\ell+1}}{\tau-1}$. Therefore, $P_{p^\ell} \subset \mathrm{B}(p, \frac{\tau^{\ell+1}}{\tau-1})$.

Suppose for contradiction there exists a point $r \in \mathrm{B}(p, \frac{\tau-3}{2(\tau-1)}\tau^\ell)$ such that $r \notin P_{p^\ell}$. Then, there exists

[1] The packing condition we give is slightly different from [2], but it is an easy exercise to prove this (more useful) version is equivalent.

*University of Connecticut reza@engr.uconn.edu
†University of Connecticut don.r.sheehy@gmail.com

a node $x^i \in T$ that it is the lowest node with $i \geq \ell$ and $r \in P_{x^i}$.

First, assume that $i = \ell$. Let y^j be the child of x^i such that $r \in P_{y^j}$. If $j < i-1$, then $x = y$ and $\mathbf{d}(p, y) > \tau^\ell$. By the triangle inequality,

$$\mathbf{d}(y, r) \geq \mathbf{d}(y, p) - \mathbf{d}(p, r) > \tau^\ell - \frac{(\tau-3)\tau^\ell}{2(\tau-1)} > \frac{\tau^\ell}{\tau-1}.$$

However, $\mathbf{d}(y, r) \leq \frac{\tau^\ell}{\tau-1}$, because y is the ancestor of r. Otherwise, when $j = i-1$, x is the nearest neighbor of y among L_i. So, $\mathbf{d}(y, p) \geq \mathbf{d}(p, x) - \mathbf{d}(x, y) > \tau^\ell - \mathbf{d}(y, p) > \tau^\ell/2$. This case as well as the case when $i > \ell$ also yield a contradiction by similar arguments (In the latter case, $y = x$). Therefore, $r \in P_{p^\ell}$. □

Theorem 2. *For a given cover tree T on a set of n points P with doubling constant ρ, one can augment T with relative lists in $\rho^{O(1)}n$.*

Proof. When constructing a cover tree, we also build a hash table to support constant time access to the list of nodes in a specific level with the same asymptotic cost. We process each list of nodes from highest to lowest. For each unvisited node u, find $\mathrm{Rel}(u)$ for the part of the tree that corresponds to the visited nodes using a method similar to [2]. Then, mark u as visited and continue until process all nodes of the cover tree. The time complexity follows from the same analysis as [2]. □

In Theorem 1, we assumed that $\tau > 3$. However, Beygelzimer et. al. [1] set $\tau = 2$, and they found $\tau = 1.3$ is even more efficient in practice. Because such cover trees may not be net-trees, we give an algorithm to increase the scale factor of a cover tree, and show when such a coarsened cover tree is a net-tree.

Theorem 3. *Given a cover tree T with $\tau > 1$ and $c = 1$, and a constant integer $k > 1$. T can be turned into a cover tree T' with scale factor $\tau' = \tau^k$ and covering constant $c' = \sum_{i=0}^{k-1} 1/\tau^i$ in $\rho^{O(1)}n$ time.*

Proof. Using Theorem 2, we augment T with relative lists in $\rho^{O(1)}n$ time. However, instead of constant 13 in the definition of relatives, we use a constant greater than $\frac{2\tau}{\tau-1}$. Then, we define a mapping between nodes of T and T'. Here, each node p^ℓ in T maps to a node $p^{\ell'} = p^{\lfloor \ell/k \rfloor}$ in T'. If $\ell = k\ell'$, then levels $\ell, \ell+1, \ldots, \ell+k-1$ of T combine into the level ℓ' in T'. Note that the level of root is ∞.

If $r^m \in T$ is the descendant at most k levels down from some p^ℓ, then by the covering property and the triangle inequality, $\mathbf{d}(p, r) \leq \sum_{i=0}^{k-1} \tau^{\ell-i} = (\tau')^{\ell'} \sum_{i=0}^{k-1} 1/\tau^i$. Because we are combining sets of k consecutive levels, it follows that each node in T' will have a node in the level above whose distance is at most this amount. It follows that T' has a covering constant $\sum_{i=0}^{k-1} 1/\tau^i$.

If $\ell = k\ell'$, then the minimum distance between points in level ℓ' of T' is equal to the minimum distance between points in level ℓ of T, which is at least $\tau^\ell = (\tau')^{\ell'}$. Thus, the points in level ℓ' of T' satisfy the packing condition.

To find a correct parent for a node $p^{\ell'}$ in T', we find the lowest ancestor q^m of p^ℓ in T such that $m' > \ell'$. Then, among relatives of q^m, we find a node r^i which is closest to p^ℓ. If $i' = \ell'$, we create a node $r^{i'+1}$ and add $p^{\ell'}$ and $r^{i'}$ as children of it. If $i' < \ell'$, create two nodes $r^{\ell'+1}$ and $r^{\ell'}$, add $r^{\ell'}$ and $p^{\ell'}$ in the child list $r^{\ell'+1}$ and set $r^{\ell'}$ as the parent of $r^{i'}$. Otherwise, find the lowest node $r^{j'}$ such that $\ell' < j' \leq i'$. If $j' = \ell'+1$, create a node $r^{\ell'}$ and add $r^{\ell'}$ and $p^{\ell'}$ as children of $r^{j'}$. If $j' > \ell'+1$, we create $r^{\ell'+1}$ as the only child of $r^{j'}$, create $r^{\ell'}$ and add $r^{\ell'}$ and $p^{\ell'}$ to the child list of $r^{\ell'+1}$. The whole construction process can be done in $\rho^{O(1)}n$.

Now, we need to show that this $r^{i'}$ is the nearest neighbor of $p^{\ell'}$ among all nodes of T' with levels greater than ℓ'. For contradiction, suppose that there exists a node $s^{t'}$ such that $t' > \ell'$, $s^{t'} \notin \mathrm{Rel}(q^{m'})$ and $s_{t'}$ is the closest node to $p^{\ell'}$ among all nodes with the level greater than ℓ'. Therefore,

$$\mathbf{d}(p^{\ell'}, s^{t'}) < \mathbf{d}(p^{\ell'}, q^{m'}) \leq c'(\tau')^{m'} < \frac{\tau}{\tau-1}(\tau')^{m'}.$$

Also, $\mathbf{d}(p^{\ell'}, s^{t'}) \geq \mathbf{d}(q^{m'}, s^{t'}) - \mathbf{d}(p^{\ell'}, q^{m'})$
$> \frac{2\tau}{\tau-1}(\tau')^{m'} - c'(\tau')^{m'}$.
These inequalities imply a contradiction. □

Corollary 4. *Given a cover tree T with $\tau > 1$ and $c = 1$ on a set of n points P, one can find a net-tree T' with scale factor $\tau' = \tau^k$, for a constant integer $k \geq 1$, and $\tau' > 1 + 2c'$, where $c' = \sum_{i=0}^{k-1} 1/\tau^i$, from T in $\rho^{O(1)}n$ time such that for each node $p^{\ell'}$ in T', $\mathrm{B}(p, \frac{\tau'-1-2c'}{2(\tau'-1)}(\tau')^{\ell'}) \cap P \subset P_{p^{\ell'}}$ and $\mathrm{B}(p, \frac{c'(\tau')^{\ell'+1}}{\tau'-1}) \supset P_{p^{\ell'}}$.*

References

[1] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *ICML*, 2006.

[2] S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics, and their applications. *SIAM Journal on Computing*, 35(5):1148–1184, 2006.