Building and Searching a Balanced, Distributed k-d Tree
with MapReduce

Russell A. Brown

ABSTRACT
The original description of the k-d tree recognized that rebalancing techniques, such as are used to build an AVL tree or a red-black tree, are not applicable to a k-d tree. Hence, in order to build a balanced k-d tree, it is necessary to obtain all of the data prior to building the tree then to build the tree via recursive subdivision of the data. One algorithm for building a balanced k-d tree finds the median of the data for each recursive subdivision of the data. A new algorithm builds a balanced k-d tree by presorting the data in each of k dimensions prior to building the tree, then constructs the tree in a manner that preserves the order of the k presorts during recursive subdivision of the data. This new algorithm is amenable to execution via MapReduce and permits building and searching a k-d tree that is represented as a distributed graph.

1. INTRODUCTION
Bentley introduced the k-d tree as a binary tree that stores k-dimensional data [3]. Like a standard binary tree, the k-d tree subdivides data at each recursive level of the tree. Unlike a standard binary tree that uses only one key for all levels of the tree, the k-d tree uses k keys and cycles through these keys for successive levels of the tree. For example, to build a k-d tree from two-dimensional points that comprise (x, y) coordinates, the keys would be cycled as x, y, x, y, x, y... for alternate levels of the k-d tree.

Due to the use of different keys at successive levels of the tree, it is not possible to employ rebalancing techniques, such as are used to build an AVL tree [1] or a red-black tree [2] when building a k-d tree. Hence, the typical approach to building a balanced k-d tree finds the median of the data for each recursive subdivision of those data. Bentley showed that if the median of n elements were found in O(n) time, it would be possible to build a depth-balanced k-d tree in O(n log n) time. However, algorithms that find the median in guaranteed O(n) time are somewhat complicated [4, 6]. An alternative approach to building a balanced k-d tree presorts the data in each of k dimensions prior to building the tree using merge sort [8, 12, 13] or heap sort [18]. The order of the k presorts is then maintained when building a balanced tree and thereby permits a worst-case complexity of O(kn log n) for building the tree [5].

The remainder of this article will use as an example a k-d tree that sorts rectangular bounding boxes [15]. This k-d tree permits a directed search of the tree by a query bounding box in order to determine which of the bounding boxes from the tree the query bounding box intersects. The principles of building and searching such a k-d tree will be discussed initially for a tree that resides in memory, then extended to a k-d tree that is distributed across multiple MapReduce partitions as a distributed graph [14].

2. BUILDING A MEMORY-RESIDENT TREE
The k-d tree will be used to search for bounding boxes that intersect a query bounding box. Hence, each node of the tree must store the \((x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})\) coordinates of a bounding box to facilitate an intersection test of that bounding box against the query bounding box, as shown for the root node in Figure 1. In addition, each node of the tree must permit a determination of which of the node’s two subtrees the query bounding box will search recursively. This determination is facilitated by storing, for each subtree, the \((x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})\) coordinates of a rectangular bounding region that just encloses all of the bounding boxes in the subtree. If the query bounding box intersects a bounding region, the query bounding box must search the associated subtree.

Figure 1: The root node \(n_r\) of the k-d tree and its two children \(n_<\) and \(n_>\) are depicted by circles. The root node stores the bounding box \(bb_r\) and the bounding regions \(br_<\) and \(br_>\) that are depicted by dashed rectangles whose sizes and locations are not represented accurately in this figure.
Given the above requirements for each node of the \( k \)-d tree, the tree may be constructed recursively as summarized below. A detailed description of the k-d tree building algorithm has been published previously [5].

First, the bounding boxes are presorted independently in \( x_{\min} \) and \( y_{\min} \) prior to building the \( k \)-d tree to create two presorted arrays of bounding boxes. In order that each bounding box may have a unique sorting key in each of \( x_{\min} \) and \( y_{\min} \), a unique name \( n_i \) is assigned to each bounding box and used to form the super keys \( x_{\min}n_i \) and \( y_{\min}n_i \) for use in the presents, where the colon represents the catenation operator.

Next, the presorted arrays are partitioned in \( x_{\min} \) at the root of the tree as follows. The bounding box \( bb_m \) at the median element \( n_m \) of the array that was presorted by the \( x_{\min}n_i \) super key is stored at the root of the tree. This median element trivially partitions the \( x_{\min}n_i \)-sorted array as shown in Figure 2. The elements of the \( x_{\min}n_i \)-sorted array at all addresses below the median address form a “less than” \( x_{\min}n_i \)-sorted subarray. The elements of the \( x_{\min}n_i \)-sorted array at all addresses above the median address form a “greater than” \( x_{\min}n_i \)-sorted subarray.

![Figure 2: The \( x_{\min}n_i \)-sorted array (\( x_{\min} \)) is partitioned trivially at its median element (\( n_m \)) to obtain the “less than” \( x_{\min}n_i \)-sorted subarray (\(<\)) and the “greater than” \( x_{\min}n_i \)-sorted subarray (\(>\)). The bounding box \( bb_m \) of the median element is stored at the root of the tree that is depicted by the circle.](image)

This median element also partitions the \( y_{\min}n_i \)-sorted array, albeit non-trivially, via a “sweep and partition” algorithm as shown in Figure 3. In order to partition the \( y_{\min}n_i \)-sorted array, the array is swept from lowest to highest address, i.e., swept in order from lowest to highest \( y_{\min}n_i \) super key, and the elements of this array are partitioned into a “less than” \( y_{\min}n_i \)-sorted subarray and a “greater than” \( y_{\min}n_i \)-sorted subarray as follows. Each element’s \( x_{\min}n_i \) super key is compared to the \( x_{\min}n_m \) super key of the median of the \( x_{\min}n_i \)-sorted array, whose bounding box is stored at the root of the tree. If the element’s super key is less than the root’s super key, the element is partitioned into the “less than” \( y_{\min}n_i \)-sorted array. If the element’s super key is greater than the root’s super key, the element is partitioned into the “greater than” \( y_{\min}n_i \)-sorted array. If the element’s super key equals the root’s super key, the element is ignored. In this manner, the \( y_{\min}n_i \)-sorted array is partitioned by the \( x_{\min}n_m \) super key of the root of the tree to create “less than” and “greater than” \( y_{\min}n_i \)-sorted subarrays that each preserve the relative \( y_{\min}n_i \)-sorted order from the initial presort.

![Figure 3: The \( y_{\min}n_i \)-sorted array (\( y_{\min} \)) is swept from lowest address to highest address and partitioned by the \( x_{\min} \) of the median element \( n_m \) to obtain the “less than” \( y_{\min}n_i \)-sorted subarray (\(<\)) and the “greater than” \( y_{\min}n_i \)-sorted subarray (\(>\)).](image)

The “less than” \( x_{\min}n_i \)-sorted and \( y_{\min}n_i \)-sorted subarrays are used to build recursively the “less than” subtree at the next level of the nascent \( k \)-d tree. The “greater than” \( x_{\min}n_i \)-sorted and \( y_{\min}n_i \)-sorted subarrays are used to build recursively the “greater than” subtree at the next level of the nascent \( k \)-d tree. At the root of each of these two subtrees, the \( y_{\min}n_i \)-sorted subarray is partitioned trivially. The \( x_{\min}n_i \)-sorted subarray is partitioned non-trivially via the “sweep and partition” algorithm that compares \( y_{\min}n_i \) super keys. The alternation of trivial and non-trivial partitioning in \( x_{\min} \) and \( y_{\min} \) continues at each level of the \( k \)-d tree until the subarrays are exhausted at the leaf nodes of the tree.

As the recursion unwinds, pointers to the node’s children are provided to each node of the nascent \( k \)-d tree. In addition, the bounding region for each node is calculated as depicted in Figure 3. For a leaf node, the dimensions of the bounding region are identical to those of the bounding box. For a non-leaf node that has one or two children, the bounding region is the region that just encloses the bounding box of the node and the bounding regions of the children. This bounding region is calculated as the minimum of the \( x_{\min} \), the minimum of the \( y_{\min} \), the maximum of the \( x_{\max} \) and the maximum of the \( y_{\max} \) of the node and of the node’s children.

3. BUILDING A DISTRIBUTED TREE WITH MAPREDUCE

A distributed \( k \)-d tree may be constructed with MapReduce using Spark. The construction proceeds recursively in a similar manner to the construction of a memory-resident \( k \)-d tree; however, arrays are not used to store the presorted bounding boxes. Instead, Spark uses a resilient distributed
Figure 4: The bounding boxes $bb_d$ and $bb_e$ of the respective leaf nodes $n_d$ and $n_e$ are identical to the bounding regions of these nodes. The bounding regions of the leaf nodes are combined with the bounding box $bb_c$ of their parent node $n_c$ as indicated by the curved, dashed arrows to calculate the bounding region $br_c$ of the parent node that is depicted by the dashed rectangle. Neither the size nor the location of $br_c$ is represented accurately in this figure.

dataset (RDD) to distribute the bounding boxes across multiple MapReduce partitions. Spark provides many methods for processing RDDs in parallel. Of particular interest are the sortByKey method \[11\] that may be used to presort an RDD prior to building the $k$-d tree and the filter method \[10\] that may be used to “sweep and partition” an RDD as shown in Figure 5.

Unfortunately, Spark provides neither a method that obtains the median element of a sorted RDD nor a method that splits a sorted RDD about its median element to create “less than” and “greater than” RDDs. However, a proposed extension to Spark that implements the drop method, which deletes a specified number of elements from an RDD, provides insight into how the necessary splitting method may be implemented \[7\]. This splitting method, named splitAt, has been implemented to iterate over the partitions of an RDD in parallel in order to count the elements in each partition \[12\] and thereby obtain the median element of the RDD. This median element is then used to split the RDD to create “less than” and “greater than” RDDs, as shown in Figure 6.

The Spark sortByKey, filter and splitAt methods permit the construction of a distributed $k$-d tree with MapReduce. In contrast to a memory-resident tree, a distributed tree does not use pointers to link each $k$-d node to its children. Instead, the tree is represented as a graph whose nodes are distributed across the partitions of an RDD \[14\]. For example, the root node of the $k$-d tree is represented as the following element of an RDD that contains all the nodes of the $k$-d tree

$$(n_r, (bb_c, n_<, br_<, n_>, br_>))$$

where $n_r$ represents the unique name of the root node of the tree, $bb_c$ represents the bounding box that is stored at the root node of the tree, $n_c$ represents the unique name of the “less than” child of the root node, $br_c$ represents the bounding region that encloses the “less than” subtree, $n_r$ represents the unique name of the “greater than” child of the root node, and $br_r$ represents the bounding region that encloses the “greater than” subtree. This RDD will be named the tree RDD (in the following discussion, the name of a specific RDD will be designated in bold face text).

The parentheses in the above representation of the root node of the $k$-d tree enclose tuples \[16\] and in particular, the outer set of parentheses encloses a two-element tuple or pair whose first element is $n_r$ and whose second element is the tuple $(bb_r, n_<, br_<, n_>, br_>)$. An RDD that comprises pairs is a pair RDD and has the special property that the two elements of each pair function as a (key, value) pair. In the case of the above-described root node of the $k$-d tree, the unique name $n_r$ is the key and the tuple $(bb_r, n_<, br_<, n_>, br_>)$ is the value. The (key, value) properties of a pair RDD permit Spark to retrieve a particular node of the $k$-d tree from the tree RDD using the unique name of that node, analogous to the manner in which a pointer to a particular node of a memory-resident $k$-d tree permits the retrieval of that node from the tree.

4. SEARCHING A DISTRIBUTED TREE WITH MAPREDUCE

Searching a distributed $k$-d is not performed in a recursive, depth-first manner such as is used to search a memory-resident tree. Instead, an iterative, breadth-first search is conducted in parallel by multiple bounding boxes via MapReduce, similar to the manner in which MapReduce executes Dijkstra’s algorithm \[14\]. An overview of the search algorithm is presented below.

![Figure 5: The $y_{min}$-sorted RDD ($y_{min}$) is filtered by the $x_{min}$ of the median element $n_{r0}$ (see Figure 6) of the $x_{min}$-sorted RDD to obtain the “less than” $y_{min}$-sorted RDD ($<$) and the “greater than” $y_{min}$-sorted RDD ($>$).](image-url)
A search of the k-d tree assumes the existence of a search RDD that contains the elements \((n_i, bb_i)\), where \(bb_i\) represents the bounding boxes that will search the tree for intersections and \(n_i\) represents the unique names of the bounding boxes. In preparation for the first iteration of the search algorithm, the search RDD is processed using the map method [10] to create a query RDD whose elements are \((n_i, (n_i, bb_i))\). For each element of this query RDD, the key is \(n_i\) and represents the unique name of the root node of the k-d tree. This key specifies that each element of the search RDD will visit the root of the k-d tree. The query RDD is joined to the tree RDD using the join method [11] that performs an inner join and returns a visit RDD that contains the elements \((n_i, ((n_i, bb_i), (bb_i, n_{<}, br_{<}, n_{>}, br_{>})))\)

The visit RDD represents the fact that each element of the search RDD visits the root of the k-d tree.

The visit RDD is subsequently processed using the flatMapValues method [11] in order to check for intersection between each \(bb_i\) and \(bb_j\). Each intersection creates an element \((n_i, n_j)\) in an intersection RDD that represents the fact that the bounding boxes \(bb_i\) and \(bb_j\) intersect the bounding box \(bb_k\), where \(bb_k\) is a subset of \(bb_i\).

The visit RDD is processed once again using the flatMapValues method in order to check for intersection between each \(bb_i\) and \(br_{<}\) as well as to check for intersection between each \(bb_i\) and \(br_{>}\). Each intersection between \(bb_i\) and \(br_{<}\) creates an element \((n_{<}, n_i, bb_i)\) in the next query RDD. Similarly, each intersection between \(bb_i\) and \(br_{>}\), creates an element \((n_{>}, n_i, bb_i)\) in the next query RDD. The bounding boxes \(bb_k\) and \(bb_{<}\) are the subsets of \(bb_i\) that will visit the “less than” and “greater than” subtrees, respectively, during the next iteration of the search algorithm. The iterative search algorithm terminates when the next visit RDD is empty, as will be the case when no intersections between \(bb_i\) and \(br_{<}\) or \(br_{>}\) are detected.

The intersection RDD from each iteration of the search algorithm is accumulated to a cumulative intersection RDD using the union method [11]. Following the final iteration of the algorithm, the \((n_i, n_j)\) pairs of the cumulative intersection RDD are reorganized using the groupByKey method [11] to create the pairs \([n_{u}, n_{v}, \ldots, n_{w}]\), where for each pair, \([n_{u}, n_{v}, \ldots, n_{w}]\) specifies the list of bounding boxes \([bb_u, bb_v, \ldots, bb_w]\) that are a subset of \(bb_i\) and that intersect \(bb_i\).

5. PERFORMANCE

The k-d tree building and search algorithms have been executed on only a local Spark node, so their performance and scalability for a cluster of Spark workers that process the multiple partitions of an RDD has not yet been measured.

6. CONCLUSION

This article presents new k-d tree building and search algorithms that are a generalization of algorithms that were published previously [3]. In contrast to those prior algorithms that build and search a memory-resident k-d tree, the new algorithms build and search a k-d tree that is represented as a distributed graph.

Source Code

The k-d tree building and search algorithms have been implemented in Scala and execute in parallel via Spark. The source code for this implementation includes the BSD 3-Clause License and is available for download.

7. REFERENCES