The R*-tree: An Efficient and Robust Access Method for Points and Rectangles

Norbert Beckmann, Hans-Peter Kriegel
Ralf Schneider, Bernhard Seeger
Praktische Informatik, Universität Bremen, D-2800 Bremen 33, West Germany

Abstract
The R-tree, one of the most popular access methods for rectangles, is based on the heuristic optimization of the area of the enclosing rectangle in each inner node. By running numerous experiments in a standardized testbed under highly varying data, queries, and operations, we were able to design the R*-tree which incorporates a combined optimization of area, margin, and overlap of each enclosing rectangle in the directory. Using our standardized testbed in an exhaustive performance comparison, it turned out that the R*-tree clearly outperforms the existing R-tree variants: Guttman's linear and quadratic R-tree and Greene's variant of the R-tree. This superiority of the R*-tree holds for different types of queries and operations, such as map overlay, for both rectangles and multidimensional points in all experiments. From a practical point of view, the R*-tree is very attractive because of the following two reasons: 1) it efficiently supports point and spatial data at the same time, and 2) its implementation cost is only slightly higher than that of other R-trees.

1. Introduction
In this paper, we will consider spatial access methods (SAMs) which are based on the approximation of a complex spatial object by the minimum bounding rectangle with the sides of the rectangle parallel to the axes of the data space.

The most important property of this simple approximation is that a complex object is represented by a limited number of bytes. Although a lot of information is lost, minimum bounding rectangles of spatial objects preserve the most essential geometric properties of the object, i.e., the location of the object and the extension of the object in each axis.

In [SK 88] we showed that known SAMs organizing (minimum bounding) rectangles are based on an underlying point access method (PAM) using one of the following three techniques: clipping, transformation, and overlapping regions.

The most popular SAM for storing rectangles is the R-tree [Gut 84]. Following our classification, the R-tree is based on the PAM B+-tree [Knu 73] using the technique of overlapping regions. Thus, the R-tree can be easily implemented, which considerably contributes to its popularity.

The R-tree is based on a heuristic optimization. The optimization criterion which it pursues, is to minimize the area of each enclosing rectangle in the inner nodes. This criterion is taken for granted and not shown to be the best possible. Questions arise such as: Why not minimize the margin or the overlap of such minimum bounding rectangles? Why not optimize storage utilization? Why not optimize all of these criteria at the same time? Could these criteria interact in a negative way? Only an engineering approach will help to find the best possible combination of optimization criteria.

Necessary condition for such an engineering approach is the availability of a standardized testbed which allows us to run large volumes of experiments with highly varying data, queries, and operations. We have implemented such a standardized testbed and used it for performance comparisons particularly of point access methods [KSSS 89].

As the result of our research, we designed a new R-tree variant, the R*-tree, which outperforms the known R-tree variants under all experiments. For many realistic profiles of data and operations, the gain in performance is quite considerable. Additionally to the usual point query,
rectangle intersection and rectangle enclosure query, we have analyzed our new R*_tree for the map overlay operation, also called spatial join, which is one of the most important operations in geographic and environmental database systems.

This paper is organized as follows. In section 2, we introduce the principles of R-trees including their optimization criteria. In section 3 we present the existing R-tree variants of Guttman and Greene. Section 4 describes in detail the design of our new R*_tree. The results of the comparisons of the R*_tree with the other R-tree variants are reported in section 5. Section 6 concludes the paper.

2. Principles of R-trees and possible optimization criteria

An R-tree is a B*-tree like structure which stores multidimensional rectangles as complete objects without clipping them or transforming them to higher dimensional points before.

A non-leaf node contains entries of the form (cp, Rectangle) where cp is the address of a child node in the R-tree and Rectangle is the minimum bounding rectangle of all rectangles which are entries in that child node. A leaf node contains entries of the form (Oud, Rectangle) where Oud refers to a record in the database, describing a spatial object and Rectangle is the enclosing rectangle of that spatial object. Leaf nodes containing entries of the form (dataobject, Rectangle) are also possible. This will not affect the basic structure of the R-tree. In the following we will not consider such leaf nodes.

Let M be the maximum number of entries that will fit in one node and let m be a parameter specifying the minimum number of entries in a node (2 ≤ m ≤ M/2). An R-tree satisfies the following properties:

- The root has at least two children unless it is a leaf.
- Every non-leaf node has between m and M children unless it is the root.
- Every leaf node contains between m and M entries unless it is the root.
- All leaves appear on the same level.

An R-tree (R*_tree) is completely dynamic, insertions and deletions can be intermixed with queries and no periodic global reorganization is required. Obviously, the structure must allow overlapping directory rectangles. Thus it cannot guarantee that only one search path is required for an exact match query. For further information we refer to [Gut84].

We will show in this paper that the overlapping region technique does not imply bad average retrieval performance. Here and in the following, we use the term directory rectangle, which is geometrically the minimum bounding rectangle of the underlying rectangles.

The main problem in R-trees is the following. For an arbitrary set of rectangles, dynamically build up bounding boxes from subsets of between m and M rectangles, in a way that arbitrary retrieval operations with query rectangles of arbitrary size are supported efficiently. The known

parameters of good retrieval performance affect each other in a very complex way, such that it is impossible to optimize one of them without influencing other parameters which may cause a deterioration of the overall performance. Moreover, since the data rectangles may have very different size and shape and the directory rectangles grow and shrink dynamically, the success of methods which will optimize one parameter is very uncertain. Thus a heuristic approach is applied, which is based on many different experiments carried out in a systematic framework.

In this section some of the parameters which are essential for the retrieval performance are considered. Furthermore, interdependencies between different parameters and optimization criteria are analyzed.

(O1) The area covered by a directory rectangle should be minimized, i.e., the area covered by the bounding rectangle but not covered by the enclosed rectangles, the dead space, should be minimized. This will improve performance since decisions which paths have to be traversed, can be taken on higher levels.

(O2) The overlap between directory rectangles should be minimized. This also decreases the number of paths to be traversed.

(O3) The margin of a directory rectangle should be minimized. Here the margin is the sum of the lengths of the edges of a rectangle. Assuming fixed area, the object with the smallest margin is the square. Thus minimizing the margin instead of the area, the directory rectangles will be shaped more quadratic. Essentially queries with large quadratic query rectangles will profit from this optimization. More important, minimizing the margin will basically improve the structure. Since quadratic objects can be packed easier, the bounding boxes of a level will build smaller directory rectangles in the level above. Thus clustering rectangles into bounding boxes with only little variance of the lengths of the edges will reduce the area of directory rectangles.

(O4) Storage utilization should be optimized. Higher storage utilization will generally reduce the query cost as the height of the tree will be kept low. Evidently, query types with large query rectangles are influenced more since the concentration of rectangles in several nodes will have a stronger effect if the number of found keys is high.

Keeping the area and overlap of a directory rectangle small, requires more freedom in the number of rectangles stored in one node. Thus minimizing these parameters will be paid with lower storage utilization. Moreover, when applying (O1) or (O2) more freedom in choosing the shape is necessary. Thus rectangles will be less quadratic. With (O1) the overlap between directory rectangles may be affected in a positive way since the covering of the data space is reduced. As for every geometric optimization, minimizing the margins will also lead to reduced storage utilization. However, since more quadratic directory rectangles support
packing better, it will be easier to maintain high storage utilization. Obviously, the performance for queries with sufficiently large query rectangles will be affected more by the storage utilization than by the parameters of (O1)-(O3).

3. R-tree Variants

The R-tree is a dynamic structure. Thus all approaches of optimizing the retrieval performance have to be applied during the insertion of a new data rectangle. The insertion algorithm calls two more algorithms in which the crucial decisions for good retrieval performance are made. The first is the algorithm ChooseSubtree. Beginning in the root, descending to a leaf, it finds on every level the most suitable subtree to accommodate the new entry. The second is the algorithm Split. It is called, if ChooseSubtree ends in a node filled with the maximum number of entries. Split should distribute M+1 rectangles into two nodes in the most appropriate manner.

In the following, the ChooseSubtree- and Split-algorithms, suggested in available R-tree variants are analyzed and discussed. We will first consider the original R-tree as proposed by Guttman in [Gut 84].

Algorithm ChooseSubtree

CS1 Set N to be the root
CS2 If N is a leaf, return N
else
Choose the entry in N whose rectangle needs least area enlargement to include the new data. Resolve ties by choosing the entry with the rectangle of smallest area
end
CS3 Set N to be the childnode pointed to by the childpointer of the chosen entry and repeat from CS2

Obviously, the method of optimization is to minimize the area covered by a directory rectangle, i.e. (O1). This may also reduce the overlap and the CPU cost will be relatively low.

Guttman discusses split-algorithms with exponential, quadratic and linear cost with respect to the number of entries of a node. All of them are designed to minimize the area covered by the two rectangles resulting from the split. The exponential split finds the area with the global minimum, but the CPU cost is too high. The others try to find approximations. In his experiments, Guttman obtains nearly the same retrieval performance for the linear as for the quadratic version. We implemented the R-tree in both variants. However in our tests with different distributions, different overlap, variable numbers of data-entries and different combinations of M and m, the quadratic R-tree yielded much better performance than the linear version (see also section 5). Thus we will only discuss the quadratic algorithm in detail.

Algorithm QuadraticSplit

[Divide a set of M+1 entries into two groups]
QS1 Invoke PickSeeds to choose two entries to be the first entries of the groups
QS2 Repeat
DistributeEntry
until
all entries are distributed or
one of the two groups has M-m+1 entries
QS3 If entries remain, assign them to the other group such that it has the minimum number m

Algorithm PickSeeds

PS1 For each pair of entries E1 and E2, compose a rectangle R including E1 rectangle and E2 rectangle
Calculate d = area(R) - area(E1 rectangle) - area(E2 rectangle)
PS2 Choose the pair with the largest d

Algorithm DistributeEntry

DE1 Invoke PickNext to choose the next entry to be assigned
DE2 Add it to the group whose covering rectangle will have to be enlarged least to accommodate it. Resolve ties by adding the entry to the group with the smallest area, then to the one with the fewer entries, then to either

Algorithm PickNext

PN1 For each entry E not yet in a group, calculate d1 = the area increase required in the covering rectangle of Group 1 to include E Rectangle
Calculate d2 analogously for Group 2
PN2 Choose the entry with the maximum difference between d1 and d2

The algorithm PickSeeds finds the two rectangles which would waste the largest area put in one group. In this sense the two rectangles are the most distant ones. It is important to mention that the seeds will tend to be small too, if the rectangles to be distributed are of very different size (and) or the overlap between them is high. The algorithm DistributeEntry assigns the remaining entries by the criterion of minimum area. PickNext chooses the entry with the best area-goodness-value in every situation.

If this algorithm starts with small seeds, problems may occur. If in d-1 of the d axes a far away rectangle has nearly the same coordinates as one of the seeds, it will be distributed first. Indeed, the area and the area enlargement of the created needle-like bounding rectangle will be very small, but the distance is very large. This may initiate a very bad split. Moreover, the algorithm tends to prefer the bounding rectangle, created from the first assignment of a rectangle to one seed. Since it was enlarged, it will be larger than others. Thus it needs less area enlargement to include the next entry. It will be enlarged again, and so on. Another problem is, that if one group has reached the maximum number of entries M-m+1, all remaining entries are assigned to the other group without considering geometric properties. Figure 1 (see section 4.3) gives an example showing all
these problems. The result is either a split with much overlap (fig. 1c) or a split with uneven distribution of the entries reducing the storage utilization (fig. 1b).

We tested the quadratic split of our R-tree implementation varying the minimum number of entries $m =$ 20%, 30%, 35%, 40% and 45% relatively to $M$ and obtained the best retrieval performance with $m$ set to 40%.

On the occasion of comparing the R-tree with other structures storing rectangles, Greene proposed the following alternative split-algorithm [Gre 89]. To determine the appropriate path to insert a new entry she uses Guttman's original ChooseSubtree-algorithm.

Algorithm Greene's-Split

[Divide a set of $M+1$ entries into two groups]

GS1 Invoke ChooseAxis to determine the axis perpendicular to which the split is to be performed

GS2 Invoke Distribute

Algorithm ChooseAxis

CA1 Invoke PickSeeds (see p. 5) to find the two most distant rectangles of the current node

CA2 For each axis record the separation of the two seeds

CA3 Normalize the separations by dividing them by the length of the nodes enclosing rectangle along the appropriate axis

CA4 Return the axis with the greatest normalized separation

Algorithm Distribute

D1 Sort the entries by the low value of their rectangles along the chosen axis

D2 Assign the first $(M+1) \div 2$ entries to one group, the last $(M+1) \div 2$ entries to the other

D3 If $M+1$ is odd, then assign the remaining entry to the group whose enclosing rectangle will be increased least by its addition

Almost the only geometric criterion used in Greene's split algorithm is the choice of the split axis. Although choosing a suitable split axis is important, our investigations show that more geometric optimization criteria have to be applied to considerably improve the retrieval performance of the R-tree. In spite of a well clustering, in some situations Greene's split method cannot find the "right" axis and thus a very bad split may result. Figure 2b (see p. 12) depicts such a situation.

4. The R*-tree

4.1 Algorithm ChooseSubtree

To solve the problem of choosing an appropriate insertion path, previous R-tree versions take only the area parameter into consideration. In our investigations, we tested the parameters area, margin and overlap in different combinations, where the overlap of an entry is defined as follows:

$$\text{overlap(E)} = \sum_{k=1}^{p} \text{area}(E_k \cap E_p) , 1 < k < p$$

Let $E_1, \ldots, E_p$ be the entries in the current node. Then

The version with the best retrieval performance is described in the following algorithm.

Algorithm ChooseSubtree

CS1 Set $N$ to be the root

CS2 If $N$ is a leaf,

return $N$

else

if the child pointers in $N$ point to leaves [determine the minimum overlap cost],

choose the entry in $N$ whose rectangle needs least overlap enlargement to include the new data rectangle. Resolve ties by choosing the entry whose rectangle needs least area enlargement, then

the entry with the rectangle of smallest area

if the child pointers in $N$ do not point to leaves [determine the minimum area cost],

choose the entry in $N$ whose rectangle needs least area enlargement to include the new data rectangle. Resolve ties by choosing the entry with the rectangle of smallest area

end

CS3 Set $N$ to be the childnode pointed to by the childpointer of the chosen entry and repeat from CS2

For choosing the best non-leaf node, alternative methods did not outperform Guttman's original algorithm. For the leaf nodes, minimizing the overlap performed slightly better.

In this version, the cpu cost of determining the overlap is quadratic in the number of entries, because for each entry the overlap with all other entries of the node has to be calculated. However, for large node sizes we can reduce the number of entries for which the calculation has to be done, since for very distant rectangles the probability to yield the minimum overlap is very small. Thus, in order to reduce the cpu cost, this part of the algorithm might be modified as follows:

[calculate the nearly minimum overlap cost]

Sort the rectangles in $N$ in increasing order of their area enlargement needed to include the new data rectangle

Let $A$ be the group of the first $p$ entries. From the entries in $A$, considering all entries in $N$, choose the entry whose rectangle needs least overlap enlargement. Resolve ties as described above.

For two dimensions we found that with $p$ set to 32 there is nearly no reduction of retrieval performance to state. For more than two dimensions further tests have to be done. Nevertheless, the cpu cost remains higher than the original version of ChooseSubtree, but the number of disc accesses.
is reduced for the exact match query preceding each insertion and is reduced for the ChooseSubtree algorithm itself.

The tests showed that the ChooseSubtree optimization improves the retrieval performance particularly in the following situation: Queries with small query rectangles on data files with non-uniformly distributed small rectangles or points.

In the other cases the performance of Guttman's algorithm was similar to this one. Thus, principally an improvement of robustness can be stated.

4.2 Split of the R*-tree

The R*-tree uses the following method to find good splits. Along each axis, the entries are first sorted by the lower value, then sorted by the upper value of their rectangles. For each sort, M-2m+2 distributions of the M+1 entries into two groups are determined, where the k-th distribution (k = 1, ..., (M-2m+2)) is described as follows: The first group contains the first (m-1)+k entries, the second group contains the remaining entries.

For each distribution, goodness values are determined. Depending on these goodness values, the final distribution of the entries is determined. Three different goodness values and different approaches of using them in different combinations are tested experimentally.

Algorithm ChooseSplitIndex

CS1 Along the chosen split axis, choose the distribution with the minimum overlap-value.

Resolve ties by choosing the distribution with minimum area-value.

The split algorithm is tested with m = 20%, 30%, 40% and 45% of the maximum number of entries M. As experiments with several values of M have shown, m = 40% yields the best performance. Additionally, we varied m over the life cycle of one and the same R*-tree in order to correlate the storage utilization with geometric parameters. However, even the following method did result in worse retrieval performance. Compute a split using m = 30% of M, then compute a split using m = 40% of M, then compute a split using m = 40% of M. If split(m1) yields overlap and split(m2) does not, take split(m1), otherwise take split(m2).

Concerning the cost of the split algorithm of the R*-tree, we will mention the following facts: For each axis (dimension) the entries have to be sorted two times which requires O(M log(M)) time. As an experimental cost analysis has shown, this needs about half of the cost of the split. The remaining split cost is spent as follows: For each axis, the margin of 2*(2*(M-2m+2)) rectangles and the overlap of 2*(M-2m+2) distributions have to be calculated.

4.3 Forced Reinsert

Both, R-tree and R*-tree are nondeterministic in allocating the entries onto the nodes. If different sequences of insertions will build up different trees, for this reason the R-tree suffers from its old entries. Data rectangles inserted during the early growth of the structure may have introduced directory rectangles, which are not suitable to guarantee a good retrieval performance in the current situation. A very local reorganization of the directory rectangles is performed during a split. But this is rather poor and therefore it is desirable to have a more powerful and less local instrument to reorganize the structure.

The discussed problem would be maintained or even worsened, if underfilled nodes, resulting from deletion of records would be merged under the old parent. Thus the known approach of treating underfilled nodes in an R-tree is to delete the node and to reinsert the orphaned entries on the corresponding level [Gut 84]. This way the ChooseSubtree algorithm has a new chance of distributing entries into different nodes.

Since it was to be expected, that the deletion and reinsertion of old data rectangles would improve the retrieval performance, we made the following simple experiment with the linear R-tree. Insert 20000 uniformly distributed rectangles. Delete the first 10000 rectangles and insert them again. The result was a performance improvement of 20% up to 50%(!) depending on the types of the queries. Therefore to delete randomly half of the data and then to insert it again seems to be a very simple way of tuning existing R-tree data files. But this is a static situation, and for nearly static data files the pack algorithm [RL 85] is a more sophisticated approach.

To achieve dynamic reorganizations, the R*-tree forces entries to be reinserted during the insertion routine. The
following algorithm is based on the ability of the insert routine to insert entries on every level of the tree as already required by the deletion algorithm [Gut 84]. Except for the overflow treatment, it is the same as described originally by Guttman and therefore it is only sketched here.

**Algorithm InsertData**

1. **ID1** Invoke Insert starting with the leaf level as a parameter, to insert a new data rectangle

**Algorithm Insert**

1. **I1** Invoke ChooseSubtree, with the level as a parameter, to find an appropriate node N, in which to place the new entry E
2. **I2** If N has less than M entries, accommodate E in N
   - If N has M entries, invoke OverflowTreatment with the level of N as a parameter [for reinsertion or split]
3. **I3** If OverflowTreatment was called and a split was performed, propagate OverflowTreatment upwards if necessary
   - If OverflowTreatment caused a split of the root, create a new root
4. **I4** Adjust all covering rectangles in the insertion path such that they are minimum bounding boxes enclosing their children rectangles

**Algorithm OverflowTreatment**

1. **OT1** If the level is not the root level and this is the first call of OverflowTreatment in the given level during the insertion of one data rectangle, then
   - invoke ReInsert
   - else
   - invoke Split
2. **OT2**

**Algorithm ReInsert**

1. **RI1** For all M+1 entries of a node N, compute the distance between the centers of their rectangles and the center of the bounding rectangle of N
2. **RI2** Sort the entries in decreasing order of their distances computed in RI1
3. **RI3** Remove the first p entries from N and adjust the bounding rectangle of N
4. **RI4** In the sort, defined in RI2, starting with the maximum distance (= far reinsert) or minimum distance (= close reinsert), invoke Insert to reinsert the entries

If a new data rectangle is inserted, each first overflow treatment on each level will be a reinsertion of p entries. This may cause a split in the node which caused the overflow if all entries are reinserted in the same location. Otherwise splits may occur in one or more other nodes, but in many situations splits are completely prevented. The parameter p can be varied independently for leaf nodes and non-leaf nodes as part of performance tuning, and different values were tested experimentally. The experiments have shown that p = 30% of M for leaf nodes as well as for non-leaf nodes yields the best performance. Furthermore, for all data files and query files close reinsert outperforms far reinsert. Close reinsert prefers the node which included the entries before, and this is intended, because its enclosing rectangle was reduced in size. Thus this node has lower probability to be selected by ChooseSubtree again.

Summarizing, we can say:

- Forced reinsert changes entries between neighboring nodes and thus decreases the overlap
- As a side effect, storage utilization is improved
- Due to more restructuring, less splits occur
- Since the outer rectangles of a node are reinserted, the shape of the directory rectangles will be more quadratic

As discussed before, this is a desirable property.

Obviously, the CPU cost will be higher now since the insertion routine is called more often. This is alleviated, because less splits have to be performed. The experiments show that the average number of disc accesses for insertions increases only about 4% (and remains the lowest of all R-tree variants), if Forced Reinsert is applied to the R*-tree. This is particularly due to the structure improving properties of the insertion algorithm.

![Diagram of R*-tree](image)

327
5. Performance Comparison

5.1 Experimental Setup and Results of the Experiments

We ran the performance comparison on SUN workstations under UNIX using Modula-2 implementations of the different R-tree variants and our R*-tree. Analogously to our performance comparison of PAM's and SAM's in [KSSS 89] we keep the last accessed path of the trees in main memory. If orphaned entries occur from insertions or deletions, they are stored in main memory additionally to the path.

In order to keep the performance comparison manageable, we have chosen the page size for data and directory pages to be 1024 bytes which is at the lower end of realistic page sizes. Using smaller page sizes, we obtain similar performance results as for much larger file sizes. From the chosen page size the maximum number of entries in directory pages is 56. According to our standardized testbed we have restricted the maximum number of entries in a data page to 50.

As candidates of our performance comparison we selected the R-tree with quadratic split algorithm (abreviation qua Gut), Greene's variant of the R-tree (Greene) and our R*-tree where the parameters of the different structures are set to the best values as described in the previous sections. Additionally, we tested the most popular R-tree implementation, the variant with the linear split algorithm (lin Gut) The popularity of the linear R-tree is due to the statement in the original paper [Gut84] that no essential performance gain resulted from the quadratic version vs. the linear version. For the linear R-tree we found n=20% (of M) to be the variant with the best performance.

To compare the performance of the four structures we selected six data files containing about 100,000 2-dimensional rectangle. Each rectangle is assumed to be in the unit cube [0,1]². In the following each data file is described by the distribution of the centers of the rectangles and by the triple (n, p_area, n_area) Here n denotes the number of rectangles, p_area is the mean value of the area of a rectangle and n_area = σ_area / p_area is the normalized variance where σ_area denotes the variance of the areas of the rectangles. Obviously, the parameter n_area increases independently of the distribution the more the areas of the rectangles differ from the mean value and the average overlap is simply obtained by n* p_area.

(F1) "Uniform" The centers of the rectangles follow a 2-dimensional independent uniform distribution (n = 100,000, p_area = 0001, n_area = 9505)

(F2) "Cluster" The centers follow a distribution with 640 clusters, each cluster contains about 1600 objects (n = 99,968, p_area = 00002, n_area = 1538)

(F3) "Parcel" First we decompose the unit square into 100,000 disjoint rectangles Then we expand the area of each rectangle by the factor 2.5 (n = 100,000, p_area = 00002504, n_area = 303458)

(F4) "Real-data" These rectangles are the minimum bounding rectangles of elevation lines from real cartography data (n = 120,576, p_area = 0000926, n_area = 1504)

(F5) "Gaussian" The centers follow a 2-dimensional independent Gaussian distribution (n = 100,000, p_area = 00008, n_area = 89875)

(F6) "Mixed-Uniform" The centers of the rectangles follow a 2-dimensional independent uniform distribution. First we take 99,000 small rectangles with p_area = 0000101 Then we add 1,000 large rectangles with p_area = 001 Finally these two data files are merged to one (n = 100,000, p_area = 00002, n_area = 6778)

For each of the files (F1) - (F6) we generated queries of the following three types:
- rectangle intersection query Given a rectangle S, find all rectangles R in the file with R \ S ≠ ∅.
- point query Given a point P, find all rectangles R in the file with P ∈ R.
- rectangle enclosure query Given a rectangle S, find all rectangles R in the file with R ⊆ S.

For each of these files (F1) - (F6) we performed 400 rectangle intersection queries where the ratio of the x-extension to the y-extension uniformly varies from 0.25 to 2.25 and the centers of the query rectangles themselves are uniformly distributed in the unit cube. In the following, we consider four query files (Q1) - (Q4) of 100 rectangle intersection queries each. The area of the query rectangles of each query file (Q1) - (Q4) varies from 1%, 0.1%, 0.01% to 0.001% relatively to the area of the data space. For the rectangle enclosure query we consider two query files (Q5) and (Q6) where the corresponding rectangles are the same as in the query files (Q3) and (Q4), respectively. Additionally, we analyzed a query file (Q7) of 1,000 point queries where the query points are uniformly distributed.

For each query file (Q1) - (Q7) we measured the average number of disc accesses per query. In the performance comparison we use the R*-tree as a measuring stick for the other access methods, i.e., we standardize the number of page accesses for the queries of the R*-tree to 100%. Thus we can observe the performance of the R-tree variants relative to the 100% performance of the R*-tree.

To analyze the performance for building up the different R-tree variants we measured the parameters insert and store. Here insert denotes the average number of disc accesses per...
insertion and stor denotes the storage utilization after completely building up the files. In the following table we present the results of our experiments depending on the different distributions (data files). For the R*-tree we also depict "# accesses", the average number of disk accesses per query.

Additionally to the conventional queries like point query, intersection query and enclosure query we have considered the operation spatial join usually used in applications like map overlay. We have defined the spatial join over two rectangle files as the set of all pairs of rectangles where the one rectangle from file1 intersects the other rectangle from file2.

For the spatial join operation we performed the following experiments:

(SJ1) file1 "Parcel"-distribution with 1000 rectangles randomly selected from file (F3)
file2 data file (F4)

(SJ2) file1 "Parcel"-distribution with 7500 rectangles randomly selected from data file (F3)
file2 7,536 rectangles generated from elevation lines
(n = 7,536, h_area = 0.0148, v_area = 1.5)

(SJ3) file1 "Parcel"-distribution with 20,000 rectangles randomly selected from data file (F3)
file2 file1

For these experiments we measured the number of disk accesses per operation. The normalized results are presented in the following table.

### 5.2 Interpretation of the Results

In table 1 for the parameters stor and insert we computed the unweighted average over all six distributions (data files). The parameter spatial join denotes the average over the three spatial join operations (SJ1) - (SJ3). For the average query performance we present the parameter query average which is averaged over all seven query files for each distribution and then averaged over all six distributions.

### Table 1

<table>
<thead>
<tr>
<th>Query</th>
<th>Spatial Join</th>
<th>Stor</th>
<th>Insert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin Gut</td>
<td>227.5</td>
<td>261.2</td>
<td>62.7</td>
</tr>
<tr>
<td>qua Gut</td>
<td>130.0</td>
<td>147.3</td>
<td>68.1</td>
</tr>
<tr>
<td>Greene</td>
<td>142.3</td>
<td>171.3</td>
<td>69.7</td>
</tr>
<tr>
<td>R*-tree</td>
<td>100.0</td>
<td>100.0</td>
<td>73.0</td>
</tr>
</tbody>
</table>

For the spatial join operation we performed the following experiments:

(SJ1) file1 "Parcel"-distribution with 1000 rectangles randomly selected from file (F3)
file2 data file (F4)

(SJ2) file1 "Parcel"-distribution with 7500 rectangles randomly selected from data file (F3)
file2 7,536 rectangles generated from elevation lines
(n = 7,536, h_area = 0.0148, v_area = 1.5)

(SJ3) file1 "Parcel"-distribution with 20,000 rectangles randomly selected from data file (F3)
file2 file1
The loss of information in the parameter query average is even less in table 2 where the parameter is displayed separately for each data file (F1) - (F6) as an average over all seven query files and in table 3 where the parameter query average is depicted separately for each query (Q1) - (Q7) as an average over all six data files.

<table>
<thead>
<tr>
<th></th>
<th>gaussian</th>
<th>cluster</th>
<th>mix uni</th>
<th>parcell</th>
<th>real data</th>
<th>uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin. Gut</td>
<td>164 3</td>
<td>216 0</td>
<td>308 1</td>
<td>347 2</td>
<td>227 2</td>
<td>206 6</td>
</tr>
<tr>
<td>qua Gut</td>
<td>112 9</td>
<td>153 9</td>
<td>121 8</td>
<td>128 1</td>
<td>144 5</td>
<td>121 1</td>
</tr>
<tr>
<td>Greene</td>
<td>123 1</td>
<td>147 1</td>
<td>115 5</td>
<td>192 4</td>
<td>144 2</td>
<td>134 8</td>
</tr>
<tr>
<td>R*-tree</td>
<td>100 0</td>
<td>100 0</td>
<td>100 0</td>
<td>100 0</td>
<td>100 0</td>
<td>100 0</td>
</tr>
</tbody>
</table>

Table 2: unweighted average over all seven types of queries depending on the distribution.

<table>
<thead>
<tr>
<th></th>
<th>point</th>
<th>insertion</th>
<th>enclosure</th>
<th>stor</th>
<th>insert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin. Gut</td>
<td>251 9</td>
<td>243 2</td>
<td>231 1</td>
<td>199 8</td>
<td>151 7</td>
</tr>
<tr>
<td>qua.Gut</td>
<td>132 4</td>
<td>132 4</td>
<td>132 4</td>
<td>126 4</td>
<td>117 6</td>
</tr>
<tr>
<td>Greene</td>
<td>140 7</td>
<td>140 9</td>
<td>140 0</td>
<td>137 7</td>
<td>121 3</td>
</tr>
<tr>
<td>R*-tree</td>
<td>100 0</td>
<td>100 0</td>
<td>100 0</td>
<td>100 0</td>
<td>100 0</td>
</tr>
</tbody>
</table>

Table 3: unweighted average over all six distributions depending on the query type.

First of all, the R*-tree clearly outperforms the R-tree variants in all experiments. Moreover, the most popular variant, the linear R-tree, performs essentially worse than all other R-trees. The following remarks emphasize the superiority of the R*-tree in comparison to the R-trees:

- The R*-tree is the most robust method which is underlined by the fact that for every query file and every data file less disk accesses are required than by any other variants. To say it in other words, there is no experiment where the R*-tree is not the winner.

- The gain in efficiency of the R*-tree for smaller query rectangles is higher than for larger query rectangles, because storage utilization gets more important for larger query rectangles. This emphasizes the goodness of the order preservation of the R*-tree (i.e., rectangles close to each other are more likely stored together in one page).

- The maximum performance gain of the R*-tree taken over all query and data files is in comparison to the linear R-tree about 400% (i.e., it takes four times as long as the R*-tree!), to Greene's R-tree about 200% and to the quadratic R-tree 180%.

- As expected, the R*-tree has the best storage utilization.

- Surprisingly in spite of using the concept of Forced Reinsert, the average insertion cost is not increased, but essentially decreased regarding the R-tree variants.

- The average performance gain for the spatial join operation is higher than for the other queries. The quadratic R-tree, Greene's R-tree and the linear R-tree require 147%, 171% and 261% of the disk accesses of the R*-tree, respectively, averaged over all spatial join operations.

5.3 The R*-tree: an efficient point access method

An important requirement for a spatial access method is to handle both spatial objects and point objects efficiently. Points can be considered as degenerated rectangles, and in most applications, rectangles are very small relatively to the data space. If a SAM is also an efficient PAM, this would underpin the robustness of the SAM. Moreover, in many applications, it is desirable to support additionally to the bounding rectangle of an object at least an atomic key with one access method.

Therefore, we ran the different R-tree variants and our R*-tree against a benchmark proposed and used for point access methods. The reader interested in the details of this benchmark is referred to [KSSS 89]. In this paper, let us mention that the benchmark incorporates seven data files of highly correlated 2-dimensional points. Each data file contains about 100,000 records. For each data file, we considered five query files each of them containing 20 queries. The first query files contain range queries specified by square-shaped rectangles of size 0.1%, 1% and 10% relatively to the data space. The other two query files contain partial match queries where in one only the x-value and in the other only the y-value is specified, respectively.

Similar to the previous section, we measured the storage utilization (stor), the average insertion cost (insert) and the average query cost averaged over all query and data files. The results are presented in table 4 where we included the 2-level grid file ([NHS84], [Hin85]), a very popular point access method.

<table>
<thead>
<tr>
<th></th>
<th>query average</th>
<th>stor</th>
<th>insert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin.Gut</td>
<td>233.1</td>
<td>64.1</td>
<td>7.34</td>
</tr>
<tr>
<td>qua.Gut</td>
<td>175.9</td>
<td>67.8</td>
<td>4.51</td>
</tr>
<tr>
<td>Greene</td>
<td>237.8</td>
<td>69.0</td>
<td>5.20</td>
</tr>
<tr>
<td>GRID</td>
<td>127.6</td>
<td>58.3</td>
<td>2.56</td>
</tr>
<tr>
<td>R*-tree</td>
<td>100.0</td>
<td>70.9</td>
<td>3.36</td>
</tr>
</tbody>
</table>

Table 4: unweighted average over all seven distributions.
We were positively surprised by our results. The performance gain of the R*-tree over the R-tree variants is considerably higher for points than for rectangles. In particular, Greene's R-tree is very inefficient for point data. It requires even more accesses than the linear R-tree and 138% more than the R*-tree, whereas the quadratic R-tree requires 75% more disc accesses than the R*-tree. Nevertheless, we had expected that PAMs like the 2-level grid file would perform better than the R*-tree. However, the over all average of the 2-level grid file performs essentially worse than the R*-tree for point data. An advantage of the grid file is the low average insertion cost. In that sense, it might be more suitable in an insertion-intensive application. Let us mention that the complexity of the algorithms of the R*-trees is rather low in comparison to highly tuned PAMs.

6 Conclusions

The experimental comparison pointed out that the R*-tree proposed in this paper can efficiently be used as an access method in database systems organizing both, multidimensional points and spatial data. As demonstrated in an extensive performance comparison with rectangle data, the R*-tree clearly outperforms Greene's R-tree, the quadratic R-tree and the popular linear R-tree in all experiments. Moreover, for point data, the gain in performance of the R*-tree over the other variants is increased. Additionally, the R*-tree performs essentially better than the 2-level grid file for point data.

The new concepts incorporated in the R*-tree are based on the reduction of the area, margin, and overlap of the directory rectangles. Since all three values are reduced, the R*-tree is very robust against ugly data distributions. Furthermore, due to the fact of the concept of Forced Reinsert, splits can be prevented. The structure is reorganized dynamically and storage utilization is higher than for other R-tree variants. The average insertion cost of the R*-tree is lower than for the well known R-trees. Although the R*-tree outperforms its competitors, the cost for the implementation of the R*-tree is only slightly higher than for the other R-trees.

In our future work, we will investigate whether the fan out can be increased by prefixes or by using the grid approximation as proposed in [SK 90]. Moreover, we are generalizing the R*-tree to handle polygons efficiently.

References:


[Hin 85] K. Hinrichs 'The grid file system implementation and case studies for applications', Dissertation No 7734, Eidgenössische Technische Hochschule (ETH), Zurich, 1985


[SK 88] B. Seeger, H.P. Kriegel 'Design and implementation of spatial access methods', Proc 14th Int Conf on Very Large Databases, 360-371, 1988