CSC6210: Advanced multidimensional search — Fall 2009
CSE department, Chinese University of Hong Kong

Lecture 14: Cost models of R-trees

By Yufei Tao (taoyf@cse.cuhk.edu.hk) — Last revision: Nov 30, 2009

Forewords

The R-tree does not have interesting worst-case guarantees on search performance. It is possible to design a dataset and a query [1] such that the query cost of an R-tree is as expensive as a brute-force scan. However, it is well-known that the R-tree is fairly efficient for practical data and queries. This phenomenon motivated people to look for a way to understand this structure in a more formal manner. An important outcome of the relevant studies is the so-called *cost model*, which is an analytical formula that estimates the query cost, based on the parameters of the query and the underlying dataset. Besides revealing the characteristics of the R-tree, such cost models are especially useful for query optimization in a database system (where the optimizer relies on accurate estimates of query cost to decide a good execution plan).

Accurate cost models are known for many types of queries including range searching, nearest neighbor search, etc. This lecture will focus on range searching because its cost model is the foundation of the models of other problems.

1 A tree-dependent model

For simplicity, we consider a 2d space Ω, where the x- and y-dimensions have a domain length of 1. Without loss of generality, we focus on the queries whose search regions q have length l_x (l_y) on the x- (y-) dimension. Furthermore, we assume:

- $A1$: $l_x \ll 1$ and $l_y \ll 1$, and
- $A2$: the underlying dataset P has a large cardinality N.

Cost model analysis usually targets a specific query distribution. We will consider that the centroid of a query is distributed uniformly in Ω. Our discussion can be generalized to higher dimensionalities and other query distributions.

Let us first deal with the following problem. Let u be a node in an R-tree. Denote by a_u (b_u) the extent length of $MBR(u)$ on the x- (y-) dimension. The goal is to calculate the probability $Pr(u)$ that a query needs to access u, or equivalently, the probability that the query rectangle q intersects $MBR(u)$. When $l_x = l_y = 0$, namely, q degenerates into a point, $Pr(u)$ is obviously $a_u b_u$. For general l_x and l_y, let us extend the x- and y-extents of $MBR(u)$ by l_x and l_y respectively, as shown in Figure 1 (note that the extension is symmetric to the centroid of $MBR(u)$). Refer to the resulting rectangle r as the *extended MBR* of u. It is easy to see that $MBR(u)$ intersects q *if and only if* r contains the centroid of q. Since the centroid is uniformly distributed in the data domain, we have

$$Pr(u) = (l_x + a_u)(l_y + b_u) \tag{1}$$

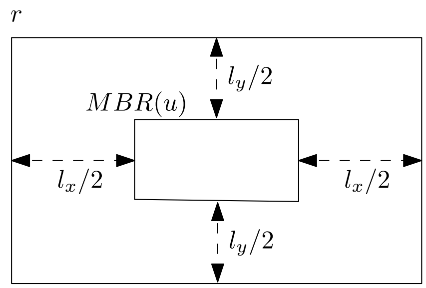

Figure 1: Extending $MBR(u)$ by l_x and l_y on the x- and y-dimensions respectively

when r completely lies in Ω. If part of r falls out of Ω, the probability is smaller. However, under the assumptions $A1$ and $A2$, there are very few nodes whose extended MBRs fall partially out of Ω. Hence, we can ignore such a case to keep our formulae simple.

In an earlier lecture, we learned that it is important to minimize the perimeter of $MBR(u)$ in constructing an R-tree, if queries are square-like, namely $l_x \approx l_y$ (which is true in practice). We are now ready to explain this in a more rigorous manner. When $l_x = l_y = l$, Equation 2 can be written as:

$$Pr(u) = l^2 + l(a_u + b_u) + a_u b_u. \tag{2}$$

It is thus clear that $Pr(u)$ depends both on $a_u b_u$ (which is the area of $MBR(u)$) and $a_u + b_u$ (which is half of the perimeter of $MBR(u)$). Hence, ideally both the area and perimeter must be minimized. If only one metric needs to be chosen, perimeter is better because a rectangle with a small perimeter usually has a small area, but not the vice versa.

Equipped with Equation 2, we are able to derive a cost model to compute the expectation $\mathbf{E}[cost(q)]$ of the cost of a query (whose search region has extent lengths l_x and l_y) on an R-tree $\mathcal{T}$. Specifically, for each node $u \in \mathcal{T}$, define a random variable X_u that equals 1 if q intersects $MBR(u)$ or 0 otherwise. Then:

$$\begin{aligned}\mathbf{E}[cost(q)] &= \mathbf{E}\left[\sum_{u \in \mathcal{T}} X_u\right] \\ &= \sum_{u \in \mathcal{T}} \mathbf{E}[X_u] \\ &= \sum_{u \in \mathcal{T}} Pr(u). \end{aligned} \tag{3}$$

2 A model for uniform data

The model of Equation 3 is difficult to evaluate because it requires the knowledge of the extent lengths of all the MBRs in an R-tree. In practice, a useful model should have as few parameters as possible. Such a model typically makes extra assumptions about the data distribution. Next, we

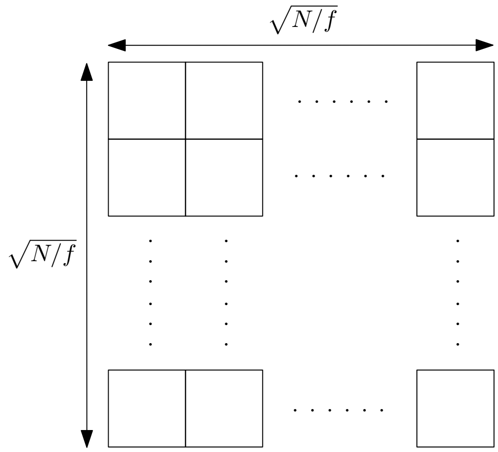

Figure 2: All leaf MBRs are aligned into a grid.

will derive a model for unform distribution, namely, the N points in our dataset P are uniformly distributed in Ω.

Besides N, the model takes another parameter f, which is the *average fanout* of all nodes in an R-tree. The value of f can be easily maintained along with the insertions and deletions on the tree (this is left as an exercise). In practice, leaf and non-leaf elements may require different amounts of storage, so the values of f may differ for leaf and non-leaf nodes. For simplicity, we ignore this complication and assume that all nodes in the tree have fanout f, except the root. The extension to account for different fanouts is straightforward. Also note that since each node (except the root) must have $\Omega(B)$ elements, it follows that $f = \Theta(B)$ for all the non-root levels.

We will carry out our analysis in a level-by-level manner, starting with the leaf nodes at level 0. Let $cost_0(q)$ denote the number of leaf nodes that need to be visited by a range query q (with extent lengths a_x and b_y on the two dimensions, respectively). Following Equation 3, we have the following about the expectation $\mathbf{E}[cost_0(q)]$:

$$\mathbf{E}[cost_0(q)] = \sum_{\text{leaf } u \in \mathcal{T}} Pr(u). \tag{4}$$

Next we will simplify the above equation. The rational is that since the data distribution is uniform, the MBRs of all the leaf nodes should be highly similar, thus casting the hope that $Pr(u)$ can be represented with the same equation for all u. It is easy to see that the number of leaf nodes equals N/f. As the x- and y-dimensions are symmetric, when assumption $A2$ holds, the leaf MBRs of a good R-tree should be fairly *regular* in the following senses:

- each MBR is a square;
- all MBRs form a $\sqrt{N/f} \times \sqrt{N/f}$ grid as in Figure 2, such that no two MBRs intersect each other, and the union of all MBRs cover the whole Ω.

Therefore, the MBR of each leaf u has the same extent length $a_u = b_u = 1/\sqrt{N/f}$ on the x- and y-dimensions. Thus, Equation 2 becomes:

$$Pr(u) = \left(\sqrt{\frac{f}{N}} + l_x\right)\left(\sqrt{\frac{f}{N}} + l_y\right) \tag{5}$$

leading to

$$\mathbf{E}[cost_0(q)] = \frac{N}{f}\left(\sqrt{\frac{f}{N}} + l_x\right)\left(\sqrt{\frac{f}{N}} + l_y\right). \tag{6}$$

The analysis can be easily extended to the i-th ($i > 0$) level. Specifically, since this level has N/f^{i+1} nodes, the MBR of each node is a square with extent length $1/\sqrt{N/f^{i+1}}$. Hence, the expectation $\mathbf{E}[cost_i(q)]$ of the number of level-i nodes accessed by a query equals:

$$\mathbf{E}[cost_i(q)] = \frac{N}{f^{i+1}}\left(\sqrt{\frac{f^{i+1}}{N}} + l_x\right)\left(\sqrt{\frac{f^{i+1}}{N}} + l_y\right). \tag{7}$$

3 Remarks

Equation 2 initially appeared independently in [2, 3]. The tree-dependent model was first discussed in [3]. The simplified model was proposed by [4], where the authors also discuss how to extend the model to predict the query cost on non-uniform data.

References

[1] L. Arge, M. de Berg, H. J. Haverkort, and K. Yi. The priority R-tree: A practically efficient and worst-case optimal R-tree. In *Proc. of ACM Management of Data (SIGMOD)*, pages 347–358, 2004.

[2] I. Kamel and C. Faloutsos. On packing R-trees. In *Proc. of Conference on Information and Knowledge Management (CIKM)*, pages 490–499, 1993.

[3] B.-U. Pagel, H.-W. Six, H. Toben, and P. Widmayer. Towards an analysis of range query performance in spatial data structures. In *Proc. of ACM Symposium on Principles of Database Systems (PODS)*, pages 214–221, 1993.

[4] Y. Theodoridis and T. K. Sellis. A model for the prediction of R-tree performance. In *Proc. of ACM Symposium on Principles of Database Systems (PODS)*, pages 161–171, 1996.