A-007

R木を用いた非矩形領域探索アルゴリズムの評価

田村　壮†　能登谷淳一‡　草苅良至‡　笠井雅夫‡

Evaluation of Non Rectangular Search Algorithm using R-Tree

Takeshi Tamura　Yoshiyuki Kusakari　Junichi Notoya　Masao Kasai

1．まえがき

データベースに画像データや空間的なデータなどに代表される多次元データを格納し，利用する局面が増大している[1]．そのような多次元データ集合を格納する各種の木構造索引が提案されているが，その一つにR木[2]があり広く利用されている．R木は格納すべき図形データ集合をいくつかの図形からなる部分集合に分割し，それぞれの部分集合中の図形を全て内包する最小の矩形（MBR）を再帰的に作成することにより構築される．R木の性能は，「どの図形データを一つのMBRに格納するか」という空間の分割戦略に依存する．R木の分割戦略に対する問い合わせ処理性能の評価，および，格納すべき図形データの分布と問い合わせ分布に従った最適分割戦略の提案は，従来主に矩形領域問い合わせに対して行われてきた[3]．しかし，R木は矩形領域問い合わせだけでなく，非矩形領域問合せに対しても適用可能である．例えば，さまざまな応用分野に出現する問い合わせパターンとして，非矩形領域の一種である超平面により分割される図形を用いた問い合わせがある．超平面分割領域を問い合わせ図形とする問い合わせは，複数の属性値の一次結合を判定基準とする問い合わせと考える事も可能である．

本研究では，R木を用いた超平面分割領域問い合わせのパラメータ分布に対する性能評価を行う．本稿では，予備実験として超平面分割領域問い合わせ図形による問合せ性能について評価を行った結果を示す．

2．超平面分割領域問い合わせ

R木は多次元空間中の図形データから，与えられた問い合わせ図形と重なりを持つ，もしくは問い合わせ図形に真に内包される図形を選択するために利用される空間索引木である．多くの場合，R木は多次元超立方体（矩形）を問い合わせ図形とする検索に用いられるが，一般的には矩形以外の形状を持つ図形を問い合わせ図形とする問い合わせにも利用可能である．応用上，用途の広い非矩形問い合わせ図形として，超平面分割領域がある．n次元超平面分割領域はデータの各属性の一次結合を利用して次式により表される．

$$Q(a_1,\cdots,a_n,C) = \left\{(x_1,\cdots,x_n) \mid \sum_{i=1}^{n} a_i x_i \le C, x_i \in [0,1]\right\}$$

ここで，超平面$\sum_{i=1}^{n} a_i x_i = C$は超平面分割の境界であり，$a_i$を係数としたデータ$x_i$の一次結合が$C$以下となる空間が問い合わせ図形である．したがって，Cとの大小関係により，格納された図形データが超平面分割領域内であるかを調べることが可能である．超平面分割領域問い合わせの応用例として，学力テストの合否判定などが挙げられる．各教科の点数をx_i，合格となる合計点数をCとし，各教科の配点の重みをa_iとすれば良い．テストの合格判定には，x_iの一次結合とCの大小関係を比較する．

R木では，MBRを利用してデータを再帰的に分割しており，葉ノードには図形データとそのMBRが格納され，非葉ノードには子ノードのMBRを全て包含するMBRが格納される．一般に図形により与えられる問い合わせに対しては，「各ノードのMBRが問合せ図形と重なりを持つかを調べ，重なりを持てばそのノードを根とする部分木へ探索を進める．探索が葉ノードに到達した場合，問い合わせ図形と葉ノードに格納された図形を比較して，重なりがあるかを調べる．」というアルゴリズムによって問い合わせを処理可能である．特に，超平面分割領域問い合わせにおいては以下のようにアルゴリズム中の図形の重なり判定を簡略化可能である．なお，これ以降は問い合わせ$Q(a_1,\cdots,a_n,C)$において係数a_iを省略して，$Q(C)$と書く．（図1）

①$Q(C)$の領域外にMBRが存在する
MBRの最小点が$Q(C)$の外部（MBR1）
②$Q(C)$の領域内にMBRが存在する
MBRの最大点が$Q(C)$の内部（MBR2）
③$Q(C)$とMBRが接する
MBRの最小点が$Q(C)$の境界上（MBR3）
④$Q(C)$とMBRが交差する
MBRの最小点が$Q(C)$の内部（MBR4）

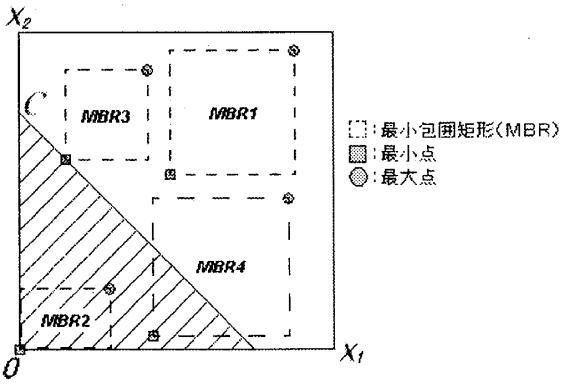

図1．超平面分割領域とMBRの関係

† 秋田県立大学　システム科学技術研究科
‡ 秋田県立大学　システム科学技術学部

表 1. R 木分割戦略のパラメータ

Parameter	Value
Dimension	2
Index Capacity	20
Leaf Capacity	20
Fill Factor	0.7
Near Minimum Overlap Factor	32
Split Distribution Factor	0.4
Reinsert Factor	0.3

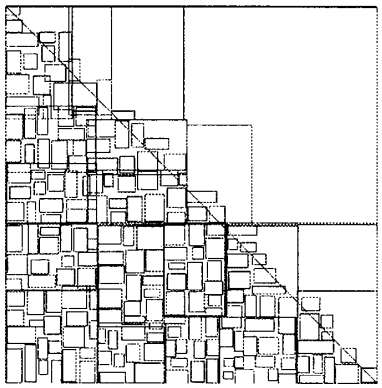

図 2. $Q(1.0)$による訪問ノード

3. 評価実験

本研究では，上記の超平面分割領域問合せを，R木を用いて処理する際のコストを実験により評価した．R木に格納するデータとして，各成分 x_i を区間 $[0,1]$ で一様分布に従って発生させた 2 次元の点データの集合を用い，R木に格納する点データ数が 1000,5000,10000 の 3 つの場合について実験を行った．また，一次結合の係数 a_1, a_2 はいずれも 1 とした．なお，R木の分割戦略としては広く用いられているR*を用い，分割パラメータは表 1 の値を用いた．

パラメータ C を 0.2 刻みで 0~2 までの範囲で固定して生成された超平面分割領域による問い合わせを行い，処理の際に訪問した合計ノード数を調べた．図 2 に格納データ数が 5000 として，$C=1.0$ とする探索を行った際に，訪問された全てのノードの MBR を示す．図 2 より問い合わせ図形に包含される MBR を持つノードと，分割超平面と交差する MBR を持つノードが訪問されていることがわかる．格納データ数の増加に伴う R 木の領域分割細分化が問い合わせ処理コストに及ぼす影響を図 3 に示す．図 3 は格納データ数が 1000, 5000, 10000 の場合の各々に対し問い合わせを行った場合の，訪れたノード数と R 木全体のノード数の比をグラフ化したものである．グラフの横軸は超平面分割のパラメータ C の値を表す．$C=0.0\sim1.0$ ではグラフは下に凸な形状を示し，$C=1.0\sim2.0$ では上に凸な形状を示した．図 4 は 10000 点データを探索したときの，全訪問ノード数と訪問葉ノード数のグラフである．全訪問ノード数と訪問葉ノード数の差が訪問非葉ノード数である．ノード訪問コストのほとんどが，葉ノードへの訪問によるものであることが図 4 からわかる．

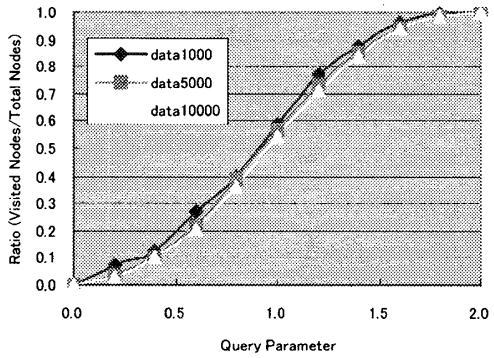

図 3. $Q(C)$ $(C\in[0,2])$ による訪問ノード数比

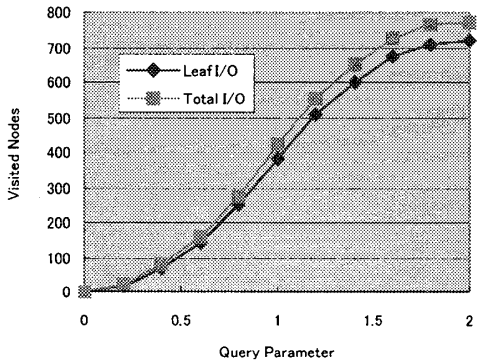

図 4. 葉ノードと全ノードの訪問数

4. まとめ

本稿では，R木に格納された一様分布で与えられる点データに対して，超平面分割領域を問い合わせ図形とする問い合わせを処理した場合のノード訪問コストの評価を行った．評価の結果，一様分布で与えられる点データ集合に対して標準的なR*の分割戦略を使用した場合，問い合わせ処理コストは，格納されるデータ数に関係なく，問い合わせ図形の体積に依存することがわかった．これにより，問い合わせ分布に適した分割戦略を考えることによって，訪れるノード数を減少させ，問い合わせ処理時間の短縮が可能と考えられる．

今後の課題として，非矩形問合せのパラメータ分布に対する，各種分割戦略を利用した R 木による問合せ処理コストの評価がある．また，R 木に格納されるデータの分布を変化させた際の非矩形問い合わせ処理コストについても調査が必要である．さらに，超平面分割領域問い合わせの処理コストを最小化する分割戦略の提案も課題である．

参考文献

[1] V. Gaede, O. Gunther, “Multidimensional Access Methods,” ACM Computing Surveys, Vol.30, No.2, pp.170-231, 1998.
[2] A. Guttman, “R-Trees: A Dynamic Index Structure for Spatial Searching,” ACM SIGMOD, pp.47-57, 1984.
[3] 大森 匡, 佐藤 龍生, 星 守, “問い合わせ分布を考慮した R 木における領域分割方式,” 電子情報通信学会論文誌, VOL.J86-D-I, NO.10, pp.746-761, 2003.