3N-6

OLAP スカイライン・キューブに対する範囲検索の実現方式と評価

宇佐美孝侑† 佐藤秀樹†

†大同大学大学院情報学研究科

1. はじめに

OLAP データキューブに対して従来の集約演算(SUM, MAXIMUM など)に代わりスカイライン演算[1]を適用したスカイライン・キューブ[2]が提案されている。[2]ではグループ化問合せ(図1(a)参照)に効率的に解答するために、事前に問合せ結果に対する実体化ビュー(materialized view)を作成する。次元属性の集合 D に対して $(2^{|D|}-1)$個の実体化ビューが考えられ、その数は|D|に対して指数関数的に増加する。そのため全ての実体化ビューを計算することは困難である。

意思決定支援のためには、スカイライン・キューブに対して次元属性の任意の範囲条件での問合せ(図 1(b)参照)が有用であり、このために本論文では R-木[3]に基づく実現方式を検討する。

次元属性値の範囲は問合せ時に指定されるため、R-木によりデータウェアハウスに蓄積されたデータのセル(次元属性値の組)に対するスカイライン演算結果への索引を管理し、範囲問合せの解の効率的な計算方法を検討する。R-木ではセルを次元属性値の組から成る多次元空間上の点に対応付け、これを R-木で管理する。

以降では、まず我々が提案した R-木の構築・探索について述べる。そして R-木の性能を実験し有用性の評価を行い最後にまとめを行う。

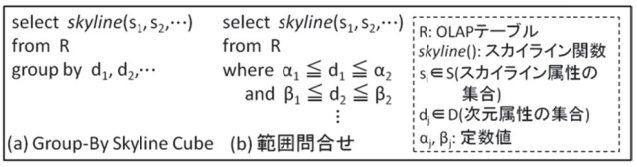


図 1. SQL 文による問合せ例

2. R-木の構築

一般に R-木のアルゴリズムには、以下の 2 つの問題点がある。

① 図形データの動的な追加・削除に備えたアルゴリズムであり、ノードの空間効率が悪い。

② 異なるノードの最小外接矩形(MBR)が重なることがある。重なった領域に対する探索には、該当するノードを全て探索することになる。

一般にデータウェアハウスに対するデータの蓄積はある一定の期間ごとに行われるため、R-木もその期間に合わせて構築することができる。その間データの変更が考える必要がないので、①の備えは不要でありノードの空間効率を高めることができる。ノードに空間的に近い点を格納させるため、空間充填曲線に従った順に多次元空間の点データをノードに記憶させる。②に関しては、MBR の重なりが少なくなるような点データのノードへの割り当てが必要であるが、これは今後の検討課題とする。

図 2(a)のように Z 曲線の順に従って葉ノードに点データを詰めていき、図 2(b)で示されるように葉ノードから根ノードへボトムアップで木を構築する。葉ノードには、(点の座標値、点に対応するセル内のデータ集合のスカイライン演算結果へのポインタ(S ポインタ))を繰り返す。中間ノードには、(下位ノードへのポインタ、下位ノードに対する MBR、MBR 内のスカイライン演算結果へのポインタ(S ポインタ))を繰り返す。

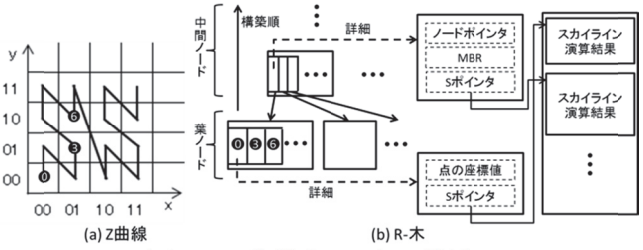


図 2. Z 曲線と R-木の構造

3. R-木の探索

問合せの範囲条件に対応する矩形領域（範囲条件で指定されない次元属性は、下限値から上限値までをその範囲とする）内の点データに対するスカイライン演算結果を R-木により探索する。探索は根ノードから開始される。①矩形領域が中間ノードに格納されている MBR を包含する場合、S ポインタが指す演算結果を取り出す。②MBR と矩形領域が重なる場合、該当するノードを再帰的に探索する。葉ノードでは、矩形領域

Implementation and Evaluation of Range Queries Processing toward OLAP Skyline Cubes
Takayuki USAMI† , Hideki SATO†
†Graduate school , Daido University

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.

が包含する点データの S ポインタが指す演算結果を取り出す。矩形領域内のスカイライン演算結果は、上記で求められた全ての演算結果を使って求める。

4. 実験評価

範囲問合せの実現には[2]の実体化ビューを使った方法も考えられる。そのため R-木と[2]の実体化ビュー(R-木の容量と同等になるように事前に計算した)を使った実験を行う。

実験の実データとして NBA 選手の成績データ(データは 21961 件、次元属性数は 6、スカイライン属性数は 17)と大リーグの選手の成績データ(データは 94259 件、次元属性数は 9、スカイライン属性数は 11)を用いる。合成データとして次元属性数を 10、スカイライン属性数を 10 とするデータを 10 万件生成する。各属性では異なる値の数を 100 として、Zipf 分布に従って値を生成する。

実験では次元属性集合からランダムに 3 属性を選び、各属性の値の範囲を 10%～100%まで 10%刻みで変化させた問合せを考える。各々の問合せを 100 回生成し、I/O 回数の平均とスカイライン演算のための比較回数の平均を求める。

図 3 は I/O 回数の結果である。[2]は実体化ビューを取り出すので I/O 回数はほぼ一定であるのに対して、R-木では範囲が狭いと狭い範囲の探索だけ済むため、R-木が優位である。

図 4 は比較回数の結果である。実データでは R-木のほうが比較回数は少なく優位である。合成データのスカイライン属性間では独立に値を生成した。そのため、合成データではスカイライン演算を行っても演算を行ったデータ集合がほぼそのままスカイライン演算結果となり、R-木では MBR の範囲ごとにスカイライン演算を事前に行う利点が生かせないことから比較回数にほぼ差がない結果となった。

5. おわりに

本研究では、スカイライン・キューブに対する範囲問合せの解を効率的に求めるための R-木の構築・探索について述べ、提案した R-木の性能を実験し有用性の評価を行った。実験の結果より I/O 回数、比較回数ともに R-木は優位であり、R-木に基づく我々の方式は有用性があるといえる。しかし、合成データでは比較回数に差がほとんどない結果となっているので、今後の課題としてスカイライン演算アルゴリズムの改良が挙げられる。

参考文献

[1] S.Borzsonyi, et al: "The skyline operator" Proc. ICDE, pp.421-430, 2001
[2] Yiu, et al: "Measuring the Sky: On Computing Data Cubes via Skylining the Measures" IEEE TKDE Vol.24, No.3, pp492-505, 2012
[3] A.Guttman: "R-trees: A dynamic index structure for spatial searching" Proc. ACM SIGMOD Conf., pp47-57, 1984

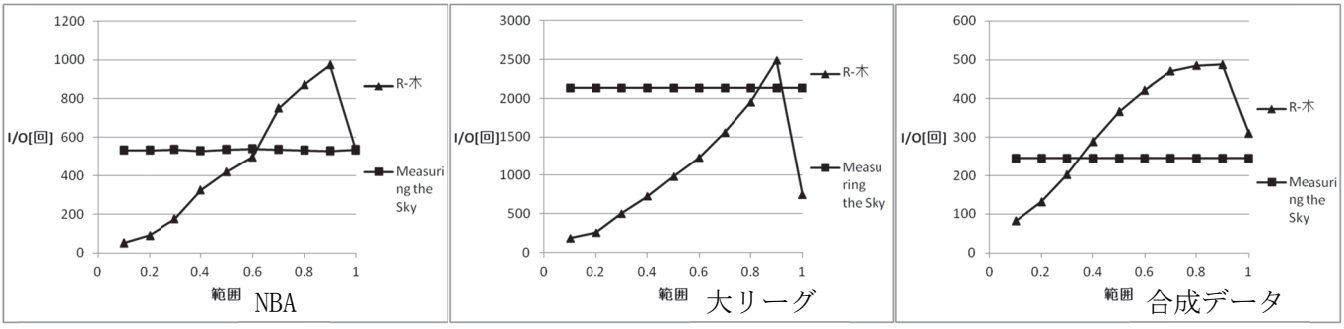


図 3. 各データによる I/O 回数の結果

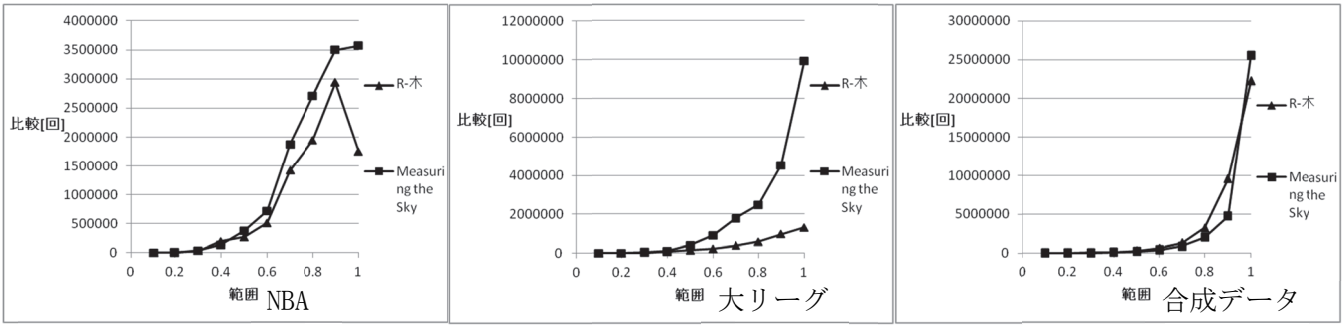


図 4. 各データによる比較回数の結果

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.