LD-5 問い合わせ分布に適応した多次元ファイル編成法 GR 木のアーカイブ環境への適用

高橋 秀和　大森 匡　星 守　高塚 寛
電気通信大学 大学院 情報システム学研究科

1　はじめに

近年のデータベースに求められる機能の 1 つに、高エネルギー物理実験 [4] や天体観測の分野 [2] における科学技術データの長大な系列 (sequence) を管理する機能がある。このようなデータは追加のみの大規模で多属性の時系列データである。従来より、このデータ管理のために、ディスクとテープロボットから構成された階層型の大容量記憶システム (Mass Storage System, MSS) が用いられている。MSS によって管理されているデータは、管理対象である大容量の実体（データオブジェクト）とその内容を記述するメタデータの 2 つから構成される。SDSS は、メタデータの管理は関係データベースでも可能であることを示した。従って、現在の問題点の 1 つは、メタデータの管理を通して、観測データ集合という大規模データの実体を、ストレージシステム上でいかにして効率良く管理するかというところにある。

一般に、観測データを表すメタデータは多属性 (仮に n 個とする) のデータであり、使用者からの問い合わせは、k 個 ($\leq n$) の属性について制約を与える。$k = n$ のような全属性上の問い合わせから、$k < n$ となるような k 個の属性上の範囲問い合わせ（**部分空間問い合わせ**）まであり得る。著者らは、このような部分空間上の範囲問い合わせが混在した状況を想定した多属性ファイル編成法の 1 つ GR 木 [1, 5] を提案している。本稿では、MSS 上で GR 木を用いてメタデータの管理を行った時のシミュレーション結果について述べる。

2　モデルと関連研究

MSS のモデル　MSS においてメタデータ管理を行ったモデルを図 1 に示す。このモデルは、フロントエンド側のディスクとバックエンド側のテープロボットから成る。ディスク側は、メタデータ管理を行う。そのため，ディスク側はメタデータに対しインデクス付けをする。このインデクスの葉ノードには、対象データに対応するメタデータの集合が格納される。すなわち，このモデルでは，メタデータには適切なクラスタリングがなされ，インデクスの葉ノードに対応する観測データ集合が 1 つのストレージブロック B にまとめられて、バックエンドのストレージに格納される。B はテープ 1 つに対応する。システムに利用者から問い合わせが与えられると、その問い合わせに該当するメタデータが検索され，次に，そのメタデータに対応する観測データを有すストレージブロックがフロントエンドへ転送される。

一般に、ディスク側では、一度アクセスしたストレージブロックをキャッシュする。さらに、CERN[4] のように、キャッシュから溢れたデータ集合に対し適切な選択処理やクラスタリングを行い、新しいテープブロックへとそれらを一時的に退避する例もある。

メタデータとインデクス　メタデータは一般にスタースキーマでモデル記述される [2]。本稿では、対象となる系列データを多属性の関係データ 1 つに絞る。従来，MSS におけるメタデータ管理は、Hilbert-R 木 [3] や B 木の組み合わせ [2], LDAP と述語キャッシュ[4] により行われる。これらは、全て多次元データの管理手法である。

GR 木　著者らは、R 木の変種として GR 木を提案している [1]。GR 木は，予め与えられた相異なる部分空間問い合わせ（問い合わせ幅や、問い合わせが行われる部分空間が異なる）の混合分布に基づいて空間を分割する R 木の 1 種である。R 木が部分空間問い合わせに対し著しく性能が低いのに比べ、GR 木は良好な性能を示している [1, 5]。

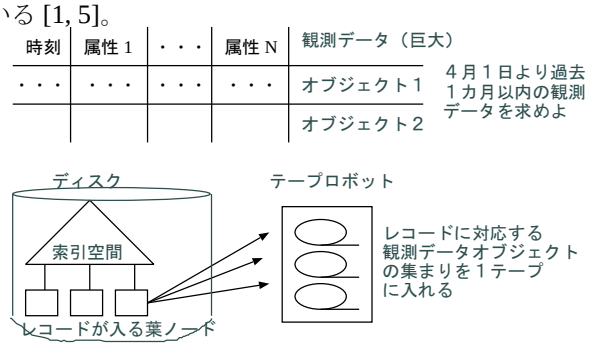


図 1: MSS のモデル

3　GR 木のアーカイブ環境への適用

以下、図 1 のモデルにおいて GR 木を用いてメタデータ管理を行った場合のシミュレーション結果を示す。実装はフロントエンドのディスクとバックエンドのテープロボットにそれぞれ相当する PC 2 台を用いた。テープ 1 つに対応して長大ファイル 1 つを用意し、必要に応じてファイル転送を行った。

3.1　追加型記憶システムの場合

単純な追加型 MSS では、データはフロントエンド (PC1) から追加され、時刻印の古いデータから順にバックエンド (PC2) へ移動する。シミュレーションは、全データのうち時刻印の新しいものから順に 50%が PC1 にまだ留まっており、残りの古い時刻印のデータが PC2 に移動した状態で開始した。この他に、フロントエンド側では、PC2 から転送して来たデータを保持するキャッシュ領域を持つ。キャッシュ領域の大きさは、データファイルの総数の 20%とした。システムは，各データファイルに対してその記憶場所 (PC1 か PC2) 及びアクセス頻度 (ヒート) についての情報を保持しており、途中、キャッシュが一杯になった時には、ヒートの低いファイルを破棄 (既にテープロボットにある場合) かバックエンドへ移動させる。

データは、区間 [0,1] の実数からなる 4 次元のデータ

10^5 点とし、点数 40%のクラスタを 2 つ、残り 20%を一様発生のノイズデータとした。問い合わせ分布は、空間を構成する 4 属性から選んだ 2 属性と 1 属性について，各々，幅 0.05 の範囲問い合わせが等確率で生じるとした。この問い合わせ分布に従って 100 個の問い合わせを系列として順に実行した。メタデータのインデクス管理方式としては、(a) GR 木を用いる場合と、(b) 通常の R 木 (VamSplit-R 木) を用いる場合を試した。

図 2-(a) が、問い合わせ系列が進むにつれて、データファイルがテープロボットとディスクとの間で転送される回数である。図を見ると、問い合わせ系列が進行するにつれて、GR 木と用いた場合の方が大幅にステージング回数が少ないことがわかる。問い合わせ 10^3 個の系列が終了した時点でのステージング回数の総数では、GR 木の場合が他方の 1/3 となった。

3.2 再編成型記憶システムの場合

次に、GR 木を用いたデータ編成を CERN の再編成型 MSS モデル [4] に適用した。システムは図 1 に従うが、フロントエンドのキャッシュ領域からデータが溢れた時の動作は文献 [4] の記述に従う。すなわち、システムは、(i) 追い出すべきデータ集合をヒートに基づいて決め、次に、(ii) 対応するメタデータ集合を、その時点までに観測された問い合わせ系列の発生分布 D に基づいて新しい GR 木に再編成する。そして、(iii) 葉ノードに応じて観測データオブジェクトを集め、テープ (に対応する長大データファイル) とし、バックエンドの PC へ転送する。

手順 (ii) の実装に際しては、[4] の手法では適切な選択や射影処理を伴うと考えられるが、今回は行っていない。また、問い合わせの処理も 3.1 節のそれとは異る。すなわち、問い合わせ Q に対し、元々ある GR 木 $R1$ を検索してアクセスすべきオブジェクト集合 O を決める。システムは、各オブジェクト $o \in O$ について、その現在の位置 (PC1 か PC2 の GR 木 $R1$ か $R2$ か) とアクセス頻度を B 木ファイルで別に保持している。結果的に、$R1$ は問い合わせに応じてメタデータをクラスタリングしテープデータのステージングを制御する役目しか担っていない。今回の実験では、アクセスすべきオブジェクト o について、$R2$ を $R1$ よりも優先して用いることとした。PC1 のキャッシュサイズは全データファイルの 20%とした。

実験は、1 次元，2 次元，4 次元の範囲問い合わせ（発生確率 1:2:1, いずれも幅 0.05) が混在して発生する状況で問い合わせ 10^3 個を順に発生させ実行した。問い合わせ系列に対する PC1 と PC2 の間のファイル転送回数を図 2-(b) に示す。GR 木を用いた場合は、VAMSplit-R 木の場合よりもファイル転送回数の総数で 1/2 程度になっている。

4 まとめ

本稿では、大規模な観測データオブジェクトの系列を大容量記憶システム (MSS) で管理する問題を論じ、メタデータをフロントエンドでクラスタリングすることで結果的にバックエンド側のテープロボットの動作効率を上げる手法を扱った。そして、観測データが多属性 (n 個の属性) であること、および、ユーザから発行される問い合わせは一般に $k < n$ 個の属性で行われることを考慮して、このような部分空間問い合わせの混在環境に適した多次元ファイル編成法 GR 木を用いたクラスタリングの効果をシミュレーションで示した。代表的な MSS モデルである追加型や再編成型で試した結果，通常の R 木をそのまま使うよりも、GR 木によりメタデータの管理を行うと、テープデータのファイル転送回数が 1/3 - 1/2 に減少した。

文献 [4] によれば、物理実験データ解析のデータアクセスパターンでは、最初に問い合わせがある制約で一度行われ、その後、その制約を満たすデータ集合の中でさらに細かく条件をつけながら多数の問い合わせが生じる。GR 木自身はこのような問い合わせ分布に向いたインデクスであり、この状況下での再編成型 MSS 上の評価を試行中である。

参考文献

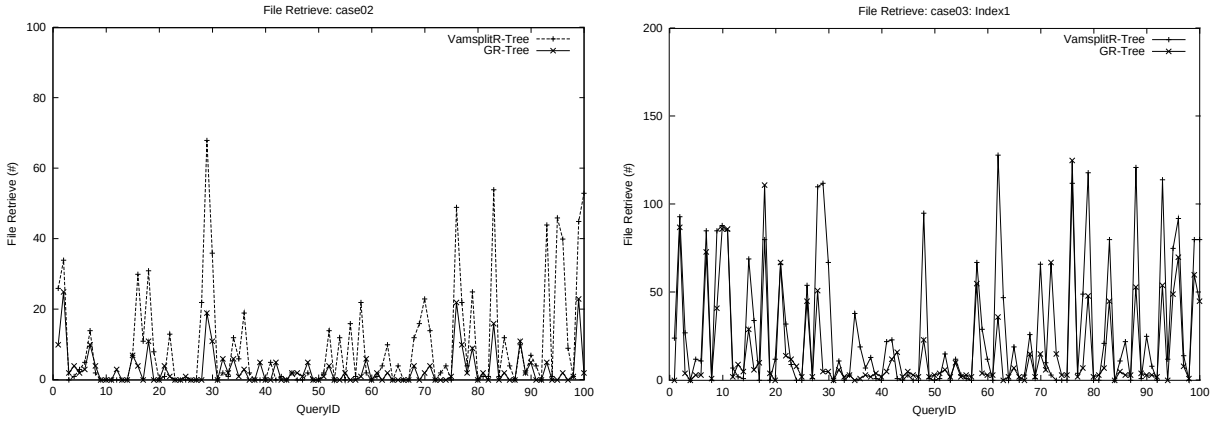
[1] 佐藤, 大森, 星, 問い合わせ分布を考慮した R 木における領域分割法. DEWS2001, 7A-5, 電子情報通信学会, 2001.

[2] Szalay et al.. *The SDSS skyserver*. ACM SIGMOD2002.

[3] Baynon et al., DataCutter: a middleware for filtering very large scientific datasets on archival storage systems, IEEE MSS2000, pp.119-134, 2000. (the whole project is reachable from http://www.cs.umd.edu/projects/adr.)

[4] K.Holtman et al.. *Towards Mass Storage Systems with Object Granularity*. IEEE Mass Storage Systems (MSS) 2000.

[5] 高橋他，問い合わせ分布に適応した多次元ファイル編成法 GR 木の評価，情処全大 64 回 6ZA-02, 2002.



(a) 追加型記憶システム　　(b) 再編成型記憶システム

図 2: 問い合わせ系列 (0,1,...,99) に対するファイル転送回数