ビットマップ索引を利用したHDF-EOSデータに対する問合せ処理機能の開発

Development of query processor for HDF-EOS files by using bitmap index structure.

渡辺 知恵美 [1]
Chiemi Watanabe[1]

[1] お茶大・理・情報
[1] none

近年，地球観測のための測定機器の高機能化, コンピュータの高性能化による計算能力の著しい発達などにより, シミュレーションや分析に使用するデータはますます肥大化し，大量のデータを個人で管理するためのデータベースシステムに対する注目が高まっている．しかし現在，科学者は未だデータベースシステムの利用を避ける傾向にある．その理由として，導入および利用のための学習コストがかかる，一度データベース化するとファイルのように気軽に他の計算機や他人にデータを移動させることが困難になる、などの理由が考えられている．

一方，地球観測分野では近年データ形式を統一化しようという意識が高まっている．HDF (Hierarchical Data Format) は地球観測データの統一ファイルフォーマットとしてNASA等ですでに採用されている形式であり，CおよびFORTRAN用のAPIを提供している．階層的に複数の種類の異なる複数のデータセットを格納し，データセット毎にメタデータを定義できるなど，データ配布だけでなく管理まで想定した豊富な機能を備えている．また，HDFはデータセット自体をBLOB(Binary Large OBject) として扱うが，地球観測データ用拡張フォーマットであるHDF-EOS形式によって，POINT, SWATH,GRIDという3種類のデータ表現形式が定義されており，それらのデータ構造を生かした部分データセットへの柔軟なアクセスが可能である．

しかしながら現段階ではHDF-EOSデータにおいてデータ構造を生かした高速かつ高機能な検索関数は提供されていない．特に，巨大なデータセットから部分セットを取り出す場合，データ値による検索を行うための索引生成機能は十分に需要が高い．

そこで我々は地球大気科学者の個人利用のデータ管理システムとして，従来のRDBMSを利用するのではなく,HDF-EOSファイルに対して個人利用のデータ管理に最低限必要な機能を備えるというアプローチで簡易データ管理システムを提供する．

その初期段階として，我々はデータモデルのための基礎的な検索機能である選択関数の実装を行った．現在，HDF-EOS用のAPIでは部分格子の抽出は提供されているものの，例えば「180 ＜ temperature ＜ 220」というようなデータ値を条件とした単純な選択関数は提供されていない．このような条件でデータを絞り込む場合は,（部分）格子を一度メモリ上に取り出した後，配列の各要素を順次確認しなければならず，巨大なデータから条件にあった部分格子だけを取り出すのに相当な時間と労力が必要であった．そこで我々はHDF-EOSファイルに索引構造を埋め込み，単純な条件検索を高速に行うためのAPI拡張ライブラリを提供した．索引構造にはビットマップ索引を利用しWAH圧縮することで，データの高圧縮と圧縮したままの索引参照およびbit演算を可能にした．また索引自体が巨大化した場合は，仮想ファイルレイヤー機能を利用し，索引だけを外部ファイルに分離することも可能である．評価実験ではSWATHデータを用い，緯度経度によるデータ検索を行い、シーケンシャルスキャンと比較したところ平均で2〜5倍程度の検索速度を得ることができた．