大規模な科学データベースからの情報検索・配信の高効率化

Study on efficient data retrieval and distribution from enormous science database

大林 信[1]; 高田 良宏[1]; 笠原 禎也[1]; 後藤 由貴[1]
Makoto Obayashi[1]; Yoshihiro Takata[1]; Yoshiya Kasahara[1]; Yoshitaka Goto[1]

[1] 金沢大
[1] Kanazawa Univ.

１.はじめに

現在、研究者が利用しやすいように、衛星データベースなどの大規模な科学データベースを各地で構築しようという動きがある。その際、問題となるのは大容量のデータをどのように効率よく検索・配信するかである。特定のデータを人手で探すことは多くの時間を費やすこととなり、研究の妨げとなる。本研究では、当グループで保管・管理されている 10 数テラバイト（TB）にのぼる、あけぼの衛星で観測された 20kHz 以下の低周波電波に関するデータを利用して、実際に大規模なデータベースを構築した後、効率良く情報を検索・配信する方法を研究することで、TB オーダーの自然科学データの効率的なデータベース化とデータ配信法について提案する。

２.研究成果

あけぼの衛星の軌道周回番号、軌道情報、観測機器の状態に関するデータをデータベースに登録するプログラムを作成し、試験的なデータベースを構築した。軌道に関するデータは 14 年分で約 1600 万件あり、それらをそのままデータベースに登録するとデータの量が約 2GB になるため、座標以外の 7 項目については実測値を 2 バイト整数に変換してから登録する方法を採用した。このことによって、データ量を約 1.5GB にすることができた。また、観測機器の状態に関するデータは、14 年分で約 9 億件に上る。一般に機器の状態の情報は離散的で、0/1 のビット表記が可能であることを利用して、データ量の削減を実現した。登録するデータの量は約 6GB となった。これらのデータベースに対して WEB ブラウザ上で検索条件を入力し、検索結果を表示させるシステムを作成した。また、データの配信方法として SOAP(Simple Object Access Protocol)を利用したものを試験的に作成した。SOAP とは XML ベースのデータを HTML など一般的なプロトコルで送信するもので、RPC(Remote Procedure Call)のメッセージング機構を実現するものである。

３.評価と今後の課題

作成したシステムを評価したところ検索に平均で約 1 分以上かかるため、検索時間の高速化について検討した。検索に用いる SQL 言語の記述法や、PHP が担当するインターフェース部の最適化にも取り組んだ。さらに、観測機器の状態の検索においては登録したビット列のパターンがある程度決まっていることを利用して、新たに検索用のテーブルを作成した。このテーブルには、通年の日付とその日に存在するビット列のパターンが登録されている。これを用いることによって、以前まで 1 年分の検索を行うには最大 6300 万件の検索をしなければならなかったが、現在は最大で約 20 万件の検索を行えばよくなり、検索時間を短縮できた。これらのことによって検索時間を平均約 2 秒に改善することができた。しかし、軌道情報と観測機器の状態の 2 つを同時に検索した場合に検索時間が平均で約 1 分かかるという問題が残っており、現在、原因と対策を検討中である。

データの配信に用いた SOAP については、SOAP メッセージによって、世界標準時(UT)からその時刻の軌道データを表示させるテストプログラムを動作させ、基本的な動作の確認を行った。しかし、結果の件数が多くなると、結果を正常に受け取ることができないなどの問題が残されており、今後の検討が必要である。