膨大な天体データを効率的に検索する方法の考察と実装

田中 昌宏* 白崎 裕治* 大石 雅寿* 水本 好彦* 石原 康秀† 堤 純平†
町田 吉弘† 中本 啓之‡ 小林 佑介‡ 坂本 道人‡
*国立天文台 †富士通 (株) ‡(株) セック

1 はじめに

世界中の天文台によって観測されたデータを有効活用するため、Virtual Observatory (VO) と呼ばれる取り組みが各国で進められている。VO では、データ配信の統一的な規格を策定し、その規格に基づく天文データ配信システムの開発が行われている。我々は、Japanese Virtual Observatory (JVO) というプロジェクトにおいて天文検索言語やシステム開発などを進めてきた [1]。これまでに VO では、統一的なメタデータ配信システム、天体データの検索プロトコル、天体データを格納するフォーマットなどの仕様が策定され、JVO を含む各国の VO の間の相互接続も実現した。今後、VO に基づく天文データ公開サーバが普及すれば、誰でも膨大な量の天体データの中から容易に検索し、目的のデータを取得することができるようになる。

2 統合天体データベース

VO によって実際の天文研究が効率的に進められるようになるであろうか。近年の天文研究では、天体の本質に迫るため、電波から X 線、ガンマ線という多波長のデータを組み合わせることが多い。また、変光星や超新星、ガンマ線バーストなど、明るさの変化が重要な場合には、時間をおいた複数の観測データが必要となる。このように、異なる観測装置によるデータが天文研究には不可欠である。一方、公開される天文データは、すばるのデータは国立天文台、ハッブル望遠鏡のデータはアメリカの機関というように、それぞれ観測をおこなった研究機関により配信されることが多く、別々のサービスとして提供される。そのため、どのサービスに目的の天体が含まれているかわからない場合には、それらのサービスすべてについて検索しなければすべてのデータを得られない。しかしこの手法は次に述べるように非効率である。第一に、すべての天体データ配信サービスにクエリを送信しなければならない。第二に、全天くまなく観測した例はわずかであり、多くの場合は天の一部の領域の観測であるため、問い合わせたサービスに目的の天体が含まていれる確率は小さい。

そこで、我々は、Web 検索サイトがあらゆるサイトのページを収集して効率的に検索ができることにならい、天体データについても、配信されている天体データを集めて「統合天体データベース」を構築し、全ての天体の効率的な検索を実現する手法を考えた。以下ではこの手法によるデータベースシステムの設計について述べる。

3 天球面インデクスによるテーブルパーティショニング

統合天体データベースの構築にはリレーショナルデータベースシステムを利用するが、登録する天体数が多いため、検索性能が問題となる。大規模な天体カタログの例として、2MASS 全天カタログは約 5 億、SDSS カタログは約 3 億もの天体のデータを含んでいる。このように、少なくとも 10 億天体のデータを検索できるデータベースが必要である。そこで、レコード数が多いデータベースを効率的に検索するための手法として、テーブルパーティショニングを用いた。天文検索では、天球座標による検索が基本であることから、天球座標によるテーブルパーティショニングをおこなった。天球座標のインデクス化の手法として、HTM (Hierarchical Triangular Mesh)[2] と HEALPix[3] の 2 種類の方式が提案されている。我々は利用実績のある HTM を用いた。HTM の手法により、天体の座標から HTM インデックスを計算し、その上位の桁によりグループ化する。今回は天球全体を $8 \times 4^6 = 32768$ の領域に分割し、psc_32768, psc_32769, ... , psc_65535 という名前のテーブルに格納した。各々のテーブルには下位の HTM インデックスをカラムに格納し、上位・下位合わせた HTM インデックスにより座標検索をおこなう。

VO では、SQL を拡張した天文検索言語 ADQL[4] を用いる。ADQL では座標検索を下記のように記述する。

```
select ra, dec, j_m
  from psc where Region('Circle 0 0 1');
```

Discussion and Implementation of an Efficient Search Method for Large Astronomical Databases

Masahiro Tanaka* Yuji SHIRASAKI* Masatoshi OHISHI* Yoshihiko MIZUMOTO* Yasuhide ISHIHARA† Jumpei TSUTSUMI† Yoshihiro MACHIDA† Hiroyuki NAKAMOTO‡ Yuusuke KOBAYASHI‡ Michito SAKAMOTO‡
*National Astronomical Observatory of Japan †Fujitsu Ltd. ‡SEC Co. Ltd.

表 1: パーティショニング性能測定結果

検索半径	天体数	経過時間 (秒)			HTM 条件数	
分角	個	Postgre SQL	独自方式	比	Postgre SQL[a]	独自方式[b]
1	2	6.46	0.04	154	32	32
10	165	3.81	0.03	127	16	16
60	6697	6.47	0.11	60	32	32
100	26720	2.02	0.31	7	4	16
180	57246	9.04	0.71	13	48	72

[a]where 句中における between でつなげた HTM 条件数
[b]union でつなげたサブクエリ条件数

この位置検索構文を HTM の範囲検索を伴う構文に置換することにより、以下のようなパーティショニングテーブル用の SQL 文を作る。

```
select ra, dec, j_m
 from ( select * from psc_63488 where
         htm_id between 0 and 65535
  union select * from psc_63488 where
         htm_id between 217088 and 218111
  union select * from psc_47104 where
         htm_id between 0 and 65535
   ...
 ) psc;
```

この構文置換プログラムは HTM 開発者によるライブラリを利用して Java で実装し、RDBMS には PostgreSQL を用いた。

4 提案手法による検索効率の測定

前節で述べた手法の性能を測定した。用いたデータは 2MASS の 5 億天体のカタログである。検索に要した時間を、検索範囲を変えて測定した結果を表 1 に示す。我々の手法により、半径 3 度という広い検索範囲でも 1 秒以下という短時間で検索できることがわかった。さらに、PostgreSQL に 8.1 版より装備されたパーティショニング機能を用いた場合と比較した結果、条件は異なるものの、7 から 150 倍高速であるという結果を得た。このように、我々の手法は大規模な天文データベースにおいても十分な性能を持つことがわかった。

5 テーブルの設計

天体カタログには、座標や明るさなどの他にも様々なデータが含まれており、その種類もカタログ毎に異なる。それらをすべて含むような統一的なテーブルの設計は困難である。そこで、統合天体データベースには、座標や明るさなどの天体データとして基本的な情報のみ保持し、さらに URL 等のデータ配信元へのアクセス情報を持つこととした。これによって、複数の天文データベースにまたがる効率的な検索と、詳細情報の取得の両方を可能にした。以上の方針で設計した統合天体データベースのカラムを表 2 に示す。

表 2: 統合天体データベースのカラム設計

category	column	description
Object	id	Object ID
	name	Object name
Position	ra	Right Ascension
	dec	Declination
	pos_err	Position Error
	htm	HTM index
Wavelength	band_name	Band name
	band_unit	Unit of band
Flux	flux	Flux value in catalog
	flux_err	Flux error
	flux_unit	Unit of flux
	flux_srch	Flux in Jy
Reference	link_ref	Link URL to reference
	org_id	ID in original catalog
	cat_id	Catalog ID

6 まとめ

天文学研究において VO サービスを利用する際の効率の問題から考えた統合天体データベース、およびその実現のために開発した効率的な検索手法とテーブル設計について述べた。実装したデータベースは、一部のデータを登録して JVO ポータル (http://jvo.nao.ac.jp/portal/) のサービスとして公開しており、一般ユーザでも利用できる。今後このデータベースに登録するデータを拡充する予定である。

参考文献

[1] 田中昌宏, 白崎裕治, 本田敏志, 大石雅寿, 水本好彦, 安田直樹, 増永良文. バーチャル天文台 JVO プロトタイプシステムの開発. 日本データベース学会 Letters, Vol. 3, No. 1, pp. 81–84, 2004.

[2] P. Z. Kunszt, A. S. Szalay, I. Csabai, and A. R. Thakar. The Indexing of the SDSS Science Archive. In *ASP Conf. Ser. 216, ADASS IX*, p. 141, 2000.

[3] Górski 他. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *Astrophysical Journal*, Vol. 622, No. 2, pp. 759–771, 2005.

[4] Yasuda 他. Astronomical Data Query Language: Simple Query Protocol for the Virtual Observatory. In *ASP Conf. Ser. 314, ADASS XIII*, p. 293, 2004.