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。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。このようなに、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構というように、それぞれ観測を所管した研究機関により配信されることが多い。それらのサービスとして提供される。そのため、そのサービスは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、全天まくも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れている確率は小さい。そこで、我々は Web 検索サイトがあればデータベースのページを収集する効率的に検索ができることになら、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ン

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。このようなに、少なくとも 10 億以上の天体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ンを用いた。天文検索では、天球座標による検索が基本であるところから、天球座標によるテーブルパーティションをおこなった。天球座標のインデッ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 \( 8 \times 4^v = 32768 \) の領域に分割し、psec.32768, psec.32769, ... psec.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、上位と下位が合わせた 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。

```sql
SELECT ra, dec, j_m
FROM psec WHERE Region('Circle 0 0 1');
```
表1: パーティショニング性能測定結果

<table>
<thead>
<tr>
<th>検索検索平均</th>
<th>天体数</th>
<th>経過時間 (秒)</th>
<th>HTM条件数</th>
</tr>
</thead>
<tbody>
<tr>
<td>分割数</td>
<td></td>
<td>PostgreSQL</td>
<td>独自</td>
</tr>
<tr>
<td></td>
<td>倍</td>
<td>SQL</td>
<td>方式</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6.46</td>
<td>0.04</td>
</tr>
<tr>
<td>10</td>
<td>165</td>
<td>3.81</td>
<td>0.03</td>
</tr>
<tr>
<td>60</td>
<td>6697</td>
<td>6.47</td>
<td>0.11</td>
</tr>
<tr>
<td>100</td>
<td>26720</td>
<td>2.02</td>
<td>0.31</td>
</tr>
<tr>
<td>180</td>
<td>57246</td>
<td>9.04</td>
<td>0.71</td>
</tr>
</tbody>
</table>

*where句におけるbetweenでつなげたHTM条件数
*unionでつなげたサブクエリ条件数

この位置検索機能をHTMの限界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ングテーブル用のSQL文を作成する。

```
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秒以下の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。このような、我々の手法は大量数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と、URLなどのデータ配信元へのアクセス情報を持つことが必要である。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を可能にするため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表を表2に示す。

6 まとめ

天文物理研究において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、一部のテーブルを登録して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、一般利用者でも利用できる。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


