C1

国立天文台電波天文データ公開利用システムの開発

文部科学省国立天文台　天文学データ解析計算センター　○大石　雅寿

Development of the Open-Use System for Radio Astronomical Data of the NAOJ

Masatoshi OHISHI, the Astronomical Data Analysis Center, the National Astronomical Observatory of Japan

ABSTRACT

We have succeeded (i) to develop the open access system for the radio astronomical data by the Nobeyama Radio Observatory, (ii) to transport its data analysis software package -- NEWSTAR -- onto the Linux and the HP-UX platforms, and (iii) to develop a new image reconstruction algorithm -- the Wavelet CLEAN -- to obtain reliable radio astronomical images through radio interferometric observations. We also investigated the data mining techniques to 1-dimensional spectral data, which human can not distinguish them into real signals or noise, resulting to extract many real signals.

Furthermore we developed a database of the numerical simulation data, a prototype for the distributed data analysis system based on the HORB, a robust method for the fast data-transfer to be combined with the above items.

1. はじめに

天文学における技術革新により、近年、大型観測装置から超大量のディジタルデータが生産されるようになってきた。これらのデータは観測後計算機で処理されるが、観測データが広く公開するシステムが十分に整備されていないために、需要があるにもかかわらずその要求に応えることができていなかった。

本研究課題では、まもなく共同利用観測を開始するすばる望遠鏡や2008年に本格観測開始を目指しているサブミリ波干渉計（LMSA）が生産する超巨大データの有効活用を図るための基礎を整備すべく、国立天文台野辺山宇宙電波観測所の電波望遠鏡が生み出す電波天文データを対象とした公開利用システムの構築を目指している。本課題では、単にデータの蓄積・ネットワーク上での公開にとどまらず、そのデータを用いた天文学を推進するために不可欠なデータ解析システム、及び、大量データを扱うために必要な情報処理技術として最近注目を浴びているデータマイニング技術の天文学への応用も行い、これらの要素を有機的に結合することによって天文学の更なる発展を目指そうとしている。

本研究プロジェクトではこれまでに、(1) 電波天文データの蓄積・ネットワーク上での公開システム、(2) 野辺山宇宙電波観測所内のみで利用可能であったデータ解析システム--NEWSTAR--をLinuxやHP-UX上へ移植する多プラットフォーム化、(3) 電波干渉計データから信頼性の高い電波強度分布図の再構築を行なうWavelet CLEANアルゴリズムの開発をほぼ終了した。さらに、これらを有機的に用いるための (4) データマイニング技術の天文学への応用研究を行い，人間には困難であった信号を機械学習により見出すこと、(5) 観測データの解釈に必要な理論シミュレーションデータのデータベース化システムの構築、(6) 大量データをネットワーク上で転送するための高速データ転送法の開発もほぼ終了した。そして、(7) データ解析システムをネットワーク上で展開するための分散システム化に向けた研究については，研究の最終段階であるので未完であるが、今後国立天文台の研究費を使用して研究を継続することとなった。

2. 各研究項目の進捗報告

2. 1　電波天文データの蓄積・公開システム

観測データのネットワークを通した公開を行なうためには、まず、観測データの蓄積（アーカイブ化、データベース化）を行なわなくてはならない。この開発については本プロジェクト開始以前より着手していたが、プロジェクト予算により、より有用性を高めるために2. 2に述べるデータ解析システム（NEWSTAR）との連携を実

現することができた。この機能の実現により、研究者はブラウザを通して本公開システムにアクセスできるだけでなく（http://nrodb.nro.nao.ac.jp/）、NEWSTARから直接公開データを入手することができるようになった。

2．2　NEWSTARのLinuxおよびHP-UXへの移植

NEWSTARは米国で開発された電波のデータ解析ツールAIPSを元にして、野辺山宇宙電波観測所の電波望遠鏡の観測データの解析用に開発されたソフトであり、ウィンドウパネルを使ったユーザーインターフェイスにより、高い操作性を実現している。しかし、基本的にNEWSTARは野辺山観測所でのデータ解析を想定して開発されたものであり、これまで対応するプラットフォームが限られていた（OSがSolaris、Sun、IRIX）。そのため観測所外のユーザーは観測データの処理を基本的には来所中に行うか、研究室へ観測所と同様の計算機環境を構築する必要がある等、使用環境には制限があった。

今回LINUXとHP-UX版NEWSTARの開発は使用可能なプラットフォームを増やすことで、ユーザーが選択できる環境の幅を広げることになった。特にPC-UNIXとしてのLINUXに対応した意義として、(1) 安価な電波データ解析システム、(2) 可搬型の解析システム、を構築できるようになったことは重要である。PCを使用できるため、現在ある資産をそのまま活用して、また新規に導入する場合も非常に安価に、研究室（や自宅）に電波データ解析システムを構築することが可能であり、さらに、持ち運び可能なノート型パソコンを活用することもできるので、データ処理環境の自由度が飛躍的に広がったといえる。

ユーザーはインターネットを通じて野辺山宇宙電波観測所のHPにアクセスすることにより、オンラインでソースおよびバイナリを入手できるようになっている。また、動作環境やバージョンアップ、インストール時・使用時のマニュアル等最新の情報を基本的にオンラインで手に入れることができる。

2．3　WaveletCLEANアルゴリズムの開発

電波干渉計観測データの画像再現時に現在世界中で最もよく用いられている方法は、Hogbom（1984）の開発したCLEANアルゴリズムによりデータをいくつかの点源（CLEAN成分）の集合に分解し合成するもので、コンパクトな電波天体に対して成果を上げてきた。しかし、有限のイメージ空間で無限に拡がる合成ビームを処理するとサイドローブから誤ってCLEAN成分を検出する危険性が生じるため、天体構造についての*a priori*な知識の助けをかりて人為的にCLEAN領域を指定する（ボックスをかける）という操作で作業を行わなければならず、画像に解析者の主観が介在して定量性にかける欠点があった。この問題を解決するために我々は、wavelet空間でCLEANを行い合成するという新しいアルゴリズムの開発することとした。

Wavelet CLEANでは、フーリエ変換と違って局所化された基底（wavelet関数）で関数を展開することから、多重分解能解析が可能となる。合成ビーム D (x, y) を基底とするwavelet関数列でダーティーイメージId (x, y) を展開し、マルチ解像度のwavelet空間でCLEAN成分の検出を行い、画像合成すると、サイドローブをCLEAN成分として抽出する可能性は著しく低下し、また多重分解能なので拡がった構造も正しくとらえることができる。

我々はこの原理に基づき、まずはプロトタイプとして、単一解像度でダーティーイメージと合成ビームの畳み込みId**D (x,y) 空間でのCLEANを行う簡易版を制作した。一次元シミュレーションでは、特にノイズが多い場合に、これまでのCLEANに比べて誤ったCLEAN成分の抽出が顕著におさえられ、真の像をより適切に再現するCLEAN成分が抽出されることが示された。これらの予備実験を基礎に簡易版wavelet CLEANの2次元アルゴリズムの製作と起動試験をおこない、CLEANでの結果と比較して簡易版wavelet CLEANの利点を調べた。Wavelet CLEAN (WLC) では、従来のCLEANと比べて各CLEAN component (CC) の位置にばらつきが起こらない。また一次元のシミュレーション結果と同様、サイドローブやノイズからCCを拾ってしまう効果を抑えていることを示している。これらの結果は論文にまとめ日本天文学会欧文研究報告誌 (PASJ) に投稿した（堀内　他2001）。

今後さらなる調整を行い、国立天文台の研究費を用いて実用版を完成させる。また拡がった構造の天体のイメージングに威力を発揮するマルチ分解能wavelet CLEANへの拡張を試みる。

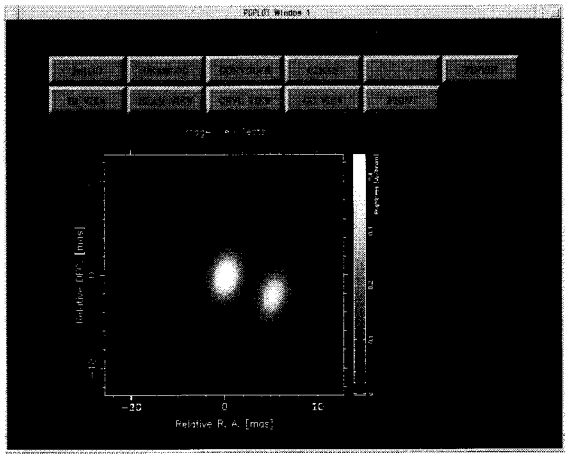

図1. WaveletCLEANの動作画面例--これはCLEANの結果の表示画面である。

2. 4 天文データマイニング

天文学の観測データに対するデータマイニングの適用をおこなった。使用したデータは国立天文台野辺山宇宙電波観測所で取得された分子雲からの電波スペクトルである。約700本の分子輝線が検出されている。そのうち、400本については分子との同定がおこなわれ、12個の分子が新たに発見されている。残りの300本については既知の分子からの輝線とは同定できず、新たな分析が待たれていた。

データマイニングの手法のうち、Decision Treeとk-th Nearest Neighbourhoodを用いて、300本の未同定分子線からノイズではなくより確からしい分子線を選び出す作業をおこなった。Decision Treeはどの属性のどのような値でよく既知のノイズと分子線が分離しているかを探し、探し出した属性値で未同定の分子線データを分類して「分子線」の可能性が高いものをピックアップするものである。k-th Nearest Neighbourhoodは既知のノイズと分子線データと個々の未同定の分子線データとの属性空間での距離を計算し、1～k番目までに近い既知のデータと同じものに未同定の分子線を分類する。

Decision Treeで300本の未同定分子線を分類した結果、300本中約200本が未知の分子線である可能性が高いという結果を得た。k-th Nearest Neighbourhoodでもほぼ同様の結果を得ることができた。ここで明らかになったことは、データの一次属性だけでは分類が難しく、目的に合った二次属性を生成することが極めて重要であるということである。また、データマイニングは大量データを扱うときの前処理として極めて有効であることが明らかになった。

2. 5 理論シミュレーションデータのデータベース化

国立天文台で共同利用されているスーパーコンピューターを使った数値シミュレーションの結果から得られたデータは、一般に膨大な容量を持ち、多くのパラメータに対するデータが得られる。それらのデータは、計算をした個人によって可視化、解析が行われることが多い。可視化や解析、観測結果との比較のためには、シミュレーションコードの特性や結果のデータフォーマットなどを考慮する必要などがあるからである。しかし、シミュレーションから得られた膨大なデータは非常に価値があり、観測との比較や新たなシミュレーションを行うにあたって、有効なデータとなりうるので多くの他の天文学研究者が利用できるようにすることが求められている。

これまでに、自らの数値シミュレーション結果を用いて、多数のパラメータにおける計算結果をデータベース化し、他のシミュレーションに応用するための例を示した。このデータベースにより、大規模シミュレーションを行ったときに知りたい結果をデータベース検索により抜き出して研究ができるようになり、今後、膨大なパラメータで行われたシミュレーションの結果の分類に有用なものであることが示された。そして、様々なシミュレーションに対応できるような検索システムの開発と新たなデータ登録を行えるようなシステムを構築した。これらはhttp://nd01.dc.nao.ac.jp/ で公開を行っている。

2. 6 分散データ解析システムの開発

2. 2で述べたデータ解析システムは汎用性が高くなったものの、新しい機能を付加するためのサポートがしづらい。これを解消する方法として、ネットワークを通してマンマシンインターフェースを起動し、サーバーのCPU能力をネットワークを通じて利用して、解析結果のみを入手する分散解析システムが考えられる。

我々は、これを実現するため、NEWSTARのマンマシンインターフェースをJavaでラッピングし、入力情報をHORBという通信プラットフォームを通じてサーバーに送り、サーバーでの解析結果を送り返すシステムの開発を行っている。概念設計は終わりコーディングを進めているが，予想外の問題が起きて2001年9月までには全ての作業は終了しなかった。しかし，残りの作業に必要な研究費は国立天文台内部で調達することができたので、完成まで作業を進めることとした。

2. 7 高速データ転送法の研究

天文学研究に使用する画像データを、観測所が建設されるような僻地からどのようにして転送するかに関する調査・研究を行っている。天文学で使用されるデータは、画像データが多く、1つのファイルで大量のデータを持つものが普通である。これを信頼性が低い回線で確実に転送する必要がある。そこで、天文画像データの圧縮の可能性と低信頼性回線によるデータ転送について検討・開発を進めた。

まず、具体的な天体画像データを、可逆圧縮後、どの程度冗長性が残っているかを調べた。その結果、流通している圧縮法以上に画期的にデータ圧縮を行なうことは困難であることが分かった。

そこでいかに安定してデータ転送を行なうかが重要であるとの結論に達し、現在、ファイル転送途中での接続切発生時の自動再接続、および、回線の信頼性保証機能が必要である。再接続はファイル分割によって、ある程度対応可能と考え、エラーチェック機能を同時に含む、自動再転送プログラムのプロトタイプを作成し，a) ファイル分割、b) エラーチェック符号計算、c) ファイル連結、d) 回線通信速度測定、e) 分割サイズ決定、の構成要素を個別に製作し、テストした後、これらのプログラム群を連続駆動するソフトを製作して、目標となるプロトタイプソフトウェアとした。

なお、本開発の結果は、東京大学を経由して特許申請した。

3. ネットワークの活用について

2. 1で述べた観測データ公開Webは既にインターネットを経由して「公開済み」データにアクセス可能であり、ネットワーク経由でデータ取得されている。また、2. 2のデータ解析システムもソースおよびバイナリがネットワークで取得可能となっている。観測データファイルの典型的な大きさは数100MBから1GB程度であり、転送時にはネットワークのバンド幅を一時的に占有する状態となる。

4. まとめ

これまでの研究開発の結果、データ蓄積・公開システム、データ解析システムの多プラットフォーム化、Wavelet CLEANアルゴリズムは一応の完成を見た。これらのうち、前者2つは既にネットワーク上で公開対象となっている。さらに理論データベースの作成、および、高速データ転送法についてもほぼ完成させることができた。これらと有機的に結合させるための分散解析システムの開発については、予期せぬ問題もあって完成させることはできなかった。しかし、当初の目標は8割方達成できたものと考えられる。今後はこれらの結果を基礎として，ネットワークをさらに活用した21世紀の新しい天文学の研究スタイルを構築できるよう今後も研究に励む所存である。

5. 研究実施体制

研究実施体制としては、プロジェクト開始当初から大きな変更はなく、国立天文台を中心として、東京大学、お茶の水女子大学が研究開発を行い、途中から高速データ転送法のテストなどに協力をお願いするため鹿児島大

学に研究に加わっていただいた。

6．参考文献

堀内，亀野，大石，伊藤 "Developmnt of a Wavelet CLEAN Algorithm for Radio-Interferometer Imaging"，2001，Publ. Astron. Soc. Japan, submitted.

Hogbom, J. A. 1984, Astron. Astrophys. Suppl., 15, 417.