

A Vision for PetaByte Data Management and Analysis Services for the Arecibo Telescope

Manuel Calimlim¹ Jim Cordes² Alan Demers¹ Julia Deneva²
Johannes Gehrke¹ Dan Kifer¹ Mirek Riedewald¹ Jayavel Shanmugasundaram¹
¹Department of Computer Science and ²Department of Astronomy
Cornell University
<http://www.cs.cornell.edu/database/escience>

Abstract

We survey the initial steps of a project to build a data management and data mining system for astronomy data generated by the Arecibo Telescope. The total amount of data that our project will have to manage will approach one Petabyte over five years. We describe some of the scientific challenges from the astronomy side, and we discuss initial thoughts on how to address these challenges through novel data mining research.

1 Introduction

The rapid growth in the generation of digital data is changing computational science in a fundamental way. Traditionally, the scope of computational problems was limited by the available processing power. But today, many problems are extremely data-intensive, and the lack of large-scale storage infrastructure creates a new bottleneck. Thus, modern data-intensive applications need high-performance computational resources *and* a system in which the computational resources are *tightly coupled* with large-scale storage.

At Cornell high-performance computing currently is being integrated with a new large-scale storage infrastructure, which will grow to one Petabyte over the next few years. On top of this system, we are building novel data management and analysis services to support research on finding new pulsars and other interesting celestial objects using data from the Arecibo Telescope.

The Arecibo Observatory (<http://www.naic.edu>), shown in Figure 1 is operated and managed by Cornell University via the National Astronomy and Ionosphere Center under a cooperative agreement with the NSF. The Arecibo telescope is the world's largest radio telescope in terms of collecting area and thus can conduct the most sensitive surveys for time-variable, pointlike objects. A multibeam feed system ALFA (Arecibo L-band Feed Array) has been installed during the summer of 2004. The feed array consists of seven dual-polarization horn feeds that provide 14 simultaneous data streams. The seven feeds point in different, but nearly contiguous directions, allowing a 7-fold speedup of the surveying rate. More details about ALFA can be found at <http://alfa.naic.edu>.

Pulsars are neutron stars, one of the endpoints of stellar evolution, which are about 10 km in radius and have 1.4 times the mass of the sun. They are, in many ways, giant nuclei. The pulsar surveys will be the deepest

Copyright 2004 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering



Figure 1: The Cornell Arecibo Telescope

(reaching to the greatest distances) ever undertaken and are expected to yield not only a large number of new pulsars (~ 1000) but also exotic objects, including millisecond pulsars spinning near the break-up speed of a neutron star; neutron stars in compact binaries with orbital periods of a few hours or less, and companion stars that are other neutron stars or black holes; and neutron stars moving rapidly owing to their birth in off-center supernova explosions. These discoveries are expected to provide numerous opportunities for followup research on the equation of state of nuclear matter, gravitation physics, and gravitational waves.

The data storage requirements for the P-ALFA surveys are very large (“astronomical”). The raw data from the P-ALFA surveys will amount to about 800 Terabytes, and it will take three to five years to acquire this data.

2 Background and Prior Research

Pulsar surveys not only discover new pulsars, but in particular aim to find exotic objects of interest, including millisecond pulsars with spin periods as fast as 1.6ms and potentially even shorter periods, and compact binaries with orbital periods of a few hours or less. For instance, compact binaries have sufficiently strong gravity that non-Newtonian gravitational theories (such as Einstein’s General Theory of Relativity) are needed to account for the orbits, which are precision monitored through continued “timing” observations of a discovered pulsar. Thus, pulsar research provides unique and important opportunities for studying extreme states of nuclear matter, gravitation physics, and gravitational waves. As an example of the physics payoff, a binary pulsar discovered in 1974 at Arecibo by Hulse and Taylor was used to infer the existence of gravitational waves in accordance with Einstein’s General Theory of Relativity [HT75]. The Hulse-Taylor binary consists of two neutron stars in an 8-hour orbit with orbital separation smaller than the Sun. For this work, Hulse and Taylor were awarded the Nobel Prize in physics in 1993.

Pulsar search procedures have evolved over the past 30 years according to our growing understanding of astrophysics and populations of compact objects such as neutron stars and black holes, but also thanks to the growth in computational and data storage capacity.

To process the raw data, we apply a filter with three parameters: (a) the temporal period; (b) the dispersion measure, which quantifies how pulses arrive differentially at different frequencies; and (c) the pulse duty cycle, which is the pulse width divided by pulse period. Figure 2 shows a single pulse from a strong pulsar in the Crab Nebula and the differential arrival times owing to dispersion. A trace is shown for each of 256 frequency channels. The slope of the locus describing the arrival time of the pulse in each channel is proportional to the

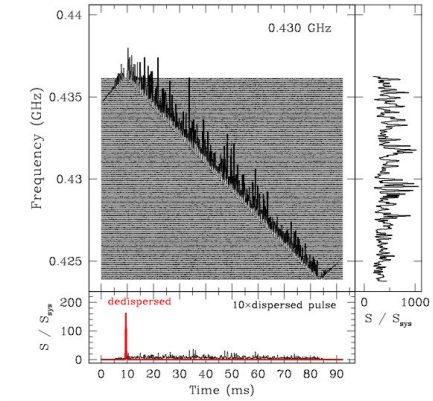


Figure 2: A single pulse from the Crab pulsar in the Crab Nebula

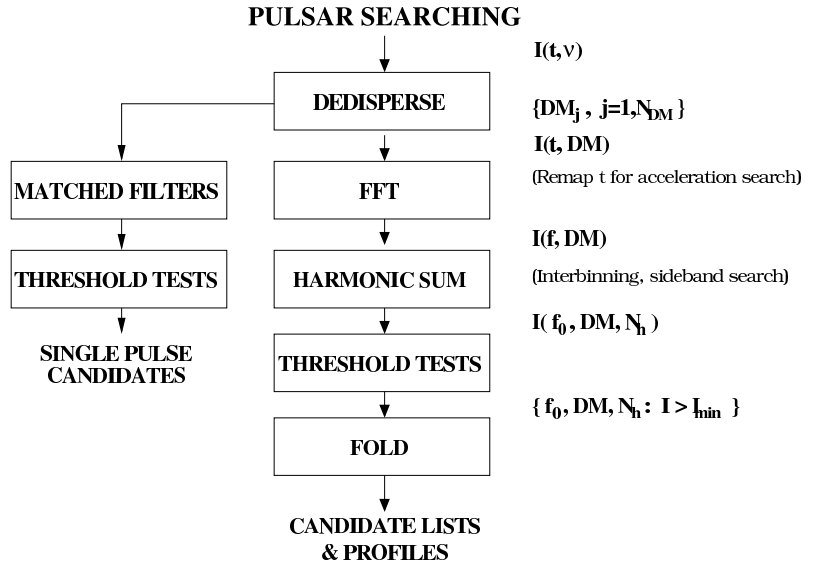


Figure 3: Flow diagram for a pulsar search algorithm

reciprocal of the “dispersion measure” (DM), a quantity that is unique to each line of sight (the endpoints of the pulse show positive slopes owing to aliasing in the sampling system). The bottom panel in Figure 2 shows the sum over frequency channels both with (“dedispersed”) and without (“dispersed”) compensation for the dispersion effect.

The main problem in pulsar surveys is that we do not know the parameters of the filter in advance. Thus, existing algorithms *search* a large space of possible parameters to find plausible settings that could mark the data as a candidate for a celestial object.

New pulsars are found largely through periodicity searches of pulse trains that are far weaker than the pulse shown. Single pulses are usually buried in noise, so a coherent analysis must be applied to long time series to bring out any periodic signal. While most data sets produce multiple candidate signals, the great majority of these are spurious signals resulting from locally generated radio frequency interference (RFI), one of the main problems of detecting real celestial signals. In order to recognize and remove RFI, we need to perform an analysis of the data products from the entire sky, taking into account that RFI signals repeat (but are episodic), while a single celestial signal should appear from only one direction on the sky. Due to the small size of existing pulsar surveys, current searches for filters could proceed brute-force without any sophisticated search strategies and high-performance algorithm implementations.

As preparation for full ALFA surveys, we have done a pilot survey of one square degree of sky using a single-beam system at Arecibo. The project has made use of pilot-Arecibo data to construct a database and develop analysis tools for investigating candidate signals from pulsars that will warrant reobservation. Reobservations are needed to robustly confirm the reality of pulsars versus artificial radio frequency interference, which is a growing problem that requires new filtering algorithms.

3 The P-ALFA Surveys

The proposed surveys for pulsars include (a) mapping of the entire Galactic plane of the Milky Way visible with Arecibo (within ± 5 deg of the midplane) where the greatest concentration of pulsars is expected; this represents a swath of sky equal to about 500 square deg, and (b) searching further out of the Galactic plane (e.g. ± 15 deg) in a shallower survey to find millisecond pulsars and binary pulsars that are expected to have a thicker population

distribution than the bulk of pulsars.

Why search this much sky? The survey payoff comes in two forms. First, the number of detected rare, exotic objects scales with the expected total number (N) of new pulsar discoveries. Based on previous surveys, having $N = 1000$ implies that we will find a handful of exotic binaries, etc. that will provide the greatest payoff in basic physics. Second, the total sample can be used to map the ionized gas and magnetic field of the Galaxy. Pulsar signals carry information about intervening magnetized, ionized gas because the signals are distorted by the consequent variable index of refraction. Techniques are well known for measuring these distortions and thus quantifying the integrated (along the path) electron density and product of electron density and line-of-sight component of magnetic field. The basic effects are known as dispersive arrival times (similar to delays in an optical fiber) and Faraday rotation of the plane of polarization of the electromagnetic waves. The challenge in using such integrated measures is to invert them into information about the gas density and magnetic field at each location in the Galaxy. Such inversion is similar to tomographic imaging. Generating tomography-based maps of gas properties is much more challenging than making the basic measurements for each line of sight. Examples to date may be found in [TC93], [CL02] and [WCK⁺03].

The raw data will be subjected to a novel multistep data analysis for identifying real celestial signals with expected (and hopefully new) signatures. The multiplicity of signals from the P-ALFA surveys drives data rates upwards and provides the opportunity for applying more intricate filtering algorithms that distinguish celestial signals from locally generated RFI. To realize the great potential of the ALFA system, new data mining algorithms are needed for tackling RFI and for detecting celestial objects of greatest interest.

The 14 signals from the “front-end” feed system (ALFA) will be signal processed in real time using a fast digital spectrometer that provides the basic “dynamic spectrum” $I(t, \nu)$, intensity vs. time and radio frequency, for each of the seven feed antennas. (The two polarization channels will be summed in real time, yielding seven distinct dynamic spectra.) Prior to the search analysis, the raw data will be filtered to remove radio frequency interference, which is expected to be nearly identical in all seven beams, whereas celestial signals will appear in at most 3 beams. As shown in Figure 3, each of the seven dynamic spectra is subjected to a detailed Fourier and statistical analysis whose results are winnowed to provide candidate periodic pulsar signals and candidate single pulse events. The dynamic spectra for each telescope pointing comprise a size equal to 7 beams \times 1024 frequency channels \times 10^7 time samples (approximately, since survey dwell times are now being optimized). The time duration for each pointing will be \sim 300s. Each dynamic spectrum will be searched for about 10^3 separate values of dispersion measure, yielding $7 \times 10^3 \times 10^7 \times 32$ separate statistical tests, where the 32 factor is the number of separate harmonic sums that are calculated, or $\sim 2 \times 10^{12}$ separate tests. Thresholds are set so that, per beam, we will have only 10 to 10^2 events above threshold that require further investigation. *Subsequent to the search analysis performed on each beam, candidate signals found in the 7 beams will be cross compared to remove low-level RFI that was missed previously, thus improving the quality of candidate lists. Additional complexity of the analysis involves optimization against potential motion of a pulsar in an orbit. To do so requires remapping the time series according to trial accelerations using typically about 200 trial values. Single-pulse candidate searches represent about 10% of the computations of the periodicity searches but add significantly to data products that need to be investigated in the meta analysis of all telescope pointings. Altogether, about 25,000 telescope pointings will be made to cover the Galactic plane with a similar number to cover portions of sky outside the plane. With the 300 s dwell time per pointing, it will take \sim 2100 hr and \sim 800 TB to acquire and store the raw data for the Galactic plane search.*

4 Data Mining Challenges

Astronomy data provides an exciting opportunity to test and develop new data mining techniques because of the wealth of information hidden in the raw data. The raw data is produced in such large quantities that *efficient* data mining and analysis becomes very challenging.

Common data reduction techniques such as subsampling, smoothing, and functional approximation are either not appropriate or need significant improvements to produce results of the same caliber as in other domains. The reason for this is the low signal-to-noise ratio. It is true that some of the signals are stronger than the noise. These include radio frequency interference from local sources as well as strong pulses from some pulsars. However, many interesting signals are *weaker* than the noise. This is the case for the majority of pulsars. The pulses they produce are weak and appear as spikes in signal intensity. Subsampling can easily miss the peaks of these spikes. Furthermore, subsampling is not advisable in noisy domains for statistical reasons — a smaller sample yields less accuracy. Careless use of smoothing and functional approximation can also cause problems.

Detection of astrophysical signals in data from Arecibo involves matched filtering, which optimizes the signal-to-noise ratio of a test statistic. Astrophysical signals are always corrupted by additive noise and generally are themselves stochastic in nature. Matched filtering requires knowledge of the functional form of the signal. For some aspects of the signals, such as the dispersion effect and periodicity mentioned before, we know these forms. However, other aspects of pulsar signals are not so well known *a priori*, such as the detailed pulse shape. In addition, we wish to discover entirely new classes of objects in the Arecibo data. These may include variations on the types of neutron stars, such as very sporadic emission, whose detection cannot rely on the periodicity, and they may include pulse shapes dissimilar to those of known pulsars. In addition, we would like to discover radio emission from other kinds of objects, such as black holes and extrasolar planets. In these cases, we know very little about the signature of the signal. For this reason, complete mining of the data requires that we search for signals with unknown structure.

A typical approach is to start with a collection F of smooth functions such as splines and find the $f \in F$ that minimizes:

$$\left(\sum_i (d(i) - f(i))^2 \right) + \lambda \int (f''(x))^2 dx$$

where d is the time-series and λ is a smoothness parameter [RS02]. Note that the second term is a roughness penalty that reduces the amount of wiggle in the approximating function by favoring functions with a smaller second derivative (acceleration). This suggests that information will be lost when we approximate a sharp spike (that may lack a second derivative or have an extremely large second derivative at some points) with a smooth function. Furthermore, because the signal is weaker than the noise, we may end up averaging the signal with the noise. This will spread out the spike (in time) and weaken its intensity.

The same can happen if we try to approximate the signal with a Fourier series. To reduce data size, the Fourier approximation (FFT) requires removing terms whose coefficients in the Fourier expansion are small. Thus the FFT is currently not used as a means of data reduction, instead it is used as a tool to discover the rotational frequency of a pulsar. Indeed, when searching for pulsars, *all* Fourier coefficients are kept and a technique called *harmonic folding* is used to discover the rotational frequency (in harmonic folding, the coefficients of several harmonics of a base frequency are added together). This yields better results than simply looking at the frequencies with highest Fourier coefficients.

To extract weak signals from the noise, it is essential for data mining to incorporate background knowledge in a principled way. Information such as “the signal is periodic” or “the signal, if present, has the following characteristics or the following functional form” must be handled automatically and efficiently. How can we use background knowledge to boost efficiency and/or accuracy of data mining techniques? Right now this is done on a case-by-case basis: Fourier series and harmonic folding for periodic signals, time warping the time-series to “normalize” it if the signal has a parametric form. Is there a more general language (short of a programming language) that can be used to specify certain background knowledge and from which algorithms can be automatically constructed? This is an open question in data mining today and astronomy data is a promising test-bed and motivator for ideas in this direction.

Closely related to the concept of mining with background knowledge is the notion of related mining tasks. Consider the following three tasks: detecting pulsars that periodically emit weak pulses, detecting pulsars that

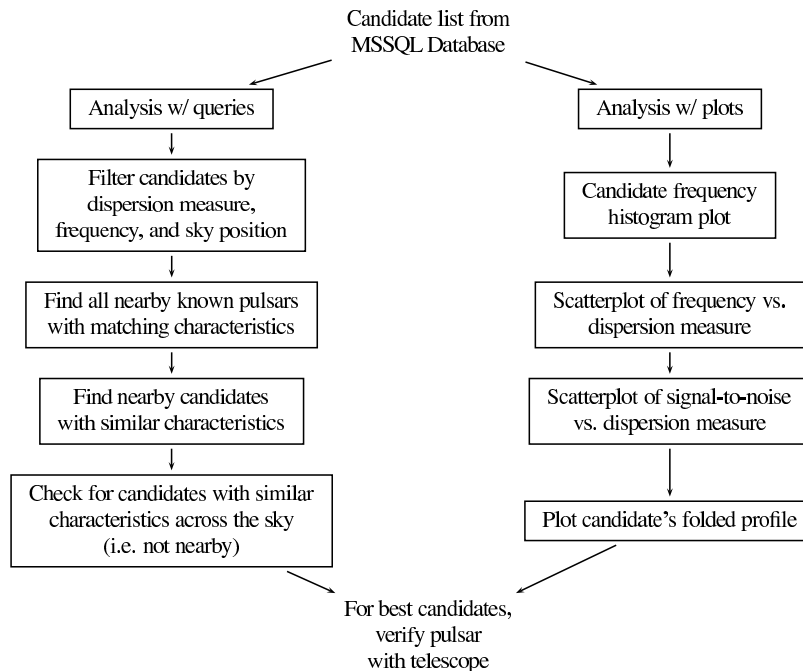


Figure 4: Pulsar candidate online analysis

occasionally emit strong signals, and detecting solar bursts. Intuitively the two pulsar-detection tasks are related to each other but not to the detection of solar bursts. The processing required for a task can be described as a sequence of steps (removal of RFI, dedispersion, FFT, harmonic folding, etc.) and each step generates intermediate data such as dedispersed time-series, lists of RFI signals, lists of Pulsar signals, and so on. Related tasks can share information and intermediate data representations to improve throughput in a way that is analogous to the sharing of buffer pages between related queries in a database management system (DBMS). For example, almost all tasks can share dedispersed data and RFI signals (which would then be removed from the time series). Other types of signals can also be shared between tasks so that they can be removed from the time series. The relationship to a DBMS is not superficial. There is a potential and need for a Data Mining Management System (DMMS) to identify related tasks, to identify shareable information and intermediate data representations, to manage this information (on disk) for optimizing future queries, and to incorporate background knowledge into the mining process.

After processing is finished, results are presented to the astronomer for further analysis. In the case of pulsars, the result would include a graph of the average pulse profile. At this point the astronomer needs to decide what is interesting, what isn't, and what deserves more investigation. The amount of results can be very large and currently results are thresholded based on signal strength. Here we have opportunities for more applications of classical machine learning and data mining. The final results can be clustered, classified, and ranked based on preferences derived from human interaction. Not only will this reduce the burden on the astronomer, but it will provide more testing data and potential benchmarks for extensions to classical mining.

5 The Current Data Management and Analysis System

The initial pilot survey that focused on a small portion of the sky is online and ready for analysis. The output of analyzing the raw signal data is divided into two categories of processed data: pulsar candidates and single giant pulses. Pulsar candidates are usually identified by the frequency of periodic radio pulses while single giant

Query :: Advanced Candidate Search

This search will find all pulsar candidates near a specified location with extra fit

1. Set parameters

Use galactic coordinates: Use equatorial coordinates:

Galactic longitude (in degrees):	<input type="text" value="54.0275"/>	Right ascension (J2000 in degrees):	<input type="text" value="10"/>
Galactic latitude (in degrees):	<input type="text" value="-0.2501"/>	Declination (J2000 in degrees):	<input type="text" value="10"/>
Search radius (in arcminutes):	<input type="text" value="5.0"/>		
Dispersion measure (pc cm ⁻³):	<input type="text" value="25.33"/>		
Dispersion measure delta (pc cm ⁻³):	<input type="text" value="10.0"/>		
Frequency (Hz):	<input type="text" value="117.591117"/>		
Frequency delta (Hz):	<input type="text" value="20.0"/>		

Figure 5: Advanced Candidate Query example

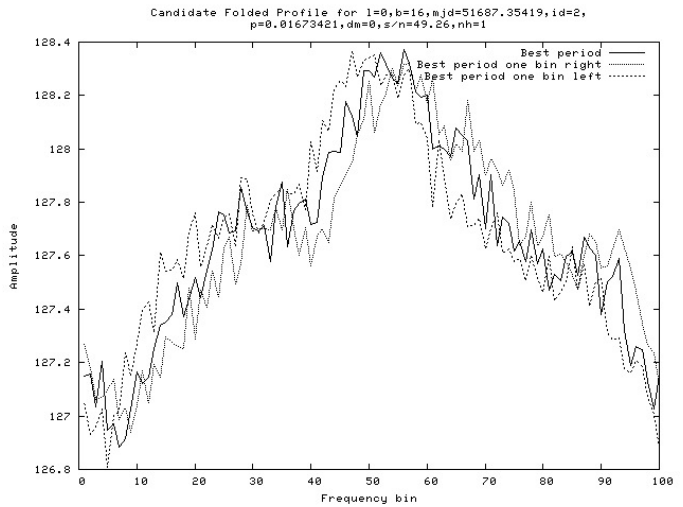


Figure 6: Candidate pulse profile example

pulses are any pulses above a certain threshold.

Most of the analysis for pulsar candidates is geared towards identifying and removing radio frequency interference (RFI). Figure 4 shows a typical workflow for online analysis of candidates. At the top of the figure, all of the candidate information is loaded into a relational database for fast searching of the data. The left branch shows some sample queries that can be made from the website to reduce the number of candidates.

First, a simple filter query can easily identify and reject some common sources of RFI. Dispersion measure is used to calculate the distance of the radio object. When the dispersion measure is close to zero, the signal is most likely terrestrial in origin. Also, the candidate list can be pruned with restrictions on frequency; removing candidates that match the frequencies of well-known sources of RFI eliminates a large number of invalid pulsar candidates.

After the obvious RFI is detected, more advanced queries are employed. Pulsar candidates are compared to the list of known pulsars to avoid detecting an object that has already been discovered. Then there is a search for matching candidates in nearby pointings since good pulsar candidates with strong signals will frequently appear in adjacent pointings. An advanced candidate search with deltas over distance, frequency, and dispersion measure is displayed in Figure 5. Finally, a search for matching candidates in far pointings is conducted; a pulsar candidate that appears all over the sky is likely RFI since a pulsar's signal should come from one direction instead of all directions.

A number of online plots are also available for the user. The user can plot a histogram of the candidate frequencies of a particular pointing versus the frequencies across all pointings. This can help to determine if an event seen in a pointing is a common RFI type event. Two scatterplots (one with frequency vs. dispersion measure and the other with signal-to-noise ratio vs. dispersion measure) are used to quickly check for trends in the data that would indicate RFI contamination. The most important plot is the folded pulse profile for pulsar candidates (see Figure 6). The folded profile helps classify the object and usually is the final determination for which objects are actual pulsars.

The best pulsar candidates are saved in a catalog. The raw data for these pointings is re-processed and re-analyzed to verify the possible pulsar. The very best candidates need further observations with the telescope to confirm the reality of new pulsars versus possible artificial radio frequency interference.

Analysis of single giant pulses is similar to the pulsar candidate analysis. Figure 7 shows the online tools available. Note that only a basic filter query is supported. Because single giant pulses are comprised of any radio pulse above a threshold, the number of giant pulses is orders of magnitude larger than the number of pulsar

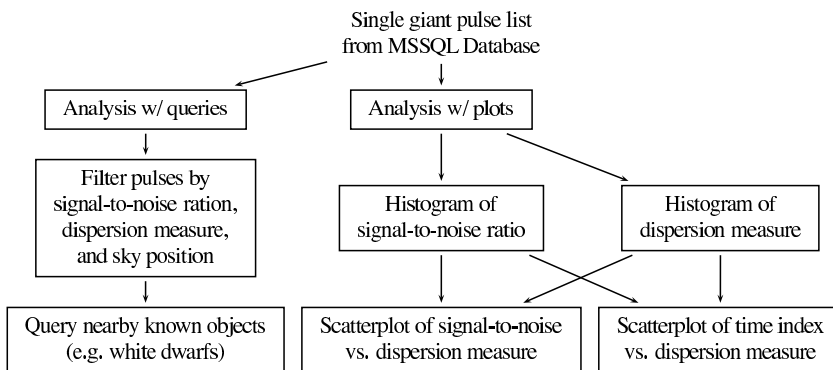


Figure 7: Single giant pulse online analysis

Pulses Query :: Parameters

1. Choose range for dispersion measure

Include: Min: Max:

2. Choose range for signal-to-noise ratio

Include: Min: Max:

3. Update SQL with new parameters

Figure 8: Candidate pulse profile example

candidates. Generally, this means that complex queries run very slowly over the single giant pulses. Special spatial indices have been added to the database to speed queries over location, but performance for the single pulses is a major issue.

Filtering the giant pulses by signal-to-noise ratio, dispersion measure, and sky position helps identify interesting single pulses, i.e., pulses with a high signal-to-noise and dispersion measure. Figure 8 shows the webpage interface that specifies the parameter ranges for searching the single pulses. Once interesting pulses are found, users can query for nearby objects of interest. These include white dwarfs, gamma ray sources, x-ray sources, and other radio sources that may explain the single giant radio pulse.

The analysis plots for the single pulses also help isolate important phenomena. Histograms for signal-to-noise and dispersion measure should reveal spikes of interest. Scatterplots with signal-to-noise vs. dispersion measure or time index vs. dispersion measure are used for recognizing which pointings have too much RFI.

The single pulse analysis can reveal pulsars that the candidate analysis misses. In fact, since so much data is captured, more exotic objects than pulsars may be found (e.g., compact binaries or neutron stars). The potential for data mining over the single giant pulses is broader and may be the most interesting area for future research.

6 Expected Results: One Example

Recall the discovery of the Hulse-Taylor binary, which consists of two neutron stars in an 8-hour orbit. In the last 30 years, only five additional binaries have been found because they are scarce and require a great deal of telescope and computer time. The orbits of double neutron star binaries (and any neutron-star/black-hole binaries) are shrinking because of energy losses associated with the gravitational waves. Consequently, we expect the Galaxy to contain binaries with orbital periods *much smaller than 8 hours*. Short-period binaries are difficult to detect because traditional search algorithms rely on the Fourier analysis of equally-spaced (or nearly so) pulses (see [Cor02] for an overview); Doppler shifts from orbital motion reduce the sensitivity of such searches unless compensating, computationally intensive algorithms are used [RCE03]. Processor speeds now are just beginning to make comprehensive searches for binaries feasible. We expect that important new binary systems will be found in ALFA surveys that will be far richer than the Hulse-Taylor binary in providing opportunities for testing gravitational physics.

7 Future Directions

As mentioned before, our goal is to mine the Arecibo data for both known classes of signals — periodic, dispersed pulsar emissions and single dispersed pulses from pulsars — and new classes of objects. The starting point is the dynamic spectrum and each kind of signal will produce a certain characteristic pattern in the frequency-time plane of the dynamic spectrum. Our aim is to develop methods that will allow us to detect (i.e. distinguish from noise) and classify all non-noise events according to their frequency-time morphology, followed by classification in terms of what kind of terrestrial interference or kind of celestial source is involved. This overall aim requires that we fully exploit the growing database by developing tools that are both open-ended and efficient.

8 Acknowledgments

Jim Gray gave us helpful advice throughout many discussions. Alex Szalay offered us generous support through usage of his software and through a visit to his group. David Lifka and his team at the Cornell Theory Center provide invaluable support to configure and run the system.

Our system is supported by a Research Infrastructure Grant Number CNS-0403340 from the National Science Foundation, and by an E-Science grant from Microsoft Corporation. Johannes Gehrke's data mining research has been supported by the National Science Foundation under Grants IIS-0084762 and IIS-0121175. Jim Cordes's research has been supported by NSF Grants AST-0206035 and AST-0206036. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

- [CL02] J. M. Cordes and T. J. W. Lazio. NE2001.I. A New Model for the Galactic Distribution of Free Electrons and its Fluctuations. *ArXiv Astrophysics e-prints*, July 2002.
- [Cor02] J. M. Cordes. Pulsar Observations I. – Propagation Effects, Searching Distance Estimates, Scintillations and VLBI. In *ASP Conf. Ser. 278: Single-Dish Radio Astronomy: Techniques and Applications*, pages 227–250, December 2002.
- [HT75] R. A. Hulse and J. H. Taylor. Discovery of a pulsar in a binary system. *Astrophysical Journal Letters*, 195:L51–L53, January 1975.
- [RCE03] S. M. Ransom, J. M. Cordes, and S. S. Eikenberry. A New Search Technique for Short Orbital Period Binary Pulsars. *Astrophysical Journal*, 589:911–920, June 2003.
- [RS02] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer-Verlag, New York, 2002.
- [TC93] J. H. Taylor and J. M. Cordes. Pulsar distances and the galactic distribution of free electrons. *Astrophysical Journal*, 411:674–684, July 1993.
- [WCK⁺03] J. M. Weisberg, J. M. Cordes, B. Kuan, K. E. Devine, J. T. Green, and D. C. Backer. Arecibo 430 MHz Pulsar Polarimetry: Faraday Rotation Measures and Morphological Classifications. *ArXiv Astrophysics e-prints*, October 2003.