Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

Efficient data mining in the X-ray sky background

John Cunniffe, Evert J.A. Meurs

Dunsink Observatory, Castleknock, Dublin 15, Ireland

Aaron Golden

IT Centre, NUI Galway, University Road, Galway, Ireland

Abstract. Abstract: The calculation of flux upper limits in the source-free (”background”) regions of photon-limited images is complicated by instrumental and statistical modeling and requires non-trivial computational effort. We propose a scheme where the properties of source-free image regions are pre-calculated and catalogued in a compact form to allow rapid estimation of flux limits and statistical properties of the background. For the case of X-ray data, we examine the expected speed-up in massive all-sky queries resulting from such a system to pre-screen coordinate requests before running a full analysis.

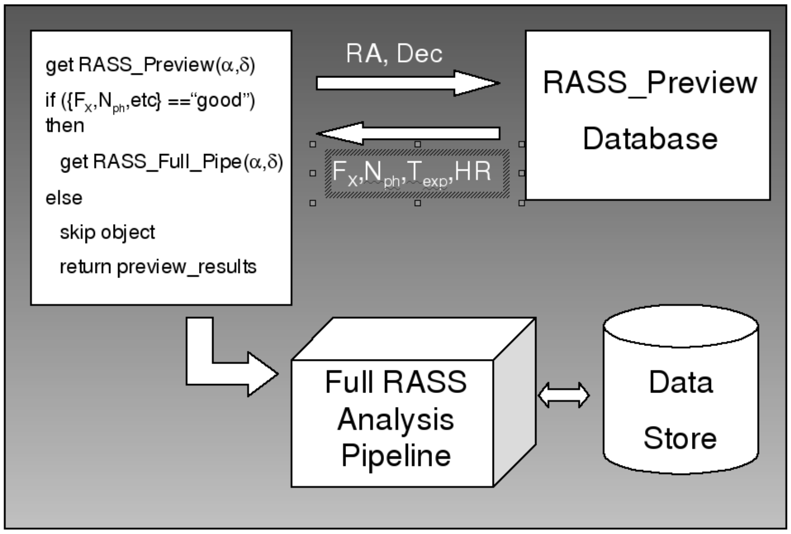
1. Introduction

We are interested in finding long term X-ray variability in classes of objects ranging from brown dwarfs to galaxies. Data-intensive searching is the only realistic method for discovery of rare transient flaring or spectral variability which may reveal important underlying physical processes. Multi-wavelength correlation offers additional insight into these processes but the most important initial criterion is to identify candidate variability ’events’.

Large databases being queried within a Virtual Observatory (VO) context allow previously unfeasible speculative searches for such rare transients. Most of the sources analysed will produce negative results so approaches which can recognise and abort clearly unproductive pipelines early will speed up the search completion. Such acceleration will become important as the size of available databases continues to rise and the demand for popular data resources creates significant queuing delays (see e.g. O’Mullane 2004).

2. Designing searches for rare transients

Correlating catalogues of source detections and comparing fluxes/count rates represents a relatively fast approach to long-term light-curve calculation. When a source is not detected in a particular observation, an upper limit to the flux can be computed using a model of the instrument, exposure time, bandpass, etc. Significant non-detections where the source has clearly dropped below the expected level may indicate an object in a pre- or post- flare state.



```
get RASS_Preview($\alpha$,$\delta$)
if ({$F_X$,$N_{ph}$,etc} =="good")
then
  get RASS_Full_Pipe($\alpha$,$\delta$)
else
  skip object
  return preview_results
```

Figure 1. RASS-Preview path during a query.

Flux upper limit or limiting magnitude calculators have been implemented using image metadata (Voisin et al. 2004). However, in the case of X-rays, the upper limit calculation process is complicated by its strong dependence on the exact photon statistics and by the varying instrument effects (vignetting, PSF, bandpass, etc) across the field of view. These factors make the creation of a simple background/upper-limit model much harder and in general the calculation of an accurate upper limit will require re-execution of the source detection pipeline to assess the possible source statistics again.

Results for such parameters are also harder to compress into representative values since even closely neighboring positions can have significantly different source and background statistics. Similarly, source morphology and temporal smapling information within an observation is harder to express in a tabular form that will satisfy general queries.

3. RASS-Preview - A RASS Model

The ROSAT All-Sky Survey (Voges et al. 1999) carried out from Jun 1990 to Jan 1991 remains the deepest all-sky map of the X-ray sky. As such it is the most useful reference point against which to compare detections from pointed mode observations from other missions/instruments (Chandra, XMM, RXTE, ROSAT pointed mode, etc) for signs of long term variation.

At present we are developing the RASS-Preview tool to allow any point on the RASS sky to be queried and parameters returned via a database request to pre-stored tables. Only in the case of interesting results from this stage of the query does the conventional analysis pipeline continue to re-compute the exact upper limits and other parameters from the raw data (see Fig 1).

The present approach is to calculate the source parameters characterising each RASS resolution element on the sky (∼6×10^8 beams/sky containing ∼10^5 source detections). Each position is described by a flux (or upper limit) in the 0.1-0.5keV, 0.5-2.4keV and 0.1-2.4keV bands, the number of photons detected in each band and a background level. This information can then be queried from a compressed database in response to each request and this database can then be used when objects are not found in the existing RASS Bright and Faint Source Catalogues of point sources.

4. Expected Speed-up

The current pipelines we are using run on a combination of software: MIDAS/Exsas (MPI), FTOOLS(Heasarc) and shell scripts. To build a database of the RASS X-ray properties of 105 objects takes ∼10 days on a single 1GHz desktop machine. The principal component in this is the time taken to preprocess and handle the ∼30 GB of RASS data rather than the calculation of each upper limit. Therefore, increasing the number of objects queried does not linearly increase the execution time. This whole process could be parallelised by partitioning the sky across many machines however what we are interested in here is developing with RASS-Preview the methods needed to make the searches as efficient as possible rather than a 'sledgehammer' approach. As the size of the X-ray data archive grows with the present generation of satellites, efficiency will become increasingly important if we still wish to carry out massive speculative all-sky queries. The significantly greater future size of the XMM and Chandra data archives in terms of number of distinguishable resolution elements needing encoding will mean that an efficient database coordinate search system (Ortiz 2003) may be needed to prevent this becoming a bottleneck.

Based on trials using a few RASS fields, we expect that queries to RASS-Preview can be processed in a few milliseconds on a desktop system compared with, at best, the several tens of seconds needed to execute a full analysis pipeline. In a search of sources in the XMM-Serendipitous Source Catalogue, greater than 90% were rejected as being more than a factor two fainter than the RASS upper limit using RASS-Preview, and therefore not worth further processing with the accurate upper limit pipeline. At present we are looking at extending the catalogued parameters to describe flux upper limits on any short period variability during the observation.

5. Future Directions

The results from RASS-Preview will allow us to assess the next steps to take in creating similar archive preview approaches and better data modeling and description tools. This is effectively creating a deeper layer of meta-data and thus allowing maximum efficiency in large scale queries of VO resources.

References

O'Mullane, W. 2004, this volume, 372

Ortiz, P. F. 2003, in ASP Conf. Ser., Vol. 295, ADASS XII, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 35

Page, C. G. 2003, in ASP Conf. Ser., Vol. 295, ADASS XII, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 39

Voisin, B. et al. 2004, this volume, 125

Voges, W. et al. 1999, A&A, 349, 389