Astronomical Data Analysis Software and Systems XII
ASP Conference Series, Vol. 295, 2003
H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, eds.

Why Indexing the Sky is Desirable

Patricio F. Ortiz
AstroGrid, Department of Physics and Astronomy, University of Leicester, Leicester, LE1 7RH, UK

Abstract.
Indexing the sky is a database-oriented term to indicate a partitioning scheme of the celestial sphere in order to achieve better performance in queries involving finding close neighbours. Several schemes have been proposed: HTM, HEALPix, "IDT: iso-declination tiles", Quadrilateralized Spherical Cube, etc., but their use has not become widespread. The scientific value of the internal indexation files is much higher though, as they keep track of the source density of catalogues and hence allow us to answer a family of questions not easily handled by a standard DB system and providing an unusual visual aid: a snapshot of the location of sources listed in any catalog. The pros and cons of adopting an VO-oriented indexation scheme are analyzed.

1. Introduction

Dividing up the sky is an old practice useful for both locating and identifying objects. The problem imposed today by the large data volume is to have the capacity to quickly locate and compare objects found in the same region of the sky from two (or more) catalogues. Multi-wavelength and time series analysis need these features. Another important issue of today's data is to describe the sky coverage of a catalogue and use that information to speed up the cross correlation process.

From the database point of view, several schemes have been adopted to optimize cross-correlation, among them: Quadrilateralized Spherical Cube (White et al. 1992), Hierarchical Triangular Mesh (HTM) (Kunszt, P et al. 2001) and iso-declination tiles (IDT) (Ortiz & Ochsenbein 2001). Others such as HEALPix (Górski et al. 1998) were created to handle data covering the whole sky, e.g., the COBE mission. These methods split the sky in zones. At a database level, either a particular index may be associated to each object, or the sky zones are used to delimit internal database boundaries which are later used to speed up cross correlation (IDT). Description of the sky coverage, at a meta-data level, helps to speed up queries involving a large number of catalogues.

2. Reasons to use a Sky Indexation Scheme (SIS)

The main reasons presented in this paper are purely based on the handling of catalogues containing positional information. Regardless of the shape and size

© Copyright 2003 Astronomical Society of the Pacific. All rights reserved.

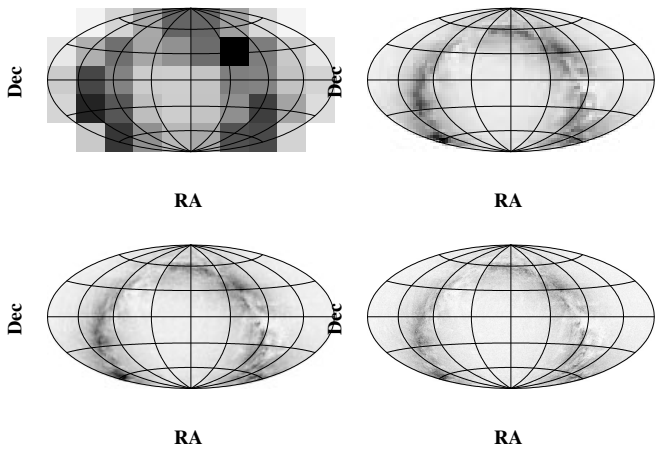

Figure 1. Star distribution in the Tycho catalogue. Clockwise from top left: 10×5 "zones", 100×50 "zones", 200×100 "zones", and 600×300 "zones"

of the zones in which the sky is partitioned, it is always possible to determine the number of elements which belong to each "tile" (**sky density**), all the above mentioned methods try to split the sky uniformly. How this information is used is the key issue. Figure 1 illustrates the point using different resolutions for the source distribution of the Tycho catalogue.

2.1. Meta-data

A direct application would be to use the sky density as a method of determining whether, within a list of catalogues, certain areas of the sky contain sources or not, that is, a sort of meta-data which would not require the examination of whole catalogues. Its main application would be to quickly rule out catalogues which have no common areas with a list of targets either taken from an existing catalogue or user generated. At $30'$ resolution ~ 160000 tiles are needed, but it would be enough to have a 1 or 0 in each tile to accomplish the task, reducing the size of the mask to $\sim$ 20Kb which could be compared in $< 0.1s$ without having to "activate" databases unnecessarily. This application of an SIS may be extremely useful in the VO era, as there is no simple way to represent the sky coverage of any catalogue and this method could contribute a uniform representation of the sky and a way to optimize the usage of resources. An interesting application is to query sites with pointed observation data (observing logs) to discover if any of the targets in a list has been observed. Regardless of the size of the list a sky-mask comparison can be faster and preselect the catalogues to be cross correlated.

2.2. Visualisation

A by-product of an SIS is to offer human users a snapshot of the sky coverage of a catalogue by providing a simple image such as the ones shown in Figure 1). Blind usage of VO facilities is a real risk, particularly when statistical algorithms will be implemented in the future; if any assumption of uniform coverage, unbiased observations, or homogeneity is not fulfilled by the catalogues used, the results of applying those methods will be meaningless.

2.3. Database Performance

An SIS should improve the performance of a cross correlation as one is assured that the search for matches is only carried out over the appropriate areas of the sky. If the DBMS allows it, the data could be arranged according to the indexation, which would also improve the I/O performance. For example, using IDT in datoz2k (Ortiz 2002) a cross correlation of Tycho-2 (2.53 million objects) within a radius of $35''$ takes 5.5 seconds on a 1.7GHz PC running Linux. Another approach is to add an extra column with a tile-ID: *PCODE* (Page 2002), index the DB on this number, and then perform a match on PCODE, which translates into a great gain over using indexation on declination only. Although they are used internally, SISs are also used across all data centres to improve their response.

2.4. Scientific Usage

There are a number of scientific questions which are very difficult to address if sky-density tools are not operational. All of them deal with the fact it is possible to select areas of the sky where the density satisfies certain criteria. The following list is by no means complete:

At a one catalogue level:

(1) Select zones and/or objects located in areas where the density satisfies certain conditions. An example would be areas where the QSO density is high enough for them to be used as astrometric calibrators.

(2) Locate areas which represent a local enhancement or depletion of objects. An example might be the location of very frequently observed zones of the sky in an observing log suitable for use in variability studies. We could impose absolute density criteria, but enhancements are often adequate.

(3) Compute the sky density for a given type of object in a catalogue and select the areas according to this newly computed density. Example 1: from a wide field image catalogue, select all the galaxies with $B-V > 1.5$ and look for density enhancements. Example 2: Locate the areas most frequently observed with an instrument in which the seeing is $< 1''$ and the moon is below the horizon during the observation to produce very long exposures or perform time variation studies.

When two or more catalogues are involved:

(4) Determine the intersecting areas between two catalogues. Example: determine which areas observed with HST have objective-prism data (images from the HST observing log vs the observing log of one or more objective-prism surveys).

(5) Find the sky areas with highest density from the combination of data in several catalogues. Example: find the most observed areas in Radio, Optical and X-ray located within 20° of the galactic plane; several catalogues from optical, radio and X-ray could be needed and finding local maxima in the resulting combination does the trick.

Due to intrinsic physical properties and the nature of our accumulated observations the tasks described above are nearly all means to discover areas of interest, either to start collecting data from an archive or to decide where to point the telescope next. However, not having this information nor the tools to explore the data space in this way would prevent astronomers from finding rich areas in which to dig.

3. Conclusions

A sky-indexation scheme provides one of the simplest ways to represent the sky coverage of a catalogue. Using an SIS will give users the chance to select areas of the sky according to their own criteria, a functionality of high scientific interest not provided by today's commercial DBMS, enriching the services provided by any data centre.

The existence of sky coverage masks for each catalogue in a data centre can make answering data-mining questions much faster by reducing the number of catalogues against which to cross correlate; a minimal time is required to know if a catalogue has any chance to have matches with a target list.

The adoption of one or more standard SISs (resolution, method, etc.) by the community, in particular in the context of VO, would provide a means of exchanging information about the properties of a catalogue at a much faster pace, particularly in the case of very large catalogues. This would be a way to include the sky coverage as part of the meta-data.

References

Górski, K. M., Hivon, E., Wandelt, B. D. 1998, in Proceedings of the MPA/ESO Cosmology Conference Evolution of Large-Scale Structure, eds. A.J. Banday, R.S. Sheth and L. Da Costa.

Kunszt, P., Szalay, A. S., Thakar, A. R. 2000, in Mining the Sky, Proceedings of the MPA/ESO/MPE Workshop, edited by A. J. Banday, S. Zaroubi, and M. Bartelmann, Heidelberg: Springer-Verlag, 631

Ortiz, P. F., Ochsenbein, F. 2001, in Mining the Sky, Proceedings of the MPA/ESO/MPE Workshop, edited by A. J. Banday, S. Zaroubi, and M. Bartelmann, Heidelberg: Springer-Verlag, 677

Ortiz, P. F. 2003, in Proceedings of "Toward an International Virtual Observatory", (http://www.eso.org/gen-fac/meetings/vo2002/), in press

Page, C. 2003, this volume, 39

White, R. A., Stemwedel, S. W. 1992, in ASP Conf. Ser., Vol. 25, Astronomical Data Analysis Software and Systems I, ed. D. M. Worrall, C. Biemesderfer, & J. Barnes (San Francisco: ASP), 379