Astronomical Data Analysis Software and Systems VII
ASP Conference Series, Vol. 145, 1998
R. Albrecht, R. N. Hook and H. A. Bushouse, eds.

Methods for Structuring and Searching Very Large Catalogs

A. J. Wicenec and M. Albrecht

European Southern Observatory, Karl-Schwarzschild-Str. 2, D-85748 Garching, Germany ESO

Abstract. Some existing (e.g., USNO-A1.0) and most of the upcoming catalogs (e.g., GSC-II[1] SDSS[2], GAIA[3]) contain data for half a billion up to several billion objects. While the usefulness of the content of these catalogs is undoubted, the feasibility of scientific research and thus the scientific value is very much dependent on the access speed to an arbitrary subset of parameters for a large number of objects. Given the data volume of such catalogs it seems impractical to build indexes for more than two or three of the parameters. One way to overcome this problem is to establish a multi-dimensional index covering the whole parameter space. This implies that the catalog is structured accordingly. The disadvantage of a multidimensional index is that the access to the main index is much slower than it would be using a dedicated structures and indexes. The most commonly used index for astronomical catalogs are the coordinates. The astrophysical importance of coordinates is limited, at least if the coordinates are not suited to the astrophysical question. However for observational purposes coordinates are the primary request key to catalogs. Here we present methods for structuring and searching very large catalogs to provide fast access to the coordinate domain of the parameter space. The principles of these methods can be extended to an arbitrary number of parameters to provide access to the whole parameter space or to another subset of the parameter space.

1. Indexing and Structuring Catalogs

Most of the existing catalogs are physically structured and sorted according to only one parameter. Mostly this parameter is the longitude of the underlying coordinate system (e.g., RA). Some newer catalogs (GSC, Tycho) are structured and sorted by regions, i.e., objects from one region are close together in the catalog. Obviously for accessing large catalogs the first approach is not very well suited. But also the second approach addresses the problem only partly, because the regions have been defined artificially. For very large catalogs the

[1]http://pixela.stsci.edu/gsc/gsc.html

[2]http://www-sdss.fnal.gov:8000

[3]http://astro.estec.esa.nl/SA-general/Projects/GAIA/

region structure should be easily re-producible from basic catalog parameters or, even better, it should be possible to calculate the correct region directly from the main access parameters. Moreover this approach can be hierarchical in the sense that subregions might be accessed by just going one step deeper in the same structure. By generalising such methods it is also possible to produce catalog structures and indexes for accessing a multi dimensional parameter space.

1.1. The Dream

Think of accessing the USNO-A1.0[4] catalog by using a zoom procedure in an online map. The first map shows the whole sky with all objects brighter than a certain limit (e.g., $mag_V \leq 5.0$). In clicking on this map only part of the sky is shown, but the limiting magnitude is now fainter. When getting to the zoom level of about the size of one CCD frame (≤ 100 sq. arcmin.) all objects in this region contained in the catalog are shown. If the upper scenario is possible then think about the implications for astrophysical applications of such a catalog, like e.g., stellar statistics.

1.2. The Reality

Since originally the catalog is structured and sorted by coordinates, any access by magnitudes means reading the whole catalog. Creating a magnitude index means creating a file of about the size of the catalog. Both is unacceptable!

1.3. How to make the Dream Reality

Create small regions of about 100 sq. arcmin sort all objects within these regions by magnitude create an histogram index of these regions, i.e., pointer to first object and number of objects within a magnitude bin. That means less than 30 pointer pairs per region. The last two steps are straightforward but how can the first step be done? There are several ways to create such regions. The HTM or QTM methods (Hierarchical Triangular Mesh, Quaternary Triangular Mesh) is used in the SDSS (Brunner et al. 1994) and GSC-II projects and have already been described at the ADASS (Barrett 1995). In the same paper the 'Bit-Interleaving' method has been proposed for storing astronomical data. Here we describe our experience with 'Bit-Interleaving' and its easy and fast implementation. Bit-Interleaving is just taking the bits of two numbers and creating a new number where every odd bit stems from the first number and every even bit from the second number, i.e., the bits are like the teeth in a zip. The resulting number may be used to sort the catalog in a one-dimensional manner. If the first 14 bits of the bit-interleaved coordinates (in long integer representation) are used as an index, this creates regions of about 71 sq. arcmin. In other words all objects having the same first 14 bits of their bit-interleaved coordinates will be assigned to one region. As already mentioned before the objects in one region may be sorted according to the rest of the bit-interleaved coordinates or, alternatively by magnitudes or any other parameter. The first will produce a hierarchy of subregions, the latter will give the possibility to access the catalog in the magnitude domain. We would like to mention here

[4]http://archive.eos.org/skycat/servers/usnoa

the possibility of building real multi-dimensional structures and indexes, like X (Berchtold et al. 1996), R (Brinkhoff et al., 1996), R* (Beckmann et al., 1990) or k-d-trees (various tree structures and comparisons given in the papers of White et al.). However some of these index-structures are hard to build and they slow down the access to the primary key.

2. Testing Bit Interleaving

We have carried out tests using the complete Hipparcos catalog containing about 120.000 objects all over the sky. All code has been written in IDL. The catalog has been accessed through a disk copy (Sun) and directly from the CD (PC). Even in the worst case, a PC with dual-speed CD-ROM drive, it took only about 15 minutes to produce the whole structure.

3. The Future

It is planned to use an advanced catalog structure/sorting for the export version of the GSC-II catalog. This structure might either be a copy of the HTM structure used in the GSC-II database or a structure produced by bit-interleaving. The export catalog will be produced in a collaboration between STScI and ESO. We will also produce a bit-interleaved version of the USNO-A1.0 catalog to test the concepts described above.

References

P. Barrett, 1995, Application of the Linear Quadtree to Astronomical Databases, in ASP Conf. Ser., Vol. 77, Astronomical Data Analysis Software and Systems IV, ed. R. A. Shaw, H. E. Payne & J. J. E. Hayes (San Francisco: ASP), http://www.stsci.edu/stsci/meetings/adassIV/barrettp.html,

David A. White - Shankar Chatterjee - Ramesh Jain, Similarity Indexing for Data Mining Applications http://vision.ucsd.edu/~dwhite/datamine/datamine.html

David A. White - Shankar Chatterjee - Ramesh Jain, Similarity Indexing: Algorithms and Performance, http://vision.ucsd.edu/papers/sindexalg/

R.J.Brunner, K.Ramaiyer, A.Szalay, A.J.Connolly & R.H.Lupton, 1994, An Object Oriented Approach to Astronomical Databases, in ASP Conf. Ser., Vol. 61, Astronomical Data Analysis Software and Systems III, ed. Dennis R. Crabtree, R. J. Hanisch & Jeannette Barnes (San Francisco: ASP), http://tarkus.pha.jhu.edu/database/papers/adass94.ps

Berchtold S., Keim D. A., & Kriegel H.-P., 1996, The X-Tree: An Index Structure for High-Dimensional Data, Proc. 22th Int. Conf. on Very Large Data Bases, Bombay, India, 28

Brinkhoff T., Kriegel H.-P., & Seeger B., 1996, Parallel Processing of Spatial Joins Using R-trees, Proc. 12th Int. Conf. on Data Engineering, New Orleans, LA

Beckmann N., Kriegel H.-P., Schneider R., & Seeger B., The R*-tree: An Efficient and Robust Access Method for Points and Rectangles, 1990, Proc. ACM SIGMOD Int. Conf. on Management of Data, Atlantic City, NJ, 322

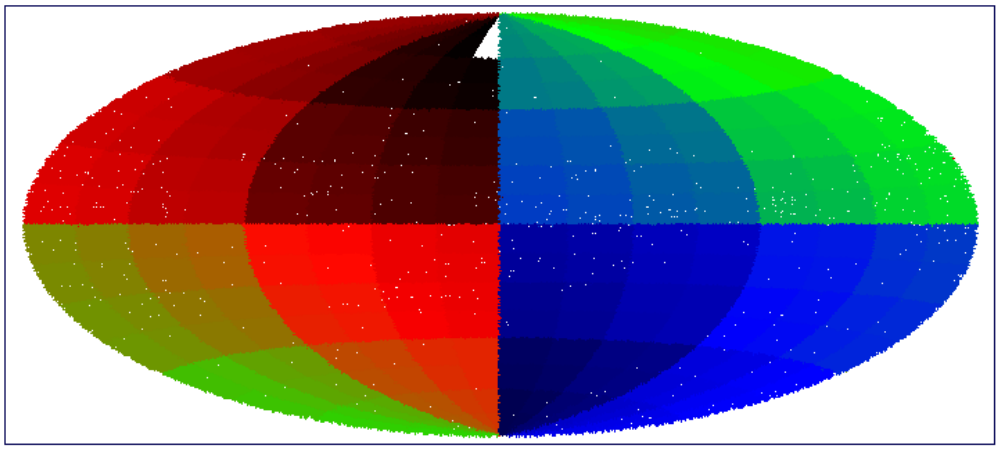

Figure 1. Tessellation of the sky using RA and DEC

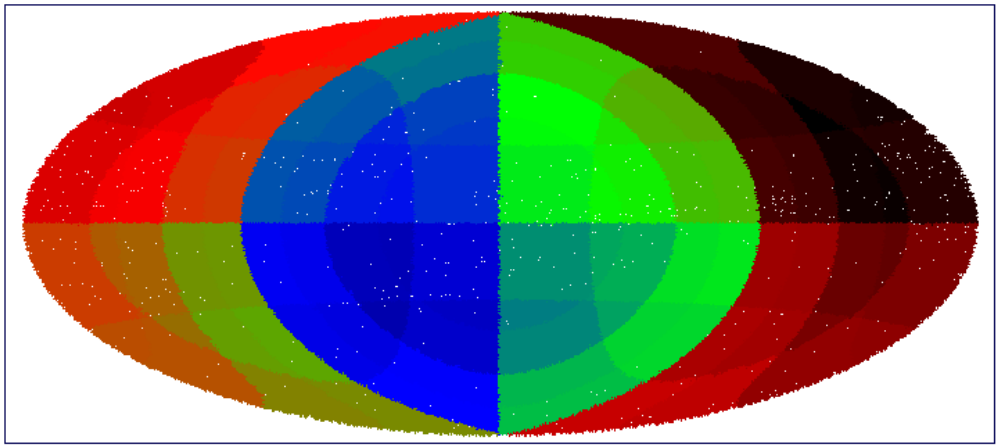

Figure 2. Illustration of a 3-D tessellation using direction cosines