CROSS-MATCHING OF VERY LARGE CATALOGS

M. V. Martynov, D. V. Bodryagin

Research Institution "Mykolaiv Astronomical Observatory"
Observatorna St., 54030, Mykolaiv, Ukraine

ABSTRACT. Modern astronomical catalogs and sky surveys, that contain billions of objects, belong to the "big data" data class. Existing available services have limited functionality and do not include all required and available catalogs. The software package ACrId (Astronomical Cross Identification) for cross-matching large astronomical catalogs, which uses an sphere pixelation algorithm HEALPix, ReiserFS file system and JSON-type text files for storage, has been developed at the Research Institution "Mykolaiv Astronomical Observatory".

Keywords: database, catalogs, virtual observatory.

1. Introduction

Cross-identification is a powerful tool which is used for solving many astrometric and astrophysical problems. Numerous programs and Web services (like CDS X-Match Service http://cdsxmatch.u-strasbg.fr/xmatch) has been developed for this in frames of Virtual Observatory (VO). Unfortunately, their functionality and options are restricted by next reasons:

– the limited list of catalogs which are available for cross-identification;

– rigorously specified object identification algorithms;

– some restrictions and difficulties in cross matching and uploading/downloading of large (hundreds of thousands or more objects) user data sets.

2. Cross-matching of the astronomical data with the software package ACrId (Astronomical Cross Identification)

This software product is a package of console scripts written in the Python programming language. There were implemented the following steps:

1) Data preparation (preprocessing and pixelation of selected catalogs);

2) Cross-identification of the objects;

3) Output and saving of the cross-matching results in the user formats.

Most of astronomical catalogs and surveys are available in two types: text (XML-format VOTable) or binary and have different patterns of records. The first stage is carried out to convert the selected digital catalogs into a common format of the JSON-type text file. This conversion allows you to describe all the different catalogs with help of uniform rules and makes easier adding other catalogs and lists of absolutely various structure, origin and type. Standard form of catalog description includes three files

1) General description;

2) List of files of the catalog (original, before pixelation);

3) Description format.

The next stage of the preparation data of "big data" type to cross-matching consists in partitioning the selected file into separate fragments. The pixelation of the sphere is frequent and efficient solution of many problems because in astronomy we have deal with data distributed on the celestial sphere. Pixelation means subdivision spherical surface on numbered fragments of equal area. Usage of pixelation allows not only to solve the problem of celestial map representation and analysis, but is also essential by constructing databases that require quick search of celestial objects. Pixelation also does possible to use several PC simultaneously for cross-matching of large surveys. There are some systems of pixelation. We used hierarchical grid with equal squares HEALPix (Hierarchical Equal Area isoLatitude Pixelization) (Górski et al., 2005). Pixelation sphere has a number of advantages over the previously used method of the sphere division using the equatorial coordinates. However, there is a technical problem when we divide sphere into pixels to place them on the HDD (if the pixel size is about half an arc minute, then their number will exceed 805 million). Not every system can create so many files. For example, the NTFS file system (OS MS Windows) allows you to create more than 4 billion files, but the creation of empty 805 million files of zero size takes about 320 GB of disk space. You can't also change the maximum number of files available without formatting the partition in the ext3 and ext4 file systems (OS Linux) and there is need more than 200GB (depending on the size of the cluster partition) for the creation of 805 million empty files. It turned out that the best choice for solving the problem is ReiserFS file system, which was designed specifically to work with a lot of small files. There are no restrictions on the maximum available number of system files and 805 million zero size files occupy only 82 GB. After the "zero" pixelation sphere, that is, after the creation of zero

size files, it is necessary to fill in their by respective objects from the catalogs. As noted earlier list of catalogs can be arbitrary because the unified files JSON-type text format were created for all catalogs before pixelation process. The actual process filling the pixels can last from several hours to several weeks depending on the computing capacity and size of the input catalog. It should be noted that this version of the data preparation for cross-identification requires high-capacity HDD for long-term storage of the results of catalogs pixelation. The advantage of this option is lack of necessity to build additional indexes to locate the data quickly, since the role of the search tree plays the file system itself in which files and folders are named and arranged so that they are playing the role of "frozen" fast path to data.

In general, the task of cross-identification astronomical objects in different catalogs is search the same source in some lists of coordinates. The main problems in this case are different epochs of observations and different limiting magnitudes of the catalogs. The first one requires taking account proper motions and their error. This leads to the fact that the use of only coordinate criterion for identification does not always yield successful results. The object may be absent in this neighborhood, or there may be several candidate objects from another catalog. Preliminary pixelation allows us to carry out a so-called "cascade" cross-identification with using more than two catalogs. This enables the construction of additional filters, which ultimately reduces the error of false identifications.

The ACrId software package has a wide range of applications in our observatory. HealPix pixelation of some basic astrometric catalogs, including the catalog of XPM (Fedorov et al., 2009), which is not in the database CDS, have performed. The results are used to calculate and the study of the proper motions of stars.

Usage the software allowed us to obtain the first version of the compiled catalog of stars with high proper motion (HPM) on the whole celestial sphere (about 968,251 objects). The distribution of the stars in the equatorial coordinates is shown in Figure 1.

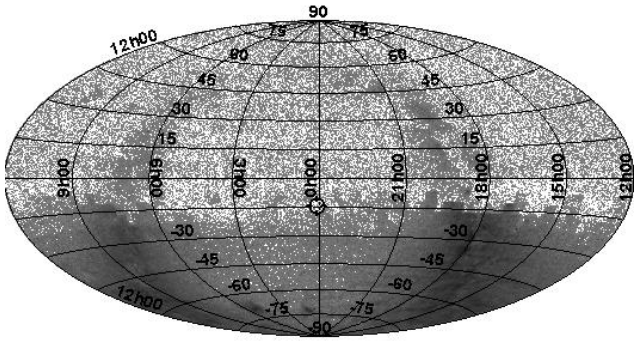

Figure 1: Distribution HPM stars on celestial sphere in equatorial coordinates.

As can be seen from the fig. 1, there are a huge number of HPM stars with proper motion more than 150mas/year on the Southern hemisphere. To validate this result, it is planned to get a second version of the catalog using the represented software with an expanded list of modern celestial catalogs and surveys.

3. Conclusion

The software package of extra-large astronomical catalogs pixelation and cross-identification "ACrId" is used to division the celestial sphere on small plots of equal area (pixels) for the following usage pixel numbering as search index. Program "ACrId" allows you to cross-matching an arbitrary number of catalogs with any format and carry out multi-stage objects cross-identification for the search common objects and determinations their parameters.

References

Górski K. M.: 2005, *AphJ*, **622**, **2**, 759.
Fedorov P.: 2009, *MNRAS*, **393**,133.
http://cdsxmatch.u-strasbg.fr/xmatch