A Cross-matching Service for Data Center of Xinjiang Astronomical Observatory (Postprint)

Authors: Hai-Long Zhang, Jie Wang, Xin-Chen Ye, Wan-Qiong Wang, Jia Li, Ya-Zhou Zhang, Xu Du, Han Wu and Ting Zhang

Date: 2024-02-01T14:50:09+00:00

Abstract

Cross-matching is a key technique to achieve fusion of multi-band astronomical catalogs. Due to different equipment such as various astronomical telescopes, the existence of measurement errors, and proper motions of the celestial bodies, the same celestial object will have different positions in different catalogs, making it difficult to integrate multi-band or full-band astronomical data. In this study, we propose an online cross-matching method based on pseudo-spherical indexing techniques and develop a service combining with high performance computing system (Taurus) to improve cross-matching efficiency, which is designed for the Data Center of Xinjiang Astronomical Observatory. Specifically, we use Quad Tree Cube to divide the spherical blocks of the celestial object and map the 2D space composed of R.A. and decl. to 1D space and achieve correspondence between real celestial objects and spherical patches. Finally, we verify the performance of the service using Gaia 3 and PPMXL catalogs. Meanwhile, we send the matching results to VO tools-Topcat and Aladin respectively to get visual results. The experimental results show that the service effectively solves the speed bottleneck problem of cross-matching caused by frequent I/O, and significantly improves the retrieval and matching speed of massive astronomical data.

Full Text

Preamble

Research in Astronomy and Astrophysics, 24:015008 (9pp), 2024 January

https://doi.org/10.1088/1674-4527/ad08e8

A Cross-matching Service for Data Center of Xinjiang Astronomical Observatory

Hai-Long Zhang[1,2,3], Jie Wang[1,3], Xin-Chen Ye[1,3], Wan-Qiong Wang[1], Jia Li[1], Ya-Zhou Zhang[1,4], Xu Du[1,4], Han Wu[1,4], and Ting Zhang[1,4]

[1] Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi 830011, China; zhanghailong@xao.ac.cn, wangjie@xao.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Received 2023 June 21; revised 2023 October 12; accepted 2023 November 1; published 2023 December 13

Abstract

Cross-matching is a key technique for achieving the fusion of multi-band astronomical catalogs. Due to the use of different equipment such as various astronomical telescopes, the existence of measurement errors, and the proper motions of celestial bodies, the same celestial object will have different positions in different catalogs, making it difficult to integrate multi-band or full-band astronomical data. In this study, we propose an online cross-matching method based on pseudo-spherical indexing techniques and develop a service combining it with a high-performance computing system (Taurus) to improve cross-matching efficiency, designed specifically for the Data Center of Xinjiang Astronomical Observatory. Specifically, we use Quad Tree Cube to divide the spherical blocks of celestial objects and map the 2D space composed of R.A. and decl. to 1D space, achieving correspondence between real celestial objects and spherical patches. Finally, we verify the performance of the service using Gaia DR3 and PPMXL catalogs, while sending the matching results to VO tools—Topcat and Aladin respectively to obtain visual results. The experimental results show that the service effectively solves the speed bottleneck problem of cross-matching caused by frequent I/O and significantly improves the retrieval and matching speed of massive astronomical data.

Key words: virtual observatory tools –astronomical databases: miscellaneous –catalogs

1. Introduction

Cross-matching calculation forms the basis for fusing multi-band astronomical observations and represents a key technique for multi-band astronomy research. It enables the fusion of astronomical data from different bands to obtain multi-band or all-band data, which benefits astronomers in revealing celestial information and making better use of various data in catalogs for scientific research \cite{Yu_{2019}}. With the rapid development of astronomical technology, many countries have built or plan to build telescopes covering multiple bands. For instance, in the radio band: Square Kilometre Array \cite{Dewdney_{2008}}, Five-hundred-meter Aperture Spherical

radio Telescope \cite{Nan_{2011}}, Robert C. Byrd Green Bank Telescope \cite{Prestage_{2009}}, and the upcoming QTT (QiTai radio Telescope) currently under construction \cite{Wang_{2023}, Zhang_{2023a}}. In the optical band: European Extremely Large Telescope \cite{Gilmozzi_{2007}}, Large Synoptic Survey Telescope \cite{Zhan_{2018}}, Large sky Area Multi-Object fiber Spectroscopic Telescope \cite{Cui_{2012}}. In other bands: Lunar-based Ultraviolet Telescope \cite{Wang_{2015}}, Cherenkov Telescope Array \cite{Acharya_{2017}}, extended ROentgen Survey with an Imaging Telescope Array \cite{Predehl_{2021}}. It is evident that astronomy has entered the big data and full-band era \cite{Cui_{2020}}, and the measurement errors of various astronomical telescopes have led to different data obtained from observing the same celestial object, causing difficulties in integrating multi-band or full-band astronomical data.

The Data Center of Xinjiang Astronomical Observatory (XAO-DC) was built in 2015 \cite{Zhang_{2022}}. Its main data sources include the Nanshan 26 m Radio Telescope (NSRT; \cite{Xu_{2018}}) and the Nanshan One-meter Wide-field Telescope (NOWT; \cite{Bai_{2020}}). It provides online retrieval services for pulsar, molecular spectrum, active galactic nuclei, and NOWT datasets \cite{Zhang_{2019}}. To facilitate astronomers in better using data from astronomical catalogs for scientific research, we have developed a cross-matching service for XAO-DC.

The features of our developed service can be summarized as follows. First, we implement pseudo-spherical sky partitioning, which divides the whole sky sphere into approximately 6×430 equal blocks to accurately locate required data and reduce unnecessary data reading, thereby reducing disk I/O overhead. Second, the service improves cross-matching speed by using pseudo-spherical index technology and parallel computing techniques, making the time consumption for cross-matching of tera-scale astronomical catalogs minimal. Third, experimental results show that our online cross-matching service achieves 4 trillion cross-matching computation results in less than one second.

The rest of this paper is organized as follows. Section 2 introduces related works on cross-matching calculations. Section 3 presents the developed cross-matching service, which is the core of this paper, in detail. Section 4 tests real astronomical catalogs and verifies the experimental results. Finally, Section 5 concludes the paper.

2.1. Related Work

Astronomical catalogs contain a variety of celestial parameters, collecting data obtained by telescopes during specific periods of astrometry. Nowadays, computer experts in many countries are studying methods for astronomical catalog cross-matching and have developed various tools and algorithms. Budavari & Szalay \cite{Budavari_{2007}} have nicely formulated cross-matching in a Bayesian framework for improving speed, providing a solid theoretical founda-

tion while improving recall and precision. Pineau et al. \cite{Pineau_{2011}} have developed an efficient and scalable cross-matching service for (very) large catalogs that supports customized cross-matching operations. VizieR \cite{Ochsenbein_{2000}}, designed by the Centre de Données de Strasbourg (CDS), includes cross-matching of astronomical observations and large catalogs, which can be performed by uploading directory files and astronomical catalogs in the tool. SIMBAD \cite{Wenger_{2000}} provides multi-source queries for small files of astronomical catalogs based on cross-matching of small astronomical catalogs, with many different options selectable during cross-matching such as source type. Xmatch \cite{Budavari_{2013}} is one of a wide range of cross-matching tools that integrates datasets from many observatories such as 2MASS, GSC, GALEX, UCAC, WISE, etc., providing functions such as download, query, and integration of astronomical tables. ARCHES \cite{Motch_{2016}} is a cross-matching service for high-energy astrophysics research that provides multi-band data with complete characteristics in the form of spectral energy distributions. Astronomers can submit their own retrieval scripts through an HTTP API, and the system will send astronomers the cross-matching results after the script is run. catsHTM \cite{Soumagnac_{2018}} uses HTM index to store hierarchical astronomical catalogs in HDF5 files, integrates DECaLS/DR5, FIRST, Gaia/DR1, Gaia/DR2, GALEX/DR6Plus7 and other datasets, and can support cross-matching between dozens of astronomical catalogs.

To speed up cross-matching calculations, Pei et al. \cite{Pei_{2011}} greatly improved cross-matching speed using Python multi-core parallel methods. Zhao et al. \cite{Zhao_{2009}} used HEALPix to divide astronomical catalogs, combined with bit operation fast indexing, and controlled the cross-matching time of large-scale astronomical catalogs within 32 minutes. Du et al. \cite{Du_{2014}} combined two partition indexing methods, HTM and HEALPix, and used thread pool technology to accelerate cross-matching time, reducing the cross-matching time of large-scale astronomical catalogs to 23 minutes and controlling medium-sized catalog cross-matching to 7 minutes. Ma et al. \cite{Ma_{2018}} proposed the E-Zone algorithm, which uses Euclidean distance for faster calculation of adjacent points and implements parallel calculation based on OpenMP. Li et al. \cite{Li_{2019}} designed a multi-band catalog unified format combined with a minimum-conflict data layout strategy to improve cross-matching parallelization, achieving 30.3% and 30.7% time reduction compared with Quad Tree Cube (Q3C) and Healpix-tree-C (H3C) at 200 million data sources. Zhang et al. \cite{Zhang_{2023b}} proposed a large-scale cross-matching framework supporting heterogeneous computing, which reduced cross-matching time to 5 seconds for small-scale catalogs, 150 seconds for medium-scale catalogs, and 260 seconds for large-scale catalogs.

2.2. The Cross-matching Based on Celestial Coordinates

Cross-matching calculations of astronomical catalogs can combine various information such as location, density, luminosity, wavelength, and so on. We choose to combine with celestial coordinates because catalogs obtained by different telescopes all contain information about the location of celestial sources. Therefore, we can determine whether two catalogs are homologous or non-homologous by comparing celestial coordinate information. As shown in Figure 1 [Figure 1: see original paper], points A and B come from astronomical catalogs A and B, respectively. When the spherical distance $d < r_1 + r_2$ (in theory), where r_1 and r_2 are the error radii of the two catalogs, the two points are successfully matched as the same object.

When implemented on the web side, we provide search radius options, allowing users to enter a matching radius according to actual needs. The output condition is that the distance between two points in the input catalog and the matching catalog is less than the search radius.

3.1. The Overall Design of the Service

We develop an online cross-matching service based on the DaCHS Virtual Observatory \cite{Demleitner_{2014}} for massive astronomical catalogs in XAO-DC. The overall structure of the service is shown in Figure 2 [Figure 2: see original paper]. The service is decomposed into (from top to bottom) a data layer, a calculation layer, and an output layer.

The data layer allows astronomers to upload astronomical catalogs that need cross-matching in two ways: via remote URL or local upload as VOTable files. We provide three methods to obtain archived astronomical catalogs from XAO-DC: Web interface, VO tools, and Python scripts. By April 2023, we have archived 20 astronomical data catalogs, including catalogs of pulsars, molecular spectra, and active galactic nuclei from NSRT and catalogs from the One-Meter Telescope of NOWT. All astronomical catalogs in XAO-DC are backed up at the headquarters of XAO and Nanshan station.

The calculation layer is the core part of the entire service, which uses parallel computing techniques for cross-matching calculations. We use celestial coordinates to calculate the angular distance between two astronomical catalogs. Theoretically, when the angular distance $d < r_1 + r_2$ for cross-matching calculation, where r_1 and r_2 are the error radii of the two catalogs, the matching of astronomical catalogs is successful. In practice, we calculate d in terms of $d <$ search radius. To improve cross-matching speed, we use a high-performance computing system built in 2016 named Taurus \cite{Zhang_{2018}}.

The output layer provides a variety of output formats such as CSV, HTML, FITS, JSON, etc. Astronomers can output and download cross-matching results according to actual scientific needs. Through the Simple Application Messaging Protocol (SAMP), the results obtained by cross-matching are sent to standard

virtual observatory tools to integrate data visualization and other related tools, supporting astronomers in customizing processing of cross-matching calculations and completing the entire process of scientific research and analysis online.

3.2. Indexing Strategy for Astronomical Catalogs

We use Q3C index technology \cite{Koposov_{2006}} to improve retrieval efficiency, which is designed for the PostgreSQL open-source database. There are several reasons for using Q3C: First, it is optimized for cone search, cross-matching, and other technologies because it uses central projection to reduce extensive trigonometric function calculations, thus reducing search time. Second, it is an open-source solution that can be downloaded from http://sourceforge.net/projects/q3c. Third, it guarantees the best I/O performance for retrieving data from the database.

As shown in Figure 3 [Figure 3: see original paper], we assume the celestial sphere is a cube, construct a quadtree on each face of the cube, and use the quadtree structure to generate two-dimensional coordinate codes (or positive integer codes). Since the initial cube has six faces, the mapping to faces can be encoded using a 3-bit binary number. This partition is easily implemented by projecting the surface center of the cube onto the sphere, and the quadtree structure can be automatically inherited by the sphere. Ultimately, the sphere is divided into several quadrilaterals by different levels of partition.

4.1. Archived Astronomical Catalogs for Cross-matching Service in XAO-DC

We have completed the archiving of observation data from NSRT and NOWT, including four datasets: pulsar dataset, molecular spectral line dataset, active galactic nuclei dataset, and NOWT dataset (see Table 1 for details of each dataset). Larger catalogs that can be matched against include Gaia, 2MASS, USNO-B, PPMXL, and more. We use a server with Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz $×$2, 256 GB memory, 4 TB $×$2 SSD, and 16 TB $×$60 SATA for online cross-matching experiments.

4.2. Use Case for Cross-matching Service

4.2.1. Input Fields

As shown in Table 2 , the following fields are available to provide input to the service. The uploaded VOTables must have exactly one pair of columns with UCDs of either `pos.eq.[ra|dec];meta.main` or `POS_{EQ}[RA|DEC]_{MAIN}`. The results of VO cone searches work well. If users have tables of their own, they must first bring them to the VOTable format. We currently do not support coordinate transformation, so users must ensure that the input coordinates match the system used in the table (for basically all of our tables, this means ICRS or FK5 J2000 to an accuracy sufficient for matching). We provide an

experimental use case, shown in Table 3 , that can be tested through the URL: http://data.xao.ac.cn/cross/q/match/form.

4.2.2. Output Result

We obtain the following matched data and corresponding parameter information, including R.A. [deg], decl. [deg], E_{raepra} [deg], etc., as shown in Table 4 for details of the cross-matching results. The cross-matching between ppmxl.main (1 billion targets) and the test catalog (4000 targets) takes less than one second. This means our online cross-matching service achieves 4 trillion cross-matching computation results in less than one second. As far as we know, Gao et al. (2008) took 407 minutes (811117 × 470992970); Zhao et al. (2009) took 32 minutes (470992970 × 100106811); Pei et al. (2011) took 10 minutes (470992970 × 100106811); Du et al. (2014) took 7 minutes (946464 × 470992970). Because there is no online platform for testing the methods implemented in the above works, it is impossible to achieve the same scale of cross-matching as in this paper.

4.2.3. The Influence of Search Radius

As shown in Figure 4 [Figure 4: see original paper], we exhibit the relationship among the number of matched objects, browser response time, and browser response size. As the search radius diminishes (from 0°.010 to 0°.001), the number of matched objects (from 12,997 to 757), browser response time (from 13,640 to 503 ms), and browser response size (from 1228.8 to 89.5 KB) decrease accordingly. This shows that the smaller the search radius, the more accurate the cross-matching results.

4.3. Use Case for Cross-matching Joint Virtual Observatory Tools

In practice, astronomers should not use a web browser for cross-matching. Instead, they should obtain a TAP client (e.g., TOPCAT or pyVO), load the table to be matched into the client, and then run a query \cite{Dowler_{2019}}. Since the backend is the same, the performance characteristics are identical to the browser service discussed above.

4.3.1. TOPCAT

TOPCAT is a browser and editor that can interactively graph tables of astronomical data in major formats such as FITS and VOTable. To facilitate astronomer data analysis, we can send cross-matching results to TOPCAT through the SAMP protocol, as shown in Figure 5 [Figure 5: see original paper].

4.3.2. Aladin

Aladin is free, interactive astronomy software that enables astronomers to interactively retrieve digitized astronomical images from astronomical catalogs of all known celestial objects, such as Simbad and VizieR, and visually compare them with DSS, PanSTARRS, and other astronomical catalogs. To facilitate astronomer data analysis, we can send cross-matching results to Aladin through the SAMP protocol, as shown in Figure 6 [Figure 6: see original paper].

5. Conclusion

In this paper, we proposed an online cross-matching method based on pseudo-spherical indexing techniques and developed a service combining it with Taurus for XAO-DC to improve cross-matching efficiency. This service supports two source table file input modes: local upload and URL. File input supports the standard VOTable format and realizes cross-matching calculations between uploaded astronomical catalogs and released astronomical catalogs in XAO-DC. The service supports HTML, CSV, FITS, JSON, and other data output modes, and integrates necessary visualization tools (such as TOPCAT, Aladin, etc.) according to the related protocols of the virtual observatory to support processing and customization of data after cross-matching. The service provides astronomers with reliable and convenient technical support intended to help them further their astronomical research.

Acknowledgments

This work is supported by the National Key R&D Program of China (Nos. 2022YFF0711502 and 2021YFC2203502); the National Natural Science Foundation of China (NSFC) (12173077 and 12003062); the Tianshan Innovation Team Plan of Xinjiang Uygur Autonomous Region (2022D14020); the Tianshan Talent Project of Xinjiang Uygur Autonomous Region (2022TSYCCX0095); the Scientific Instrument Developing Project (grant No. PTY-Q2022YZZD01); the China National Astronomical Data Center (NADC) of the Chinese Academy of Sciences; the Operation, Maintenance and Upgrading Fund for Astronomical Telescopes and Facility Instruments, budgeted from the Ministry of Finance of China (MOF) and administered by the Chinese Academy of Sciences (CAS); and the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01A360). This work is supported by the Astronomical Big Data Joint Research Center, co-founded by the National Astronomical Observatories, Chinese Academy of Sciences.

ORCID iDs
https://orcid.org/0000-0003-0380-6395 Jie Wang
https://orcid.org/0000-0001-6448-0822 Xu Du

References

Acharya, B. S., Agudo, I., Al Samarai, I., et al. 2017, *Science with the Cherenkov Telescope Array* (Hackensack, NJ: World Scientific)
Bai, C.-H., Feng, G.-J., Zhang, X., et al. 2020, *RAA*, 20, 211
Budavari, T., & Lee, M. A., 2013 Xmatch: GPU Enhanced Astronomic Catalog Cross-Matching, Astrophysics Source Code Library, ascl:1303.021
Budavari, T., & Szalay, A. S. 2007, *ApJ*, 679, 301
Cui, C., Tao, Y., Li, C., et al. 2020, *A&C*, 32, 100392
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, *RAA*, 12, 1197
Demleitner, M., Neves, M. C., Rothmaier, F., & Wambsganss, J. 2014, *A&C*, 7, 27
Dewdney, P. E., Hall, P. J., Schilizzi, R. T., & Lazio, T. J. L. W. 2008, *Proc. IEEE*, 97, 1482
Dowler, P., Rixon, G., Tody, D., & Demleitner, M. 2019, IVOA Recommendation
Du, P., Ren, J., Pan, J., & Luo, A. 2014, *SCPMA*, 57, 577
Gilmozzi, R., & Spyromilio, J. 2007, *Msngr*, 127, 3
Gao, D., Zhang, Y. X., & Zhao, Y. H. 2008, *AcASn*, 49, 348
Koposov, S., & Bartunov, O. 2006, *adass XV*, 351, 735
Li, B., Yu, C., Li, C., et al. 2019, *PASP*, 131, 054501
Ma, X., Du, Z., Sun, Y., et al. 2018, in *Computational Science–ICCS 2018: 18th Int. Conf., Wuxi, China, June 11–13, Part III 18* (Berlin: Springer), 473
Motch, C., Carrera, F., Genova, F., et al. 2016, arXiv:1609.00809
Nan, R., Li, D., Jin, C., et al. 2011, *IJMPD*, 20, 989
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, *A&AS*, 143, 23
Pei, T., ZHANG, Y., PENG, N., & ZHAO, Y. 2011, *SSPMA*, 41, 102
Pineau, F.-X., Boch, T., & Derriere, S. 2011, *adass XX*, 442, 85
Predehl, P., Andritschke, R., Arefiev, V., et al. 2021, *A&A*, 647, A1
Prestage, R. M., Constantikes, K. T., Hunter, T. R., et al. 2009, *Proc. IEEE*, 97, 1382
Soumagnac, M. T., & Ofek, E. O. 2018, *PASP*, 130, 075002
Wang, J., Wu, C., Qiu, Y., et al. 2015, *P&SS*, 109, 123
Wang, N., Xu, Q., Ma, J., et al. 2023, *Sci. China-Phys. Mech. Astron.*, 66, 289512
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, *A&AS*, 143, 9
Xu, Q., Li, L., & Wang, N. 2018, *Proc. SPIE*, 10700, 107002W
Yu, C., Li, B., Xiao, J., Sun, C., & Fan, D. 2019, *ExA*, 47, 359
Zhan, H., & Tyson, J. A. 2018, *RPPh*, 81, 066901
Zhang, H., Demleitner, M., Wang, J., et al. 2019, *AdAst*, 2019, 5712682
Zhang, H., Wang, J., Demleitner, M., et al. 2022, *A&C*, 39, 100578
Zhang, H., Wang, J., Tang, K., et al. 2018, in *2018 Int. Conf. on Sensing, Diagnostics, Prognostics, and Control (SDPC)* (Piscataway, NJ: IEEE), 705
Zhang, H. L., Zhang, Y. Z., Zhang, M., et al. 2023a, *RAA*, 23, 125023
Zhang, Y., Yu, C., Sun, C., et al. 2023b, *MNRAS*, 519, 6381
Zhao, Q., Sun, J., Yu, C., et al. 2009, in *Algorithms and Architectures for Par-*

allel Processing: 9th Int. Conf., ICA3PP 2009, Taipei, Taiwan, June 8–11 (Berlin: Springer), 604

Note: Figure translations are in progress. See original paper for figures.