Xinjiang Astronomical Observatory Online Cross-Identification Service Postprint

Authors: Zhang Hailong, Nie Jun, Zhao Qing, Ye Xinchen, Wang Jie

Date: 2017-09-26T00:00:00+00:00

Abstract

The massive astronomical catalog online cross-matching service of the Xinjiang Astronomical Observatory Data Center supports two catalog data input modalities: remote URLs and local uploads of VOTable format files containing UCD information, enabling cross-matching with astronomical data previously published by the data center. The resulting cross-matches can be transmitted via the SAMP protocol to standard Virtual Observatory tools for data visualization and related processing, and the service supports multiple data output formats including HTML, CSV, FITS Table, and JSON. The cross-matching speed for massive astronomical catalogs has been substantially enhanced through parallel computing technology and pseudo-spherical sky partitioning technology.

Full Text

Preamble

Vol. 14 No. 3 Jul. 2017 ASTRONOMICAL RESEARCH AND TECHNOLOGY 53-1189 1672-7673

Online Cross-Matching Service of Xinjiang Astronomical Observatory

Hai-Long Zhang[1,2], Xin-Chen Ye[1], Jie Wang[1], Jun Nie[1]

[1] Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi, Xinjiang
[2] Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Nanjing, Jiangsu
[3] Tianjin University of Science and Technology, Tianjin

Abstract

The online cross-matching service for massive catalog data at the Xinjiang Astronomical Observatory (XAO) Data Center enables remote cross-identification

of astronomical data published by the center. The service accepts two input methods for catalog data: local file upload and remote URL. Results from cross-matching can be sent to standard Virtual Observatory tools via the Simple Application Messaging Protocol (SAMP) for data visualization and further processing. The service supports multiple output formats including VOTable, FITS Table, and others. Through parallel computing technology and pseudo-spherical sky partitioning techniques, the cross-matching speed for massive catalog data has been significantly improved.

Keywords: Data center; Virtual Observatory; Cross-matching; Catalog data

1. Introduction

With the rapid development of information technology and manufacturing technology, astronomy has entered the era of all-sky surveys across multiple bands. Multi-wavelength observational data from different astronomical facilities are growing exponentially. How to effectively fuse massive astronomical data and study the characteristics of celestial objects across wavelengths has become an urgent problem in current astronomical research. Cross-matching computation forms the foundation of multi-wavelength astronomical data fusion and is a prerequisite for multi-wavelength astronomical studies. Cross-matching operations are typical data-intensive computations.

In recent years, computer scientists from multiple countries have conducted systematic research on cross-matching and proposed effective solutions. Turing Award winner Jim Gray, who served as the chief scientist for cross-matching issues at the US Virtual Observatory[1], pointed out that solving cross-matching problems must rely on parallel computing technology [1]. He designed a cross-matching service for the US Virtual Observatory based on Microsoft SQL Server' s zone algorithm [2-5], which integrated datasets from multiple observatories for the Sloan Digital Sky Survey (SDSS) data access platform. However, due to hardware performance limitations a decade ago and implementation constraints of the MS SQL Server database system, the data scale for cross-matching was relatively small. AstroGrid[2] provides a simple cross-matching service on its website [6-8], but it is inefficient and cannot achieve large-scale data cross-matching. Currently, major astronomical data centers such as VizieR, Simbad, Aladin, and NED provide their own cross-matching services, but with restrictions on the number of sources per matching operation.

China has built numerous astronomical scientific facilities in recent years, and cross-matching technology has become essential in observation target selection and data processing. The core of the spectral confirmation process for the Guoshoujing Telescope (LAMOST), a major national scientific project, is cross-matching. In this context, Chinese researchers have achieved results in efficient cross-matching. For instance, a fast cross-matching algorithm based on KD-Tree and Python was proposed, with efficiency suitable for medium data volumes from hundreds of thousands to millions of records [9-11]. Multi-core parallel methods have significantly improved cross-matching speed, and Bayesian hypothesis testing methods have been applied to radio catalog cross-matching with good results in SWIRE and ATLAS CDF-S catalog identification.

Cross-Matching Principle

For two points A and B on a sphere with coordinates (α_A, δ_A) and (α_B, δ_B) from catalogs A and B respectively, the angular distance d between them can be calculated as follows:

When |α_A - α_B| $\leq$ 180°: $\Delta\alpha$ = α_A - α_B

Otherwise: $\Delta\alpha$ = 360° - |α_A - α_B|

According to the spherical cosine theorem: cos(d) = sin(δ_A)sin(δ_B) + cos(δ_A)cos(δ_B)cos($\Delta\alpha$)

When the angular distance between two points is very small, the following approximation holds: d^2 $(\Delta\alpha \cos \delta)^2 + (\Delta\delta)^2$

The criterion for successful identification is: $d^2 \leq$ (r_A + r_B)^2

where r_A and r_B are the error radii of the two catalogs. When the distance between two points satisfies this condition, they can be considered successfully matched as corresponding objects.

[Figure 1: see original paper] Cross-match principle

GAVO Framework and IVOA Standards

The Xinjiang Astronomical Observatory Data Center is built upon the framework of the German Astrophysical Virtual Observatory (GAVO)[7]. GAVO' s implementation follows the standards and protocols of the International Virtual Observatory Alliance (IVOA)[8] and represents one of German astronomers' contributions to extending and utilizing the Virtual Observatory. The main functions of the Virtual Observatory include: enabling or improving the publication and retrieval of astronomical data such as astrometry and time series through well-defined standards and protocols; using standard data retrieval and query methods to help astronomers easily discover, access, and understand data; ensuring proper description, access, and use of relevant observational data; and providing standard Virtual Observatory software to assist astronomers in obtaining and analyzing data.

Pseudo-Spherical Partitioning Techniques

Pseudo-spherical indexing essentially divides the celestial sphere into blocks using specific geometric shapes. During index construction, all blocks on the celestial sphere are systematically coded based on encoding or coordinate information, dividing the sphere into N equal-area or variable-area portions and sorting the codes. The most widely used pseudo-spherical indexing methods currently are: Hierarchical Triangular Mesh (HTM)[9], HEALPix (Hierarchical Equal Area isoLatitude Pixelisation)[10], and Q3C (Quad Tree Cube)[11].

Q3C is a new pseudo-spherical indexing method specifically designed as an indexing scheme for cross-matching and other spatial searches in the open-source PostgreSQL database. Its source code is available from the project website[12]. Q3C' s sky partitioning method differs from HTM and HEALPix, using quadrilateral division on a pseudo-sphere to partition the celestial sphere. The method imagines the sphere as a cube, constructs a quadtree on each face of the cube, and uses the quadtree structure to generate two-dimensional coordinate codes. Since the initial cube has only six faces, a 3-bit binary number can encode the mapping relationship with the faces, and the quadtree structure is automatically inherited by the sphere. The sphere' s surface is ultimately divided into multiple quadrilateral faces.

This partitioning has two advantages: (1) The mapping between the sphere and the cube surface is simply a central projection[13], so the calculation method is simpler than HEALPix, using fewer trigonometric operations. The region uses a quadtree structure and special table query acceleration algorithms, maintaining good efficiency even with deep hierarchical partitioning. (2) Q3C' s partitioned sky areas are not completely equal, making it a non-equal-area partition, which differs from HEALPix.

The Q3C indexing method maps each point on the sphere to an integer, laying the foundation for creating spherical indexes and enabling fast searches on the sphere. It ensures that the values representing pixels near a certain point are continuous. To effectively utilize the index, each query first calculates the corresponding block based on the pre-matched right ascension and declination to obtain the appropriate partition position. Once a small sky region is determined, data meeting the conditions can be quickly retrieved from the database.

[Figure 2: see original paper] Q3C Sphere Segmentation
[Figure 3: see original paper] Q3C Cone Query

Since Q3C indexing technology is designed for the open-source PostgreSQL database and optimized for cone search and cross-matching, and because it uses central projection to reduce extensive trigonometric calculations, thereby improving retrieval efficiency, this work ultimately selected Q3C as the indexing scheme after testing.

Implementation

Data Server Configuration

The Xinjiang Astronomical Observatory data server configuration is shown in Table 1. The server handles data archiving, downloading, and various computation-related tasks. The cross-matching service is one of many services provided by the data server.

Configuration information of data servers

Component	Specification
OS	Linux
CPU	Intel® Xeon® E5-2692 v2 @ 2.20 GHz
Memory	300 GB
Chipset	Intel Corporation C600/X79 series
Network	IB Card 56 Gbps
Storage	SATA

Cross-Matching Data Source Formats

The Xinjiang Astronomical Observatory Data Center cross-matching service URL is: http://data.xao.ac.cn/cross/q/match/form. The service name is "XAO DC Custom Uploading Crossmatcher." The service can be accessed through the Xinjiang Astronomical Observatory Data Center link.

The service accepts two methods for uploading catalog data: local files and remote URLs. For local VOTable files, direct upload is supported. For files on remote servers that meet the format requirements, the URL can be provided directly. The file format can be referenced at http://data.xao.ac.cn/static/cross_{match}.

The cross-matching service accepts files that must strictly conform to the VOTable[14] format, which must specify the right ascension and declination fields using Unified Content Descriptors (UCD). The format is as follows:

```
<?xml version='1.0'?>
<VOTABLE version="1.3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xmlns="http://www.ivoa.net/xml/VOTable/v1.3"
         xmlns:stc="http://www.ivoa.net/xml/STC/v1.30">
  <RESOURCE name="crossMatchCatalog">
    <TABLE name="cross_{match}" nrows="5">
      <DESCRIPTION>Only RA & DEC needed in the table.</DESCRIPTION>
      <FIELD datatype="double" name="ra" ucd="pos.eq.ra;meta.main"/>
      <FIELD datatype="double" name="dec" ucd="pos.eq.dec;meta.main"/>
      <DATA>
        <TABLEDATA>
          <TR>
            <TD>336.5396994</TD>
            <TD>-29.9669121</TD>
          </TR>
          <TR>
            <TD>340.8337065</TD>
            <TD>-34.8434972</TD>
          </TR>
```

```
        <TR>
          <TD>340.8296062</TD>
          <TD>-34.4649278</TD>
        </TR>
        <TR>
          <TD>340.8304808</TD>
          <TD>-34.4992970</TD>
        </TR>
        <TR>
          <TD>340.0254577</TD>
          <TD>-30.8180950</TD>
        </TR>
      </TABLEDATA>
    </DATA>
  </TABLE>
 </RESOURCE>
</VOTABLE>
```

The TABLEDATA fields in the example represent the positions of sources to be cross-matched, requiring at least the right ascension and declination of each source. In actual matching, the format can be modified as needed.

Cross-Matching Web Interface

The Xinjiang Astronomical Observatory Data Center cross-matching service[15] currently only supports newer browsers with compatibility mode disabled. The web interface consists of several components: the left menu contains links and basic information about the cross-matching service; the upper right section provides service instructions and information about allowed upload files; the "Tables available for ADQL" link allows viewing all table information supported by the data center; and the "service info" link provides specific information about the cross-matching service.

The service interface includes the following fields: Local file, Remote URL, Target Table, Search radius, Table Limit to, and Output format. Local file and Remote URL represent the two input source options, supporting local VOTable file uploads and remote VOTable URLs. Target Table allows selecting catalogs from those published at the data center for matching. Search radius should be determined based on the error radii of the two catalogs being matched, with reference to the formula in Section 2. Table Limit to represents the number of output rows displayed in the browser after successful matching, which should not be set too high as large data returns severely affect browser response time. Users can adjust this value as needed. Output format represents the data format for successful matches, supporting HTML, Text, VOTable, JSON, FITS Table, and CSV formats.

[Figure 4: see original paper] Cross-match page

After determining all parameters, clicking the "Quick Plot" button executes the cross-matching. Results can be sent to standard Virtual Observatory tools [14] such as TOPCAT via SAMP for data reprocessing and visualization. Users can select fields and point styles for plotting.

[Figure 5: see original paper] Cross-match Results
[Figure 6: see original paper] Results visualization

2. Determining Search Radius

The search radius should be determined based on the error radii of the two catalogs being matched. The criterion is given by the formula in the Cross-Matching Principle section. When the distance between two points satisfies $d^2 \leq (r\_A + r\_B)^2$, they can be considered successfully matched.

Performance Comparison

Recent domestic research results in cross-matching are detailed in references [11-13, 15]. Since the methods provided in these papers did not offer online testing platforms, data were directly cited from the literature for comparison. Specific test results are shown in Table 2.

Cross-match results comparison

Method	Catalog Size	Time	Rate (records/min)
Gao et al. (HTM, 10 cores)	50,000	10.3 s	290,000
Pei et al. (HTM, 6 cores)	20,000	4.2 s	285,000
Pei et al. (Q3C, 29 cores)	20,000	1.4 s	857,000
Zhao et al. (HTM, 8 cores)	20,000	3.8 s	315,000
HEALPix	20,000	5.1 s	235,000
Q3C (29 cores)	20,000	1.3 s	923,000

Experimental Results Analysis

For cross-matching at the 10,000-20,000 record scale, the time returned is approximately 0.001 ms, which greatly accelerates the matching speed. Comprehensive analysis of Table 2 shows that the online cross-matching platform implemented in this work is significantly more efficient than comparable domestic results.

Due to server load limitations, matching time may reach the second level when source catalog data exceeds 100 million records or data files exceed 20 MB, depending on upload time and network bandwidth. During local server testing, browser return times for both operations were on the order of seconds. Test files are provided for peer evaluation, named cross_{match} and cross_{{{match}}_{{20000}}}.

Conclusion

This work implements the first online cross-matching platform in China. The Xinjiang Astronomical Observatory Data Center online cross-matching service supports both local upload and VOTable file input as source catalogs. All catalogs published at the data center can serve as target catalogs for cross-matching. Through pseudo-spherical sky partitioning technology and parallel computing, cross-matching speed has been greatly improved. The service provides a search radius option for peer testing and supports multiple output formats for identification results. Cross-matching between 10,000-level and 100-million-level catalogs consumes time on the millisecond scale.

Acknowledgments

We thank members of the Astronomy Science Cloud project for their support in constructing the Xinjiang Astronomical Observatory Data Center. Data preprocessing and testing were completed on the Taurus high-performance computing system at Xinjiang Astronomical Observatory.

References

1. Nieto-Santisteban M A, Thakar A R, Szalay A S. Cross-matching very large datasets [C/OL]. https://esto.nasa.gov/conferences/nstc2007/papers/Nieto-Santisteban_{{{Maria}}*{{A10P2}}}*{NSTC}-07-0074.pdf.
2. Gray J, Szalay A. A spatial data library [J/OL]. https://arxiv.org/ftp/cs/papers/0701/0701163.pdf.
3. Fekete G. Using table valued functions in SQL Server 2005 to implement a spatial data library [J/OL]. https://arxiv.org/ftp/cs/papers/0701/0701163.pdf.
4. Gray J, Nieto-Santisteban M A, Szalay A S. The zones algorithm for finding points-near-a-point or cross-matching spatial datasets [J/OL]. https://arxiv.org/ftp/cs/papers/0701/0701171.pdf.
5. Gray J, Szalay A S, Thakar A R, et al. There goes the neighborhood: Relational algebra for spatial data search [J/OL]. https://arxiv.org/ftp/cs/papers/0408/0408031.pdf.
6. Budavári T, et al. Cross-matching multiple spatial observations and dealing with missing data [J/OL]. https://arxiv.org/ftp/cs/papers/0701/0701172.pdf.
7. Joins S R. Technical report of AstroGrid [J/OL]. http/wiki.astrogrid.org/bin/view/AstroGrid/SpatialInde
8. Zhao Q, Sun J, et al. A paralleled large-scale astronomical cross-matching function [C]//International Conference on Algorithms and Architectures for Parallel Processing. Springer Berlin Heidelberg, 2012: 604-614.
9. Gao D, Zhang Y, Zhao Y. The realization of cross-identification based on huge multi-wavelength catalog data [J]. Publications of the National Astronomical Observatories of China, 2011, 8: 186-193.
10. Gao D, Zhang Y, Zhao Y. A system integrated with query, cross-matching and visualization [J]. Astronomical Research & Technology, 2012, 9: 186-193.
11. Gao D, Zhang Y, Zhao Y. The development and application of the China-VO cross-matching tool [J]. Acta Astronomica Sinica, 2013, 54: 348-358.

12. Pei T, Zhang Y, Peng N, et al. The application of multi-core parallel computing in massive catalog cross-matching [J]. Scientia Sinica Physica, Mechanica & Astronomica, 2014, 44: 102-107.
13. Fan D, Budavari T, Norris R P. Matching radio catalogs with realistic geometry: Application to SWIRE and ATLAS [J]. Monthly Notices of the Royal Astronomical Society, 2015, 451: 1299-1305.
14. Zhang H, Demleitner M, Wang N. Using the Xinjiang Astronomical Observatory pulsar data archive [J]. Astronomical Research & Technology, 2016, 13: 473-480.
15. Zhao Q. Research on efficient astronomical cross-matching for massive data [D]. Tianjin: Tianjin University, 2014.

---

[1] http://www.usvao.org/
[2] http://www.astrogrid.org/
[3] http://vizier.u-strasbg.fr/
[4] http://vizier.u-strasbg.fr/
[5] http://aladin.u-strasbg.fr/
[6] https://ned.ipac.caltech.edu/
[7] http://www.g-vo.org/
[8] http://www.ivoa.net/
[9] http://www.skyserver.org/htm/
[10] http://healpix.jpl.nasa.gov/
[11] https://sourceforge.net/projects/q3c/
[12] https://sourceforge.net/projects/q3c/
[13] http://adsabs.harvard.edu/full/2006ASPC..351..735K
[14] http://www.ivoa.net/documents/VOTable/20130920/index.html
[15] http://data.xao.ac.cn/cross/q/match/form

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

Source: ChinaXiv —Machine translation. Verify with original.