中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

多波段、多信使天文数据高效融合关键技术研究与应用

作者姓名：许允飞

指导教师：崔辰州 研究员 中国科学院国家天文台

樊东卫 副研究员 中国科学院国家天文台

学位类别：理学博士

学科专业：天文技术与方法

培养单位：中国科学院国家天文台

2020 年 12 月

Research and Application of Key Technologies for Efficient Fusion of Multi-Wavelength and Multi-Messenger Astronomical Data

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Natural Science

in Astronomical Technology and Method

By

Xu Yunfei

Supervisor: Professor Cui Chenzhou

Associate Professor Fan Dongwei

National Astronomical Observatories, Chinese Academy of Sciences

December, 2020

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

天文学研究已迈入时域天文观测和多信使天文观测的时代。一系列新兴观测设施产生的海量数据流给数据处理和挖掘带来了全新挑战。通过对时域天文观测和多信使天文观测的数据处理需求进行调研，本文总结当前亟需解决的技术挑战是在海量数据的基础上实现暂现源的实时交叉证认、多信使事件电磁对应体的高效搜寻，以及暂现源的随动多信使观测证认等方面。

面向这些挑战，本文提出了一种多波段、多信使海量数据高效融合解决方案，其关键技术包括海量星表高效检索方法、在线交叉证认及置信度计算以及异构多波段图像的高效组织、检索及可视化。本文就这些关键技术展开了研究，并针对爱因斯坦探针卫星（EP）暂现源多波段交叉证认及引力波电磁对应体高效搜寻这两个实际项目进行了应用验证。本文主要开展的工作及创新点包括：

1. 针对十亿级以上体量海量星表检索存在的瓶颈，提出了多层级覆盖天区空间索引方法及基于天球划分的星表分表分区策略。通过基于关系型数据库构建的多波段星表数据库，测试了该索引方法结合分表分区策略与其它主流星表检索方法间的效率对比。结果表明本文提出的方法对大范围天区检索效率有较大幅度提升，综合检索效率也优于同类型方法。

2. 将本文提出的多层级覆盖天区空间索引方法和星表分表分区策略应用至海量多波段星表间的位置匹配，并针对多波段星表由于空间定位精度不同带来的多对象匹配问题，提出了并行化的贝叶斯推断星表交叉证认置信度计算方法。通过与 XMM-COSMOS 多波段星表的对比测试，结果显示本方法在多波段星表交叉证认的效率及准确率上均有提升。

3. 针对多波段、多信使联合观测对异构图像数据的统一组织、高效检索及可视化的需求，本文提出了基于层次渐进模式标准（HiPS）的多波段图像组织、检索和可视化框架。其中的主要创新点是提出了一种新的图像数据转换至 HiPS 标准数据集的方法，并实现了该方法的并行化，大幅度提高了海量图像数据标准化组织的效率。以此方法为基础，进而实现了图像数据的高效检索和获取。本文还在万维望远镜软件平台中实现了三维场景下 HiPS 标准数据集的沉浸式可视化。通过与万维望远镜原有数据组织方式下的可视化效率进行对比，本文的实现

方法在内存占用及加载时间上均有优势。

4. 应用本文所提出的方法，针对爱因斯坦探针卫星的暂现源证认需求构建了一系列工具。包括从 EP 观测数据中提取观测源，判别该观测源是否为暂现源，并对该暂现源进行多波段交叉证认。这些工具的核心是 EP 暂现源多波段参考数据库，涵盖了多个波段的星表数据。其中 X 射线星表用于暂现源的判别，其它波段星表用于暂现源的多波段交叉证认。

5. 基于本文提出的关键技术构建了引力波随动观测规划系统。该系统针对地面光学观测网络在引力波电磁对应体搜寻中面临的挑战，解决了如何在引力波事件的定位天区中高效搜寻宿主星系、如何排序宿主星系的观测优先级、如何从观测数据中高效识别电磁对应体等关键问题。该系统已应用于 Ligo/Virgo 引力波探测网络 O3 运行期间的引力波电磁对应体的搜寻，指导地面观测网络对 53 起引力波事件开展了随动观测。

本文属于应用基础性研究范畴，提出的一系列关键技术将为 EP 发射后的暂现源观测及多信使观测的科学发现提供有力支撑。

关键词：时域天文学，多信使天文学，数据融合，交叉证认，虚拟天文台

Abstract

Astronomy research has entered the era of time-domain and multi-messenger observation. The massive data streams, generated by a series of emerging observation facilities, have brought new challenges to data processing and mining. By investigating the data processing requirements of time-domain and multi-messenger observations, several technical challenges that need to be solved: to achieve real-time identification of transients, efficient search of electromagnetic counterparts of multi-messenger event, and follow-up multi-messenger identification of electromagnetic transients.

The efficient fusion of multi-wavelength and multi-messenger massive data is a solution to these challenges. The key technologies include: efficient retrieval methods for catalogs, catalogs cross-matching and confidence evaluation, as well as the efficient organization, retrieval and visualization of heterogeneous multi-wavelength images. This thesis describes these key technology researches, and launches two practical projects, i.e. Einstein Probe Satellite (EP) transients multi-wavelength cross-matching, and efficient search of gravitational wave electromagnetic counterparts. The main work and innovations carried out in this thesis include:

1. Aiming at the bottleneck in the data retrieval of catalog with volumes above one billion level, we propose a multi-order sky region coverage spatial index method and a catalog sharding strategy based on the division of the celestial sphere. With a multi-wavelength catalog on a relational database, we compare the efficiency among the index method combined with the catalog partitioning strategy, and other catalog retrieval methods. The results show that the method proposed in this thesis can greatly improve the retrieval efficiency of a large area on the sky. The comprehensive retrieval efficiency is better than other similar methods.

2. We apply the multi-order sky region coverage spatial index and the catalog sharding strategy proposed in this thesis to the massive multi-wavelength catalogs' cross-matching, and propose a parallelized Bayesian inference calculation method for cross-matching confidence evaluation. Through the comparison test with XMM-COSMOS

multi-wavelength catalogs, the results show that this method has improved the efficiency and accuracy of multi-wavelength catalog cross-matching.

3. Multi-wavelength and multi-messenger joint observation generates massive multi-source heterogeneous image data, which requires unified organization, efficient retrieval and visualization. In order to satisfy these requirements, this thesis proposes a heterogeneous image organization, retrieval and visualization framework based on Hierarchical Progressive Survey Scheme (HiPS). We find a new method to convert image data to HiPS standard data set, and implement its parallelization which greatly improves the efficiency of the mass observation image standardization organization. Based on this method, efficient retrieval and acquisition of image data are realized. This thesis also implements the immersive 3D visualization for HiPS standard dataset on the WorldWide Telescope (WWT) software platform. Comparing with the original data organization of WWT, our method performs much better in memory occupation and loading time.

4. Applying the method proposed in this article, a series of tools are built for the EP satellite's transient identification, e.g. extracting the observation source from EP observation data, judging whether the observation source is a transient, performing multi-wavelength cross-matching of the transient and generating its spectral energy distribution. These tools' core is the EP transient Multi-wavelength reference database, which covers catalogs of multiple wavelengths. In this database, the X-ray catalogs are used for the identification of transient sources, and the other band catalogs are used for the multi-wavelength cross-matching of the transient.

5. Based on the key technologies proposed in this article, a gravitational wave follow-up observation planning system is also implemented. The system addresses the challenges in the search for gravitational wave electromagnetic counterparts by the ground-based optical observation network. It solves key issues, e.g. how to efficiently search for host galaxies in the location of gravitational wave events, how to prioritize the observation of host galaxies, and how to identify electromagnetic counterparts from observation data efficiently. The system has been applied to search electromagnetic counterparts of gravitational waves, during the operation of LVC O3, and guided the ground observation network to carry out follow-up observations of 53 gravitational wave events.

The methods described in this thesis are part of the applied basic research. The key technologies will strongly support the scientific discovery of transient observation and multi-messenger observation of the EP satellite.

Keywords: Time Domain Astronomy, Multi-Messenger Astronomy, Data Fusion, Cross-matching, Virtual Observatory

目 录

图形列表

表格列表

第 1 章　引言

1.1　时域天文学与多信使天文学带来的挑战

天文学是观测驱动的科学，而其中最能推动学科发展的是天文发现。历史上天文学所获的诺贝尔奖，大多数都是授予重大的天文发现。宇宙远非宁静，而是充满了各种剧烈的爆发和变化。这种变化主要表现为两类：一类是暂现源（或爆发源），即突然出现在某一或多个波段的新天体，在一定时间后消失（时标短者仅有秒量级，长者则可达数月或年）。暂现源往往产生于天体辐射的剧烈爆发。另一类则是已知天体的辐射的变化。对时变天体的观测和研究极大地丰富了人类对宇宙及其基本物理规律的认知和探索。这一研究领域称为“时域天文学”。对时域的探索是现今最为活跃并快速发展的天文学研究方向，涉及天文学研究的方方面面，包括从太阳系到宇宙学，以及从恒星演化到极端相对论现象。

目前，国内外时域天文学的研究正处于蓬勃发展阶段，各国都在系统地部署时域研究计划，丰富的时域数据快速积累。一系列新兴巡天项目产生的 PB 级的数据流给数据的处理和挖掘带来了全新挑战，表 1.1列出了国内主要巡天项目的日均和年均数据增长量。可以看出，这些挑战中最关键的一个方面是实时（或接近实时）的大数据流处理分析。大型综合巡天项目通过在大片天区搜寻变化天体，由于时域观测的现象是暂现的，现象的检测以及及时恰当的随动观测非常重要。这需要从望远镜获取数据的同时就进行实时的数据处理，并与同一天区的存档数据进行对比，进而自动可靠地检测、分类暂现源并且根据优先级安排及时的随动观测。上述过程中高效的数据检索、获取、证认服务非常重要。

表 1.1 国内主要巡天项目的日均和年均数据增长量

项目名称	波段	日新增数据量	年新增数据量
FAST	射电	~15TB	~5PB
EP	X-Ray	~17GB	~5TB
LAMOST	光学	~22GB	~8TB
CSST	光学	~1TB	~360TB
GWAC/SVOM	光学/X-Ray/γ-Ray	~4.9TB	~1.74PB
天籁计划	射电	~4TB	~1.5PB

同时，天文学业已进入多信使观测的时代。2016 年 2 月 12 日，aLIGO 正式宣布首次直接探测到了人类历史上第一个引力波信号 GW150914。一年之后的 2017 年 8 月 17 日，LIGO-Virgo 引力波探测器网络捕获了一个来自两个致密星遗迹（“中子星”）旋近的引力波信号。仅仅在引力波网络观测到这个信号后的 1.7 秒，命名为 GRB170817A 的伽玛射线暴被费米伽马射线暴监视系统探测到。在这次的事件中，引力波和伽马射线触发器生成了发送给天文界的警报，发起了一场随动观测运动，最终探测到了这一事件宿主星系 NGC4993 的衰退的电磁波信号。在此之后，LIGO-Virgo 引力波探测器网络经过了两次升级，大幅度提升了观测灵敏度和观测效率，仅在第三次观测运行期间（O3，2019 年 4 月至 2020 年 4 月，其中 2019 年 10 月暂停观测一个月）就发现 53 例引力波事件 (gra)。可以预见，随着更多探测器的投入使用和空间引力波探测器的发射，结合电磁波和引力波的多信使观测将会蓬勃开展。

1.2 时域天文观测与多信使天文观测的科学数据融合需求

时域天文学与多信使天文学面临的新挑战对天文科学数据融合提出了新的需求。其中时域天文观测的暂现源多波段证认要求在特定时间内完成暂现源搜寻、判定与多波段证认，而未来将涌现的多信使事件则要求高效证认电磁对应体以及时开展随动观测。针对这些暂现源及多信使事件的随动观测都需要在分钟级甚至秒级的时间限制内做出反应，即实现暂现源候选体进行快速证认。具体包括对不同波段观测数据的高效获取，给出准确的交叉匹配结果并生成能谱，同时给出对应波段的观测图像数据作为参考。接下来就时域天文观测和多信使天文观测的一些实际案例对其需求进行简单介绍。

1.2.1 爱因斯坦探针卫星暂现源的多波段证认

预计将于 2022 年发射的爱因斯坦探针卫星（Einstein Probe, 简称 EP）具有超大视场、高灵敏度、全天观测、快速指向能力和数据下传等方面的优势, 特别是其大视场和高灵敏度, 为高能天体物理的各项暂现事件的探测提供了一个理想的观测平台。此外，在多信使观测领域，X 射线是非常理想的探测引力波电磁对应体辐射的波段。对于双中子星（Binary Neutron Star，BNS）并合事件，在中心天体为长时间存在的超大质量中子星的情况下，任何方向的观测者都将会看到较强的 X 射线辐射 (高鹤 等, 2001)。目前国际上在轨运行的 X 射线探测器视场

太小，或者灵敏度较低，或能段偏高，都不适合探测引力波事件电磁对应体。EP 是采用了 micro-pore optics(MPO) 龙虾眼聚焦成像技术的极宽视场 X 射线巡天卫星，其视场为 3600 平方度，接近 1 ／ 11 全天。未来 EP 与引力波探测器开展联合观测时，EP 的一次观测即可覆盖引力波事件定位天区，若存在引力波电磁对应体，则能够在很大概率上捕捉到其爆发的 X 射线波段信号, 因此，发现引力波电磁对应体是 EP 的主要科学目标之一。

由于 EP 首次采用了最先进的 X 射线 MPO 龙虾眼聚焦成像技术，结合气体探测器，探测灵敏度比现有设备（如 MAXI）高近两个数量级。对一个天区持续曝光几百千秒将能达到 ROSAT 全天巡天 RASS 的灵敏度。其巡天能力 Grasp（探测有效面积 × 视场）比 eROSITA 高一个多数量级。且对 X 射线源的定位精度较高，分辨率为几十角秒。基于其高灵敏度和大视场，EP 能够较大概率上在引力波事件定位天区范围内捕捉到电磁对应体的 X 射线辐射。EP 的科学目标需求中要求 3 分钟内完成对暂现源的证认，且引力波电磁对应体观测要求较高的时效性，如何实现暂现源的快速证认是一项难度较大的挑战。为此，EP 科学应用团队需要研发暂现源判定算法并构建多波段参考数据库，对 EP 观测源进行筛选，找出暂现源候选体并进行多波段证认，从而指导进一步的随动观测。

1.2.2　多信使事件的高效电磁对应体证认

2018 年 5 月在马里兰大学举行的首届多信使天体物理信息基础设施研讨会，对未来十年多信使天文学的基于和重大挑战及学界需求进行了深入的讨论。与会者对如何通过建设新一代信息基础设施（Enhanced Cyberinfrastructure）来帮助天文学研究迈入多信使天体物理学时代提出了种种设想。尤其对未来多信使天文观测的场景进行了有趣的畅想。如在 2020 年代早期，LIGO 和 VIRGO 探测器将在更高灵敏度下与 KAGRA 协同运行，该联合网络将能够在 200Mpc 距离内探测到双中子星并和事件的前 100s 的信号，即其预合并阶段的信号。因此，可以设想以下一系列事件：

1.2.2.1　在合并之前检测到双中子星的旋近

LIGO + VIRGO + KAGRA（LVK）在合并前 100s 识别出距离为 80Mpc 的双中子星发出预合并警报。LVK 数据分析系统自动生成引力波事件暂现源警报并给出 100 平方度以内的天区定位，而后将警报和数据分发给全球范围的观测

机构和科学家。当多个合作观测机构接收到警报后，将通过基于代理的方法进行协商，以分治的方式分别观测该天区，并在多个光学波段中成像。与此同时，在轨的 SVOM 卫星也将用其高能观测载荷指向该天区进行观测。在 BNS 合并的一秒内，SVOM 将检测到明亮的硬 X 射线暂现源，并基于对原始 LVK 警报的修订发布引力波电磁对应体报告。仅仅一秒钟之后，LVK 通过完整的引力波信号实现精确定位并更新警报，并注意到它与 SVOM 观测到的爆发一致。几秒钟之后，观测网络中的一台程控自主望远镜进一步观测报告了 V 12 mag 光学对应体的存在。多个合作光谱观测设施也注意到了明亮的 BNS 光学对应体，并协同在光学和近红外波段观测。在接下来的几个小时里，随着地球自转和望远镜的优先级转移，光谱观测被移交给其他望远镜，使得观测在覆盖范围上能够无缝衔接。自动成像设施将报告红色千新星的出现，并触发太空望远镜在 1 微米至 5 微米波段观测红色千新星。BNS 理论/建模小组的观测科学家指出，对于出现的千新星缺乏相应的毫米波长观测，并广播呼吁观测。随后的 ALMA+ 观测记录了千新星的第一次早期观测，并对 BNS 外向流提供了新的约束。在几天的时间尺度上的后续建模产生了对千新星红外演变的预测，该预测可以用于设计随后的 NIR/IR 光谱观测活动，以进一步测试核状态方程模型 (Allen 等, 2018)。

1.2.2.2 来自潮汐瓦解事件的高能中微子

全强度或扩展的 KM3NET 与 IceCube-Gen2 的联合操作将对 TeV-PeV 中微子暂现群的观测具有较高灵敏度。可以设想以下场景：通过多信使观测信息基础设施，证认了 KM3NET 和 IceCube-Gen2 之间的中微子重合，估计误报率为 3 次/年，其推导得到的联合定位天区为 0.4 平方度，并将中微子重合警报分发给感兴趣的观测机构，多台望远镜对该事件的光学对应体进行协同搜索。IceCube 观测小组的观测科学家对警报进行了注释，并重新分发警报。警报中的定位包括了 Abell 星系团及其若干组成星系，包括中央型 cD 星系。中微子与星系团间存在联系的可能性，触发了使用 eROSITA 任务进行深度成像和 X 射线观测的程序。随着地球的自转，全球自动观测设施接力进行光学成像观测。12 小时后，在 eROSITA 数据中识别出一个新的 X 射线源，叠加在星团中矮星系的核区域，这个 X 射线源及其可能的宿主星系将被更新在警报中并发布。自动化多信使天体物理系统利用 LSST 观测中的星系光学存档数据对警报进行注释，显示出 0.1 mag 的随机变率。进行核高能中微子暂现源观测的相关科学家指出，更新后的位

置与星团中一个成员星系的星系核一致，并要求对该星系的核区域进行深入的近红外 IFU 光谱分析。几乎同时，自动观测设施对警报进行了更新，称自第一个观测阶段以来，目标星系核的亮度增加了 1 mag。X 射线、光学成像和近红外光谱数据共同确定了作为中微子重合事件的高置信度源的潮汐瓦解事件（Tidal Disruption Event, TDE）的存在，使得对 TDE 中强子加速过程有了新的约束。

1.2.2.3　超大质量双黑洞证认

通过多信使联合观测证认旋近超大质量双黑洞（Super-Massive Black Hole Binary，SMBHB），可以得到对 SMBHB 演化及其与更大的星系宿主环境间关系的新的理解。脉冲星计时阵（Pulsar Timing Arrays, PTA）观测可能识别出多个表征 SMBHB 早期旋近期的连续波源。这些观测能够提供粗略的定位天区（高达几千平方度），并将提供对频率的直接测量，但在质量、距离和偏心率具有简并性。识别这种系统的主星系可以通过红移测量来打破质量/距离简并，从而实现对双星进行精确质量测量，进而直接与宿主星系的属性进行比较。一旦 PTA 检测到连续波，就可以根据给定的质量/距离比率来限制目标星系，从而创建宿主候选星表; 利用光学、X 射线、射电和其他设施进行多波段搜索可以共同识别这些星系中的 SMBHB 标记（例如周期性变化，扰动星系系统，AGN）。使用 VLBI 和谱线进行深度搜索可识别运动双星或目视双星; 当识别出宿主时，可以测量 MBH-宿主关系、研究环双星盘演化，以及解析 SMBHB 旋近机制。PTA 还可以通过引力波记忆效应观测到并合事件。值得注意的是，这里描述的 SMBHB 的 MMA 所涉及的时间尺度较慢，因为该系统可以持续数月到数十年，这取决于它们被发现到时所处的旋近阶段。

1.2.2.4　银河系超新星

由超新星预警系统（Supernova Early Warning System, SNEWS) 协调的全球 MeV 中微子观测网络已经为发现下一个银河系超新星保持了 20 年的持续观测。下一代探测器的完工，特别是 Hyper-Kamiokande 和 Watchman，能够将灵敏度扩展到 M31，其超过用于探测核心坍缩超新星的预期灵敏度两倍多。随着先进的引力波观测设施正在投入使用，通过在银河系超新星初始激波爆发电磁信号之前若干分钟（type Ibc）或数小时（type II）内观测到引力波和 MeV 中微子，大大提高了三重事件探测的实现前景。随后对引力波和中微子信号的观测和解释

将产生对核心崩塌物理学、核物理学、基础物理学以及激波和激波加速的天体物理学的深入了解。

1.3 研究目标

针对以上时域天文观测和多信使天文观测面临的挑战和实际需求，为了能充分发挥观测设施的潜力，使得科学价值最大化，需要多波段和多信使数据资料的强力支持。中国虚拟天文台团队承担的国家自然科学基金天文联合基金重点项目“面向时域天文学的虚拟天文台核心能力建设与科学应用”对上述需求进行了总结，具体包括以下几个方面：

暂现源的实时交叉证认：发现暂现源时，为了更好地了解暂现源的物理性质，需要实时或者准实时（一分钟以内）数据的交叉证认，从电磁波的伽马射线、X 射线、紫外、光学、红外、射电等一直到引力波、中微子的探测结果。已有天文观测的星图、星表、光谱数据等，以及非电磁波的探测结果需要实时地完成与新发现暂现源的交叉证认，反馈证认的数据并给出置信度评价。

多信使事件的观测证认：当多信使天文观测设施发出事件警报（引力波、中微子）时，这个警报通过天体物理学多信使天文台网络（Astrophysical Multimessenger Observatory Network, AMON）或者伽马射线协调网络（Gamma-ray Coordinates Network, GCN）等传递到其他时域观测系统时必然会有一定的时延，也许是几个小时甚至几天。需要通过暂现源警报消息包去查询已有时域观测的历史观测记录，证认在暂现源事件发生的时刻其他时域观测项目是否在对应天区有观测数据。如果有，需要快速提取出来提供研究人员进一步分析。

暂现源的多信使观测证认：这是与上面需求相对应的，时域观测项目发现的暂现源需要去相应的其它观测计划中去寻找对应体或者探测结果。比如，爱因斯坦探针卫星发现的暂现源需要查看 LIGO 的归档数据库中是否有观测，观测结果如何。

多波段、多信使天文数据的融合是上述需求的有效解决方案。数据融合的概念在不同的学科中含义不同：如汇合、组合、协同、综合等。对该定义较为精确的描述是“数据融合是一个多级多侧面的加工过程，包括对多个源的数据和信息的自动化探测、互联、相关、估计和组合处理。”天文中的数据融合技术是将多个独立的天体信息通过位置或物理特征整合成一个整体，这些信息包括天体

的基本物理属性、测光、光谱、图像等。天体的基本物理属性、测光、光谱这些信息一般以天体为单位记录在星表中，本文将其概括为星表数据。星表数据融合的主要方法是交叉证认，图像数据融合则是将不同波段或不同时间观测图像进行空间位置或像素级的匹配，以完成进一步的分析。以往的交叉证认和图像融合的研究主要集中在方法探索上，并未针对时域观测和多信使观测的实际需求开展大数据规模下的高时效性数据融合研究。同时，传统的天文数据融合大多是为满足传统科研课题的需要而不是时域天文学科学计划甚至科学工程的需要，对于方法的工程实践性方面没有严肃的考虑。为此，开展多波段、多信使海量数据的高效融合研究将提升虚拟天文台面向时域天文研究和多信使天文研究提供数据服务的核心能力。

1.4　主要研究内容

针对以上研究目标，本文的研究内容集中在解决海量多波段数据融合所面临的技术难题上。

1.4.1　海量星表高效检索方法

实现暂现源在不同波段对应天体间的交叉证认，首先要计算出该暂现源位置误差范围内各波段天体的位置匹配关系，即当不同波段天体的位置关系满足某一条件时，它们才有可能互为对应体。直观的方法是对各个波段的星表进行空间检索，通过球面几何计算得出各波段星表在暂现源位置误差范围内的天体，并计算它们的距离。对于百万数量级以下体量的星表，采用常规的空间索引如 R 树、生成搜索树等均可以实现对星表条目的快速空间检索及距离计算。

但现代巡天产生的星表越来越庞大，尤其光学星表的体量已达数十亿级。这种庞大的体量使得星表检索的效率极低。一种可行的方式是将天球基于某一粒度进行划分，落入该划分内的天体则考虑它们互为对应体。通过将多波段星表同样按照该划分进行分表存储，则可避免对大规模星表进行扫描和数据读写。天球划分粒度需要考虑不同波段天体观测的误差范围和实际检索场景的误差范围条件。为此，本文将在多波段星表高效空间索引和分表策略方面展开攻关，实现海量星表百毫秒级的空间检索。

1.4.2 在线交叉证认及置信度计算

不同波段观测数据的定位误差各异，一个 X 射线源误差范围内可能涵盖数百个光学源，为此必须引入概率方法，给出目标暂现源与各波段天体间匹配的置信度。在完成基于天球划分的多波段天体位置匹配后，得到的是可能具备对应关系的多波段天体的集合。为此还需要将它们一一对应地进行置信度计算。这使得置信度计算时间会随天体数目和星表数量呈指数级增长。

为实现多波段数据的在线获取，就必须将交叉证认置信度计算的时间花费降低到一个可容忍的程度。为此本项目将研究实现星表交叉证认置信度计算的优化和并行化，大幅提高算法运行效率，使其能够在线运行，以此应对未来大规模暂现源证认的时效需求。

1.4.3 异构多波段图像的高效组织、检索及可视化

暂现源在不同波段对应的图像数据可以直观的展示出暂现源的物理特性，将多波段图像叠加融合，通过像素级的对比可以做进一步分析。但是现阶段不同观测设施对其图像数据的组织和管理方式各异，用户没有统一的方法对多波段图像进行检索和获取。且多波段图像数据格式、投影方式各异，常规方法下实现多波段图像的匹配及叠加分析需要繁杂的人工操作，如格式转换、投影转换、图像裁剪、像素对齐等，无法满足大数据量快速检索的需求。

本文将针对光学/红外、射电、高能波段观测图像数据的特点，研究实现在同一标准框架下的多波段图像组织、检索与可视化。其核心是图像投影、文件分幅及像素匹配的标准。在该标准框架下，多波段图像数据在入库时自动完成格式及投影转换，并按照空间位置切分成幅，赋予空间索引，以实现对图像的空间检索。并在像素尺度上保持各个波段图像的位置匹配，从而实现多波段图像的统一可视化功能。

1.5 论文章节安排

第一章，主要介绍本论文工作的背景。描述已经迈入时域天文观测和多信使天文观测领域的天文界面临的主要技术挑战，并通过时域天文领域和多信使天文领域中的具体需求引出本文的主要研究目标和研究内容。

第二章，针对本文的主要研究内容，介绍海量多波段数据的主要特征，并对

多波段数据高效融合涉及的关键技术方法，包括星表交叉证认、多波段图像数据的组织、检索与可视化、现有的天文数据查询服务等进行了描述。

第三章，详细介绍本文提出的海量星表数据检索方法及其实现。海量星表高效检索是实现多波段星表高效融合的基础，本章首先介绍了海量星表索引的基本概念和主要方法，之后介绍本文提出的多层级覆盖天区空间索引的主要原理和对应的空间检索方法。针对海量星表的检索瓶颈，本章还提出了基于天球划分方法的数据库分表分区策略及其联合检索策略。最后介绍本章提出的方法在关系型数据库中的实现，并通过一系列测试将该方法与其它空间索引进行了对比。

第四章，主要介绍多波段星表交叉证认及置信度计算的解决方案。本章首先介绍基于上一章提出的海量星表高效检索方法如何应用于多波段星表间的位置匹配，之后针对多波段星表由于空间定位精度不同带来的多对象匹配问题，在前人方法的基础上提出基于贝叶斯推断的交叉证认置信度计算方法，并实现了该方法的并行化。最后，基于 XMM-COSMOS 多波段星表对本章提出的解决方案进行了测试和验证。

第五章，首先详细介绍海量多波段图像数据的组织、检索及可视化方案。基于当前主要时域及多信使观测设施的观测图像发布方式，我们提出一种能够适用于当前主流观测数据的统一组织、检索及可视化框架，能够高效实现多波段图像数据的配准、拼接、裁剪、组织管理、检索及可视化。该框架在万维望远镜平台上进行了实现，并对其性能进行了对比测试。

第六章，应用本文所提出的方法，针对爱因斯坦探针卫星的暂现源证认需求构建了一系列工具。包括从 EP 观测数据中提取观测源，判别该观测源是否为暂现源，并对该暂现源进行多波段交叉证认并生成能谱。这些工具的核心是 EP 暂现源多波段参考数据库，涵盖了多个波段的星表数据。其中 X 射线星表用于暂现源的判别，其它波段星表用于暂现源的多波段交叉证认。

第七章，介绍针对引力波随动观测的实际需求，基于本文提出的关键技术构建的引力波随动观测规划系统 GWOPS。该系统针对地面光学观测网络在引力波电磁对应体搜寻中面临的挑战，解决如何在引力波事件的定位天区中高效搜寻宿主星系、如何排序宿主星系的观测优先级、如何从观测数据中高效识别电磁对应体等关键问题。

第八章，对本文提出的关键技术和解决方案进行总结，并对未来的进一步优

化和改进进行展望。

第 2 章　天文数据融合研究现状

多波段、多信使天文观测是对宇宙射线，中微子，引力波和全波长范围内电磁辐射的全球协调观测。电磁波观测得到的天文数据根据波段可划分为射电、红外、可见光、紫外、X 射电、γ 射线数据。在多信使天文观测范畴，除了电磁波观测手段，还包括了引力波观测、中微子观测、宇宙射线观测等。正如对首例发现电磁对应体的引力波事件 GW170817 那样，多波段多信使联合观测有望为天体物理机制提供最关键的数据和信息。多波段观测和引力波观测设备将产生呈指数增长的数据量。但单个观测设备只能获取有限波段范围内的资料。为了充分利用多波段、多信使的观测数据，迫切需要天文数据融合技术。天文数据融合旨在高效准确的获取同一天区或同一目标源在不同时间、不同波段、不同信使的相关数据，从而实现对天体物理事件的快速分析。本章将主要对当前天文数据融合研究进展进行综述介绍，针对当前研究存在的不足引出本文提出的主要关键技术和创新点。

2.1　多波段、多信使天文数据

在数据形态上，多波段数据主要可分为星表数据和图像数据。星表数据主要记录了天体的属性信息，包括其位置、定位误差、流量、光度、类型，一些星表还提供了天体的光变和光谱数据。不同波段的星表数据差异主要体现在数据量和定位误差上。光学星表的体量最大，能够达到数十亿级条目，射电及高能波段星表的数据量相对较小，从数十万到数百万不等。定位误差方面，光学波段能达到微角秒级，而射电和高能波段的定位精度可能只有数十角秒甚至角分级。大型星表的数据一般存储在关系型数据库中，天文科学数据领域主要采用的关系型数据库管理系统为开源的 PostgreSQL(PostgreSQL)，它支持多种空间索引，并对基于位置的查询有很好的支持。也有部分星表采用了商业数据库管理系统，如斯隆数字巡天 (Sloan Digital Sky Survey, SDSS) (Fukugita 等, 1996) 采用了微软的 SQL Server(Szalay 等, 2002)。部分体量较小的星表如 Swift BAT 105-Month Hard X-ray Survey (Oh 等, 2018)，直接采用 FITS 文件组织。这种方式的优点是便于下载及程序分析，缺点是不适用于空间查询。

图像数据主要指通过观测设施成像系统所产生的基于像素记录天体信息的数据。它包含了比星表数据更丰富的信息，但处理分析的难度也更大。不同观测波段所产生的图像数据差异较大，对多波段图像进行分析能够提取更多关于暂现源的信息。多波段图像数据格式以 FITS 文件为主，此外还有 HDF 格式。各个波段的图像数据量均较大，一般为数百 TB。对于多波段图像数据的组织和管理主要基于文件系统，部分数据发布机构提供了 FTP 下载，用户可以基于观测时间或观测任务的名称通过目录检索找到对应的图像数据。一些巡天项目如 Pan-STARRS (Vaccarella 等, 2008) 也基于 IVOA (International Virtual Observatory Alliance，国际虚拟天文台联盟) 的 SIA (Simple Image Access，简单图像获取) (Dowler 等, 2015) 协议构建了图像数据检索的服务。此外还有一些巡天项目如 SDSS、AllWISE (Cutri 等, 2013) 等为了便于用户浏览数据，将图像数据基于 IVOA HiPS (Hierarchical Progressive Survey，层次渐进模式) (Fernique 等, 2015) 标准进行了发布，用户可以在虚拟三维天球上无缝浏览所有数据。多波段图像数据总体上是异构的，用户现阶段没有一个统一的方式去便捷的获取不同波段的图像数据。

针对多波段星表数据和图像数据的特征，其融合手段有所区别。对于多波段星表数据的融合，首要方法是交叉证认。多波段图像数据的融合，则需要实现统一的图像组织、管理、查询框架。而这两者的基础是实现高效的天文数据查询检索方法。

2.2 星表交叉证认

美国虚拟天文台将交叉证认定义为“在物理层面可信地鉴定不同时间不同波段的多次观测数据是否同源”。其基本思路是在异构的星表间进行一对一的源匹配。由于不同仪器的观测差异，同一天体在不同星表上的坐标可能略有不同。实现星表间的交叉证认，首先要计算目标位置误差范围内各星表中天体的位置匹配关系，即当不同星表天体的位置关系满足某一条件时，它们才有可能同源。常用的交叉证认方法是计算不同星表天体间的距离，当距离小于一定阈值时则认为二者互为对应体。此外，不同波段的观测数据的定位误差不同，光学和近红外波段观测的定位精度能够达到微角秒级，而高能波段、射电波段的定位精度在角秒，极端情况下会达到角分尺度。如 X 射线波段中，Chandra 的平均定位精度为 0.5 角秒 (Weisskopf 等, 2000)，XMM-Newton 的平均定位精度为 7 角秒 (Jansen

等, 2001)，而 ROAST 的 2RXS 星表定位精度在 29 角秒左右，极端情况下会达到 1 角分 (Boller 等, 2016)。即一个 X 射线源在位置关系上可能匹配多个光学源。因此，对于不同波段天体之间的关联，不仅要基于位置关系进行匹配，还要进一步给出它们互为对应体的置信度。

2.2.1　星表交叉证认主要方法

理论上，任何两个星表间的交叉匹配都应该将第一星表的每个源与另一个星表的所有源进行比较 (Riccio 等, 2016)。显然这种方式的计算量过大，时间效率太低。因此，前人提出了多种针对该问题的优化解决方案，其中较为常用的如条带算法 (Zone Algorithm)(Gray 等, 2007)，分层三角网格（HTM）(Kunszt 等, 2001)，同纬度等面积像素网格（HEALPix）(Gorski 等, 2005) 和四叉树立方体（Q3C）(Koposov 等, 2006) 。这些方案的基本思路是对天区进行固定的划分，通过简单的坐标计算可以得出每个天体所在的天区，再计算落在同一或相邻天区中的天体间的对应关系，从而大幅度降低计算量。

条带算法通过对天球进行条带划分并赋予唯一编号，使用该编号及赤经对星表数据进行聚集索引，这样可以将坐标邻近的数据也在硬盘上邻近存储，从而提高数据的存取效率。该算法最初是在单个 Microsoft SQL Server 上实现。Nieto-Santisteban 等 (2007) 等实现了条带算法在多个 SQL Server 上的并行化，使用 8 个节点在 20 分钟内实现了 SDSS DR3 和 2MASS 的交叉证认。Wang 等 (2013) 开发了 ZoneMatch 算法，该算法将星表中的源按赤经值排序，并在每个区域中执行二分查找。该算法能够在单 GPU 环境下几秒钟内完成百万规模星表的交叉证认。Budavari 等 (2013) 提出的多 GPU 条带算法实现 Xmatch，在 4 分钟内完成了 4.5 亿条和 1500 万条星表的交叉证认。近年来还涌现出一些基于条带算法的优化算法。如 Becla 等 (2007) 提出的 OptZones 算法是条带算法的一种改进形式，它采用了 LSST 的特定假设，即每个条带的相邻集最多包含三个条带。其主要优点是可以快速计算条带的相邻集，并且不需要像原始条带算法中那样预先计算和维护条带的相邻关系。Ma 等 (2018) 提出了 Euclidean-Zone 算法，该算法使用欧氏距离来进行更快的相邻点计算，并基于 OpenMP 实现了并行化计算。Fan 等 (2013) 采用星表天区覆盖信息来滤除无关天体，大幅度提高了原始条带算法的计算效率。

HTM 天球划分方法将球面细分为形状和大小几乎相等的三角形 (Kunszt 等,

2001)。该方法从八个三角面开始（北半球和南半球各四个）迭代渐进地拟合球体，每一个三角面在下一级划分为四个较小的三角面，并赋予唯一编号。因此，HTM 划分方法特别擅长不同分辨率（从角秒级到半球级）的天体搜索。Mi 等 (2011) 使用 HTM 将星表划分为分层的三角形网格，并基于 MapReduce 中的定向联接算法设计了一个交叉证认系统。(Soumagnac 等, 2018) 同样采用 HTM 索引将星表分层存储在 HDF5 文件中，该方法支持数十个星表间的交叉证认。

HEALPix(Fernique 等, 2015) 将天球划分面积相等的等边四边形，其分区策略与 HTM 基本相同，主要区别在于 HEALPix 的分区基于四边形，从十二个基础四边形开始迭代渐进拟合球体。Pineau 等 (2011) 使用 HEALPix 作为分区方案在 30 分钟内采用普通服务器完成了 2MASS（4.7 亿条目）与 USNOB1（10 亿条目）的交叉匹配。Zhao 等 (2009) 在 SQL Server 上使用 HEALPix 实现了并行交叉证认，在 32 分钟完成了 4.7 亿条目和 1 亿条目的星表交叉。采用 HEALPix 进行分区索引会产品块边缘问题，它是指不同星表中的同一天体由于位置误差和索引计算误差落在不同的 HEALPix 分区中。该问题的常用解决方案是使用适当尺寸的边框来扩展 HEALPix 分区。(Du 等, 2014) 采用 HEALPix 和 HTM 索引相结合的方式来减少块边缘问题，并采用线程池以加速交叉证认。对于大规模星表间的交叉证认，也有一些研究基于 GPU 来加速 HEALPix 索引的计算。(Jia 等, 2015) 采用索引循环联接方法，利用 HEALPix 索引在七个节点的 CPU-GPU 集群上实现了十亿级星表的交叉证认。但是，这种方法通常会产生在群集中样本多次发送的问题。为了克服这些问题，(Jia 等, 2016) 采用多分配单联接（MASJ）方法进行了改进，执行效率是先前的算法的 2.69 倍。

Q3C(Koposov 等, 2006) 的划分策略是将立方体渐进拟合为天球。Q3C 采用了四叉树的数据结构，因此它的索引计算比 HTM 和 HEALPix 简单得多。其特有的索引查找表可以实现快速的节点查找，当划分粒度很细时，它能够比 HTM 和 HEALPix 更快计算出天体所在的天球划分索引。值得一提的是，Q3C 在天文领域广泛使用 PostgreSQL 数据库上提供了一个开源的插件，这使它能够非常便捷的应用于自有星表交叉的证认。Landais 等 (2013) 在很大程度上受 Q3C 启发，基于 HEALPix 方法构建了一个 PostgreSQL 插件 HEALPiX-Tree-C（H3C）。它具有与 Q3C 相同的功能和实现环境，但其核心是 HEALPix 算法。

2.2.2　交叉证认的置信度计算

如前所述，不同波段的观测数据对天体的定位精度不同。因此，在进行多波段星表的交叉证认的同时，必须给出天体匹配的置信度。Sutherland 等 (1992) 提出了基于概率计算的 Likelihood Ratio（LR）方法进行多波段天体的交叉证认。LR 方法适用于两个星表间的匹配，将其中一个作为主星表 A，再计算出星表 B 中条目是星表 A 的对应体的概率。Budavári 等 (2008) 在 LR 方法的基础上引入了贝叶斯推断框架，使之能够同时对多个星表进行匹配，同时还可以加入其它物理属性作为独立变量计算匹配的置信度。Fan 等 (2015) 将该模型扩展为可用于交叉证认射电星表的显式几何模型，该模型包含了除点坐标以外的天体的物理属性。Pineau 等 (2017) 开发了基于贝叶斯统计框架，用于基于显式简化星表模型的多星表交叉证认。Salvato 等 (2018) 提出了一种新的基于贝叶斯统计的算法 NWay，其特色是不仅可以给出天体匹配的置信度，还可以给出具体某一天体在另一星表中存在对应体的概率。

尽管在星表交叉证认方面已经取得了巨大的进步，但仍有一些问题值得进一步研究。星表交叉证认的过程类似于数据库中的联接操作，但是由于还需要计算天体之间的距离误差，因此大幅度提高了其计算复杂度。而且，随着新一代观测设施的投入使用，现有方法已无法应对大数据规模的挑战。因此，迫切需要及时且有效的，具有极高可扩展性的算法、技术和工具来应对海量级星表交叉证认带来的挑战。

2.3　多波段图像数据组织、检索与可视化

在本文中，多波段图像融合是指实现不同观测设施的异构观测图像提供统一的组织管理、检索与可视化。

多波段图像融合的主要过程包括图像配准、图像拼接、点源匹配和创建层次细节（Level of Details, LOD）图像。它不仅可以使得多波段图像数据更易于获取和对比分析，还可以使图像数据兼具高分辨率和大视野的优势。不同波段的天文图像数据格式各异，主要有 FITS、CASA、HDF 等。其组织管理方式包括文件系统、关系型数据库、对象存储系统等。同时不同波段图像的检索获取方式也各不相同。

各个波段的数据有自身常用的集成处理软件，如光学波段的 IRAF(Tody,

1986)，高能波段的 HEASoft(HEASARC, 2014)、CIAO (Fruscione 等, 2006) 以及射电波段的 CASA (McMullin 等, 2007)。它们将原始观测数据生成为图像数据产品。在实际的研究应用中，对多波段图像的融合，尤其是将多波段图像在全天尺度上的匹配，还需要需要这些图像进行基于空间位置的图像拼接镶嵌。

MOPEX(Makovoz 等, 2005) 主要用于图像镶嵌，可以高效实现覆盖数平方度的天区的图像镶嵌，在将输入图像插值到公共网格之后，可以将它们组合为单个的镶嵌图像。

Montage(Laity 等, 2005) 可以根据用户指定的尺寸、图像角度来实现图像镶嵌，且镶嵌结果符合 WCS 投影和坐标系，同是它具有背景建模和校正功能。Berriman 等 (2004) 研发了 Montage 适用于大规模图像镶嵌处理的网格化版本。它最大程度地利用了 Montage 架构已有的并行化功能，将投影转换作业添加到任务池中由多个处理器去执行，从而实现图像投影转换的并行运行。

SWarp(Bertin, 2010) 是由法国 TERAPIX 中心研发的图像坐标配准程序，它可以将不同坐标系统下的 FITS 图像根据 WCS 标准进行重采样和拼接。由Gwyn (2008) 研发的 MegaPipe 图像处理 pipeline，将加拿大-法国-夏威夷望远镜（CFHT）的多个相机观测的图像拼接输出为单个图像。SWarp 被用在该 pipeline 中做图像坐标修正和拼接。

yourSky(Jacob 等, 2002) 是一种用户友好的天文图像镶嵌软件，该软件可以对指定的区域、数据集、分辨率、坐标系、投影、数据类型和图片格式实现即时进行镶嵌。它是可以处理图片镶嵌构造的所有方面，包括镶嵌请求的管理，输入图像和输出镶嵌数据缓存的管理，以及并行化的图像镶嵌构造等。同一研究团队又提出了 yourSkyG(Jacob 等, 2006) 作为 yourSky 的升级版。yourSky 只能使用本地多处理器实现并行计算，而 yourSkyG 能够在以诸如 Information Power Grid（IPG）之类的计算网格上实现并行处理，从而能够以高吞吐量在网格上构造多个镶嵌图。

Drizzle 是用于合并来自哈勃深场（Hubble Deep Field）的不规则采样数据而研发的图像镶嵌算法。它保留了测光和分辨率，可以根据每个像素的统计显着性对输入图像进行加权，并且消除了几何畸变对图像形状和测光的影响 (Fruchter 等, 2002)。该方法可以归类为空间自适应滤波器，其利用局部邻域中的像素的线性组合来对期望位置处的像素进行降噪或重构。Drizzle 算法是面向微弱信号源

的基础上开发的，因此，它不太适合高信噪比的未解析对象。Takeda 等 (2006) 提出了一种 Super-Drizzle 算法，该算法利用内核回归框架 (Takeda 等, 2007) 使其适用于不规则采样的数据，提高了 Drizzle 算法的重建质量。Fruchter (2011) 提出了一种从欠采样数据中创建波段限制图像的新方法 iDrizzle，由于在频域中进行了迭代信号提取和低通滤波，因此在小尺度范围，iDrizzle 在对欠采样特征的像素化进行去卷积方面要比 Super-Drizzle 具有更好的性能。Wang 等 (2017) 提出了 fiDrizzle 算法，该算法对 iDrizzle 的有效性和计算速度进行了改进。

上述这些图像处理算法已在个人计算机和计算集群上广泛使用，或将其集成到工作流和 pipeline 中以创建新的数据产品。但是多波段数据融需要的是自动高效镶嵌拼接整个天球的多波段的图像，因此，不仅需要高效的坐标、投影转换和镶嵌算法，还需要研发统一的框架来对数据进行有效的组织管理和高效检索。

IVOA 的 HiPS 标准为多波段图像融合提供了思路。它基于 HEALPix 天球划分，对图像进行切分与重新组织，并赋予统一的索引编号，每个索引编号对应天球中固定的区域。但是 HiPS 更多用于多波段数据的可视化，将其用于对于多波段图像数据组织管理和检索还需要进一步的研究。

2.4　多波段数据融合服务

很多知名的天文科学数据中心都发布了自己的天文数据查询检索和交叉证认工具，其中较为知名的包括 VizieR、SIMBAD、STILTS、TOPCAT、CDS-Xmatch、ARCHES。

VizieR(Ochsenbein 等, 2000) 是在斯特拉斯堡天文数据中心（CDS）研发和运营的在线天文星表数据库，它是迄今为止收录最为完整的天文星表库的数据库，用户可以在线访问该数据库，并获取星表的文档。该数据库以自证数据库的形式进行组织。用户可通过 VizieR 的查询工具选择需要的星表，根据给定的条件调取数据记录并格式化下载，同时还可使用 VizieR 的交叉证认功能对小型星表或上传的星表进行快速的交叉证认。

SIMBAD(Wenger 等, 2000) 是由 CDS 开发和维护的包含太阳系外超过 270 万个天体的基本信息、相关出版物参考以及部分观测结果的参考数据库。SIMBAD 提供多种天体检索模式，如天体名称、坐标或星等。与 VizieR 相似，SIMBAD 支持较小数据集的交叉证认。

Starlink Tables 基础工具集（Starlink Tables Infrastructure Library Tool Set, STILTS）(Taylor, 2006) 是一个命令行工具包，用于对星表执行多种处理操作操作。在交叉证认方面它支持多种交叉模式，包括设置全局或逐行最大角度间隔作为同源阈值，也可以二维或三维笛卡尔空间中位置接近程度作为同源判定条件。TOPCAT(Taylor, 2011) 是基于 STILTS API 的交互式图形化星表查看和编辑器。同 STILTS 一样，它提供了灵活多样的交叉证认方式。

CDS 交叉匹配服务（CDS-Xmatch）(Boch 等, 2012) 是一种新的数据融合和数据管理工具，可用于有效交叉证认大数据的星表（所有的 VizieR 表或 SIMBAD 数据库），它同样可支持用户上传的大星表间的交叉证认。用户可以通过 CDS-Xmatch 的 Web 界面或直接按照 UWS 模式提交交叉证认的作业。处理结果将存储在用户的个人存储空间中，并由 iRODS 进行备份。

ARCHES(Motch 等, 2016) 是用于高能天体物理研究的交叉证认服务。它旨在以光谱能量分布（Spectral Energy Distribution，SED）的形式向国际天文学界提供特性完整的多波段数据。用户可以通过 ARCHES 提供的 HTTP API 提交自己的检索脚本，脚本运行完成后系统会发送给用户交叉结果的下载方式。

上述这些多波段数据检索和交叉证认服务采用的方法和交互模式各异，每种服务的使用均需一定的学习成本。且这些主要面向小规模数据的获取及研究，尚无法满足时域天文观测和多信使天文观测时代的海量数据高效融合的需求。

2.5 本章小结

针对本文的主要研究内容，本章对海量多波段天文数据高效融合涉及的关键技术方法进行了调研。现有多波段数据融合服务都支持小规模星表的交叉证认，但没有能够应用于时域天文观测多信使天文观测的高效数据融合服务，其主要瓶颈在于时效性。不同波段的观测图像数据的文件组织方式、检索系统、可视化方式也各不相同，科学用户需要在不同的数据中心分别寻找对应的数据下载，并用本地软件进行不同波段图像间的配准、镶嵌、融合等工作，这无疑大幅降低了诸如暂现源证认等时效性要求极高的研究工作效率。本文的主要目标是针对这些瓶颈探索行之有效的解决方案和关键技术，将其实现并实际应用。

第 3 章　海量星表高效检索方法及实现

星表包含了大量的天体信息及数据。现阶段不同巡天设施产生了一系列星表，随着观测精度的不断发展，星表的规模越来越大。光学波段较为代表性的如斯隆数字化巡天第 14 次数据的测光星表（Sloan Digital Sky Survey Data Release 14, SDSS DR14）包含 1,231,051,050 条天体。盖亚天文卫星（Gaia Astrometry Satellite）第二次数据发布的星表包含了 1,692,919,135 条记录 (Lindegren 等, 2018)。泛星计划（Panoramic Survey Telescope and Rapid Response System，Pan-STARRS）的第二次数据发布的星表包含了 10,723,304,629 条记录，也是迄今为止发布数据量最大的巡天数据集 (Flewelling, 2018)。随着设备精度的进一步提高，未来的巡天项目将提供更多更大的星表，典型的如 Vera C. Rubin Observatory （旧名 Large Synoptic Survey Telescope，LSST）建成后其发布的星表总量将高达 200 亿，平方千米阵（Square Kilometre Array，SKA）以及我国的 FAST 项目也都会产生体量巨大的星表。实现海量星表中的高效数据检索是实现多波段星表交叉证认的前提，因此本章将首先介绍海量星表检索的主要方法，之后介绍我们提出的一种海量星表索引方法和分治策略，并在关系型数据库中进行了实现和性能测试。

3.1　海量星表的索引构建

星表是结构化的数据，因此将其存储在数据库中能够实现便捷的数据维护。在天文观测领域，常用的数据库类型是关系型数据库，它采用模型关系来组织数据。如 SDSS 采用了微软的 SQL Server 关系型数据库维护其发布的星表和光谱目录。其他主要关系型数据库在天文领域也得到了广泛的应用，如 Oracle、DB2、Sybase、MySQL、MariaDB、PostgreSQL 等。

3.1.1　关系型数据库索引

实现关系型数据库的高效数据检索主要依赖于数据索引。它是指对数据库某一列或多个列的值进行预排序的数据结构。通过使用数据索引，可以让数据库系统不必扫描整个表，而是直接定位到符合条件的记录，这样就大大加快了查询速度。常见的数据索引模型包括哈希表模型、线型数组模型、二叉树模型、B 树模型、B+ 树模型。散列表索引的原理是通过一个散列函数，将键值映射为一个

整数，每个整数对应一个桶。桶中存放 [键值, 数据记录指针] 的数组。在一般情况下只需要一次磁盘 I/O 就可以找到数据记录指针所在的块，然后在内存中扫描数组。合适的散列函数使每个桶分到的记录数相当，因此可以提高数据查询的平均时间。二叉树模型、B 树模型、B+ 树模型的原理都是通过树状结构的数据组织来缩短数据查找的路径。以在天文星表存储中常用的 B+ 树模型为例，它对索引列建立多路搜索树，其中非叶子节点不存储数据，只存储索引指针，以存放更多索引并降低树高度。其叶子节点同样包含索引字段并存储了指向具体数据的指针连接。其示例结构如图 3.1所示。

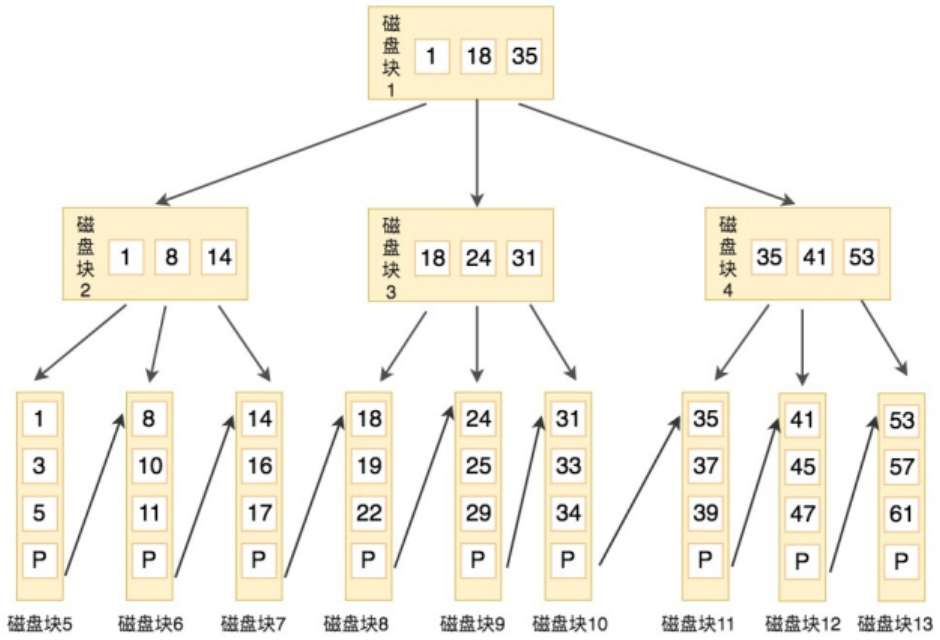


图 3.1 B+ 树模型索引结构

如查找图 3.1中关键字 16，查找过程如下:

- 与根节点的关键字（1,18,35）进行比较，16 在 1 和 18 之间，得到指针 P1（指向磁盘块 2）
- 找到磁盘块 2，关键字为（1,8,14），因为 16 大于 14，所以得到指针 P3（指向磁盘块 7）
- 找到磁盘块 7，关键字为（14,16,17），然后找到了关键字 16，进而获得其对应的数据

由于 B+ 树的叶子节点不存储数据，能够存储更多的索引，因此相比其他树状索引深度较浅，和磁盘的读写交互更少，从而有着更高的查询效率。

从数据存储方式上划分，索引可以分为聚簇索引和非聚簇索引。对于聚簇索引的存储引擎，数据的物理存放顺序与索引顺序是一致的，即：只要索引是相邻的，那么对应的数据一定也是相邻地存放在磁盘上的。对于非聚簇索引的存储引擎，表数据存储顺序与索引顺序无关，叶结点包含索引字段值及指向数据页数据行的逻辑指针，其行数量与数据表行数据量一致。

3.1.2　星表数据库空间索引

针对星表的最主要检索方式是空间查询，即将一个或多个空间范围的描述作为查询条件。该描述可以是以某个坐标为中心并指定半径的原型区域，也可以是多个空间坐标构成的多边形。位于该空间范围的天体即纳入检索结果。以常用的锥形检索为例，如果目标星表以赤经 RA、赤纬 DEC 作为空间坐标字段，则使用 Haverisne 公式进行查询的语句为：

```
SELECT *
FROM TARGET_CATALOGUE
WHERE 2 * ASIN( SQRT(SIN(($DEC_T-DEC)/2)
    * SIN(($DEC_T-DEC)/2) + COS($DEC_T) * COS(DEC)
    * SIN(($RA_T - RA)/2) * SIN(($RA_T - RA)/2))) <= $SR_T
```

其中 RA DEC 为条目的位置坐标，$RA_T, $DEC_T 和 $SR_T 为目标圆锥参数，整个查询过程需要进行复杂的三角函数计算。使用空间索引来帮助实现空间查询，则可以大幅降低查询的计算量。

同其它种类的数据库索引一样，空间索引的作用是提升空间查询的效率。空间索引一般包括两类，一类是直接对空间坐标构建索引，以 2 列或更多列的坐标数据作为索引键构建索引树。另一类则是将空间坐标转换为单一值，针对该值构建索引数，实现对空间位置的间接查找。下面将分别介绍这两种方式的代表，KD 树索引和基于 HEALPix 的空间索引，这两种索引也是本文提出的空间索引方法的基础。

3.1.2.1　基于 KD 树的空间索引

KD 树是 K-dimension tree 的缩写，是对数据点在 k 维空间（如二维 (x，y)，三维 (x，y，z)，k 维 (x，y，z..，k)）中划分的一种数据结构，主要应用于多维空间关键数据的搜索。KD 树是一种空间划分树，即将整个空间划分为特定的几个部分，然后

在特定空间的部分内进行相关搜索操作。如 K 维空间数据集 $T=\{x_1,x_2,...,x_N\}$，其中 $x_i=(x_i^{(1)},x_i^{(2)},...,x_i^{(k)})^T,i=1,2,3,...,N$，其构建 KD 树的流程如下：

1. 构造根节点：选择 $x^{(1)}$ 为坐标轴，将 T 中所有对象以 $x^{(1)}$ 为中位数，将空间以垂直 $x^{(1)}$ 轴划分为两个空间，并从根节点生成左右两个深度为 1 的子节点。左子节点的坐标均小于切分点，右子节点均大于切分点。

2. 迭代构造节点：对深度为 j 的节点，选择 $x^{(l)}$ 为切分的坐标轴，$l=j(modk)+1$，以该节点继续将空间划分为两个子区域。直至子区域没有对象存在时停止，从而形成 KD 树的区域划分。

以星表的二维坐标（赤经、赤纬）空间为例，KD 树按照二分规则将其进行划分。其构建过程如下:

1. 首先确定划分域：分别计算赤经赤纬的方差，取方差较大的字段作为划分域。

2. 确定域位数据点：将划分域字段数据排序取中值，则域位数据点为该中值对应的坐标点。并将二维平面以通过该坐标且垂直于划分域轴进行分割。

3. 子空间递归分割：针对第二步中分割得到的空间，分别递归执行步骤 1、2，完成对所有坐标在二维平面中的分割。

在执行以上过程时，依次将每次迭代的域位数据点作为根节点构建平衡二叉树，即完成 KD 树的构建。其时间复杂度为 $O(NlogN)$。如图 3.2所示：

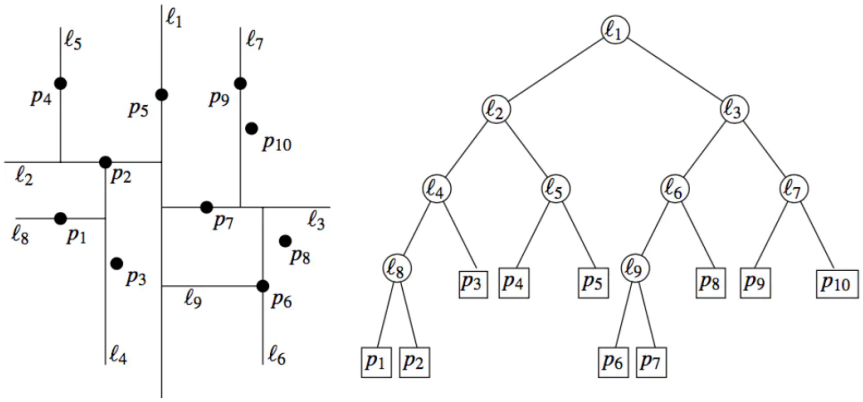


图 3.2 KD 树构建示意

针对 KD 树索引的空间查询，主要是采用最邻近搜索方法。首先从根节点出发，对 KD 树作深度优先遍历，以预设半径或最邻近节点距离为半径画圆（多维空间为超球面），将圆外的子树全部忽略。继续深度优先遍历，并与原最邻近节

点对比距离以更新最近邻，并继续画圆排除子树。重复以上过程，直至不更新最近邻。将所有距离小于预设半径的节点插入结果队列即得到检索结果。基于 KD 树检索的时间复杂度为 $O(2\sqrt{N})$。

3.1.2.2　基于 HEALPix 的空间索引

HEALPix 是一种球面划分方法，它将天球划分为等面积的区域，并可以递归划分各个子天区面积，直至一个像素点。对这些划分得到的区域进行编号，即实现了对天区的索引。HEALPix 对天区的划分，根据层级的不同粒度也不同。在第零层划分时，HEALPix 将天球划为 12 个基础子天区，从北至南分为 3 行，每行均匀分布 4 个子天区。每个子天区又可以细分为 2^{2K} 个子网格，其中 K 为划分的层级，如图 3.3。依次对这些划分得到的网格进行编号，即得到对应网格的索引。索引由三部分组成，分别为网格所在的 12 个基本子天区的编号 $faceindex$，索引所在层级 k，以及该网格在基本子天区中细分后的网格集合中的编号 $pindex$。通过这三个参数，可以计算出该网格在该细分层级中的一个唯一编号 $hpx, p \in [0, 12\times 2^{2k})$。HEALPix 的索引编号方式有 NESTED 和 RING 两种，其排序方法不同：NESTED 对各个子天区进行依次排序，依次递进；RING 则是对整个天区进行环形划分，从北天极点开始逆时针编号。

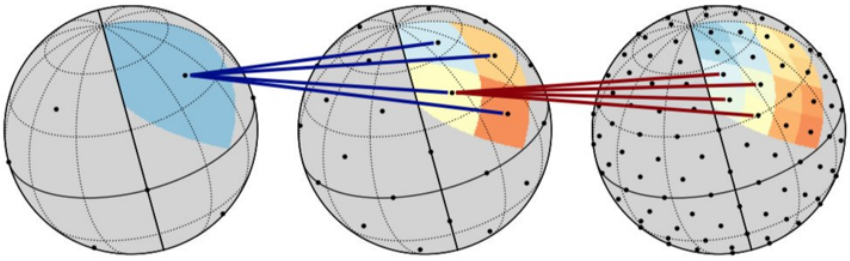


图 3.3 天球的 HEALPix 细分。此处显示的球面均划出了十二个基础层的 HEALPix 四边形，并用不同的灰度级阴影显示。从最左边起依次显示了层级 $k = 0, 1, 2$ 的 HEALPix 网格，以说明 HEALPix 的层级结构，其中每个网格按每个连续的顺序分为四个自相似的子网格。

NESTED 的 HEALPix 编号方式以特定顺序枚举所有网格。在层级为 1 时，共有 48 个子网格（12×4），其编号依次为 0 至 47。如图 3.4，在 NESTED 编号方案中，编号为 M 的网格 4 个子网格的编号为 $(M\times 4)+3$，$(M\times 4)+1$，$(M\times 4)+2$，$(M\times 4)$。相应的，网格 N 的父网格的编号为 $N/4$。

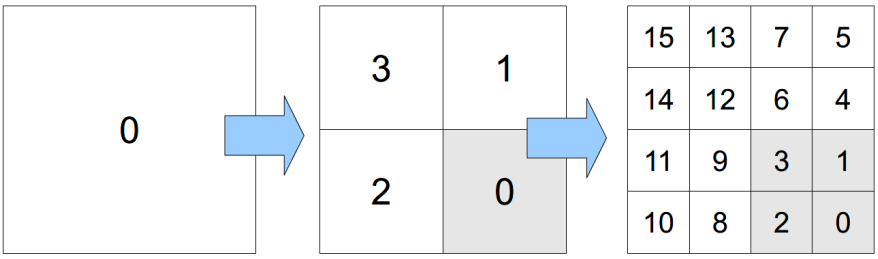


图 3.4 NESTED 方式下的 HEALPix 网格层级细分。自左向右分别是层级 0-2.

在 NESTED 模式下，HEALPix 的基础层网格划分如图 3.5, 注意图中为体现红色网格 0 是网格 3 的临接网格，将其位置画在了网格 1 的左上方，所以其 x、y 轴顺时针转动了 90°。

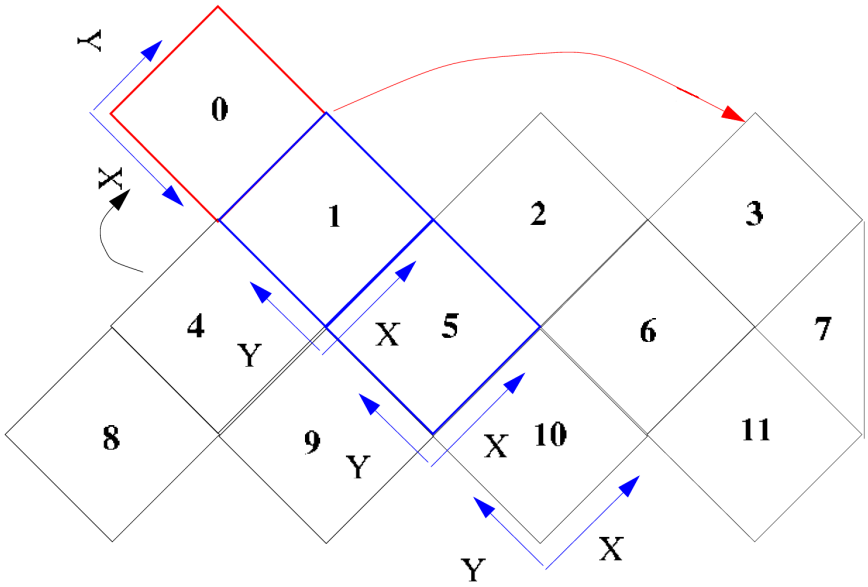


图 3.5 展开平面后的 NESTED 模式下的基础层 HEALPix 网格，图中标出了每个网格的 x、y 轴指向，子网格的编号按 x、y 轴方向依次增加。

在不同的划分层级下，HEALPix 对天球划分的网格数目不同。同时不同的星表天体条目也不同，针对星表构建 HEALPix 索引即是将天体划入对应的 HEALPix 网格中，为了计算的便捷，每个网格中的天体数目在 100 条左右为宜，因此要根据星表的条目数确定 HEALPix 划分的层级。HEALPix 在各层级对天球划分的网格数及网格面积如表 3.1所示。

确定划分层级后，接着根据每个天体条目的坐标计算其所在的 HEALPix 平面坐标系下的索引 i 和 j，之后再将它们转换为 HEALPix 在对应层级 k 上的网格编号 hpx。具体流程如下：

表 3.1 HEALPix 在各层级对天球划分的网格数及网格分辨率。

层级 k	总网格数	网格分辨率（网格边长）
0	12	58.6°
1	48	29.3°
2	192	14.7°
3	768	7.33°
4	3,072	3.66°
5	12,288	1.83°
6	49,152	55.0′
7	196,608	27.5′
8	786,432	13.7′
9	3,145,728	6.87′
10	12,582,912	3.44′
11	50,331,648	1.72′
12	201,326,592	51.5″
13	805,306,368	25.8″
14	3.22×10^{9}	12.9″
15	1.29×10^{10}	6.44″
16	5.15×10^{10}	3.22″
17	2.06×10^{11}	1.61″
⋮	⋮	⋮
29	3.46×10^{18}	3.93×10^{-4}″

首先将赤经 RA 和赤纬 Dec 转换为球面坐标，其计算公式如 3.1。

$$\begin{aligned} \theta &= \frac{(90 - Dec) \times \pi}{180} \\ z &= cos\theta \\ \phi &= \frac{RA \times \pi}{180} \end{aligned} \tag{3.1}$$

当 z 的绝对值小于等于 2/3 时，该位置在天球的赤道带上，即基础层网格号 $findex \in [4,5,6,7]$，则其索引号 hpx 计算公式如 3.2, 其中 k 为层级，I 函数表示向下取整：

$$\begin{aligned} i &= I\left(\left(2 - \frac{3z}{2}\right)2^k\right) \\ j &= I\left(\frac{2^{k+1}\phi}{\pi} + \frac{\left(i - 2^k + 1\right) mod 2}{2}\right) \\ hpx &= 2^k\left(i - 2^k - 1\right) + j + 2^{k+1}\left(2^k + 1\right) \end{aligned} \tag{3.2}$$

当 z 的绝对值大于 2/3 时，该位置位于天球的两极条带上，即 $findex \in [0,1,2,3,8,9,10,11]$, 则 hpx 计算公式如 3.3。

$$\begin{aligned} i &= I\left(3 \times 2^k\sqrt{1 - z}\right) \\ j &= I\left(\frac{2i\phi}{\pi} + \frac{1}{2}\right) \\ hpx &= 2i\left(i - 1\right) + j \end{aligned} \tag{3.3}$$

将 hpx 作为一个字段插入至星表中，此时我们已经将二维的空间位置转换成了一维。接着对其构建 B+ 树生成星表的簇索引，这样在天区位置上临近的条目，其在数据库中的存放位置也相邻，从而加快了临近查找的速率。

基于 HEALPix 的空间检索的基本思路是采用 HEALPix 网格拼出搜索区域，如图 3.6。由于网格不能完整拟合圆盘，所以还需要在查询时对第一步检索出来的结果做过滤计算，采用笛卡尔坐标的线性计算公式进行过滤。完整的检索 SQL 语句如下，其中 $HP_LOW 和 $HP_HIGH 代表检索区域的起始索引，一次检索可能包含多个起始索引。X、Y、Z 表示星表中天体在单位球面上的笛卡尔坐标。X_T、Y_T、Z_T 为查询区域中心点坐标，$COS_SR 为查询半径的余弦。

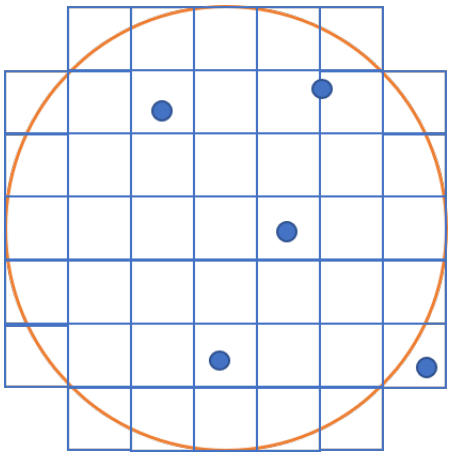


图 3.6 用 HEALPix 网格拟合的锥形检索区域，图中网格为 HEALPix 在某一层级的网格。图中右下方的天体在检索范围外，但仍被检索网格包含，因此还需进一步计算以过滤类似的天体。

```
SELECT *
FROM TARGET_CATALOGUE
WHERE HEALPIX_IX BETWEEN $HP_LOW AND $HP_HIGH;
AND X*$X_T+Y*$Y_T+Z*$Z_T >= $COS_SR
```

3.2　基于多层级覆盖天区的空间索引

从上述的介绍中可以看出，基于 KD 树的索引构建时间复杂度为 $O(NlogN)$，对于数十甚至数百亿级的星表来说，索引构建的花费是巨大的，(Kalpakis 等, 2001) 尝试利用 Informix 数据库为 USNO-A1.0 星表建立索引, 结果显示索引的建立非常耗时, 其中为整个星表建立树状索引居然用时 25 天。现有的主要星表数据库主要采用的是基于天区划分的空间索引构建方式，如 SDSS、Pan-STARRS 采用了 HTM 空间索引，法国 CDS 数据中心主要使用 HEALPix 索引。

但仅仅基于天球划分方法构建空间索引存在一系列问题。如上述介绍的 HEALPix 空间索引构建方法，其实质是将二维坐标映射至一维数据再构建树状索引，并没有解决树状索引占用空间随星表体量大幅度增长的问题。此外，现有的主要天球划分都通过渐进迭代的方式拟合天球，在不同层级下对天球的划分分辨率是固定的。这就导致 HEALPix 索引只能反映该天体在固定层级的天区位置，当空间检索的区域较大时，用于检索的 HEALPix 网格将大幅增加，每个

网格均需要与星表索引进行对比，导致检索效率大幅降低。因此也无法用较深的层级去划分天区，进而不能很好的拟合检索区域，导致漏源。此外，在进行星表交叉证认时，若星表的层级划分不同，交叉证认前还需进行层级的转换，将增加额外的计算量。

针对上述问题，本文提出一种基于多层级覆盖天区的空间索引方法，充分利用天球划分方法中层级间的迭代属性，构建能够表征多个 HEALPix 层级的空间索引，进而改善大范围空间检索的效率. 由于 HEALPix 天球划分具有较多的开源实现，故本方法主要基于 HEALPix 天球划分，实际上也可移植至 HTM 天球划分方法。

3.2.1 多层级覆盖天区索引构建方法

如前文所述，HEALPix 天球划分方法首先将全天划分为 12 个基础网格，并逐层级细分，每层级细分为 4 个子区域。根据这个特性，可以将 HEALPix 的父网格和子网格间关系映射为四叉树，将基础网格作为根节点，依次细分构建子节点，直至目标层级，而星表中天体在目标层级的网格为叶子节点。我们将该四叉树称为多层级覆盖天区四叉树（Multi-Order Coverage Tree，MOC-Tree ）。MOC-Tree 可以根据星表中天体的空间分布建立索引，记录该天体从层级 0 至目标层级的相对空间位置，这里目标层级最高可达 29 级，在实际应用中应根据星表空间分辨率、天体定位误差和星表规模来确定。对于十亿级的光学星表，目标层级可定为 14 级，即 3.22×10^{9} 个网格，分辨率 12.9″。一棵的 MOC-Tree 模型示意如图 3.7。

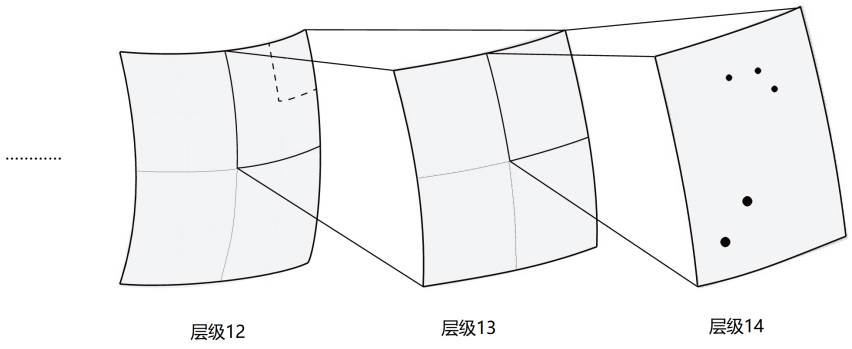


图 3.7 某天体的的 MOC-Tree 模型，图中列出了最后三个层级的层次关系。

MOC-Tree 对 HEALPix 的索引编号规则进行了改进，使之能够在存储不同层级下的相对位置。如图 3.8所示，作为根节点的 0 层的索引编号统一为 0, 其在层级 1 的 4 棵子树索引编号为 0、1、2、3，索引编号为 2 的节点其在层级 2 子

树索引编号分别为 20、21、22、23，索引编号为 20 节点其子树索引编号分别为 200、201、202、203，……依次类推，可以看出，MOC-Tree 的索引编号位数即为其层级数，每一位的编号代表该网格与其父网格的相对位置。

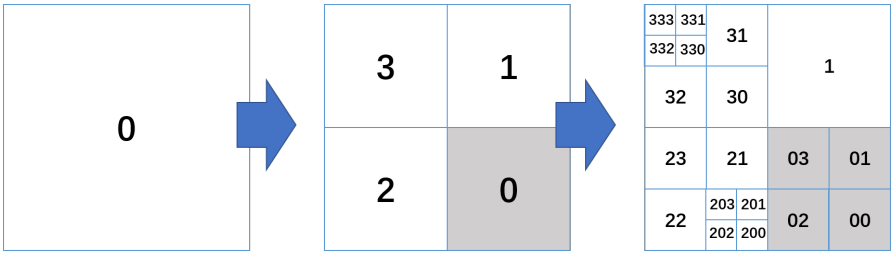


图 3.8 MOC-Tree 的索引编号规则。

整个天球区域是由 12 棵 MOC-Tree 组成的森林。为了解算的便捷，每棵 MOC-Tree 只在根节点编号上作区分，子节点不做区分，即根节点的编号为 $F \in [0, \cdots, 11]$。从原始 HEALPix 索引号 hpx 可以便捷的转换为 MOC-Tree 索引号。首先要把 hpx 转换至基础网格为 0 时对应的原始 HEALPix 索引号，称之为 hpx_0，如原始索引号为 121 的 HEALPix 网格，其所在位置为基础网格 7 的层次 2 中，其在基础网格 0 对应位置的原始索引号 hpx_0 为 9。如公式 3.4。

$$hpx_0 = hpx - \left(2^{2k} \times F\right) \quad \ldots (3.4)$$

之后从目标层次 k 开始，如公式 3.5, 依次计算该网格在层次 k 的编号 N_k。最终得到的 MOC-Tree 编号形如 $KN_0N_1N_2...N_k$。

$$N_k = I\left(\frac{hpx_0}{2^k}\right) \quad \ldots (3.5)$$

MOC-Tree 的逻辑结构如图 3.9, 它在内存中的存储方式如图 3.10。其中左侧为 MOC 节点索引编号，以数组形式存储叶子节点，指向右侧实际星表条目链表。根据构建索引时确定的划分层级，一个 MOC-Tree 节点可能包含多个星表条目，这些星表条目间以链表进行关联。

在 MOC-Tree 模式下，星表中每个天体的索引设计为 4 个字节共计 32 位的二进制编码，共包括 15 个字段。其中 14 个字段各 2 位，表征该天体在不同层级上的相对位置。最前面一个字段为 4 位，表征该天体所处的 HEALPix 基础网格。如 **10 10** 01 10 00 11 00 00 10 11 00 01 10 11 00 11 表示的 MOC-Tree 索引

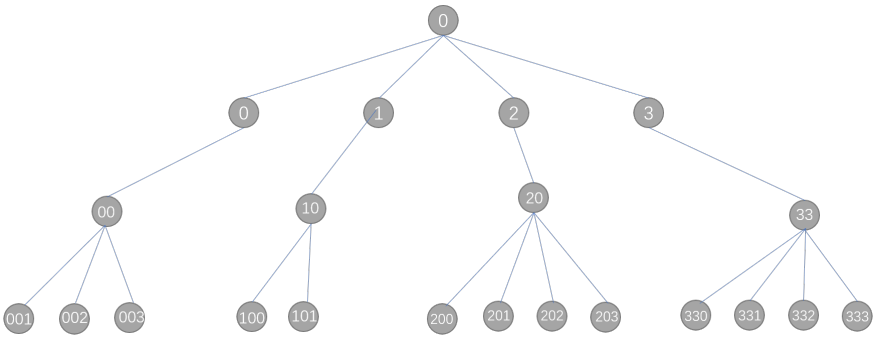


图 3.9 构建的 MOC-Tree 逻辑结构。

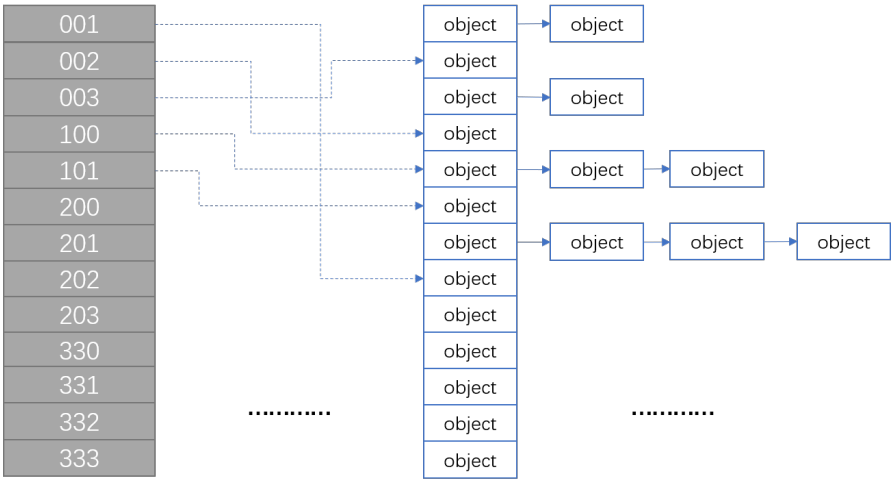


图 3.10 MOC-Tree 物理存储结构。左侧为 MOC 节点索引编号，以数组形式存储叶子节点，指向右侧实际星表条目链表。根据构建索引时确定的划分层级，一个 MOC-Tree 节点可能包含多个星表条目，这些星表条目间以链表进行关联。

号为 12030023012303。其中粗体展示的前 4 位字段表示该天体位于编号 10 的 HEALPix 基础网格上。14 个字段可表示最大层级为 14。对于一些小规模的星表或定位误差较大的如 X 射线、射电星表，其星表划分最大层级可以较小，因此可只利用这 14 个字段中的部分字段。

3.2.2　基于多层级覆盖天区的空间检索方法

HEALPix 多层级覆盖天区索引方法的一个主要目标是提高空间检索的效率。对于天球划分方法索引，实现空间检索的途径是用天球划分网格拟合目标空间范围，具体到 HEALPix 方法上，是以低于星表索引划分的层级的网格拟合查询区域。如图 3.6所示，图中共计 45 个网格，则在查询时需扫描 45 次索引。当检索范围较大时，检索时间会大幅增长。

基于多层级覆盖天区的检索方法有效解决了该问题。对应于星表所构建的 MOC-Tree 索引，该检索方法对目标检索区域同样构建 MOC-Tree，以不同层级分辨率的网格去拟合目标天区，如图 3.11。该流程具体如下：

1. 针对目标检索天区生成与星表索引最大层级一致的 HEALPix 网格。

2. 对生成的网格进行归并，当存在 4 个临接网格时则将它们合并为一个上一层级的网格。

3. 迭代上一步骤，直至不存在临接网格。

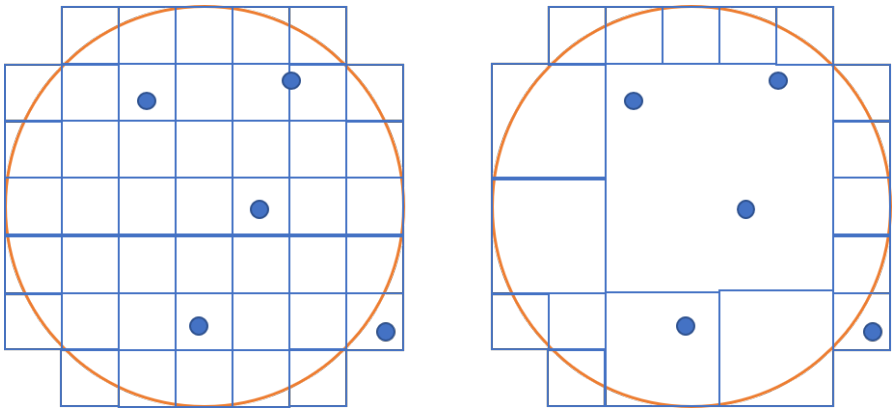


图 3.11 MOC-Tree 以不同层级分辨率的网格去拟合目标天区。左图为 HEALPix 索引下的检索区域拟合，右图为 MOC-Tree 拟合的对应天区，其网格数大幅减少。

如图 3.12所示，左边为检索目标的多边形区域，右边为 MOC-Tree 拟合该区域的网格，分别包含了层级 3HEALPix 编号为 73、74、75 的网格；层级 4 的 291、384、1407 号网格；层级 5 的 1226、5973 号网格。对这些网格构建用于检

索的 MOC-Tree 编码。用于检索的 MOC-Tree 编码规则与星表索引有一些区别，检索的 MOC-Tree 只记录该网格从 0 层到其所在层级的网格相对位置编号，其他字段在编码前填 1，因为 HEALPix 基础网格的编码为 0000-1011，所以不会出现 1100、1101、1110、1111，且填充位数为偶数。故填充 1 能够区分字段是填充字段还是表征 HEALPix 基础网格的字段。如图中层级 3 编号为 73 的网格，其编码为 11 11 11 11 11 11 11 11 11 11 11 00 01 01 00 01，表示该网格位于基础网格 1 中的 101 网格。

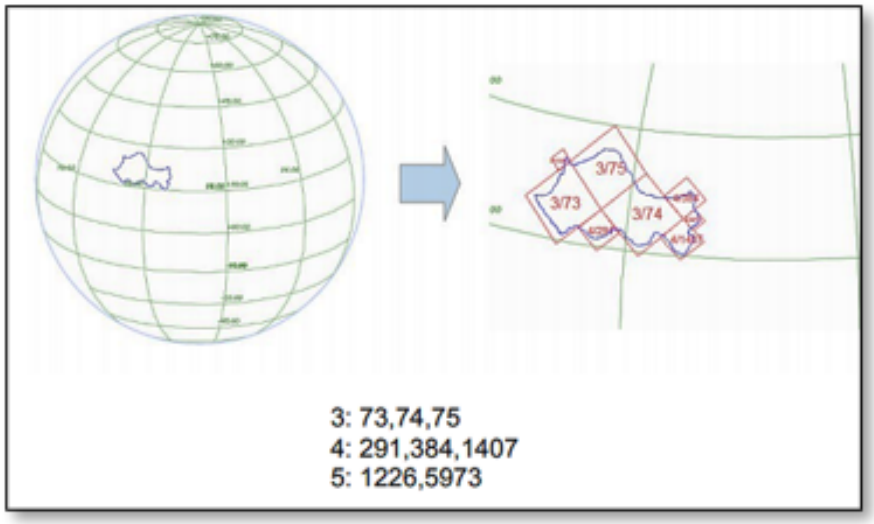


图 3.12 基于多层级覆盖天区的空间检索方法示意。左边为检索目标的多边形区域，右边为 MOC-Tree 拟合该区域的网格。

检索过程的比较可通过位运算快速实现，具体流程如下：

1. 针对检索条件的一个网格，计算其编码非填充位的位数 p

2. 将星表索引编码右移 32-p 位，左边填充 1

3. 将检索网格编码与星表索引编码做异或运算，结果为 0 则说明符合该检索条件，即目标天体位于检索网格内。若不全为 0，检索网格在高位处先出现 1，则说明检索网格编码大于目标天体编码；反之说明检索网格编码小于目标天体编码。

4. 针对检索条件中的其他网格重复上述过程。

在对星表索引做检索的过程中，采用二分法进行对比查找，故上述检索过程的时间复杂度为 O(logN)。

对于百亿级星表，最大层级 HEALPix 网格中包含的天体数目仍较大，数量

在数十至数百个不等。当检索范围小于最大层级空间分辨率时，还需要在最大层级网格中对目标天体作进一步筛选。可以对最大层级网格中的天体构建 KD 树索引，以进一步加速该过程。

3.3　海量星表分治策略

随着观测精度和观测深度的不断发展，多波段星表的数据容量会越来越大，包含的天体信息会越来越多。相应地，星表检索操作的开销也会越来越大。当星表规模到达十亿级规模以上，无论怎样升级硬件资源，单台服务器的资源（CPU、磁盘、内存、网络 IO、事务数、连接数）总是有限的，最终星表数据库所能承载的数据量、数据处理能力都将遭遇瓶颈。

针对海量星表数据检索出现的效率瓶颈，其基本解决思路是“分治”。关系型数据库的分治思路包括数据库的分表、分区和读写分离。其中分表是把一张表按一定的规则分解成 N 个具有独立存储空间的实体表，系统读写时需要根据定义好的规则得到对应的子表再对其进行操作。分区是指将一张表的数据分成 N 个区块，在逻辑上始终是一张表，但底层是由 N 个物理区块组成。读写分离的基本的原理是增设从数据库，让主数据库处理事务性增、改、删操作，而从数据库处理查询操作，并用数据库复制把事务性操作导致的变更从主数据库同步到从数据库中。从多波段、多信使海量数据融合的实际需求出发，海量星表数据的读事务操作要远远大于写事务操作，因此读写分离并不太适用于海量星表检索的应用场景。本节将主要讨论海量星表的分表和分区策略。

关系型数据库常用的分表策略主要有三种，分别是查询切分、范围切分、映射切分。

1. 查询切分：将原数据表主键与分表的对应关系记录在一个单独的表中。其优点是实现简单，且对应方法可以随时修改，但缺点也很明显，引入了新的查询单点，当数据量过大时同样产生瓶颈。

2. 范围切分：根据主键区间或其他属性区间进行分表。其优点是单表大小固定且实现简单。缺点是关联操作受限。导致原本一次查询能够完成的业务，可能需要多次查询才能完成。

3. 映射切分：根据某种规则将主表中数据按照某种映射规则划分到子表中。其优点是可以通过映射策略规避关联操作的限制，缺点是实现较为复杂。

从上述分表策略的介绍不难看出，映射切分较为适用于海量星表的分治。其核心问题是采用何种映射方法，并解决在分表粒度不一致时，多表关联操作 join 查询无法实现的问题。分区则是在分表的基础上对数据表本身作进一步的划分，将其物理上划分成文件块，而逻辑上仍为单张数据表，由数据索引指向文件块中具体的数据条目，这样可以最大程度的利用磁盘 IO 性能，提高检索效率。我们从 HEALPix 天球划分方法中获得启发，通过将全天星表根据 HEALPix 的网格进行划分，同时结合基于多层级覆盖天区的空间检索方法，提出了一种针对海量多波段星表的分表策略，很好的解决了上述问题。该策略主要包括分表分区方法和检索划分方法两部分。

3.3.1 海量星表分表分区

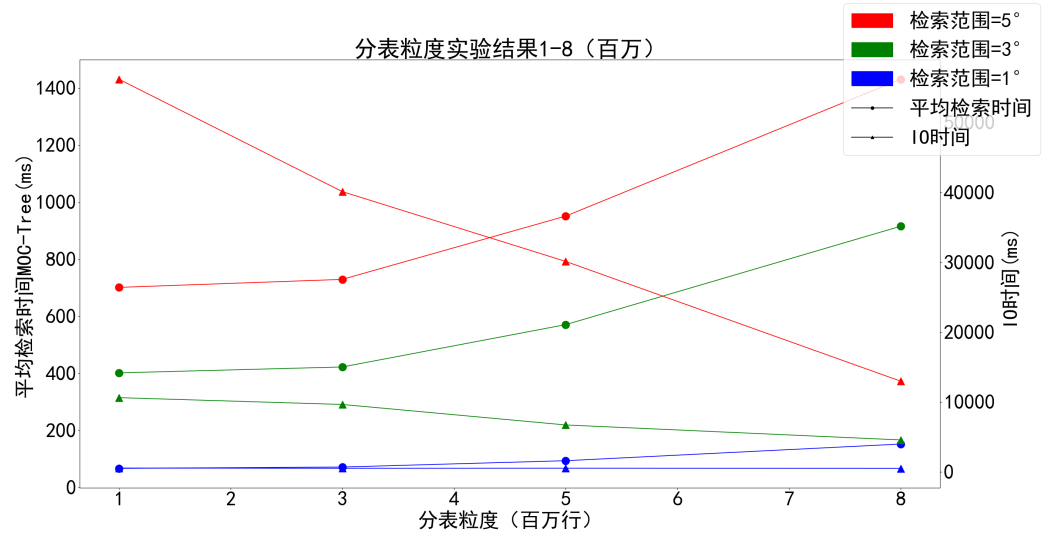
在常用的关系型数据库如 MySQL、PostgreSQL 中，完整的数据表均对应三个文件，分别是数据文件、索引文件以及表结构文件。星表的分表实际上是拆分上述三个文件，并能够在检索层面上重新整合。我们基于 HEALPix 天球划分对星表进行分表，即将星表中的条目根据其空间位置映射至不同的 HEALPix 分区网格中。由于 HEALPix 天球划分的递归性和空间一致性，对于不同量级的多波段星表可以采用不同层级的划分方式。虽然这样造成了不同星表的分表的粒度不同，我们通过上一节介绍的基于多层级覆盖天区的空间检索方法仍能完成多表的联合检索。

3.3.2 分表分区粒度

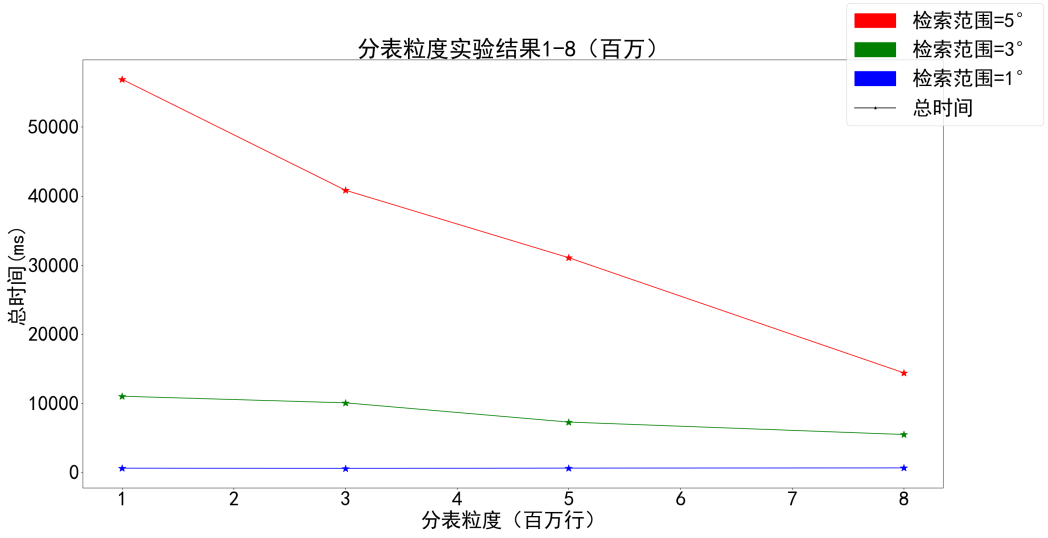
对海量星表分表分区首先要确定划分粒度。分表粒度的主要依据是数据库计算节点的数据服务能力，需要由具体硬件资源、网络环境而定。分表粒度并非越细越好，分表过多，子表条目过少，会造成总表映射分配时间增加，也会造成计算资源的浪费。此外，分表的主要目标是通过横向划分增加计算资源在检索事务上的应用，其本质是增加数据库的并发检索能力。而在基于天球划分索引的星表中，针对每一个划分网格的检索都可以看成是一次并发。因此并发检索的效率与空间检索的面积息息相关，进而常用的检索场景也是确定一个星表分表粒度的主要因素。

我们基于典型的服务器环境（4 核 3.0GHz 主频，16GB RAM，SSD 存储），利用 SDSS DR14 测光星表（共计 1,231,051,050 行）在不同检索场景下进行了一

系列分表实验，分别验证了分表粒度在 100 万至 800 万，与 1000 万至 1 亿量级的检索效率。检索条件是以固定空间位置为中心，1°、3°、5° 不同半径的锥形检索。具体结果如图 3.13、 3.14。



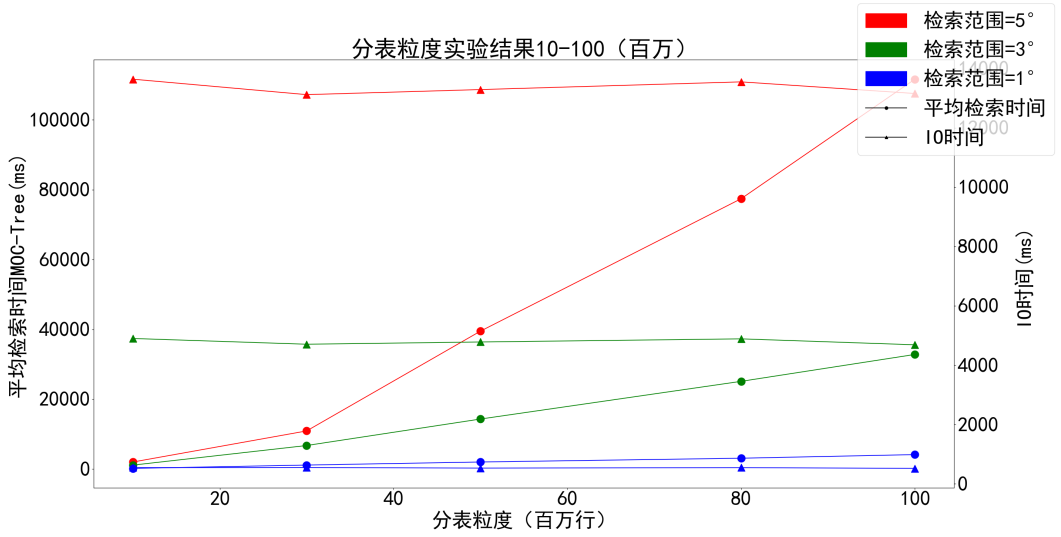
(a)



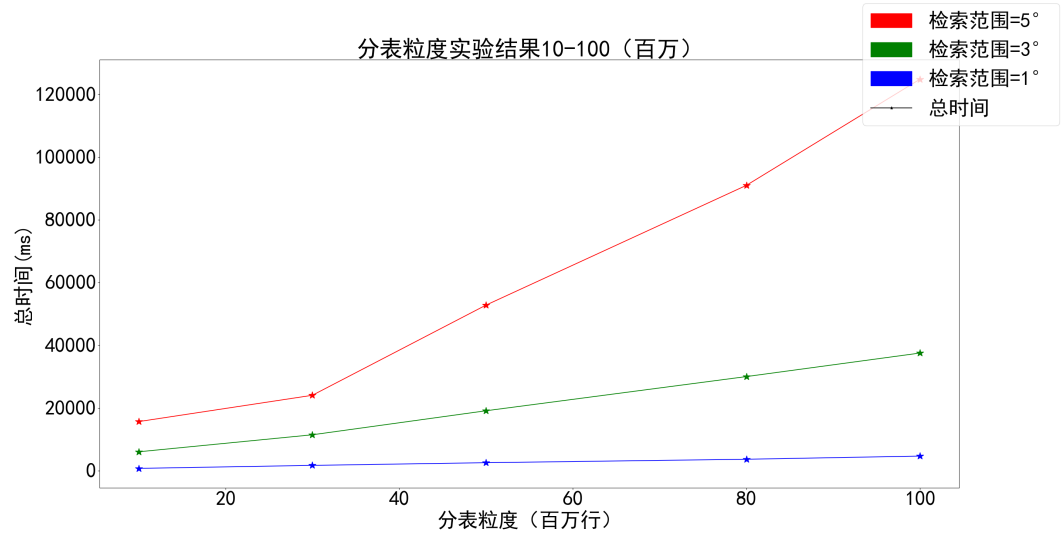
(b)

图 3.13 基于 SDSS 数据的 100 万至 800 万条条目分表粒度实验结果，检索半径分别为 1 度、3 度、5 度。(a) 为分表粒度在 100 万至 800 万条目间的平均检索时间及 IO 时间。(b) 为分表粒度在 100 万至 800 万条目间的总时间消耗。

从结果中可以看出，当空间检索半径较小时（1°），平均检索时间随子表条目数的增大平缓增长，IO 时间基本无变化。这是因为当检索半径较小时，所有



(a)



(b)

图 3.14 基于 SDSS 数据的分 1000 万至 1 亿条目表粒度实验结果，检索半径分别为 1 度、3 度、5 度。(a) 为分表粒度在 1000 万至 1 亿条目间的平均检索时间。(b) 为分表粒度在 1000 万至 1 亿条目间的总时间消耗。

检索结果较大概率来自同一个子表，不涉及检索结果组合的问题。

当空间检索半径较大（3°），子表条目在 100 万至 800 万间时，平均检索时间随子表条目增加增幅较快，而 IO 时间呈下降趋势。子表条目在 1000 万至 1 亿条间时，平均检索时间线性增长，IO 时间基本无变化。总体时间花费在 100 万至 800 万的子表条目间呈平缓下降趋势，而在 1000 万至 1 亿条目间呈线性上升趋势。即当检索范围较大时，检索结果可能来自多个子表，需要归并检索结果，故子表条目越少，IO 时间越高。当子表条目达到一定量级，不再需要归并检索结果，则 IO 时间趋于平稳。但子表条目越大时，平均检索时间越长。综合来看，在检索范围较大时，分表条目在 500 万至 1000 万以内的较为适宜。

当空间检索半径极大时（5°），检索时间随子表条目数的增长而显著增长，而 IO 时间随子表条目增长线性下降，当子表条目在 1000 万以上时，IO 时间趋于稳定。原因同样是因为当检索范围极大时，需要从不同的子表读出结果归并，当子表条目数越多，需归并的次数越少。从总时间花费来看，子表条目数从 100 万至 800 万间时总时间支出呈线性下降趋势，1000 万至 1 亿条时总时间支出呈线性增长趋势。故分表条目在 1000 万左右较为适宜。

综上，在多波段星表融合的检索场景下（检索半径 1° 至 5°），对于十亿级的星表分表，子表条目数在 1000 万为宜。

此外可以看出，IO 时间在整个检索时间支出中占比较高，因此提高数据库节点的磁盘 IO 性能能够较好的提升检索效率。在数据库分表后，各个子表应根据磁盘 IO 性能进行分区。当前主流的高性能服务器采用 SSD 作为磁盘存储。SSD 主要由 NAND Flash 单元构成存储空间，每个 NAND Flash 单元包含多个 Block，每个 Block 又包含多个 Page，如图 3.15。由于 NAND 的特性，其存取都必须以 page 为单位，即每次读写至少是一个 page，通常每个 page 的大小为 4KB 或者 8KB。而常用的关系型数据库的基本存储单位为 8KB（PostgreSQL）或 16KB（MySQL）。因此在实际部署时，需要根据 SSD 的平均随机读速率，具体的 page 尺寸，结合使用的关系型数据库的基本存储单元尺寸来具体确定分区粒度。如某型号 SSD 随机读速率为 150MB/s, page 尺寸为 8KB，使用 PostgreSQL 数据库，则分区文件应保持为 8KB 的整数倍并小于 150MB。

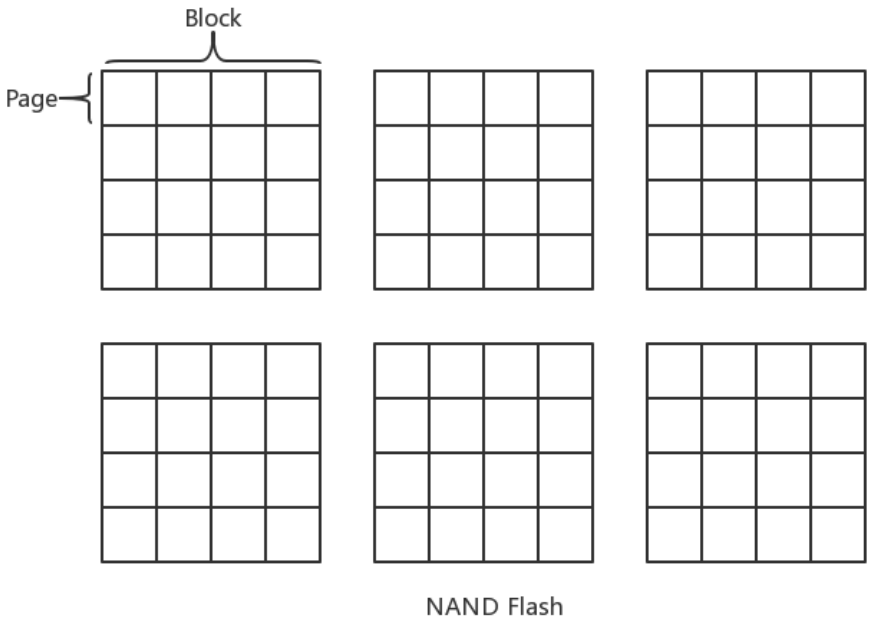


图 3.15 典型 SSD 存储单元结构。

3.3.3 基于 HEALPix 天球划分的分表分区流程

如图 3.16展现了对一个典型星表的分表分区流程。在确定分表粒度后，提取星表中的每一个条目的位置坐标（RA，Dec），通过公式 3.2、3.3确定该天体所在的 HEALPix 网格，其中转换计算所需的层级 k 根据分表粒度确定，如公式 3.6。其中 N 为星表总条目数，S 为子表条目数。

$$k = I\left(lb\left(\frac{N}{12 \times S}\right)\right) - 1 \quad \ldots (3.6)$$

针对上述 HEALPix 网格构建数据文件、索引文件以及表结构文件，形成子表，并分布至数据库节点。再由数据库存储引擎对各子表进行分区，构建分区文件及映射关系文件。

3.3.4 分表联合检索策略

对多波段星表分表带来的一个主要问题是多表关联操作 join 空间查询无法实现，以图 3.17为例，星表 1 和星表 2 分别以自增 ID 进行范围切分，子表条目分别是 3 和 6。对这两个星表进行 join 查询能够得到 ID 间的对应关系，即星表 1 的子表 1、2、3 对应星表 2 的子表 1、2。但对位置坐标字段进行联合则无法实现，因为无法得知星表 1 中各天体位置分布在星表 2 中哪个子表。

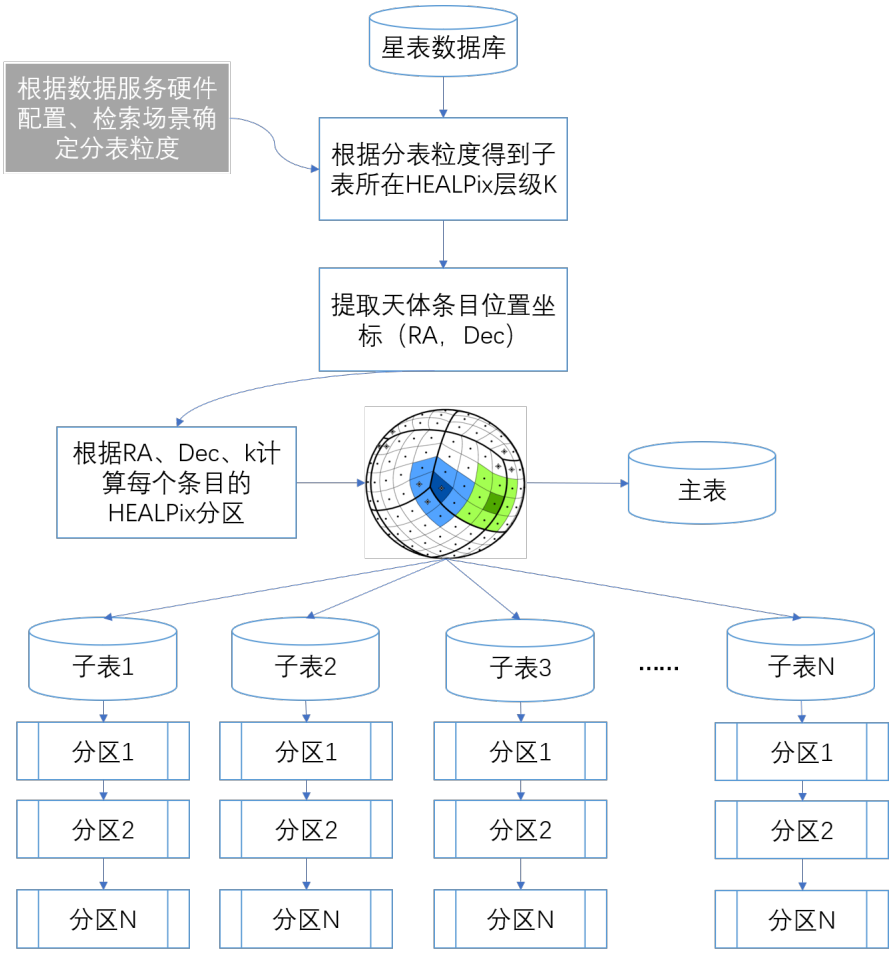


图 3.16 基于 HEALPix 天球划分的星表的分表分区流程

而基于 HEALPix 天球划分的分表的一个特性是子表间的空间位置字段存在对应关系。这是由于在分表映射时，各子表映射到的 HEALPix 网格存在归并关系，即每个星表的划分粒度与 HEALPix 天球划分层级 K 相关，而不同划分粒度可以根据其对应层级进行子表间的归并。

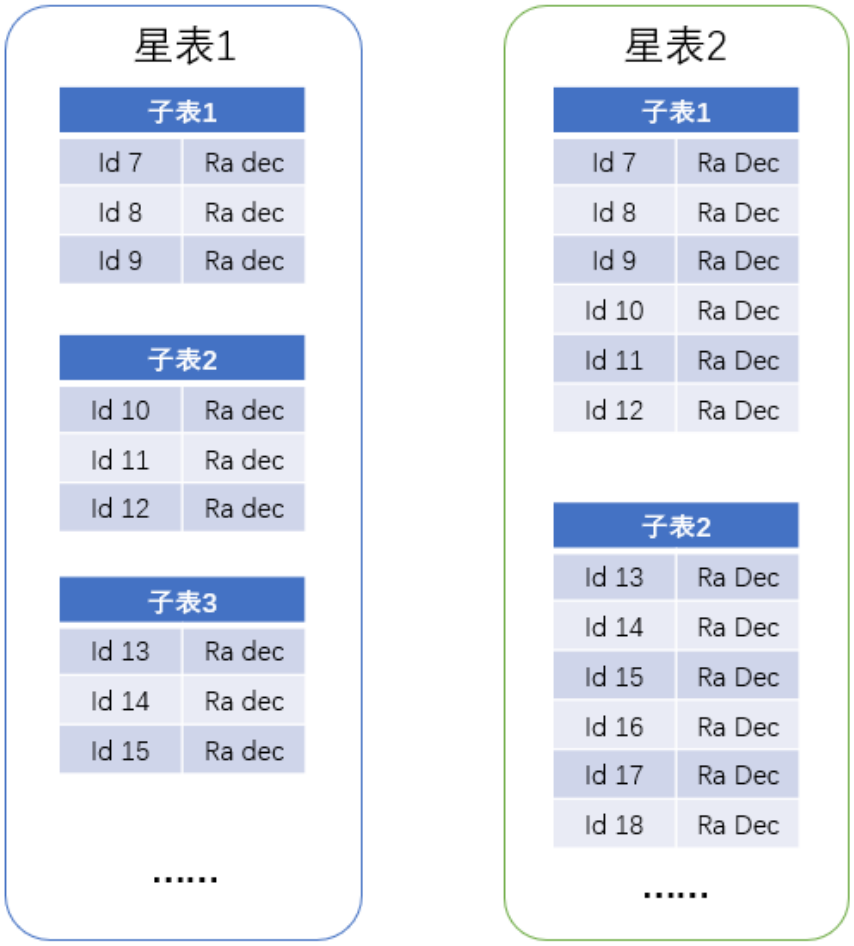


图 3.17 分表后无法直接多表关联空间查询示例

在进行联合检索时，首先将检索条件转化为多层级覆盖天区的 MOC-Tree，查询 join 语句左侧表符合查询条件的条目，根据总表索引得出其覆盖的子表，进而得出各子表所在的 HEALPix 索引号。将这些索引号归并至 join 语句右侧表分表的层级，找出右侧表对应的子表，再以左表检索结果匹配右表，得到联合检索的结果。其流程如图 3.18。

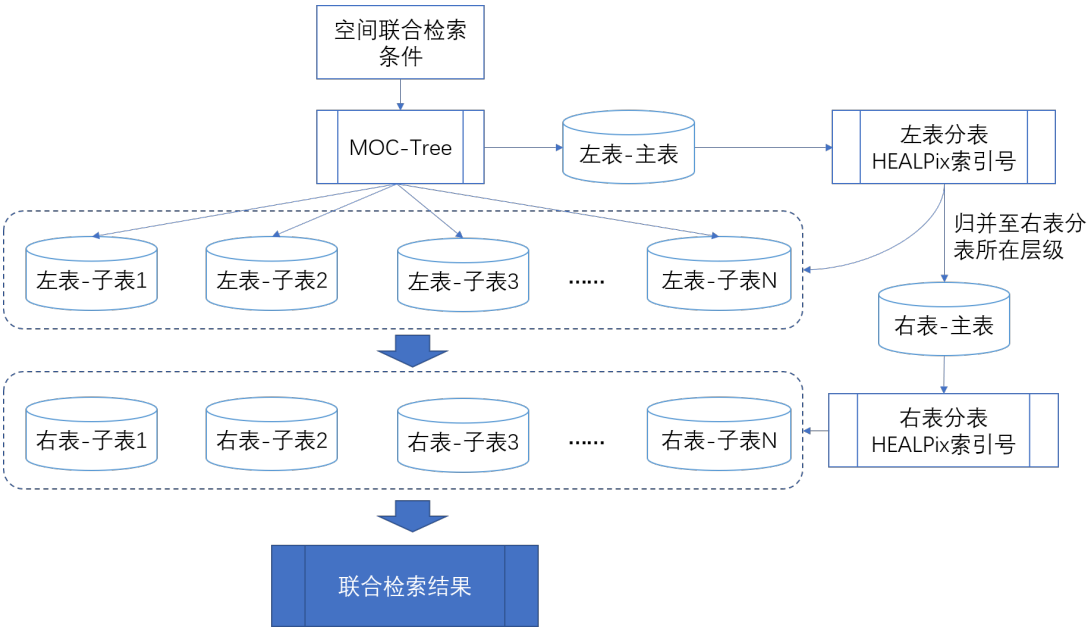


图 3.18 基于 HEALPix 天球划分的分表后联合空间查询流程

3.4　多波段星表数据库设计实现及测试

为了验证本章所研发的基于多层级覆盖天区的空间索引、海量星表分治策略的实际性能，我们基于开源数据库引擎 PostgreSQL 构建了一个多波段星表数据库，并在该数据库上实现了本章所介绍的方法和策略，同时开展了一系列性能测试，并与 Q3C、HEALPix 索引方法进行了对比。

采用 PostgreSQL 作为多波段星表数据库的引擎的主要原因在于其优秀的可扩展性和自定义能力。在 PostgreSQL 中实现本章中提出的方法，需要构建自定义函数。PostgreSQL 的自定义函数也称为存储过程，是存储在数据库服务器上并可以使用 SQL 语句调用的一组过程语句 (声明，分配，循环，控制流程等)。一般用来进行非 SQL 内置计算功能或较为复杂的计算逻辑。PostgreSQL 中实现的自定义函数包括完成从天体坐标到 MOC-Tree 编码的转换、将检索区域转化为多层次盖天区网格、分表 HEALPix 映射、多表联合查询流程等。这些自定义函数能够在 PostgreSQL 的创建表、创建索引、空间查询的 SQL 语句被自动调用，大幅度降低了实现验证的工作量。

PostgreSQL 还可以扩展索引接口，本文提出的基于多层级覆盖天区的空间索引可以通过该特性实现。其基本原理是将 MOC-Tree 编码作为一个新的数据类型，为其定义操作符类，实现 MOC-Tree 编码间的大于、等于、小于的操作运算，这些运算根据前文的描述均采用位运算实现，能够加速检索过程。

3.4.1 多波段星表数据库的部署

多波段星表数据库的体系结构如图 3.19所示。该数据库主要包含了射电、光学、x 射线的多波段星表。此外还包含了 Ligo/Virgo O1 至 O3 期间发布的引力波事件定位天区数据。多波段星表的原始数据以 CSV 文件存储，引力波事件定位天区数据采用 FITS 文件存储。具体星表星系如表 3.2所示。其中光学星表条目数量巨大，其他波段数据体量则相对较小，可用于测试方法在不同体量星表上的表现。

表 3.2 多波段星表信息

星表名称	波段	条目数量	文件尺寸
SDSS DR14 测光星表	光学	1,231,051,050	235GB
2MASS	红外	470,992,971	200.2GB
FIRST 14	射电	946,432	2.3GB
Chandra CSC 2.0	X 射线	317,167	1.5GB
XMM DR6	X 射线	468,440	1.7GB

我们开发了一个基于 Python 的程序 load merge 将 CSV 数据和 FITS 转换为 PostgreSQL 的数据文件、表结构文件和索引文件。基于 PostgreSQL 的自定义函数，这些星表被分表存储在 5 个安装了 PostgreSQL 10.1 版本的虚拟机上, 具体的软硬件配置见表 3.3、 3.4。所有星表的主表都存储在 Master 节点上，其他 4 个节点用于存储分表。其中对光学/红外波段的 SDSS、2MASS 星表进行了分表分区，其他波段星表由于体量较小，未作分区处理。在 PostgreSQL 上安装了 Q3C 和 HEALPix 插件，分别提供基于 Qube Tree 和原始 HEALPix 索引功能，用于对比测试。

表 3.3 多波段星表数据库软件环境

类型	名称	版本
数据库管理软件	PostgreSQL	10.1
操作系统	Ubuntu	14.04
检索插件	Q3C	2.0.0
检索插件	PG_HEALPix	1.0.0
软件运行环境	Python	3.6

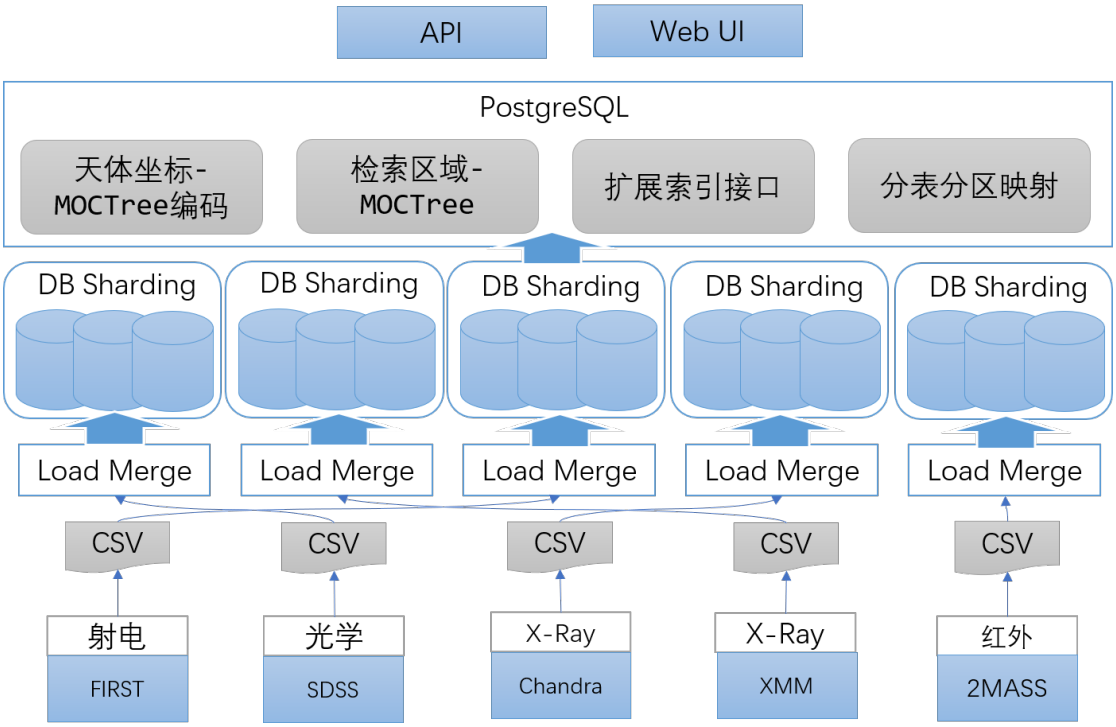


图 3.19 多波段星表数据库的总体架构

表 3.4 多波段星表数据库硬件环境

类型	型号
CPU	i7-8700 6 核 3.2GHz
RAM	16GB DDR4
磁盘存储	500GB SSD
网络带宽	1000 Mbps

3.4.2 测试结果分析

基于上述多波段星表数据库针对本章提出的方法进行了一系列测试与验证，具体包括索引构建时间、索引体积、空间检索效率。

3.4.2.1 索引构建时间与索引体积

针对多波段星表数据库中的各个星表分别构建空间索引，包括 MOC-Tree 索引、Q3C 空间索引、HEALPix 空间索引，具体索引构建时间和体积如表 3.5所示。

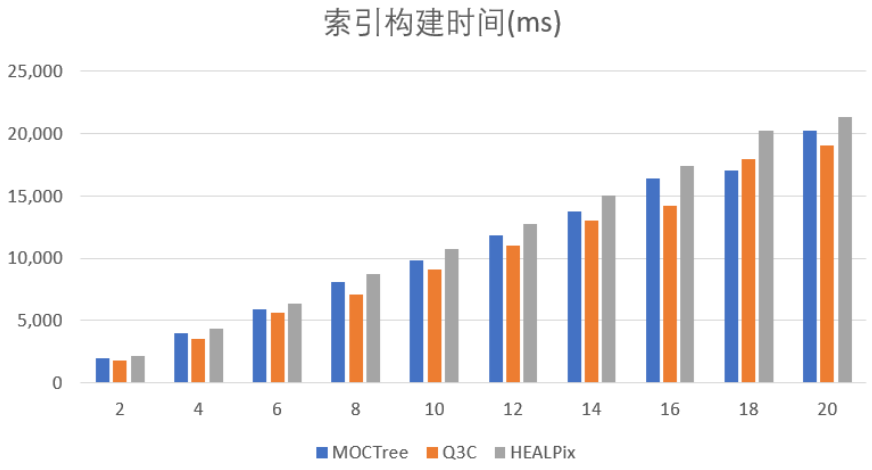
表 3.5 多波段星表索引体积及构建时间对比

星表名称	星表条目	索引类型	索引构建时间（ms）	索引体积（B）
SDSS	1,231,051,050	MOCTree	12,983,738	28,847,389,230
		Q3C	11,845,984	35,562,253,312
		HEALPix	14,137,743	27,651,407,872
2MASS	470,992,971	MOCTree	1,979,280	10,311,903,920
		Q3C	1,882,208.34	12,738,298,393
		HEALPix	2,303,973	9,837,366,273
FIRST	946,432	MOCTree	992	17,377,837
		Q3C	891	21,282,816
		HEALPix	1,048	17,648,593
Chandra	317,167	MOCTree	514	5,538,179
		Q3C	372	7,151,616
		HEALPix	597	5,572,648
XMM	468,440	MOCTree	763	8,173,839
		Q3C	689	10,543,104
		HEALPix	801	8,273,937

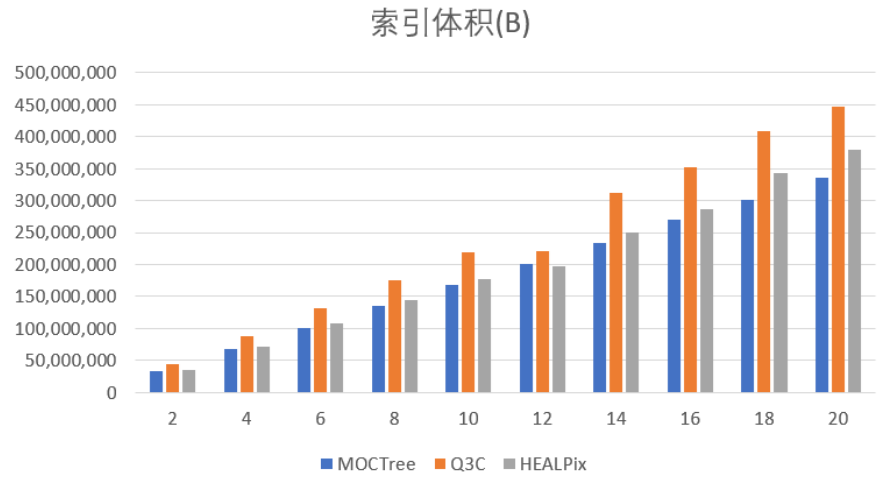
从对比中可以看出，HEALPix 构建索引时间最长，Q3C 采用了四叉树归并算法计算索引，计算效率较高，故花费时间最短。MOC-Tree 与 HEALPix 一样需要对每个条目计算索引，但主要采用位运算，故时间花费低于 HEALPix，但仍高于 Q3C。

在索引体积上，HEALPix 索引的 B 树结构占用空间最小，Q3C 的四叉树占用空间最大。MOC-Tree 的各个子表采用了线性数组构建索引，同时最大层级中的条目构建了 KDTree，占用空间介于 Q3C 和 HEALPix 之间。

此外我们基于 SDSS 星表分别对 2 百万至 2 千万量级的条目进行了不分表测试，对比了索引构建时间和体积对比。结果如图 3.20。结果与多波段星表测试得到的体积和构建时间一致，均呈线性增长趋势。其中 Q3C 的索引体积因四叉树结构随星表条目增大而增长明显。HEALPix 与 MOC-Tree 则增长幅度较一致。



(a)



(b)

图 3.20 2 百万至 2 千万条目星表在不同索引方法下的构建时间（a）和索引体积（b）对比。

3.4.2.2　空间检索效率

采用不同检索方式对数据库中各个星表进行空间查询测试。其中上亿级条目的星表在使用 MOC-Tree 索引的同时也进行了分表操作：SDSS 按照 HEALPix 层级 2 分成了 129 个子表，2MASS 按照 HEALPix 层级 1 分成了 48 个子表。检索条件为以固定坐标为中心，1°、3°、5° 半径范围的锥形检索。在 100 并发下重复检索 30 次取平均值。测试结果如表 3.6 所示。

表 3.6 多波段星表在不同检索范围下的锥形检索时间及 IO 时间对比

星表名称	星表条目	索引类型	1° 查询时间（ms）	1°IO 时间（ms）	3° 查询时间（ms）	3°IO 时间（ms）	5° 查询时间（ms）	5°IO 时间（ms）
		MOCTree-分表分区（129）	201.80	513.20	1321.20	5312.20	2012.40	13642.20
SDSS	1,231,051,050	Q3C	2981.69	511.60	24849.33	5039.60	78135.87.80	11649.10
		HEALPix	3393.77	512.90	27626.84	5066.20	97663.68	11623.00
		MOCTree-分表分区（48）	192.30	513.10	1254.30	5260.60	1969.30	12784.20
2MASS	470,992,971	Q3C	2489.60	511.20	19595.30	5114.20	44356.45	10231.90
		HEALPix	3999.37	513.00	25099.17	5113.20	68081.07	10232.20
		MOCTree	69.50	221.30	391.60	635.60	691.60	1763.80
FIRST	946,432	Q3C	77.10	220.10	593.10	634.70	1659.20	1762.30
		HEALPix	89.90	221.30	688.30	635.40	1871.20	1762.90
		MOCTree	65.30	124.30	139.20	514.30	309.20	783.20
Chandra	317,167	Q3C	77.20	123.60	187.60	515.60	744.10	782.90
		HEALPix	89.10	122.90	219.10	515.90	831.20	783.60
		MOCTree	69.20	153.40	183.80	583.40	400.50	892.10
XMM	468,440	Q3C	77.60	152.20	201.60	582.20	698.20	891.60
		HEALPix	83.40	152.80	239.80	582.60	765.20	892.50

从测试结果可以得出，在十亿级星表（SDSS，2MASS）上，MOC-Tree 索引结合分表分区在检索效率上有较大优势，在不同的检索范围尺度上的检索时间均大幅低于 Q3C、HEALPix 索引方法。在未分表的百万量级星表上，MOC-Tree 索引也优于其他两者，尤其在 3° 和 5° 检索范围上，其检索效率有较大优势。在 3° 检索范围上，MOC-Tree 查询时间比 HEALPix 低 37.7%，比 Q3C 低 27.3%。在 5° 检索范围上，MOC-Tree 查询时间比 HEALPix 低 59.6%，比 Q3C 低 54.8%。这是由于 MOC-Tree 的多层级覆盖天区检索方法降低了进行大范围空间查询时的查询次数。即空间检索范围越大，MOC-Tree 的性能优势越明显。

3.5　本章小结

海量星表高效检索是多波段星表交叉证认的基础，也是面向未来的大型巡天项目所产生的超大规模数据管理的关键技术之一。本章首先介绍了基于关系型数据库构建的星表检索技术，并对常用的空间索引方法进行了详细的说明。针对天文领域常用的基于天区划分的空间索引存在的问题，本文提出了解决方案——基于多级覆盖天区的空间索引 MOC-Tree。MOC-Tree 充分利用了天球划分方法中层级间的迭代属性，能够通过索引号表征天体在所有天球划分层级中的相对位置，结合专用的空间检索方法，能够减少空间检索时的索引扫描次数，进而提高大范围天区检索的时间效率。

此外，单服务器上的资源无法满足海量星表的检索开销，必须考虑横向扩展方案，为此本章还提出了针对 MOC-Tree 索引的海量星表分表分区策略，并针对分表的 join 空间查询提出了分表联合检索方法。基于 SDSS 星表对不同的分表分区粒度进行了一系列空间检索实验，得出结论为在多波段星表融合的典型检索场景下，海量星表的子表条目数在 1000 万为宜，分区文件应保持为 8KB 的整数倍并小于 150MB。

本章还设计并部署了多波段星表星表数据库对提出的方法和策略进行了测试，包括了不同空间索引构建时间、体积以及检索效率的对比。在索引构建时间和体积上，MOC-Tree 介于 HEALPix 索引和 Q3C 四叉树索引之间，但在检索效率上有着很大的提升，尤其是在大范围空间检索时，检索时间大幅低于其他两者。

第 4 章　多波段星表交叉证认及置信度计算

星表交叉证认是多波段星表数据融合的最主要方法。由于观测仪器、数据采集和校准方法的差异，同一天体在不同星表中的坐标会有所差异，这就需要将不同星表中的天体进行匹配。对于观测精度较高的如光学、红外波段星表，主要采用基于位置的方法进行交叉证认。对于多波段星表，由于不同波段观测定位误差差异较大，则还需要给出交叉证认的置信度，这就要采用基于概率的方法。

对于海量星表间的交叉证认，还存在着计算效率的问题。星表交叉证认的计算量是随着星表条目和参数数量的增加呈指数级增长，这对于动辄百万，甚至数十亿级的星表体量无法忍受的。这就要求采用并行处理策略提高计算效率。

本章针对以上问题展开了研究。在基于多层级覆盖天区空间索引的基础上，实现了分表结构下多波段星表的高效交叉证认，并基于贝叶斯方法计算各星表间天体匹配的置信度，最终形成一个以主星表条目为基准的融合星表。该过程的总体流程如图 4.1所示。

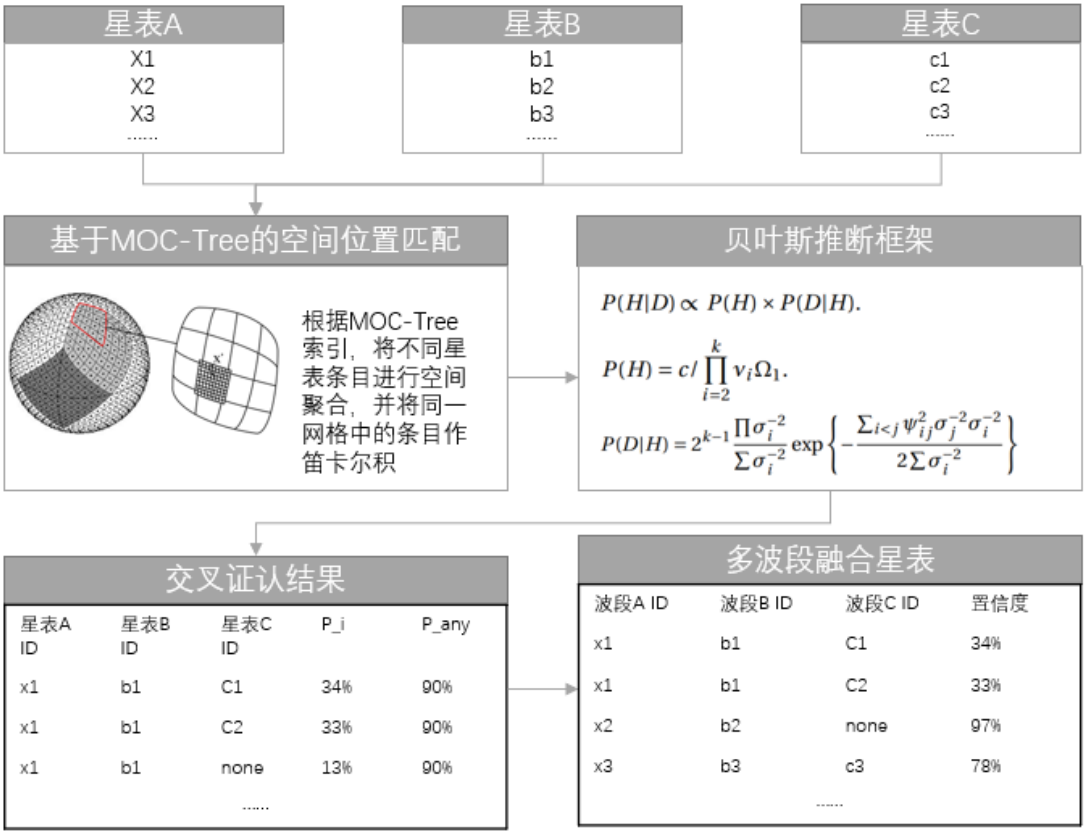


图 4.1 多波段星表交叉证认及置信度计算流程。

4.1 基于位置的星表交叉证认

基于校准误差的距离阈值是识别两个天体是否同源最直观的方法。通常，两个对天体之间的角距离 d 的计算公式如 4.1，其中两个天体 O_1 和 O_2 的坐标为 (ra_1, dec_1) 和 (ra_2, dec_2) 处。

$$d = \arccos\left(\sin\left(dec_1\right)\sin\left(dec_2\right) + \cos\left(dec_1\right)\cos\left(dec_2\right)\cos\left(ra_1 - ra_2\right)\right) \quad \ldots (4.1)$$

当这两个天体足够近时，角距离 d 可近似计算为：

$$d = \sqrt{\left(\left(ra_1 - ra_2\right) \times \cos\delta\right)^2 + \left(dec_1 - dec_2\right)^2} \quad \ldots (4.2)$$

其中 $\delta = (dec_1 + dec_2)/2$。

当角距离 d 小于阈值 γ 时，则认为两者为同一天体，通常 γ 取值为：

$$\gamma = 3 \times \sqrt{r_1^2 + r_2^2} \quad \ldots (4.3)$$

其中 r_1 和 r_2 分别是天体 O_1 和 O_2 对应星表的误差半径。

理论上两个星表的交叉证认应对所有的星表条目进行上述的距离计算，但这样将会产生极大的计算量，因此大多数观测计划及数据中心均开发了基于天球划分的交叉证认策略，如 HEALPix、HTM、Q3C 等，这些方法的原理前文已经介绍过，在此不再赘述。

4.2　基于概率的星表交叉证认

当误差半径在微角秒量级时，这种基于距离的交叉证认方法是有效的。但是在多波段多信使观测时代，不同观测设备的定位误差各异，一个 X 射线源误差范围内可能涵盖数百个光学源，而引力波定位天区更可高达数百平方度。为此需要引入概率方法，给出目标源与各波段天体间匹配的置信度。

最常用的基于概率的交叉证认方法是 LR 方法Sutherland 等 (1992)。LR 方法一般用于两个星表 A、B 间的匹配，将其中一个作为主星表，考虑两者的天体密度、位置坐标（及相对误差）和星等分布，给出来自星表 B 的源是星表 A 中某个源的对应体的可能性与该源是背景中的源的可能性（似然）的比值。之后根据不同的因素计算出一个似然比的阈值，当两个源的似然比在该阈值之上则认为它们是相关联的，即同一天体。该方法也可用于多个星表之间的交叉匹配，即重复该过程实现星表 A 与 C、D 等星表的匹配。如果这些星表均来自相似波长的图像且均具有足够深度，那么对于 A 中的绝大多数源在 BCD 等星表中的对应体是一致的，对于部分源的对应体的关联还可以基于能谱形状做进一步验证。LR 方法已经成功应用在 XMM-COSMOS(Brusa 等, 2007)，CDFS(Luo 等, 2016)，Chandra-COSMOS(Marchesi 等, 2016)，XXL(Georgakakis 等, 2017) 等项目上。这些巡天项目上均采用 LR 方法将 X 射线源分别与光学、近红外、中红外数据进行了独立的匹配。对于 LR 方法未能给出唯一匹配结果的情况，对可用的辅助数据按照可靠性进行排序，将深度较深和分辨率较高的数据排在前面，以选择出正确的对应体。

与数据驱动的 LR 方法相反，基于贝叶斯推断的交叉证认方法使用模型作为先验，因此可以应用于小样本和区域数据之间的交叉证认。这是该方法的优势，但也存在模型分布的假设与真实情况不一致的问题。对于观测深度已经足够在小样本量的情况下构建可靠的经验星等分布模型的星表的交叉证认中，贝叶斯

方法已被广泛应用。贝叶斯方法的另一个优点是可以采用许多先验，且每个先验之间是独立的，因此可以根据其位置、星等、颜色等采用贝叶斯推断来计算不同星表间天体匹配的概率。

(Budavári 等, 2008) 在贝叶斯算法中引入了形式体系，使之能够实现多个星表同时交叉匹配。该方法可利用天文观测获得的贝叶斯因子与其他物理因子组合在一起进行交叉证认的的计算，有效提高了匹配的准确性。例如Roseboom 等 (2009) 等通过计算一定半径范围内每个天体的光度红移和能谱拟合来搜索与亚毫米源相对应的对应体。

多波段星表交叉证认的另一个难点在于，对覆盖数百平方度天区的星表与其它波段星表进行交叉证认时，用于查找对应体的多波段星表往往是由不同观测数据拼接而成，故在不同天区其观测深度不同，从而影响视场中源的实际星等分布，进而影响实际对应体的确定。

4.3 多层级覆盖天区空间索引在星表交叉证认中的应用

无论是采用基于距离还是基于概率计算的交叉证认方法，都与星表数据库采用的索引方式息息相关。这是因为大多数星表的覆盖天区是不同的，需要通过星表检索的方式找出各个星表在某同一区域的天体条目。以星表 A 与星表 B 进行交叉证认举例说明，星表 A 体量较小，仅覆盖 200 平方度，而星表 B 覆盖整个北天区。则应使用星表 A 的覆盖天区作为检索条件，找出星表 B 中对应天体，之后再进行交叉证认运算。这样可以避免星表 A、B 间的全表对比，从而大幅度降低计算量。在这个过程中，星表 A 称为主星表，交叉证认的结果以主星表的条目为基准。

4.3.1 基于天球划分方法的交叉证认实现

基于天球划分方法的空间索引可以实现上述过程。以采用了 HEALPix 空间索引的星表为例，首先比较星表 A 和星表 B 的 HEALPix 索引是否在同一层级，并将层级较高的星表索引降级转换至较低层级。这是因为星表中的 HEALPix 索引层级一般是由星表的定位误差决定的，较低层级的每个 HEALPix 网格覆盖面积较大，即表明该星表的定位误差范围也较大，故应以其覆盖的天区范围为准进行查询。如果反之，较高层级的网格无法覆盖低层级网格，会造成漏源。转换之后对星表 A 的所有 HEALPix 索引构建数组，匹配找出星表 B 中对应的索引号的

条目。由于此时的 HEALPix 索引的网格覆盖范围与星表 A、B 的最大误差范围一致，因此即可认为落在同一索引号网格中的天体可能互为对应体。以上流程如图 4.2。从上述描述中可以看出，该过程其实是星表 A 与星表 B 的 left join 联合查询过程。

图 4.2 基于 HEALPix 的交叉证认实现示例。

上述过程实际上是基于位置方法对两个星表间天体的匹配关系进行了第一次筛选。即使天体落入同一 HEALPix 网格内，由于定位误差的差异，还需要进一步计算它们之间互为对应体的概率。

从上一章的测试结果可以得出，大覆盖范围的空间检索中，HEALPix 索引的查找效率很低。此外，在交叉证认过程中还需要对不同星表的索引层级进行转

换计算，对于大规模星表该步骤花费时间较大，进一步提高了时间消耗。采用其他天球划分的索引方法如 HTM、Q3C 均存在同样问题。

为此，我们提出了基于多层级覆盖天区空间索引的位置匹配方法，实现海量星表的高效位置匹配，作为星表交叉证认的第一步工作。

4.3.2 多层级覆盖天区空间索引下的位置匹配

多层级覆盖天区空间索引在一定程度上解决了上述问题。首先，它不需要对多个星表进行层级变换操作，因为多层级覆盖天区索引的 MOC-Tree 编码方式记录了天体在每个层级的相对位置，默认 32 位 MOC-Tree 索引也记录了每个天体从 0 层至 14 层的所有相对位置。即使两个星表的定位误差范围不同，也可以通过截断对比的方式找出落在较小层级的同一网格中的天体。如星表 A 的误差范围使其记录的天体条目在 HEALPix 索引第 k_1 级，则其 MOC-Tree 索引的前 $4+2k_1$ 位为有效位数。星表 B 定位误差范围较小，其 HEALPix 索引达到层级 $k_2(K_2 > k_1)$，则其 MOC-Tree 索引的前 $4+2k_2$ 位为有效位数。在进行对比时，只需对量星表条目中 MOC-Tree 索引的前 $4+2k_2$ 位进行异或运算，结果为 0 则说明对应的天体落入层级 k_1 的同一 HEALPix 网格内，即它们可能互为对应体。其具体运算过程如下：

1. 从星表的元数据表中得到各个星表的 MOC-Tree 索引有效位数。

2. MOC-Tree 索引有效位数较小说明该星表定位误差范围较大，将其作为主星表，设其有效位数为 P。

3. 以主表作为左表进行索引对比，将两者的索引号均右移 $32-p$ 位，并在之前空出的位中填充 1。

4. 对两者索引号作异或运算，结果为 1 则说明该天体位于同一 HEALPix 网格，存在位置匹配关系。

对于进行了分表分区的大规模星表，需要做进一步处理才能实现星表间的位置匹配。在上一章中介绍过，本文基于 HEALPix 天球划分构建了星表的分表分区策略，对于数亿级的星表按照 1000 万左右的子表条目粒度进行分表。因此，对于分表后的星表间的位置匹配，主要存在以下三种情况：

1. 只有一个表实施了分表。将未分表的星表作为主星表 A，并将其索引数组按照已分表的星表 B 的 HEALPix 天球划分层级分组，再将分组后的索引与星表 B 的各子表分别进行，最后将结果合并，其流程如图 4.3。

2. 两表均实施了分表，且分表粒度不同。将分表所在 HEALPix 层级较低的星表作为主星表 A（分表依据的 HEALPix 层级较低说明其条目较少，覆盖天区也较小），将其各子表对应的 HEALPix 网格归化至星表 B 分表所依据的 HEALPix 层级，此时星表 A 的一个子表对应星表 B 的多个子表，再采用情况 1 的方法分别进行位置匹配并合并结果。

3. 两表均实施了分表，分表粒度相同。将覆盖天区较小的星表作为主星表，根据各子表所在的 HEALPix 网格分别进行 Left Join 操作，最后合并结果。

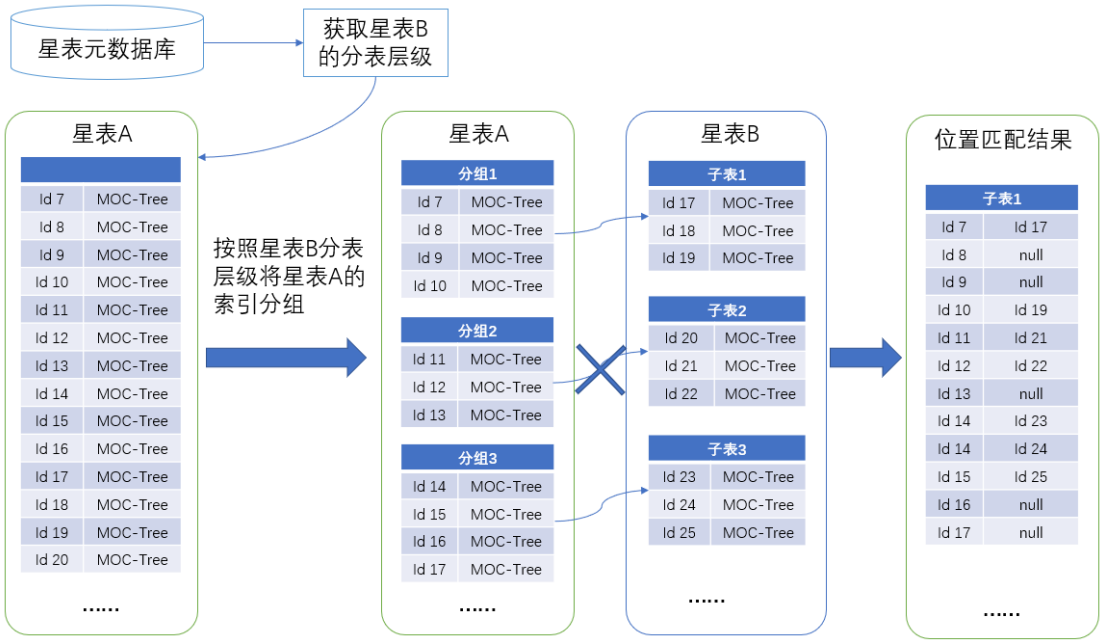


图 4.3 只有一个表实施了分表时星表间的位置匹配流程

通过上述方法可高效完成星表间的位置匹配。对于多星表间的位置匹配，我们采用的策略是以主星表为基准，依次对其他星表进行 Left Join 操作，再将所有的匹配结果以主星表条目为基准合并。其耗费时间与参与位置匹配的星表数目和星表体量相关。

在同一 HEALPix 网格中可能包含各星表的多个天体，为此还需要对其进行进一步的关联。由于此时完成关联的天体的距离关系已经在主星表的定位误差范围以内，所以基于位置的交叉证认已经无法再给出更精确的天体匹配，我们需要进一步计算这些天体间互为对应体的概率。

4.4 基于贝叶斯推断的交叉证认置信度计算

完成基于位置关系的第一轮匹配后，接着对这些形成关联的天体进行它们互为对应体的置信度计算。对位于每个 HEALPix 网格中的 k 个星表的天体作笛卡尔积，得到 n 个匹配组合的集合 $E = \{e_1, e_2, e_3, ..., e_n\}$，其中一个匹配组合 $e_i = \{x_1, x_2, x_3, ..., x_k\}$ 表示 K 个来自不同星表的天体间互为对应体。$X_i, i \in [1, k]$ 包含了至少 ID、RA、DEC、Position_Error 字段。

在进行交叉证认的置信度计算之前，首先需要从元数据表中获取每个参与的星表的覆盖天区和星表条目数。以 N_i 表示参与交叉证认的第 i 个星表的条目数，以 Ω_i 作为该星表的覆盖天区面积，则该星表的天体密度表示为 $\upsilon_i = N_i/\Omega_i$。假设主星表的一个天体在所有参与交叉证认的 k 个星表中均存在对应体，则存在 $\prod_{i=1}^{k} N_i$ 种可能的关联。给定一个假设 H，对于 k 个覆盖相同天区且源分布均匀的星表，它们中的天体在空间位置上一致的概率可写作公式 4.4。

$$P(H) = \frac{N_1}{\prod_{i=1}^{k} N_i} = \frac{1}{\prod_{i=2}^{k} N_i} = \frac{1}{\prod_{i=2}^{k} \upsilon_i \Omega_i} \quad \ldots (4.4)$$

P(H) 即为星表中天体互为对应体的先验概率。但在实际应用中，采用的星表的覆盖天区并不一定相同。因此还需引入一个先验完整性因子 C 来给出更可靠的对应关系概率。即公式 4.5：

$$P(H) = \frac{c}{\prod_{i=2}^{k} \upsilon_i \Omega_i} \quad \ldots (4.5)$$

设坐标集合 $D = \{x_1, x_2, ..., x_k\}$ 对应 k 个不同星表中的观测坐标。基于贝叶斯公式，可得出在纳入观测数据 D 后，K 个星表中天体同源的后验概率 $P(H|D)$ 与先验概率 $P(H)$ 的关联可由公式 4.6表示：

$$P(H|D) \propto P(H) \times P(D|H) \quad \ldots (4.6)$$

贝叶斯因子可以作为似然函数。给定假设 H_i，表示一个匹配组合 e_i 中的所有天体均同源；假设 H_0 为该匹配组合 e_i 中所有天体均不同源。通过计算 H_i 和 H_0 的先验概率和后验概率，可得到贝叶斯因子，如公式 4.7：

$$\begin{aligned} \frac{P(H_i|D)}{P(H_0|D)} &\propto \frac{P(H_i)}{P(H_0)} \times \frac{P(D|H_i)}{P(D|H_0)} \\ B &= \frac{P(D|H_i)}{P(D|H_0)} \\ P(H_i|D) &= \left[1 + \frac{1 - P(H_i)}{B \times P(H_i)}\right]^{-1} \end{aligned} \tag{4.7}$$

当 B 数值远大于 1 时，则认为假设 H_i 成立，即该匹配中所有天体同源；若 B 小于 1，则假设 H_0 成立，即该匹配均不同源。如果 B 接近 1，这可认为两种假设均存疑，还需要更多证据检验。

本文以 (Budavári 等, 2008) 中基于星表观测误差符合球形正态分布 (Breitenberger, 1963)，结合公式 4.7得到的贝叶斯因子作为似然函数 4.8：

$$P(D|H) = 2^{k-1} \frac{\prod \sigma_i^{-2}}{\sum \sigma_i^{-2}} exp\left\{ -\frac{\sum_{i<j} \phi_{ij} \sigma_j^{-2} \sigma_i^{-2}}{2 \sum \sigma_i^{-2}} \right\} \quad \ldots (4.8)$$

其中 σ 为星表的观测精度误差，以角秒表示。ϕ_{ij} 指该组匹配中第 i 个星表和第 j 个星表中对应天体的角间距。

采用贝叶斯因子作为似然函数的优势在于，当得到观测证据时，可将根据其得到的贝叶斯因子与旧贝叶斯因子相乘，得到新的贝叶斯因子作为先验概率。从而可以不断加入新的数据和证据对检验进行假设。

接下来通过计算给出一系列置信度估计的计算。首先给出对于主星表中的某个源 i，它在其他星表中是否有对应体的概率 P_{any}，后验概率 $P(H|D)$ 是通过先验概率 4.5和似然函数 4.8基于贝叶斯公式 4.6计算得到的。针对所有可能的匹配组合计算该后验概率 $P(H_i|D), i \in [1, n]$，并将其归一化。结合假设 H_0 的后验概率 $P(H_0|D)$，可以得出 P_{any} 的计算公式 4.9:

$$P_{any} = 1 - \frac{P(H_0|D)}{\sum_i P(H_i|D)} \quad \ldots (4.9)$$

P_{any} 的值越接近 1，则表明主星表中的该天体在其他星表中存在至少一组对应体。如果 P_{any} 很小，则表明主星表中该天体在其他星表中很有可能没有对应体。但是对于实际运用中该阈值的设定，应根据参与交叉证认的星表来决定。为此我们对每个参与交叉证认的星表制作了一个副本，将其位置坐标进行了偏移，

其偏移量远大于星表本身的定位误差。将该副本与原星表进行交叉证认，并计算各组匹配的概率，从而得出该星表匹配的误报率 $cutoff$。即当该星表作为主星表进行交叉证认置信度计算时，对 P_{any} 设定的阈值应大于该 $cutoff$ 值。

之后针对每个匹配组合计算其置信度 $P_i, i \in [1, n]$，即第 i 组匹配组合对于所有匹配组合的相对概率，如公式 4.10。

$$P_i = \frac{P(H_i|D)}{\sum_{i>0} P(H_i|D)} \quad \ldots (4.10)$$

如果某组匹配组合的 P_i 结果较高，则说明该组匹配的天体互为对应体的可能性较大。

在对交叉证认得到的每组天体匹配组合中，P_{any} 和 P_i 值只会出现三种情况：P_{any} 和 P_i 均较大：一般来说，对于主星表中的各个天体，当其 P_{any} 值大于 95%，且有某组匹配组合的 P_i 大于 95%，则认为该组匹配组合是非常可靠的。P_{any} 较大，存在多个匹配组合的 P_i 值相接近：说明主星表的该天体在其他星表中存在对应体，但现有数据无法分区出对应体，需要更多证据。P_{any} 和 P_i 均较小：P_{any} 较小是指其接近 $cutoff$ 值。此时说明现有数据无法支持主星表的天体在其他星表中存在对应体。

4.5 置信度计算的并行化处理

由于需要对于落入同一 HEALPix 网格中的不同星表天体作笛卡尔积，并对所有结果执行置信度计算，当星表规模较大时，其计算量程指数级增长。为此需要进行并行化计算实现加速。由于这些匹配组合均位于同一个 HEALPix 网格，这就为并行处理提供了基础的条件。在上一章的多波段星表数据库的实现上，我们针对不同的星表规模间的交叉证认置信度计算采用了不同的并行处理方式。

当参与交叉证认的星表均未分表时。采用多线程的方式实现并行化，每个 HEALPix 网格中的置信度计算作为一个线程，并发数与星表部署的数据库节点 CPU 核心数相关。在本文构建的多波段星表数据库中，数据库节点为单 CPU6 核心，故能够实现 6 个 HEALPix 网格的并行计算。

当参与交叉证认的部分星表或全部实施了分表时。这种情况下，在进行空间匹配时已经对分表联合检索的粒度匹配进行了处理，在位置匹配完成后直接生

成进程执行置信度计算，在进程内部再按照上一种实施线程并行，即其并发数为与星表分表所处数据库节点数与 CPU 核数的乘积。

4.6　测试与验证

基于上一章构建的多波段星表数据库，我们对本章提出的多层级覆盖天区空间索引下的星表间位置匹配方法进行了性能测试，并对匹配的置信度计算进行了验证。COSMOS(Cosmic Evolution Survey，宇宙演化巡天) 天区 (Scoville 等, 2007) 提供了理想的测试平台。COSMOS 天区是指哈勃空间望远镜的 COSMOS 巡天对 10h 00m 28.6s +02° 12’ 21.0”(J2000) 附近一片 1.4° × 1.4° 区域所进行的巡天，其他观测设施在多个波段对这块区域进行了均匀且深入的观测。其中 XMM-Newton 对该区域进行了 X 射线波段的观测 (Cappelluti 等, 2007)，Brusa 等 (2007) 采用 LR 方法将 XMM-Newton 的观测结果与 I 波段 CFHT/Megacam 星表 (McCracken 等, 2007) 进行了交叉证认。随后Brusa 等 (2010) 采用 LR 方法结合人工检查，使用近红外 (McCracken 等, 2009) 和中红外 (Ilbert 等, 2010) 波段星表，分别对该交叉证认结果进行了改进，得到了 XMM-COSMOS 多波段星表。在最近几年，Chandra 对 COSMOS 视场也进行了观测，获得了更深入和均匀的观测结果 (Marchesi 等, 2016)。基于 Chandra 观测数据更好的定位能力，XMM-COSMOS 多波段星表中的匹配关联得到了验证和更新。

首先以 XMM-COSMOS 的 XMM 星表为主星表，分别与多波段星表数据库中的 SDSS、2MASS、FIRST、Chandra 星表进行交叉证认，并对比了采用 Q3C 和 HEALPix 索引方式的时间花费。其结果如表 4.1所示。

从对比结果可以看出，对于 COSMOS 天区的覆盖范围尺寸（1.4° × 1.4°），MOC-Tree 发挥了其高效检索的优势，Left Join 操作的执行时间在不同体量的星表上均小于其他两种索引方式。尤其在数亿级的 SDSS 星表与 2MASS 星表中，基于分表分区处理，其位置匹配时间远小于未作分表分区处理的 Q3C 和 HEALPix 索引方法。

XMM-COSMOS 的 XMM 星表共包含 1848 个点源，在后继的 Chandra 观测中对该星表中 1281 个源的对应体进行了确认，128 个源的对应体进行了修正。我们主要通过这 1409 个源在光学波段（SDSS）和红外波段（2MASS）来验证交叉证认的准确性。在 COSMOS 天区中，XMM 星表的平均误差范围为 1.8 角秒（最

表 4.1 XMM-COSMOS 的 XMM 星表与多波段星表位置匹配的时间花费

星表名称	星表条目	索引类型	位置匹配时间（ms）
		MOCTree-分表分区（129）	419.40
SDSS	1,231,051,050	Q3C	5843.30
		HEALPix	6566.80
		MOCTree-分表分区（48）	394.20
2MASS	470,992,971	Q3C	4116.60
		HEALPix	5384.30
		MOCTree	173.20
FIRST	946,432	Q3C	198.60
		HEALPix	224.10
		MOCTree	131.40
Chandra	317,167	Q3C	158.20
		HEALPix	195.40

小 0.1 角秒，最大 7.33 角秒），SDSS 和 2MASS 的误差范围均设定为 0.1 角秒。在交叉证认的结果中，获得了 93.1% 的正确率，共计 1312 个正确匹配，其中 1259 个源的对应体与 XMM-COSMOS 原始星表中的对应体一致，53 个源的对应体与基于 Chandra 修正后的对应体一致，高于 LR 方法的 90.9% 的准确率。

4.7 本章小结

本章在上一章提出的海量星表高效检索方法的基础上，实现了多波段星表的高效交叉证认和置信度计算。首先基于多层级覆盖天区空间索引实现星表天体间的位置匹配，匹配的天体落入误差范围较大的星表所在 HEALPix 层级的网格中。在这个阶段采用了 MOCTree 索引的位置匹配方法，相较基于天球划分索引的匹配方式省去了层级归化的计算步骤，且采用位运算的比较方式，大幅度提升了位置匹配的效率。对于实施了分表的海量星表间的位置匹配，也针对不同情况分别实现了匹配策略。

由于不同星表存在定位误差范围差异，还要计算各星表间天体位置匹配组合的置信度，通过采用基于贝叶斯推断的置信度计算方法，以星表定位误差和源密度计算先验概率，并以星表观测误差符合球形正态分布而得到的贝叶斯因子

作为似然函数，针对以主星表条目为基准的各组匹配组合分别计算置信度（后验概率）P_i，并给出主星表每个条目存在对应体的概率 P_{any}。通过 P_i 和 P_{any} 的组合判断各组匹配组合的可靠性。此外，对于海量星表的交叉证认置信度计算采取了并行化处理，提高了其计算效率。

最后，采用了 COSMOS 天区作为测试区域，以 XMM-COSMO 的 XMM 星表作为主星表，与多波段星表进行了交叉证认。结果表明，基于多层级覆盖天区的位置匹配方法具有较高的时间效率。在交叉证认的准确率方面，本章采用基于贝叶斯推断方法也高于 LR 方法。

第 5 章　多波段图像高效组织、检索与可视化框架实现

不同波段的图像数据格式各异，主要有 FITS、CASA、HDF 等。其组织管理方式包括文件系统、关系型数据库、对象存储系统等。不同观测设施的数据中心提供的图像数据获取方式也各不相同，一些提供了按照观测任务组织的图像数据，另外一些则允许用户获取自定义天区和视场的存档观测图像。此外，不同波段的图像的叠加分析及可视化需要进行格式转换、投影转换、图像配准与拼接等一系列繁琐复杂的操作。而多波段、多信使联合观测的数据融合需求要对多源异构图像数据进行统一组织管理和可视化。为此本章将主要探索海量多波段图像数据的组织、检索与可视化方法，针对上述问题提出了一个异构多波段图像统一组织、检索与可视化框架，并在万维望远镜软件（Worldwide Telescope, WWT）上实现了这一框架。

5.1　天文图像数据组织方式

大型巡天及时域观测计划都会发布自己的观测图像数据，这些数据往往数据量巨大，按照不同的组织方式存储在数据库、文件系统或对象存储系统中。天文图像数据在存储格式上是比较多元化的，比较主流的是用 FITS 格式，射电方面开始流行 CASA，此外还有 HDF5 格式。除了 HDF5 是支持并行 I/O 之外，其他格式都不是为并行存储设计的。

大天区的图像数据经常要划分为成千上万个 FITS 文件，处理时又要读回、拼接、叠加，计算量和处理时间远远超出可实际操作范围，其中系统 I/O 是极大的瓶颈。澳大利亚 ICRAR 的 MWA 及 SKA SDP 技术团队对此的解决方案之一是，预先计算留出 FITS 文件内容各帧所在的空间，“并行”将各帧写入，提高了存储效率，但是并不能从根本上解决问题。尤其是 CASA 这样的打包文件，根本无从预留空间。对此，ICRAR 提出了一种方案，底层采用 HDF5 进行并行存储，而在其上提供一层接口，实现所有 CASA 的读写函数。此时存储的细节就对于 CASA 系统透明了。基于同样的思路，也可以支持 FITS 等格式。相似的，荷兰射电天文学院 (ASTRON) 和 LOFAR 项目考虑到数据多样性、超长字节数、并行 I/O、分布式文件系统等，也在探索用 HDF5 进行 LOFAR 射电数据封装。

5.1.1 基于文件系统的图像组织方式

不同格式的天文数据均存储在分布式文件管理系统中，以实现对海量图像的组织管理。业界目前比较主流的分布/并行存储系统有 Haystack、GFS、TFS 等等。

Facebook 的 Haystack(Beaver 等, 2010) 方案是一个典型的分布式存储架构，技术的核心是将海量的小尺寸文件堆砌到一个大文件中，减少系统级的元数据访问压力，而对每一个真实文件的访问则是采用接口实现，接口会在索引文件或者数据库中获取真实文件在大文件中的偏移量以及数据体大小。目前基于 Haystack 的思想已经有了若干开源的实现，比如 BFS[1] 和 SeaweedFS[2]。

Google 在更早之前的 2003 年提出了它的分布式文件系统解决方案，提供的也是海量数据的分布式服务。GFS 由单 master 和多个 Chunk server 组成。其思想与 Haystack 类似，但是 GFS 块的大小与 Haystack 的不同，每个块的大小是 64MB，并且拥有唯一的 ID，块和块副本都是以文件的形式在服务器中存储。GFS 提供了常见的文件系统的接口，比如目录分层管理等。不同于传统文件系统，GFS 没有能够列出目录下所有文件的每目录数据结构，也不支持同一文件或者目录的别名（例如，Unix 语境中的硬链接或者符号链接）。GFS 将其名称空间逻辑上表现为全路径，作为一个类似主键的存在，方便查找。目前广泛得到使用的 Hadoop Distribute File System（HDFS）(Shvachko 等, 2010) 就是参考 GFS 的开源实现，并且重点面向大数据的读写应用。

阿里巴巴为了解决其在淘宝系统上的海量数据管理问题，开发出了 TFS（Taobao FileSystem）(Fu 等, 2014)，是一个高可扩展、高可用、高性能、面向互联网服务的分布式文件系统，主要针对海量的非结构化数据，它构筑在普通的 Linux 机器集群上，可为外部提供高可靠和高并发的存储访问。TFS 为淘宝提供海量小文件存储，通常文件大小不超过 1M，满足了淘宝对小文件存储的需求，被广泛地应用在淘宝各项应用中。它采用了 HA（High Availability，高可用）架构和平滑扩容，保证了整个文件系统的可用性和扩展性。同时扁平化的数据组织结构，可将文件名映射到文件的物理地址，简化了文件的访问流程，一定程度上为 TFS 提供了良好的读写性能。其设计思路与 Google GFS 类似。

[1] https://github.com/Terry-Mao/bfs

[2] https://github.com/chrislusf/seaweedfs

纵观这些分布式文件系统，有一个大概相似的设计思路，主要包含一个底层文件存储系统（Haystack Store，Chunk Server 等）以及一个目录（元数据、中枢）服务（GFS Master，Haystack Directory 等）。

5.1.2　基于天球划分的图像组织方式

文件系统能过实现海量图像数据的存储和并行处理，但是对于图像检索存在一定的缺陷，需要采用其它手段来弥补这一点。除了将文件访问地址存放在数据库中，另外一种思路是通过天球划分将图像文件与天球区域对应起来，通过位置检索实现对图像的检索。常用的基于天球划分的图像组织方式包括三角分层网格和基于 HEALPix 的层次渐进模式。

5.1.2.1　三角分层网格

三角分层网格 (Hierarchical Triangular Mesh，HTM) 被成功应用于 SDSS 的观测图像组织和管理。HTM 通过对天区递归的多层次三角划分将观测数据与天区一一对应，它从一个八面体开始，依次递归划分，将球面三角投影至天球上，如图 5.1所示。将平面图像转换至球面需要应用投影方法，HTM 采用了 TOAST（Tessellated Octahedral Adaptive Subdivision Transform）投影将观测图像转换至 HTM 网格中。在进行 TOAST 投影转换时，输入图像需为等角矩形，TOAST 会将其转换成 8 个三角形组成的正方形，再将其贴至这 8 个三角形对应的天球区域 (Szalay 等, 2007)。图 5.2展现了这一过程.

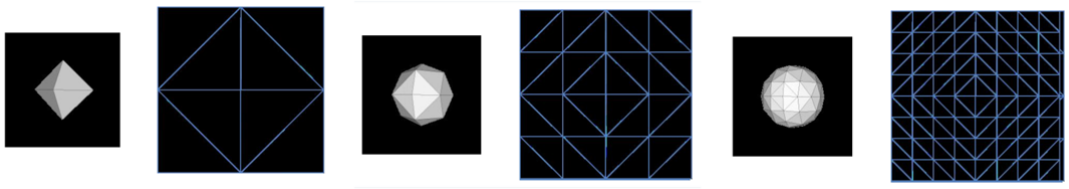
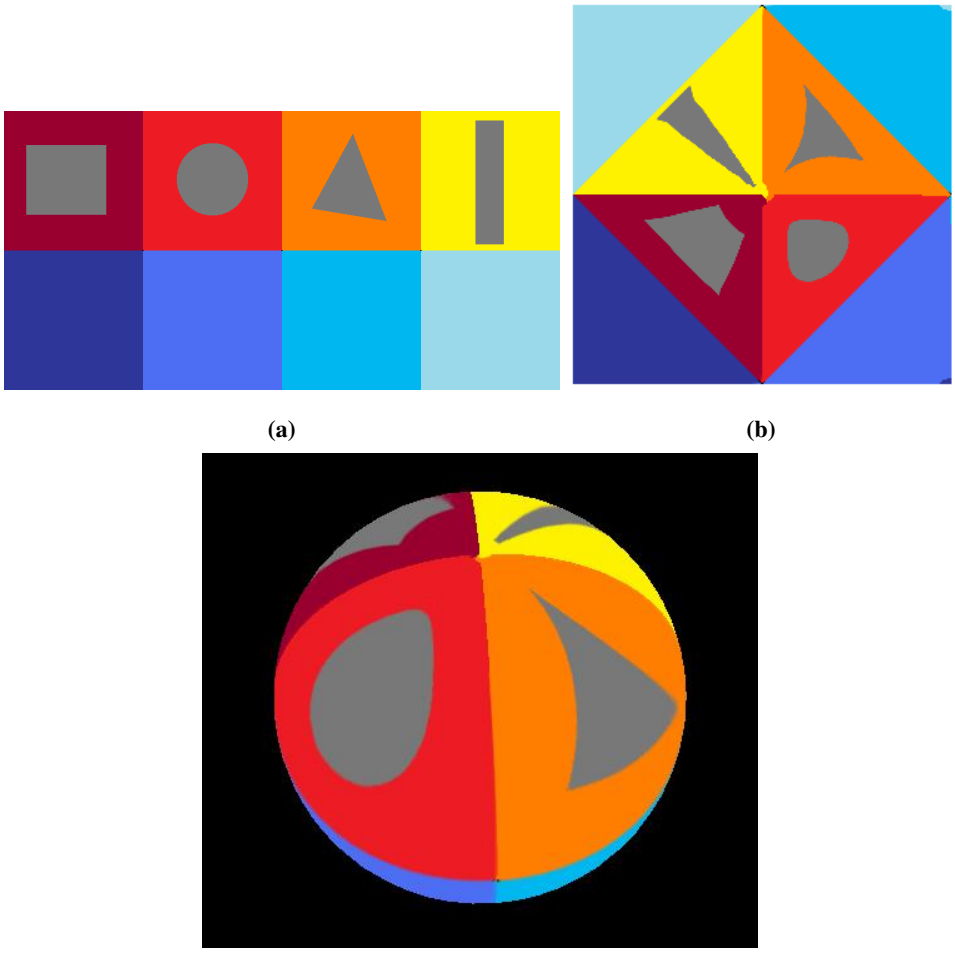


图 5.1 HTM 通过三角形网格递归切分的方式拟合球体，图中分别展示了在层级 0、1、2 级时 HTM 对天球的拟合情况

通过 HTM 编号可以找到图像数据对应的天球区域。如图 5.3，HTM 将基础的 8 个球面三角形按南北方向分别命名为 N0-N3 及 S0-S3。再在三角形各边取中点，并分别连接构成新的三角形。对这些三角形顶点进行编号，每个三角形的顶点被命名为 0、1、2，各边对应的中点被命名为 0’、1’、2’。子三角形通过在其父三角形编号后添加 0、1、2、3 进行命名。此时，子三角形的编号会越来越长，



(a) (b)

(c)

图 5.2 TOAST 投影方式示例 (a) 等角矩形的原始图像 (b) 投影转换为 TOAST 投影的正方形图像，从图案中可以看出投影前后的位置对应关系 (c) 投影后图像在天球上的贴图

其长度也表征其所在的层级。该编号最长可以达 64 位，最大可划分至 31 层级。而 25 层级已经可以满足精度要求，它可以表示单位球面上约 0.02 角秒的宽度。

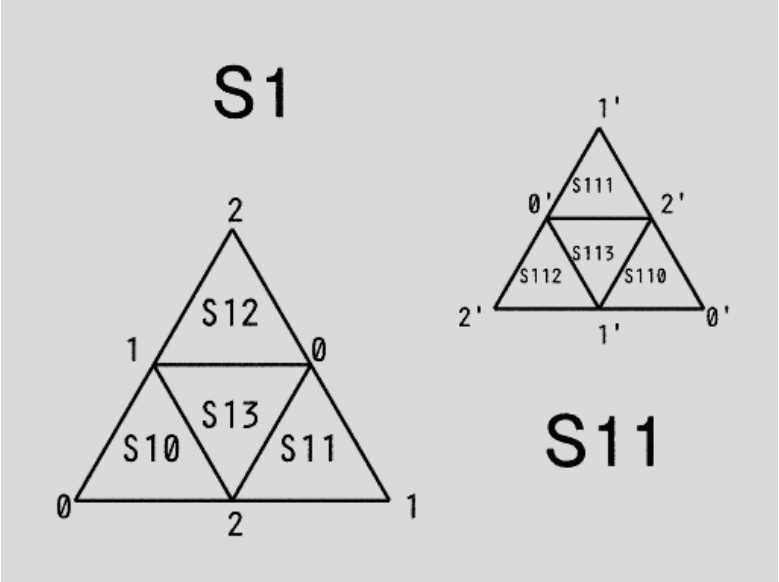


图 5.3 HTM 通过三角形网格的编号模式

5.1.2.2　层次渐进模式

层次渐进模式（Hierarchical Progressive Survey，HiPS）是一种海量天文数据管理的解决方案，它由 CDS 提出，并于 2017 年成为 IVOA 正式标准。同 HTM 一样，HiPS 以天球划分作为文件组织方式，不同的是 HiPS 采用了 HEALPix 天球划分方式作为标准，通过将天文图像数据投影至 HEALPix 天球网格后再切片存储。在第三章已经详细介绍过 HEALPix 天球划分及对各天区的编号策略，在此就不再赘述。

HiPS 将切片后图片数据存储为不同的的文件格式，并依照其 HEALPix 编号组织在不同的目录中，因此，可以将基于 HEALPix 空间索引的检索方法应用在 HiPS 文件的检索上。在 HiPS 标准中，数据存储在服务器端，客户端通过 HTTP/HTTPS 协议与服务端交互，这些客户端可以是网页，也可以是不同操作系统上的应用程序，这就使基于 HiPS 的应用有着非常好的可扩展性。HiPS 对图像数据可视化也有很好的支持。HiPS 对原始数据作数据切片时，可以基于 HEALPix 的层级划分实现多分辨率图层。在支持 HiPS 标准的客户端上，用户可以自由缩放浏览巡天数据，既能宏观地了解整个数据集的全貌，也可观察某一天体的细节，如图 5.4。

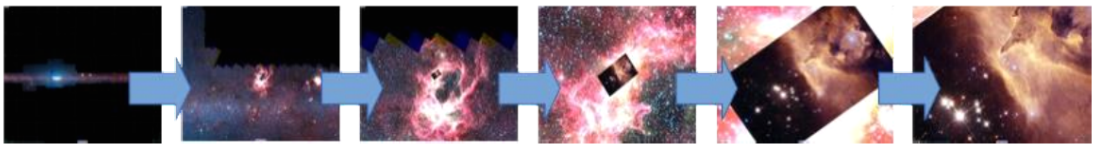


图 5.4 HiPS 的层次渐进可视化模式，通过缩放可以观测巡天数据的宏观尺度和天体细节

5.2 天文图像数据获取方式

当前主要的天文观测计划大多提供了图像获取方式，以方便科学用户能够快速的从存档数据中找到符合其研究需求的部分。不同的观测设施提供的图像数据获取方式不同，主要可分为两种：

1. 观测图像检索：根据用户检索的空间坐标，直接提供覆盖该坐标范围的存档观测图像。这种方式一般用于天体查询，如用户输入 M31，则返回覆盖 M31 坐标的对应观测图像。该类图像数据检索多见于定点观测类设备，如 Chandra、Swift、XMM-Newton 等。

2. 拼接图像获取：用户提供所需图像的中心点坐标、图像尺寸、视场范围，由观测设施后台服务根据这些条件从多副图像中拼接生成对应的图像返回给用户。这种方式常见于大型巡天观测计划，如 Pan-STARRS、SDSS。

IVOA 于 2009 年正式制定了 Simple Image Access （SIA）标准，SIA 标准即涵盖了以上两种不同的图像获取方式。用户只需提交以矩形描述的天区即可返回一系列满足空间位置需求的图像列表。返回结果以 VOTable 组织，包含了每幅图像的访问链接。图像可能是一幅完整的观测图像，也可能是由多副图像拼接得出，格式可以是 FITS、PNG、JPG 或其他图像格式。IVOA 于 2015 年又提出了 SIA 2.0(Dowler 等, 2015) 标准，加入了对多维图像获取的描述，并以 ObsCore 数据模型为基础对图像数据进行描述，该模型主要通过物理轴（空间，光谱，时间和极化）描述数据产品。

5.2.1 图像检索方法

对于天文观测图像数据检索的主要方法是构建基于 R-Tree 的空间索引。观测图像数据检索的一个重要的应用场景是获取包含目标源的图像，即查找某一空间位置是否有对应的观测图像。R-Tree 是一种用于处理多维数据的数据结构，用来访问二维或者更高维区域对象组成的空间数据。R-Tree 是平衡树，包含两类结点：叶子结点和非叶子结点。每一个结点由若干个索引项构成。对于叶子结

点，索引项形如 (Index，Obj_ID)。其中 Index 表示包围空间数据对象的最小外接矩形 MBR（Minimum Bounding Rectangle），Obj_ID 标识一个空间数据对象。对于一个非叶子结点，它的索引项形如 (Index，Child_Pointer)。Child_Pointer 指向该结点的子结点。Index 仍指一个矩形区域，该矩形区域包围了子结点上所有索引项 MBR 的最小矩形区域。

对图像数据构建 R-Tree 索引如下，首先基于图像的四角坐标构建多边形，并基于此多边形 n 生成其 MBR。将重叠的 MBR 合并，生成新的 MBR，再将独立存在的 MBR 并入与其距离最近的 MBR。将生成的 MBR 作为原有 MBR 的父节点，重复该过程，直至最终生成最后的根 MBR，此时完成 R-Tree 的构建。以如图 5.5的观测数据的 R-Tree 构建为例，首先合并重叠部分，生成的 MBR 的为 R16(R1,R2), R17(R3,R4), R18(R6,R9,R10), R19(R7,R8), R21(R12, R13, R14). 计算单独存在的 MBR 与此时的 MBR 间的距离，将其并入最近的 MRB 中，合并后的 MBR 为 R20(R11, R15), R17(R3, R4, R5); 重复上述过程，生成的 MBR 为 R22(R16, R17), R23(R18, R19), R24(R20, R21)。最终生成的 R-Tree 如图 5.6所示。在进行基于位置的图像检索时，从根节点起依次检查每个 MBR 是否包含此位置，如果包含，则检查子节点，直至抵达叶子节点。需要注意的是，为了保证检索效率，设定每个节点中包含的 MBR 不超过 3 个。则生成的 R-Tree 的层数为 $log3N$，其中 N 为图像总数。基于此方法构建的 R-Tree 检索的时间复杂度为 $O(log(N))$。

5.2.2　图像配准与拼接方法

拼接图像获取是在检索得到一系列观测图像之后，再将其处理为以目标为中心点、自定义尺寸和视场的图像返回给用户。这个过程涉及到图像的配准和拼接。具体包括如下几个步骤：

1. 对每幅图进行特征点提取
2. 对特征点进行匹配
3. 进行图像配准
4. 把图像拷贝到另一幅图像的特定位置
5. 对重叠边界进行特殊处理
6. 按需求尺寸进行裁剪

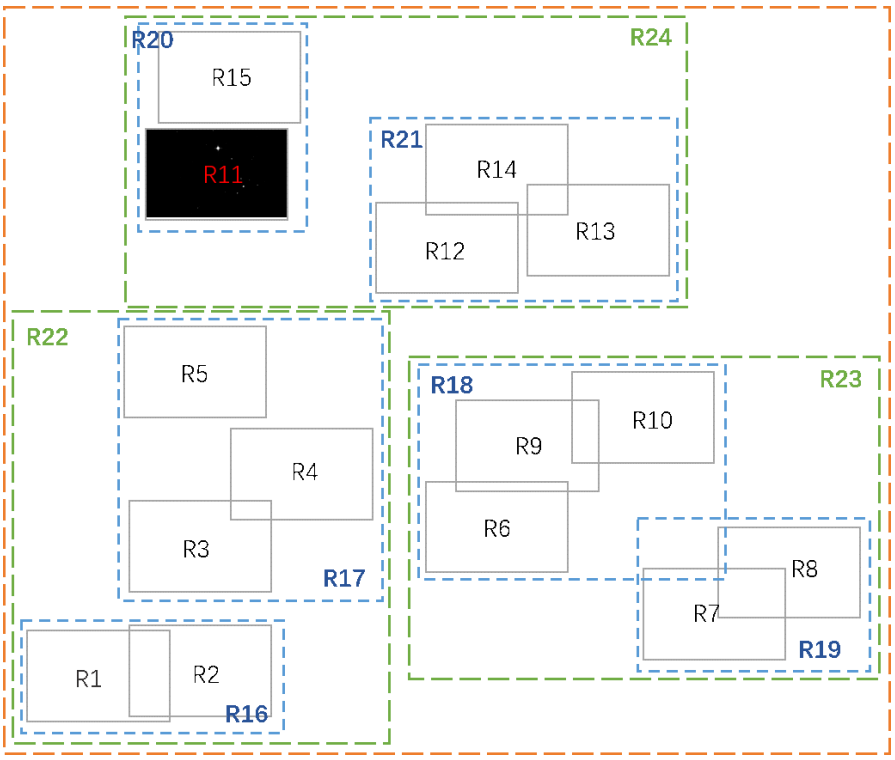


图 5.5 基于观测数据的四角坐标生成 MBR，并递归聚集生成父 MBR，R11 为包含了检索位置的图像数据，检索从根节点开始，对比命中的 MBR 依次为 R24，R20，R11

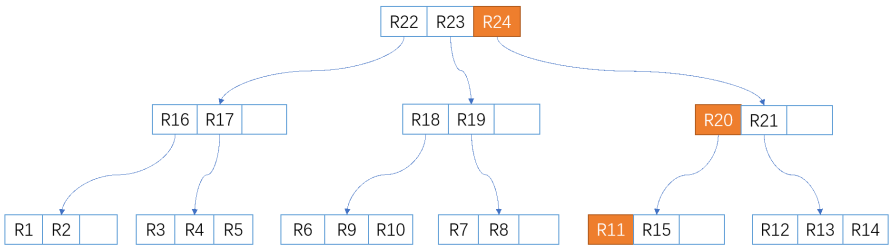


图 5.6 基于观测数据的四角坐标生成 MBR，并递归聚集生成父 MBR，构建 R-Tree

5.2.2.1　图像特征点提取

图像的特征点提取是图像配准和拼接的基础。天文领域图像数据特征点提取较为常用的算法有 SIFT 算法 (Lowe, 2004) 和其改进型 SURF 算法 (Bay 等, 2008)。SIFT (尺度不变特征转换，Scale-invariant feature transform) 是一种计算机视觉的算法。它用来探测与描述图像中的局部性特征，并在空间尺度中寻找极值点，以提取出位置、尺度、旋转不变量作为特征点。这些特征点与图像的大小和旋转无关，对于图像的对比度、噪声、些微视角改变的容忍度较高。SIFT 算法主要包含四个步骤，分别是建立尺度空间，常用方法为构建图像的高斯金字塔；在尺度空间中检测极值点，并进行定位；为特征点赋值方向，此时每个特征点包含了位置、尺度和方向三个信息；计算特征描述子, 构建特征描述向量。其流程如图 5.7

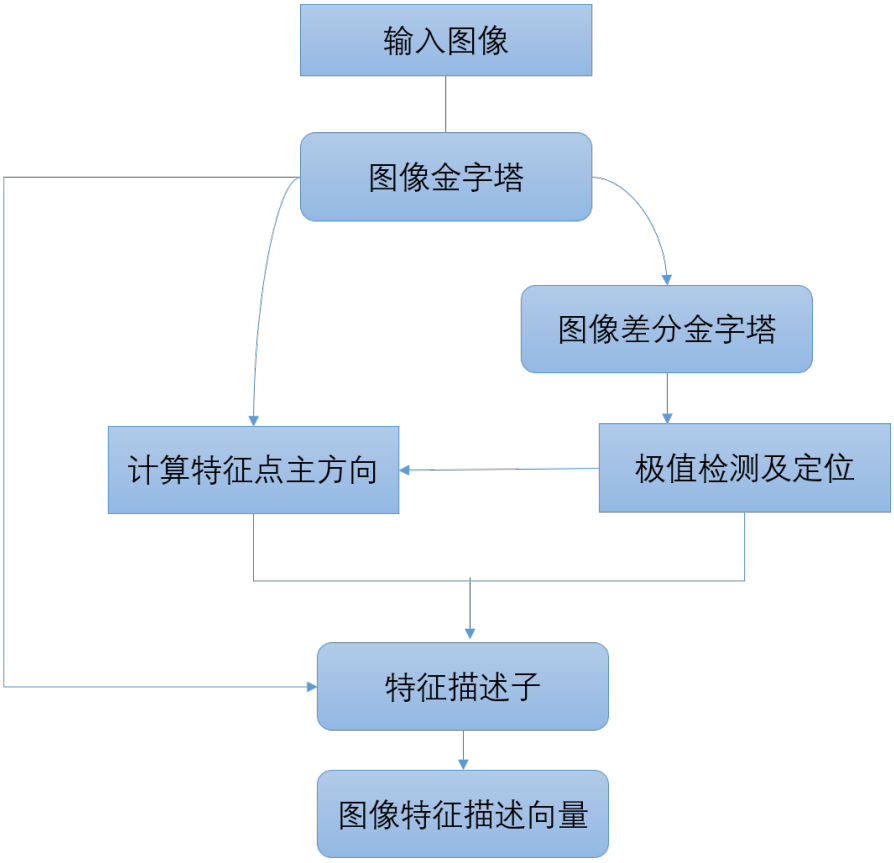


图 5.7 SIFT 提取图像特征点流程

SURF（Speeded Up Robust Features，加速稳健特征）算法是对 SIFT 算法的改进，大幅度提升了算法的执行效率，为算法在实时图像分析中的应用提供了可能。SURF 主要对 SIFT 的特征的提取和描述方式进行了改进，采用了更为高效的方式完成特征提取和描述，具体实现流程如下：

1. 构建 Hessian 矩阵（Hessian Matrix）：生成生成图像稳定的边缘点（突变点），用于特征提取。

2. 构建尺度空间：同 SIFT 一样，常用方法为构建高斯金字塔。

3. 特征点定位：将经过 Hessian 矩阵处理的每个像素点与二维图像空间和尺度空间邻域内的 26 个点进行比较，初步定位出关键点，再经过滤除能量比较弱的关键点以及错误定位的关键点，筛选出最终的稳定的特征点。

4. 特征点主方向分配：通过统计特征点圆形邻域内的 harr 小波特征，以最大值方向作为该特征点的主方向。

5. 生成特征点描述子：在特征点周围沿特征点主方向取一个 4×4 的矩形区域块。

6. 特征点匹配：通过两个特征点间的欧式距离确定匹配度，欧氏距离越短匹配度越好，此外还加入了 Hessian 矩阵迹的判断。若两个特征点的矩阵迹符号相同，则这两个特征具有相同方向上的对比度变化，反之则说明这两个特征点的对比度变化方向相反，即欧氏距离为 0。

SURF 对 SIFT 的主要改进是积分图在 Hessian 矩阵中的应用，提高了滤波运算速度，此外采用了降维的特征描述子（64 维），维度是 SIFT 特征的描述子的二分之一，降低了运算复杂度。

5.2.2.2 图像拼接与融合

得到多幅图像的匹配点集后即可进行图像的配准，这个过程主要是计算特征点集间的变换矩阵，将多张图像转换至同一坐标下。完成映射后，将多幅图像在该坐标系上进行叠加，即实现图像匹配。

完成图像匹配后，还需要对图像叠加部分进行图像融合。当对图像拼接和融合的实时性较高时，不会采用过于复杂的融合算法，一般采用的方法包括线性融合算法和中值滤波算法。

线性融合算法对图像叠加区域进行双线性插值，即渐入渐出的加权平均算法，其运算速度取决于重叠部分的面积，当面积较大时速度较慢，且信号损失较多。适合于原始观测图像尺寸较小时的图像融合。

中值滤波算法采用中值滤波器来处理图像重叠部分像素值的突变。即利用中值滤波去除高于某个阀值的点，消除像素值的突变，保持像素值的连续性。中值滤波法的优点在于能够增强细节，保持较高频的图像信息，从而突出变化目

标，且能够在 GPU 中运行，能够大幅提高计算效率。

5.3　天文图像数据可视化

数据可视化是对数据产品的二次加工，解决的是如何展示数据的问题，以方便寻找数据特征、展示研究成果等。直观的如嫦娥工程的月球测绘图像全月面拼接，NVST 的图像合成视频，巡天观测的全天图像拼接、球幕展示，多波段数据合成彩色图像等。这是一个涉及面极广泛的领域，也吸引了包括计算图像技术、艺术设计等多个专业的研究人员。

当前天文图像可视化主要关注的是大数据条件下全天图像的拼接、展示及在全天图像之上的观测数据可视化。各大天文数据中心分别实现了不同的数据可视化方法，包括基于球面的矩形、三角形或菱形逐层划分方法。较为常见的如美国 MAST 的 SkyView 系统 (McGlynn 等, 1998)，法国 CDS 的 Aladin 系统 (Bonnarel 等, 2000)，也有部分数据中心直接使用了类似 Google Map 的地图 API 进行图像数据可视化。

全天图像拼接，所要做的是将望远镜多次观测的图像合成为一张完整的大图，这也是计算机图像处理的一个经典领域。就天文领域而言，一个比较简单的办法就是先画出一个球面网格，再将计算好 WCS 的图像扭曲后贴到网格上。这其中又有一个图像 WCS 归算的问题。

现在天文用的 CCD 都比较大，一个图像就能达到数百 MB 甚至上 GB。对于一个有数万个观测图像而言的巡天观测，很难将其直接合成为一张完整的大图。并且这样一张大图的文件大小也必定相当惊人，无法方便地进行传播。SDSS 和 CDS 分别采用 HTM 和 HiPS 把这些图片分层保存，将全天网格进行金字塔式分层，最底层是原始大小的图像，每向上一层则做一次缩略图。只在最底层的一个区域范围内对重复观测进行叠加、合成。每层的每张图的大小都是一样的，但是视场越向上越大。从而形成一个自顶向下完整的全天图像。这样的方式尤其适合通过网络进行查看，每次只载入所需视场的缩略图即可，直至最底层，减少了网络数据传输，大大加快浏览速度。

分层存储大大方便了图像的浏览，但是却进一步增加了图像所占用的空间。因而部分天文学家也在考虑使用 JPEG2000 这样的存储格式解决类似的问题。一方面，JPEG2000 有极好的压缩比。另一方面，JPEG2000 的算法核心是小波分析，

本身就可以逐层对图像信息进行压缩，并分片存储。对视场较大的情况只需要解压顶层几个区域的数据即可。这样，JPEG2000 自身就可实现上述分层存储的功能，同时还大大减少了空间占用。基于以上优点，JPEG2000 在医学领域已经得到了一定范围的应用。但是，JPEG2000 目前仍有许多专利限制，相关软件的收费也相当惊人，图像压缩速度也需要进一步提高。

除了图片分层存储、显示，Aladin、SkyView、万维望远镜（Worldwide Telescope，WWT）(Rosenfield 等, 2018) 等系统，还提供了图层叠加的数据可视化方法。比如在最底层显示光学波段的全天图像，其上显示一层多边形表示某次观测的范围，在其上又可以在用一层用来显示观测范围内的目标坐标并在旁边显示文字、编号等信息。WWT 甚至还可以通过动态改变图片透明度的方式来同时显示两套图像，给图像直观对比带来了极大的方便。

虽然 Aladin、SkyView、WWT 等都可以分层显示多边形、文字等信息，但是都需要通过编程实现，没有一个统一的图形、信息描述手段可以在多个平台上无缝展示。WWT 本身定义了一个 WTML 格式并部分兼容 Google Map 的 KML 格式，但是这两者都不能直接在 Aladin 及 SkyView 上显示。这就需要针对不同的平台进行开发，增加了开发者的工作量。

5.4 基于层次渐进模式的海量图像组织、检索与可视化框架

虽然现阶段已经存在一系列图像数据的检索和获取方式，但异构的图像数据组织方法与图像获取方式并未给科学用户带来良好体验，尤其是在获取不同波段不同来源的图像数据进行交叉分析时，数据的检索和处理均较为繁杂。比如对比分析某一暂现源在光学波段和 X 射线波段的图像，要首先在 SDSS 数据库中检索找到该区域数据的存放位置，然后下载该数据，通过 DS9 加载浏览分析。再去 XMM-Newton 的图像 FTP 中的一系列观测文件中找到对应的图像，下载数据，进行裁剪，配准，然后叠加对比分析，整个操作流程非常繁琐。这种异构的数据组织、获取及可视化方式无法满足多波段、多信使联合观测的高效数据融合需求。

为此本文基于层次渐进模式（Hierarchical Progressive Survey，HiPS）标准的图像组织方式，提出一种针对多波段海量图像数据统一组织管理、高效检索，能够对全天尺度图像进行沉浸式可视化的框架标准，并在万维望远镜软件平台中

实现了该框架。其中的关键技术包括如何海量图像数据高效转换至 HiPS 标准数据集、层次渐进模式下的多波段图像检索和获取方法、以及多波段图像的三维可视化。

5.4.1　多波段图像数据转换至 HiPS 标准数据集

实现对多波段图像的统一组织管理，首先要对多波段图像进行标准化。本文采用 HiPS 标准实现多波段图像的标准化，这就要求将不同来源的图像数据转换至 HiPS 标准数据集。HiPS 标准数据集的每块数据都与 HEALPix 天球划分的网格一一对应，原始图像数据需要通过数据投影转换至 HEALPix 投影后，再进行图像匹配和拼接，最后还要根据不同层级下的 HEALPix 网格进行切分，按照层级组织归档存储。HiPS 标准的提出者 CDS 发布的 HiPSGen 可以完成以上这些操作。

但是 HiPSGen 只能对小批量数据集进行转换，虽然它采用了多线程来加快处理速度，但是无法实现在计算集群上的并行化，因此完全无法满足海量多波段数据的处理需求。为此本文提出了一种并行写入方法，能够将图像投影转化、拼接与切分的步骤合并，同时可以运行在多节点计算集群上。我们使用中国探月工程公开发布的嫦娥 2 号全月表面 7m 分辨率的正射影像数据分别对两种方法进行了测试，结果表明本文提出的方法大幅提高了转换效率。

5.4.1.1　基于图像匹配和拼接的转换方法

HiPSGen 主要采用了基于图像匹配和拼接的方法将图像文件转化为 HiPS 标准文件。其主要步骤为投影转换、图像拼接和图像裁剪。HiPSGen 只支持 FITS 文件作为输入。在读取文件时，会首先读取 FITS 头中 WCS 关键字以确认原始图像的投影。需要注意的是，WCS 关键字包括 25 种投影，但 HiPSGen 仅支持”SIN”, ”TAN”, ”ARC”, ”AIT”, ”ZEA”, ”STG”, ”CAR”, ”NCP”, ”ZPN”, ”SOL”, ”MOL”,”TAN-SIP”,”FIE” , ”TPV”, ”SIN-SIP”，对于一些数据常用的 Mercator 投影则需要做中间投影转换，如转换至 STG，即 Polar Azimuth 投影。该转换可采用开源地理数据处理库 GDAL 实现。

之后进行图像所在 HEALPix 网格的计算。首先需要根据图像分辨率确定生成 HiPS 标准数据的最大层级。以最大层级 10 为例，将图像文件按照 1024×1024 的尺寸切分成为若干块，然后根据这些子图像块的四角坐标计算出其覆盖

在 HEALPix 层级 10 的编号。然后再依次计算出这些子图像块的中心点 RA、Dec 坐标以及四个角的 RA、Dec 坐标，将这些信息写入到中间文件中，用于下一步的图像切分。

根据计算出的 HEALPix 网格编号生成临时文件夹，文件夹以网格编号命名。对原始图像按照索引中间文件中的四角位置进行切分，并把切分后的子图像存入对应的编号名的文件夹中。

对同一编号文件夹中的子图像进行图像匹配和拼接。图像匹配时的特征点提取采用了前文提到的 SURF 算法，具有较快的速率。对图像拼接后再对叠加部分进行像素融合，得到该编号对应的 HiPS 标准数据。

最后还需要对图像做归并得到 10 层级以上各级的图像。按照如图 5.8的 HEALPix 编号升降级方式计算出各图像父网格的编号，将父编号相同的图像根据其自身编号确定相对位置进行合并。如此递归直至层级 0，完成原始数据至整个 HiPS 标准数据集的转换。

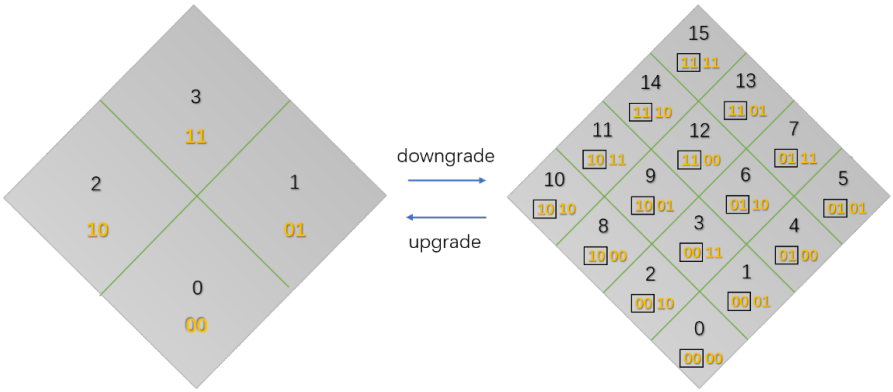


图 5.8 HEAPix 在 NESTED 模式下，层级 1 至层级 2 的索引变换，黑色数字表示该网格索引号，其下方金色数字表示该索引号二进制表示。降级时只需在原索引号的二进制表示后添加对应子网格的相对父网格的二进制编号即可。

上述整个流程串行执行，且步骤较多，在图像裁剪和匹配拼接及融合过程时间花费较大。基于该方法使用 BASS 巡天图像数据（2.2TB）制作 HiPS 标准数据集共花费约 720 小时。

5.4.1.2 基于像素并行写入的转换方法

对于大型巡天项目产生的海量观测数据，图像匹配和拼接方法由于时间耗费较大不能胜任。为此我们基于 HEALPix 网格的层级递归原理，提出了一种直接写入像素的 HiPS 标准数据集转化方法，并实现了并行处理。

首先根据原始图像的分辨率确定生成 HiPS 标准数据的最大层级，并计算所有图像在最大层级上的 HEALPix 网格。HiPS 标准数据集的常用尺寸为 512×512，生成每个 HEALPix 网格的对应空白图像，再向其中写入像素。因此其核心操作是计算出空白图像每个像素在对应原始图像中的位置，取出该位置的像素作为切分后的图像的对应位置像素。通过 HEALPix 网格编号的层级递归特性可以实现该操作。

以层级 10 的某个网格的空白图像为例，若其尺寸 S 是 512×512，根据 HEALPix 的层级特性，其各个像素点对应相同天区层级 19 中的各个网格。即原始层级 K 中的尺寸为 $S \times S$ 的图像中每个像素对应着更高层级 k_{high} 中对应的网格，其关系如公式 5.1。

$$k_{high} = k + log_2 S \quad \ldots (5.1)$$

根据网格的 HEALPix 编号 hpx 和层级 k 得出该网格的中心点坐标 RA 和 Dec，计算公式如 5.2。再根据原始文件 FITS 头中的的 WCS 信息计算出这个坐标在原始图像上的 XY 坐标，取其周围四个点的值进行线性插值，将该插值作为像素值填入空白图片的对应像素位置。如果该值为空，则设为空。通过这种方式，在进行图像切分的同时，完成了将原始图像的投影转换，同时省去了图像匹配拼接和像素融合的步骤。

$$\begin{aligned} p &= \frac{hpx + 1}{2} \\ i &= I\left(\sqrt{p - \sqrt{I(p)}} + 1\right) \\ j &= hpx + 1 - 2i(i-1) \\ z &= 1 - \frac{i^2}{3 \times 4^k} \\ \phi &= \frac{\pi}{2i}\left(j - \frac{1}{2}\right) \\ RA &= \frac{180 \times \phi}{\pi} \\ Dec &= 90 - \frac{180 \times arccosz}{\pi} \end{aligned} \quad (5.2)$$

5.4.2 基于 Spark 的海量多波段图像 HiPS 标准数据集处理集群

由于 HEALPix 划分的各个网格间空间位置相互独立，以上过程可以通过分治方式进行加速。我们在并行计算环境下采用嫦娥 2 号 7m 分辨率遥感影像数据对上述方法进行了实验，取得了很好的效果。

公开发布的嫦娥 2 号 7m 分辨率正射影像分为极地区域和赤道区域两个部分。极地区域的数据覆盖北纬 70 度至北极点，南纬 70 度至南极点的数据，数据投影为 Polar Azimuth 投影，赤道区域数据为北纬 70 度到南纬 70 度之间的数据，采用 Mercator 投影。总数据量为 747.9GB，由于体量较大，数据被分幅为 844 个分块，每块数据在 1GB 左右。数据采用 Tiff 格式存储，每副数据均另附了 prj 文件和 tfw 文件，用于记录该副数据的投影和地理坐标。对这些数据文件进行合并和格式转换，将 tiff、prj 和 tfw 文件合并为单一的 FITS 文件，作为输入源文件。

我们采用 Spark 搭建了并行计算环境。Spark 是一个基于内存计算的开源的集群计算系统，属于大数据计算框架 Hadoop 生态系统中的一部分。它基于 Map-Reduce 的分治思想，但将中间输出结果保存在内存中，从而在迭代计算过程中不需要读写分布式数据库。简化版的 Spark 基础架构如图 5.9所示，其中最底层是 HDFS 分布式文件系统，用于管理存储资源。之上是资源管理与调度系统 YARN，用于管理计算资源，它将集群内的所有计算资源抽象成一个资源池，包含 CPU 和内存两个维度。Map-Reduce 是 Spark 的分治算法实现，即首先将数据映射（Map）至不同的任务上进行计算，再规约（Reduce）合并计算结果。RDD（Resilient Distributed Datasets，弹性分布式数据集）是 Spark 的核心数据结构，可以让用户显式地将数据存储到磁盘和内存中，并能控制数据的分区。

图 5.10中分别是基于 Spark 构建的海量多波段图像 HiPS 标准数据集处理集群的结构，以及其中运算的主要流程。原始图像数据存储在 HDFS 中，它是一个分布式文件系统，它为以流式存取大文件而设计，适用于几百 MB、GB 以及 TB 级文件尺寸，并且一次读多次写的场合。数据进入 Spark 计算节点后以 RDD 模式组织，计算在内存中流式进行，被任务管理器分别分配到不同节点进行处理。其分配的方式为基于 HEALPix 天区划分。在整个计算过程中需要进行两次划分，具体步骤如下：

1. 生成索引文件之前进行第一次划分，按照 HEALPix 第 3 层级将原始图像分为 768 组，分别生成索引中间文件并根据中间文件生成空图像。为了节省计算

内存，这些空图像只有一个地址名，为其在最大层级的 HEALPix 网格编号。

2. 对 768 组原始图像按照其分幅进行第二次划分，并将第一次划分的各组空图像地址名拷贝至每个二次分组里。

3. 执行像素定位计算，并写入像素值至对应的空图像地址，此时生成 512×512 的 HiPS 标准文件，迭代直至所有像素完成计算，此时得到最大层级的 HiPS 标准数据。

4. 迭代组合归并 HiPS 标准数据，得到从 0 层至最大层级的所有数据，形成 HiPS 标准数据集。

表 5.1 基于 Spark 的 HiPS 标准数据集处理集群配置

类型	详情
节点数	9 个
节点 CPU	$1\times 3.0GHz$
节点逻辑核数	4 核
节点内存	8GB
Spark 版本	2.2.0

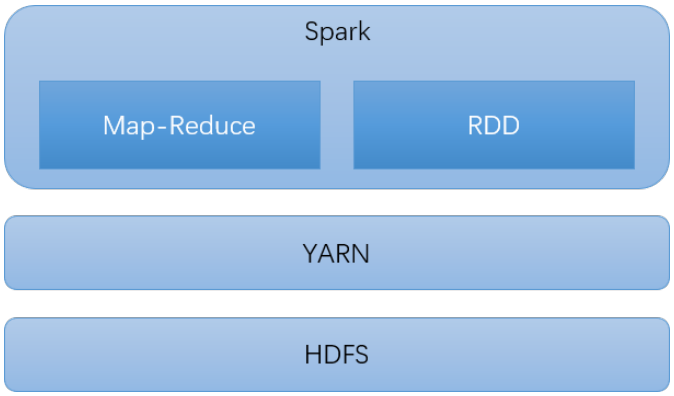


图 5.9 基于 Spark 的并行计算环境架构

这在一些定点观测项目中，除了图像数据外，还有很多附加数据是通过具体观测进行关联的，因此需要保留 HiPS 标准数据集与原始图像间的关联。在 HiPS 标准数据集生成的过程中，原始数据和标准数据集之间的对应关系存入 PostgreSQL 数据库，用于在数据检索时能够同步获取原始图像。生成的 HiPS 标准数据集存储在 HDFS 文件系统中。

在实际测试中，采用该集群对嫦娥 2 号数据进行处理共花费 4 小时 32 分钟，

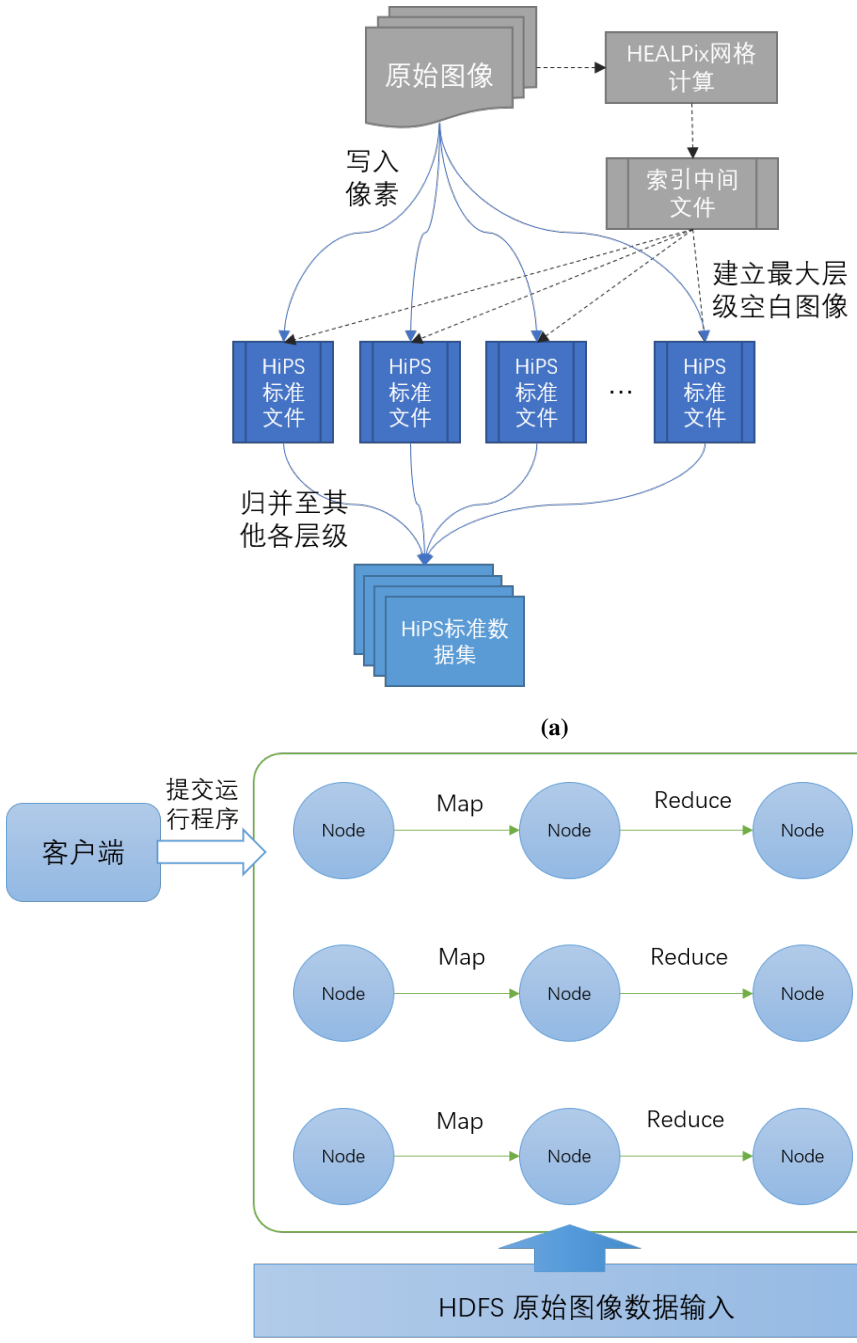


图 5.10 Spark 上采用像素并行写入方式将原始图像转换至 HiPS 标准数据集 (a) 基于像素并行写入的转换方法，其中虚线部分的指向为计算原始图片对应写的 HEALPix 网格，在 Spark 计算中属于 Map 阶段，实线指向部分为向各网格划分得到的空白图像中写入像素，为 Reduce 阶段 (b) Spark 处理框架的结构。原始图像数据存储在 HDFS 中作为输入，从客户端发起执行任务，由各个 node 并行执行，结果存入 HDFS 和 PostgreSQL 数据库中。

而同样数据在八核 3.0GHz、64GB 内存的计算服务器上采用 HiPSGen 需要 152 小时才能完成转换。这种差异主要来自两方面的因素：一方面是前文所提到过的算法计算效率。基于像素并行写入方式在进行 HiPS 标准数据集图像切分的同时，完成了原始图像投影转换，且省去了图像匹配拼接和像素融合的步骤，其计算量大幅减小。另一方面来自计算时的 CPU 和内存利用率。虽然 Spark 并行处理集群的内存总量和单节点测试服务器的内存总量相当，但单节点服务器上的计算任务只能按照 CPU 总线程数（8 线程）进行并发，每个并发所独占的内存为 8GB，但实际计算只需要 1GB 运行内存（与图片本身大小相当）。Spark 集群具有更多的 CPU 核心数，能够并发更多任务，每条任务均基于内存计算进行，不仅具有 IO 时间上的优势，且内存利用率较高。因此大幅降低了 HiPS 标准数据集转换的时间花费。该方法在图像平滑性上仍需要进一步改进。当不同时期的观测图像亮度没有统一的情况下，基于并行像素写入方法生成的 HiPS 图形会出现亮度不均的情况。主要是因为在并行计算时缺乏对图形整体亮度的归一化，导致不同观测时期的亮度差异较大。但如果原始观测图像在预处理时经过亮度平滑或归一化，则不会出现该问题。

5.4.3　层次渐进模式下的多波段图像检索和获取方法

完成 HiPS 标准数据集之后，将其部署在发布服务器上，并提供给用户图像检索和拼接图像数据获取功能。

HiPS 标准数据集是以文件系统的方式组织的，将其所有图像数据集存储为目录集合。这些目录的结构遵循层级-网格（图像）的层次结构，层级目录名的前缀为 Norder，图像文件名的前缀是 Npix。为了避免目录过大，在 Norder 下使用子目录前缀 Dir 进行了细分，每个子目录最多包含 10000 副图像。

我们通过 Nginx 反向代理的方式将其部署在分布式服务器上。当用户进行图形检索时，会得到符合检索条件数据的访问地址，通过 HTTP 访问该地址得到对应的数据。HiPS 的 URL 形如 $http://\{domain\}/\{data\_name\}/Norder\{0\}/Dir\{1\}/Npix\{2\}$. 其中，$\{domain\}$ 为提供 HiPS 服务的域名，$\{data\_name\}$ 为具体的 HiPS 数据名，$Norder\{0\}$ 表示该网格的数据所在的层，$Dir\{1\}$ 为数据所在的文件夹，$Npix\{2\}$ 为网格在该层级上的索引，所有 {} 中都是变量，具体与指向的 HiPS 标准数据相关。

前文介绍了 R 树作为空间索引方式为图像数据实现检索，但是 HiPS 标准数

据集已经完成了图像分幅和裁剪处理，每幅图像均不重叠，且图像紧邻排列，因此使用 R 树索引会造成树高过深或每层节点过多，大幅降低检索效率。

由于 HiPS 标准数据集是基于 HEALPix 天球划分进行文件组织的，各个图像的文件名对应 HEALPix 网格编号，因此可以采用 HEALPix 空间索引实现图像文件的查找。

当用户进行点源查找时，将该点源坐标转化为该数据集的 HEALPix 最大层级下的网格编号，通过访问对应该编号的 HiPS 标准数据地址即可获取该数据。

当用户需要获取自定义区域和视场的图像时，则首先将该区域转化为对应视场的 HEALPix 网格集合，再分别请求这些网格对应的图像进行拼接和裁剪，由于 HiPS 标准数据集不再需要图形配准和融合，该过程能够实时完成。

5.4.4 层次渐进模式下多波段图像数据可视化的实现

层次渐进模式支持图像数据的沉浸式可视化，随着视场的缩放，可以展示全天球上不同分辨率的图像，从而既能够在宏观尺度上查看数据，也可获得某一特定区域的细节。此外，基于 HiPS 标准数据集的空间位置一致性，还可以叠加不同波段的图像做对比分析。我们在万维望远镜软件中实现了层次渐进模式下的数据可视化，其关键技术是 HEALPix 网格绘制和层次细节模型的构建，以及将 HiPS 标准数据作为纹理在三维场景中进行渲染。

5.4.4.1 HEALPix 网格绘制

万维望远镜是一款三维数据可视化软件，它将天文图像数据作为纹理贴在三维天球上，并通过三维图形处理管线渲染到屏幕上。在三维场景中，所有的物体都是由三角形组成的，原因是三角形可以拟合成几乎任何形状。因此，要创建 HEALPix 网格的三维天球，需要首先定义形成天球的三角形，即确定这些三角形在三维空间中的顶点坐标。

首先依次计算出 HEALPix 在层级 0 时的 12 个基础网格的顶点坐标。这里要定义两个参数：N_{side} 和 $faceindex$。N_{side} 与层级相关，是天球总网格数除以基础网格数 12 后的平方根。$faceindex$ 是 12 个基础网格的编号，范围为 0 至 11。HEALPix 基础网格的排列方式如图 5.11所示，总计三行四列，根据其行列上的次序给以两位编码用于后继计算，例如第二行和第三列的 $faceindex$ 编码为 12，十进制为 9。每个基础网格的所有子网格均继承其父网格的 $faceindex$。

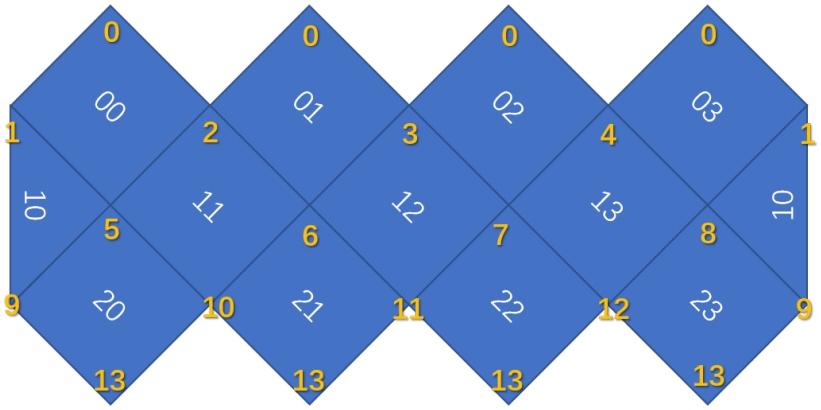


图 5.11 HEALPix 在层级 0 时基础网格排列方式，并给予两位编码用于后继的计算。

当层次递进到到下一个层级时，当前层级的每个网格都分为四个子网格，编号如公式 5.3：

$$tile_{index} = parent_{tile_{index}} \times 4 + 2y + x \quad \ldots (5.3)$$

在实际计算中，每个基础网格均被分为四个四边形，这是为了让三角形面片尽量多以能够在较低层级时就能很好的拟合球形。如图 5.12所示，以计算层级 0 一个基础网格的顶点坐标为例，它分成四个单位边长的四边形，每个四边形的右下角坐标的位置为 $\left(\frac{1+x}{2N_{side}}, \frac{1+y}{2N_{side}}\right)$，其中 x，y 分别是四边形的行索引号和列索引号。之后在此基础上将四角坐标增加或减少 $dc = \frac{1}{2\times N_{side}}$，得到基础网格上各个点在平面坐标下的的坐标。由于 HEALPix 网格的层级递归特性，其他各层级网格的顶点均采用同样的方法进行计算。

为了更精细的模拟天球，需要对层级 3 之前的各层的网格的顶点进行差值，即增加顶点使之能够更好的拟合一个球形，为此，我们添加了一个 $step$ 变量用于在网格的各条边上增加顶点。当层级小于等于 3 时，$step$ 设为 4，层级大于等于 4 时，$step$ 设为 1。当 $step$ 不是 1 时，即四边形具有额外的点，则需要增加或减少另一个偏移量 $\frac{1}{step\times N_{side}}$。据此可以总结出 HEALPix 各层级网格顶点坐标的计算公式如 5.4，其中 $i \in [0, 4step - 1]$。

随后需要将这些平面坐标转换为天球坐标。天球上各网格中心的坐标由 $(z \equiv \cos\theta, \phi)$ 定义，其中 $\theta \in [0, \pi]$ 是从北极测量的余纬度，$\phi \in [0, 2\pi]$ 是向东测量的经度。通过使用这些相对坐标，再结合它们所位于的基础网格，可以获得天球球坐标系中每个顶点的 z 和 ϕ。

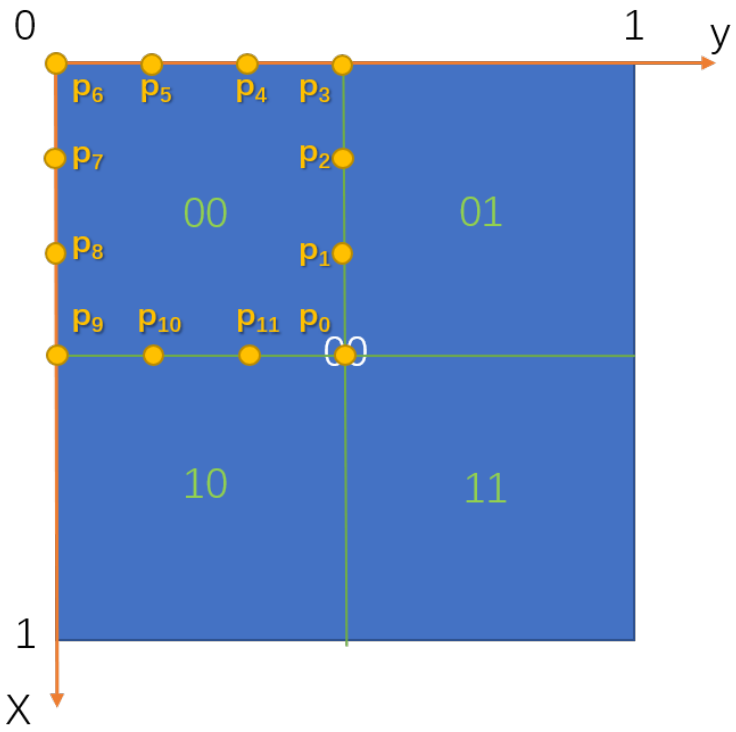


图 5.12 编号 0 的基础网格的四个子四边形，图中 $p_0 - p_{11}$ 是其中一个子网格包含了插值点的顶点，*step* 设为 3

$$
\begin{aligned}
\left(P_{i_x}, P_{i_y}\right) &= \left(\frac{1+x}{N_{side}} - \frac{i}{step \times N_{side}}, \frac{1+y}{N_{side}}\right) \\
\left(P_{(i+step)_x}, P_{(i+step)_y}\right) &= \left(\frac{x}{N_{side}}, \frac{1+y}{N_{side}} - \frac{i}{step \times N_{side}}\right) \\
\left(P_{(i+2step)_x}, P_{(i+2step)_y}\right) &= \left(\frac{x}{N_{side}} + \frac{i}{step \times N_{side}}, \frac{1+y}{N_{side}}\right) \\
\left(P_{(i+3step)_x}, P_{(i+3step)_y}\right) &= \left(\frac{1+x}{N_{side}}, \frac{y}{N_{side}} + \frac{i}{step \times N_{side}}\right)
\end{aligned}
\tag{5.4}
$$

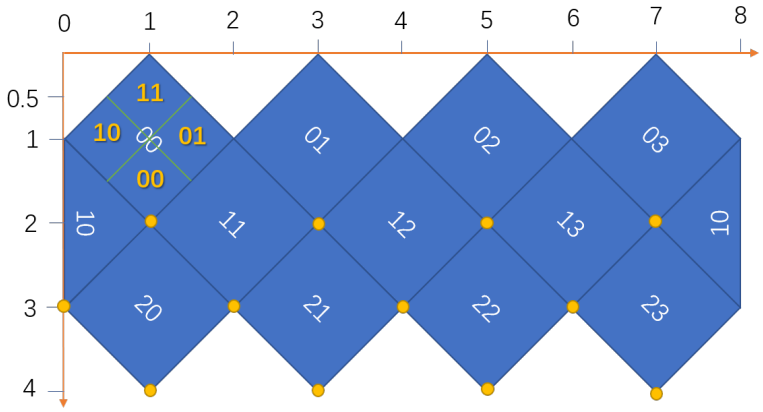


图 5.13 在平面坐标系下基础网格顶点的坐标，图中黄色圆点是 HEALPix 层级 0 的基础网格最下方的顶点。

如图 5.13所示，在平面坐标系下，将 HEALPix 层级 0 的基础网格的下顶点的坐标设为(2,1),(2,3),(2,5),(2,7),(3,0),(3,2),(3,4),(3,6),(4,1),(4,3),(4,5),(4,7)。当四边形位于北半球 ($face \in [0,1,2,3]$) 的 4 个面中时，坐标的计算公式为 5.5：

$$
\begin{aligned}
z &= 1 - \frac{(2 - p_x - p_y)^2}{3} \\
\phi &= \frac{\pi}{2}\left(2face + 1 + \frac{p_x - p_y}{2 - p_x - p_y}\right)
\end{aligned}
\tag{5.5}
$$

当四边形位于南半球 ($face \in [8,9,10,11]$) 四个面时，其坐标计算公式为 5.6：

$$
\begin{aligned}
z &= \frac{(p_x + p_y)^2}{3} - 1 \\
\phi &= \frac{\pi}{2}\left(2face - 15 + \frac{p_x - p_y}{p_x + p_y}\right)
\end{aligned}
\tag{5.6}
$$

当四边形位于赤道上四个面，即 ($face \in [4,5,6,7]$)，坐标计算公式为 5.7：

$$\begin{aligned} z &= \frac{2\left(p_x + p_y - 1\right)}{3} \\ \phi &= \frac{\pi}{2}\left(2face - 8 + p_x - p_y\right) \end{aligned} \tag{5.7}$$

最后使用公式 5.8将 z, ϕ 转换为笛卡尔坐标系的坐标，即可获得三维空间中每个顶点的坐标，其中 $z = \cos\theta$。

$$\begin{aligned} x &= \sqrt{(1-z)(1+z)}\cos\phi \\ y &= \sqrt{(1-z)(1+z)}\sin\phi \\ z &= z \end{aligned} \tag{5.8}$$

通过以上计算即得到 HEALPix 基础网格在三维空间中的顶点，它们构成的三角形网格在三维场景中构成天球。

5.4.4.2 层次细节模型

上述计算得到的是 HEALPix 基础网格的顶点，还需要迭代创建逐次渲染各子层级的网格。这个过程就是层次细节（level of details，LOD）实现。如图 5.14，每个网格都在下一层及细分为 4 个网格，但是并非一直细分下去，递归的终结在于判断当前网格的下一层次是否在视场内，同时该网格在三维场景中的显示尺寸是否足够大。同时满足这两个条件才会绘制该网格。

在三维场景中，天球上的某个区域是否显示在屏幕上取决于该区域是否位于视场中。在实现上，视场是由 6 个面组成的立方体。为了模拟真实世界的视野，通常使用透视投影在屏幕上进行三维场景投影，因此将该立方体转化为一个平头锥体，如图 5.15所示，当网格位于视锥体中时，则对其进行渲染，否则将其剔除。

视锥体的视场由以下 4 个参数确定：$(left, bottom, -near)$ 和 $(right, top, -near)$ 指定近剪切面左下角和右上角坐标。$near$ 和 far 指定视点到近、远剪切面的距离。

位于视锥体内部的点 (x, y, z) 均被投影到 $z = -n$ 的近剪切面上，对应的坐标为 $\left(x', y', z'\right)$，其中 $l \le x' \le r$ 和 $b \le y' \le t$。

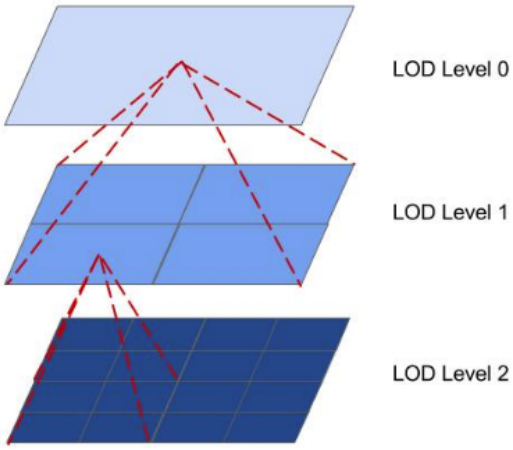


图 5.14 层级 0 中的 HEALPix 网格的一个网格中层次细节模型，从上自下分别展示了层级 0-3。

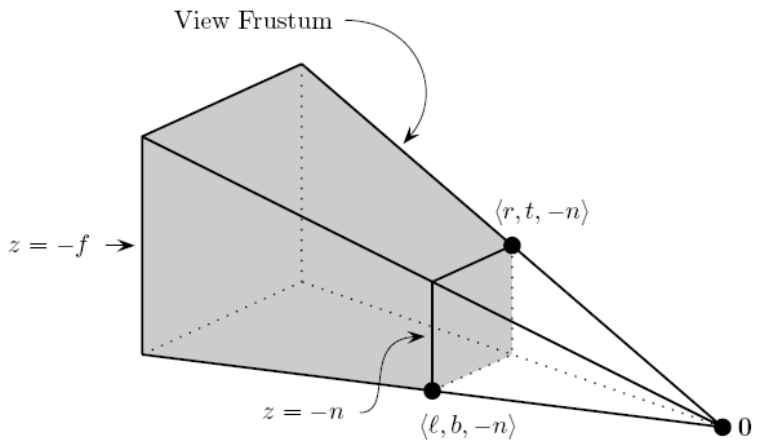


图 5.15 视锥体的结构。其中 O 是视点，$(r, t, -n)$ 是右上角的坐标，$(l, b, -n)$ 是左下角的坐标，它们都位于近剪切面。

以各网格中心点为原型，网格对角线为直径构建边界球，通过该边界球确定网格与视锥体的位置关系。计算边界球球心到视锥体每个面的距离，如果小于球体的半径，则该网格将与该面相交。如果大于球体半径且为正，则网格位于该面的正面。如果为负，则网格位于面的背面。距离计算公式如 5.9，其中 C 是球体的中心坐标，N 是视锥体各个面的法线向量，D 是从坐标系原点到视锥体各个面的距离。

$$distance = (C \cdot N) + D \quad \ldots (5.9)$$

如果网格位于视锥体内部，则还需要确定其尺寸是否足够大以适合显示在屏幕上。从网格的四角坐标可以变换得到其屏幕坐标，进而得到其在屏幕上四个边的长度。当最长边小于预定值 L 时，则该网格被认为太小而不应显示在屏幕上。L 是一个经验值，当它较小时，屏幕上将显示更多的网格，从而显示更加精细，也会占用更多内存。反之则显示精度降低，占用内存更少。

5.4.4.3 HiPS 标准数据集载入至三维场景

完成 HEALPix 网格的构建和层次细节模型的渲染后，需要将 HiPS 标准数据集的图像作为纹理贴图到对应的网格中。如前所述，在 HiPS 数据组织中，图像数据是根据 HEALPix 的网格和层级结构进行分组的。例如，将全天观测的多波段图像按第层级 3 切分成 768 个图像，每个 HEALPix 网格对应一个图像。在 HEALPix 的 RING 方案和 NESTED 方案中，HiPS 标准采用 NESTED 作为其索引方案 (ano)。

定义索引 $p_n \in \left[0, 12N_{side}^2 - 1\right]$，并定义 $p_n' = \left(p_n \mod N_{side}^2\right)$，其中 p_n' 表示每个基础网格内的子网格的索引号，其二进制表示形式是 $...b_3b_2b_1b_0$。给定网格分辨率参数 N_{side}，每个基础网格上的子网格位置由两个索引 x 和 $y \in \left\{0, N_{side} - 1\right\}$，其中 x 索引沿东北方向运行，而 y 索引沿西北方向运行。从 p_n' 的二进制表示可以得到 x 和 y 的值，分别是 p_n' 的偶数位和奇数位组合，即 $x = ...b_2b_0$，$y = ...b_3b_1$ (Gorski 等, 2005)。当层次结构被细分时，可以通过拼接父层级和新层级的 XY 索引来获得新的 p_n'。

在构建 HEALPix 网格的层次细节模型过程中，已根据用户的视野计算出要显示的 HEALPix 层级和网格。请求与编号相对应的图像并将其添加到下载队列

中。从 HiPS 服务器下载图像后，将它们作为纹理渲染到相应的图块上，即完成了 HiPS 数据集的加载与显示。

具体到万维望远镜中，可以通过添加配置文件来导入 HiPS 标准数据集的信息。万维望远镜的配置文件是一个 XML 文件，其格式如下：

```
<ImageSet Generic="False" DataSetType="HiPS_SkyMap"
 BandPass="Visible" Name="DSS2 Blue (XJ+S)"
 Url="http://alasky.u-strasbg.fr/DSS/DSS2-blue-XJ-S/Norder{0}/Dir{1}/Npix{2}"
 BaseTileLevel="0" TileLevels="9"
 FileType=".jpg"
 Projection="Healpix"
>
</ImageSet>
```

主要信息包含在 ImageSet 元素中，其中 DataSetType 属性表明它是一个 Skymap 类型的 HiPS 数据。Url 属性是 HiPS 数据的下载地址，其格式如上一节所述。BaseTileLevel 属性表示数据加载是从第几层开始，由于 WWT 的递归的数据加载方式，所有的数据都是从第 0 层开始加载，即便大部分 HiPS 并不提供 0-2 层的数据。TileLevels 属性表示该 HiPS 的总层数。FileType 属性表示 HiPS 数据的文件格式。Projection 属性表示该数据为 HEALPix 投影。

我们将上一节介绍的嫦娥 2 号全月面 7m 分辨率正射影像 HiPS 标准数据集载入至万维望远镜，实现了良好的可视化效果，如图 5.16。

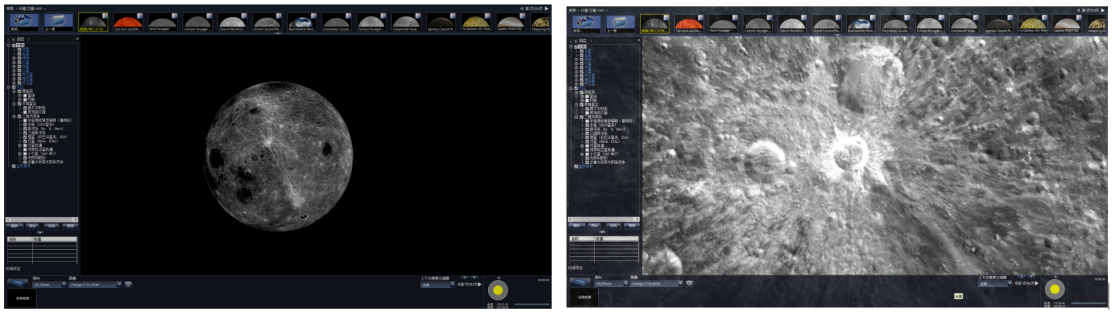


图 5.16 在万维望远镜中可视化嫦娥 2 号全月面 7m 分辨率正射影像 HiPS 标准数据集

值得一提的是，引力波事件定位天区数据的组织方法是以 HEALPix 网格为基础的，它将全天划分为 HEALPix 层级 8 至 10 的网格，之后在每个网格中填充引力波事件在当前网格中发生的概率。因此，我们可以采用 HiPS 标准数据集在万维望远镜中的可视化方法来对引力波定位数据进行展示，结合多波段图像数

据，可以直观的显示出其他波段的历史观测数据在引力波事件定位天区中的情况。在万维望远镜中对首个发现电磁对应体的引力波事件 GW170817 的定位天区的可视化如图 5.17。

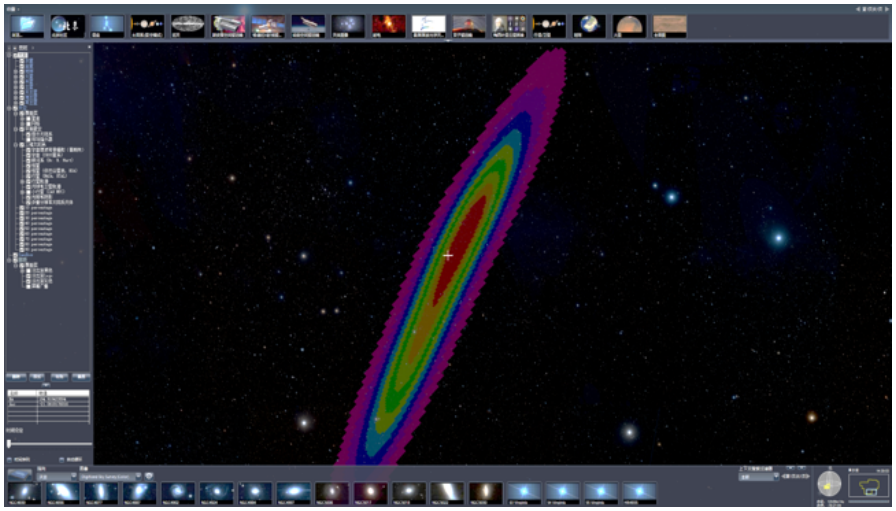


图 5.17 引力波定位区域数据在万维望远镜中的可视化

5.4.4.4 渲染效率对比

万维望远镜软件平台的原生数据集组织方式是 TOAST（Tessellated Octahedral Adaptive Subdivision Transform，八面体自适应细分变换）(Rosenfield)，这是一种基于 HTM 天球划分方式的图像组织方式。基于它与 HiPS 标准数据集的相似性，我们对比了两者在万维望远镜中的渲染效率。

参与对比测试的数据集共包括 12 组，其中 6 组为 HiPS 标准数据集，另外 6 组为 TOAST 数据集。结果如表5.2所示，其中 $MaxOrder$ 表示数据的最高层级。需要说明的是，由于图像切分方法的不同，HiPS 标准数据集和 TOAST 数据集的最大层级并不一致，大多数 TOAST 数据集的分辨率（256×256）低于 HiPS 标准数据集的分辨率（512×512）。表中的 Ti 和 Tm 分别表示层级为 0 时和层级最大时图像填充整个屏幕所需的时间。在测试环境中，显示器的分辨率为 2560×1440。Mi 和 Tm 则分别是当层级为 0 和层级最大时图像填充整个显示屏幕时万维望远镜所消耗的系统内存。

从结果中可以看出，在常用切片图像分辨率下（512×512），HiPS 标准数据集在渲染时的内存消耗略大于 TOAST，其主要原因是 HiPS 数据最常用的图像切分尺寸是 TOAST 数据的两倍。因此我们还测试了图像切分尺寸为 256×256

的 HiPS 标准数据集的渲染性能，此时其内存消耗小于 TOAST 数据。该结果表明，减小 HiPS 数据集图像切分的尺寸后，其内存消耗会得到显著改善。

HiPS 标准数据集在最大层级数据加载速度方面有较大优势。这是由于 HiPS 标准数据集具有更高的切分图像分辨率，因此能够以更少的图像来填充整个屏幕。如果应用了不同的 HiPS 或 TOAST 切片分辨率，以上测试结果可能会有所不同，因此在将原始图像数据转换至 HiPS 标准数据集时需要对可视化时的加载速度和内存消耗进行取舍。

表 5.2 万维望远镜软件原生数据集（TOAST）和 HiPS 标准数据集的加载速度及内存消耗对比。表头中 $MaxOrder$ 表示数据的最高层级，Ti 和 Tm 分别表示层级为 0 时和层级最大时图像填充整个屏幕所需的时间。Mi 和 Tm 分别是当层级为 0 和层级最大时图像填充整个显示屏幕时万维望远镜所消耗的系统内存，除 PLANCK 外，所有 TOAST 数据块的分辨率均为 256×256，HiPS 数据块的分辨率为 512×512。

Data name	Data type	Ti	Tm	Max order	Mi	Mm	Tile Size
Digitized Sky Survey	TOAST-Skymap	2.28s	7.73s	12	851920KB	825584KB	256×256
	HiPS-Skymap	2.25s	8.60s	9	1327612KB	1488840KB	512×512
WISE All Sky	TOAST-Skymap	1.90s	9.58s	7	821484KB	821832KB	256×256
	HiPS-Skymap	2.10s	3.86s	8	1551080KB	1221216KB	512×512
2Mass: Imagery	TOAST-Skymap	2.23s	11.86s	8	864448KB	821052KB	256×256
	HiPS-Skymap	2.30s	4.83s	9	1335512KB	1306756KB	512×512
Moon: Lunar Reconnaissance Orbiter WAC Global Morphologic Map	TOAST-Planet	2.12s	16.66s	10	817988KB	840520KB	256×256
	HiPS-Planet	1.65s	1.53s	6	1413816KB	23987602KB	512×512
Jupiter: PIA07782, Cassini's Best Maps of Jupiter	TOAST-Planet	1.43s	4.46s	4	846060KB	831084KB	256×256
	HiPS-Planet	2.30s	0.98s	3	1756340KB	1083836KB	512×512
PLANCK	TOAST-Skymap	0.71s	0.54s	6	990892KB	1000896KB	256×256
	HiPS-Skymap	0.63s	0.59s	3	757152KB	712348KB	256×256

5.5　本章小结

多波段、多信使联合观测对海量多源异构图像数据提出了统一组织、高效检索及可视化的需求。本章首先对当前主流的天文图像组织方法、数据获取方法和可视化技术进行了总览，之后详细介绍我们提出的海量多波段图像组织、检索和可视化框架。该框架主要基于层次渐进模式标准，其关键技术主要包括海量图像数据高效转换至 HiPS 标准数据集、层次渐进模式下的多波段图像检索和获取方法、以及全天尺度多波段图像进行沉浸式可视化。

我们提出了一种新的图像数据转换至 HiPS 标准数据集的方法，并在 Spark 并行计算集群上进行了实现。该方法利用了 HEALPix 网格的层级递归原理，通过像素并行写入实现了图像投影转换过程中直接生成 HiPS 标准文件，省去了传统的图像匹配拼接和像素融合的步骤，大幅度提升了 HiPS 标准数据集的处理效率。本文利用嫦娥 2 号全月面 7m 分辨率正射影像对本方法进行了测试，效果大幅领先于 CDS 发布的 HiPSGen 工具。

基于 HiPS 标准数据集的图像文件组织和命名方式，我们实现了层次渐进模式下海量图像文件检索和图像获取。其基本思路是在将用户的检索目标转化为 HEALPix 的天区网格。由于 HiPS 标准数据集的图像文件是与 HEALPix 天区网格是一一对应的，采用 HEALPix 空间索引即可实现查找。且 HiPS 标准数据集使用文件系统管理，并通过 HTTP 服务发布，通过简单拼接即可得到每块数据的访问地址。由于在将图像转换为 HiPS 标准数据集的过程中，原始图像与 HiPS 标准数据的文件对应关系也存入关系型数据库，因此也可通过该方式检索得到原始图像文件。

我们在万维望远镜中实现了上述框架，并实现了 HiPS 标准数据集的三维可视化。其关键技术是通过计算 HEALPix 天球网格在三维场景中的顶点坐标实现 HEALPix 网格的绘制，并根据视野场景和图像位置关系构建了层次细节模型，最后利用 HiPS 标准数据集的组织方式实现了图像文件在三维场景中的纹理渲染。我们还对比了 HiPS 标准数据集和万维望远镜原生数据集在三维场景中的渲染效率和内存消耗，结果证明，在相同图像切分分辨率下，HiPS 标准数据集在渲染时的内存消耗及加载速度上均有优势。

第 6 章　EP 暂现源判别及多波段证认

爱因斯坦探针卫星（Einstein Probe，EP）是我国新一代时域天文项目，其任务目标是通过监测 X 射线天空，发现和探索宇宙中的极端剧变天体和事件，研究天体的活动性。它将以高于现有设备一个数量级的探测灵敏度，在软 X 射线波段开展快速巡天监测，系统性的探测和研究各种时标上的 X 射线暂现源和天体的 X 射线时变。EP 的核心科学目标是：

- 发现宇宙中的 X 射线剧变天体；监测已知天体的活动性，探究相关现象的性质及物理机制；
- 发现和探索宇宙中沉寂黑洞的耀发；测绘黑洞的分布，进一步理解其起源、演化及物质吸积过程；
- 探寻来自引力波源的 X 射线信号，以增进对极端致密天体及其合并过程的认知。

EP 的暂现源系统性巡天还将发现宇宙更为遥远（早期）的伽马射线暴，以追踪照亮宇宙“黑暗时代”的第一代恒星，并利用其作为“灯塔”探索早期宇宙的黎明和再电离时期，以及第一代星系和早期星际介质。EP 也将探测超新星爆发瞬时的 X 射线爆发辐射。此外，爱因斯坦探针还将积累海量的全天时域巡天数据，可以用来开展大样本天体的 X 射线时变监测普查，研究包括恒星、致密天体、活动星系核等天体的活动性及其物理过程和起源 (袁为民 等, 2018)。

从 EP 的科学目标中可以看出，对 X 射线暂现源的证认是 EP 科学发现的前提工作。暂现源的证认过程包括对暂现源信号的探测，与已知源的比对从而确认其暂现源本质，并与多波段的数据进行交叉证认，从而了解其多波段的辐射特征。其中多波段数据能够对暂现源物理特性的分析起到至关重要的作用。比如，通过多波段数据可以了解该暂现源对应体的类型、距离、周围环境等信息，并了解暂现源的多波段能谱分布和时变特征，从而对研究其物理起源和性质提供极大帮助。例如，一个 X 射线暂现源在光学波段的对应体为一颗恒星，则该暂现源可能是一个恒星耀发事件；若在光学和射电波段的对应体是一个活动星系核，那么该暂现源就可能就是这个活动星系核的 X 射线耀发等。

针对 EP 观测暂现源的多波段证认的需求，我们基于本文提出的关键技术构

建了 EP 暂现源判别工具与多波段证认参考数据库，分别实现了 EP 观测源提取、暂现源判别、多波段数据融合等一系列流程。

6.1 EP 宽视场 X 射线望远镜观测源提取

EP 包含两个有效载荷，宽视场 X 射线望远镜（Widefield X-Ray Telescope, WXT）用于全天监测；随动观测 X 射线望远镜（Follow-up X-Ray Telescope, FXT）用于暂现源和爆发源的随动观测及机遇目标的观测。其中 WXT 的观测视场极大，超过 3600 平方度，是搜寻全天 X 射线暂现源的利器。由于 EP 宽视场 X 射线望远镜独特的龙虾眼聚焦成像光学结构，其成像特征不同于以往的 X 射线望远镜，基于 EP 科学仪器团队的模拟，其成像会在 CMOS 靶面上形成一个十字形结构（如图 6.1），且由于仪器支撑结构的遮挡，十字臂还会出现断裂，采用传统的图像识别方法提取图像中的 X 射线源，会将交叠的十字臂误判为源，其部分断裂的十字臂也会被误判为亮源，需要人工复核以提高准确率。为此，在本工作中首先基于深度学习网络研发了目标检测算法，实现 EP 数据 X 射线源的快速准确提取。

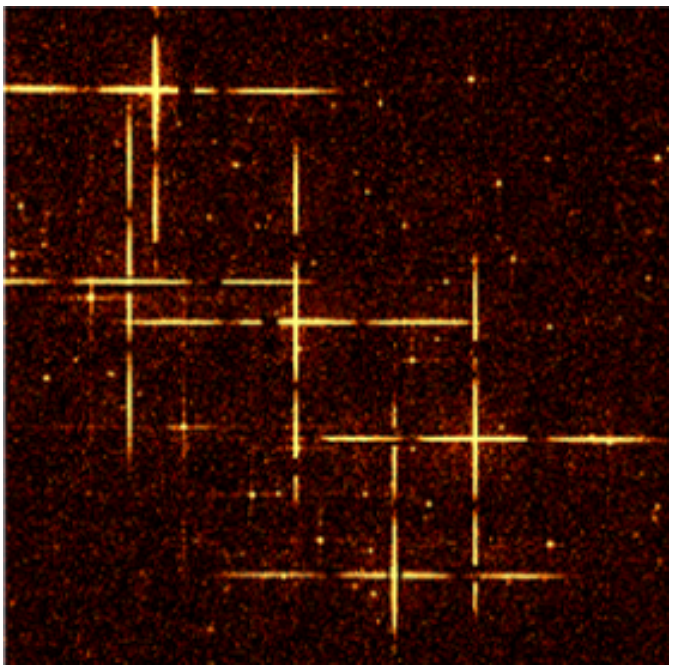


图 6.1 EP 模拟数据图像，500ks 曝光，其中亮源呈十字状，且有断臂暗纹

针对 EP WXT 数据提取观测源，主要是剔除假源并获取真正观测源的准确位置和流量。我们采用的思路为通过残差卷积神经网络的深度特征提取能力，获

取 EP 观测源的深度图像特征，基于该特征值实现提取。在实际操作中，通过下采样的方式降低图片尺寸，同时将多个像素聚合，并基于神经网络进行迭代计算，以均方误差作为损失函数实现训练。在进行源的判断时，通过训练得到的神经网络权值计算每个聚合像素的响应值，当值大于某一预定阈值时则认为该聚合像素为 X 射线源，而该数值即为源的流量。

网络训练所用的训练样本为 EP 科学仪器团队开发的暂现源仿真器生成的模拟数据。暂现源仿真器使用随机抽样的方法生成不同的观测仿真图像，根据 Geant4 模拟仿真给出的 PSF 数据进行抽样叠加到实际的观测视场，在模拟时保持暂现源位置、亮度（0.3 至 10mCrab）的随机性，并根据星表添加已知源。所有源的位置、亮度都记录在样本数据库中。通过该方式生成 10 万幅模拟数据，尺寸为 100×100。其中 9 万幅作为训练数据集，1 万幅作为验证数据集。

我们采用残差神经网络构建训练模型，基于模拟数据对方法进行了实验。基于经典残差神经网络 ResNet50(He 等, 2016) 构建了两个网络模型 EPNet25 和 EPNet100（结构如图 6.2），分别输出 25×25 和 100×100 的特征图。EPNet25 主要用来对亮度进行预测，EPNet100 用于预测源的位置，最后将两个模型的输出结果融合。模型优化过程中使用优化器 Adam 来更新和计算网络参数，使参数逼近或达到最优值，从而最小化损失函数。学习率开始时设置为 0.001，每当误差平稳时学习率除以 10。在测试中，模型给出了优秀的识别结果，精确率 92.59%，召回率 96.26%，流量预测平均相对误差 5.33%。

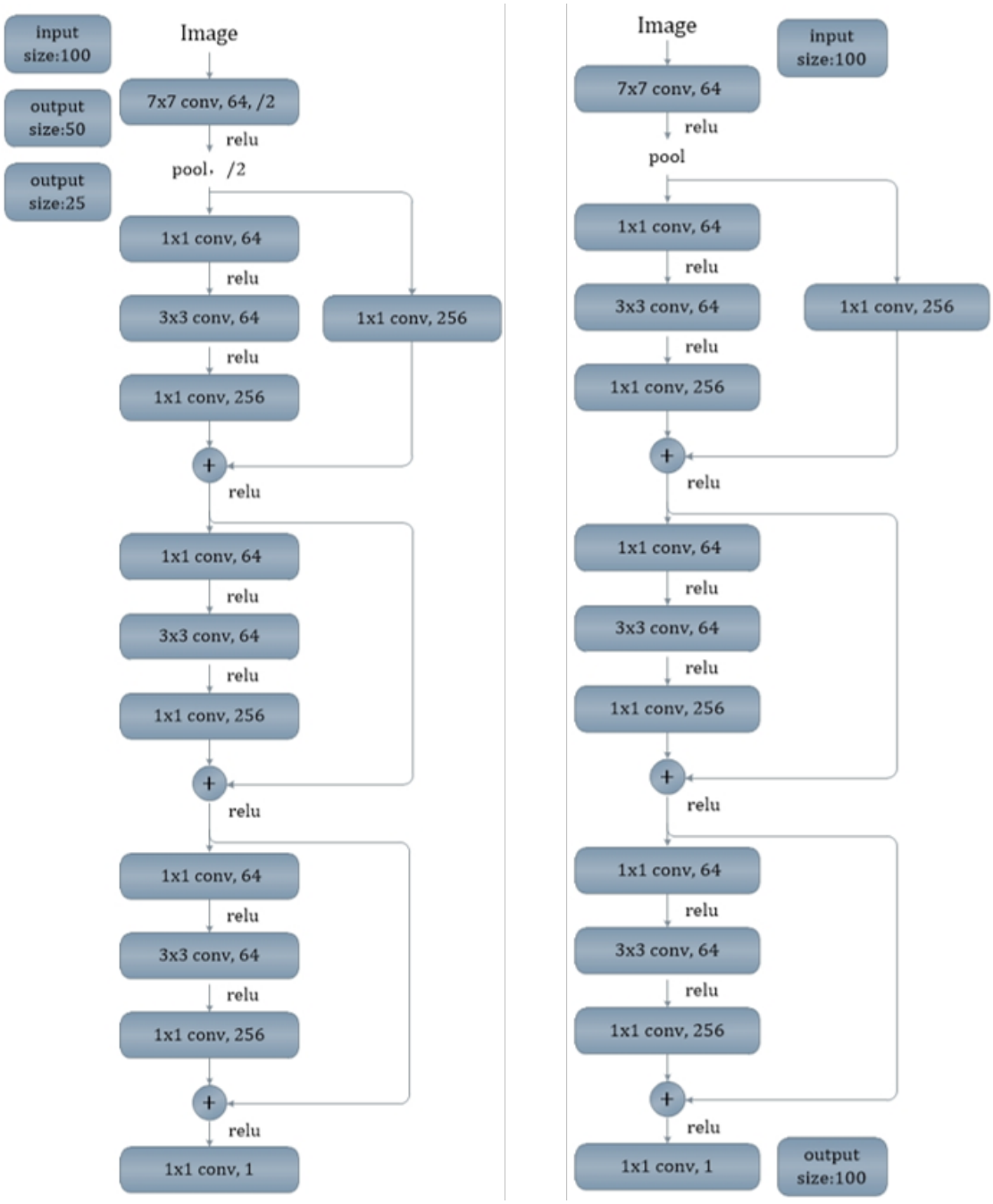


图 6.2 构建的残差卷积神经网络结构，自上而下由三个残差单元组成，每个单元包含了 3 层卷积，对应不同尺寸的卷积核，前两个卷积生成 64 组特征图，最后一个卷积生成 256 组特征图，最后通过 1x1 卷积输出每个像素的识别结果。左图为 EPNet25 的结构，右图为 EPNet100 的结构。两者的主要区别在于 EPNet25 的第一个卷积单元对数据进行了两次下采样，将 100×100 降维为 25×25

6.2　EP 暂现源判别

对识别出的 EP 观测源要进一步判别其是否为暂现源。通过与现有 X 射线源表进行对比剔除已知源，从而得出变源或暂现源。本工作采用的已知源表如表 6.1。对这些已知源表采用本文第四章介绍的关键技术进行交叉证认实现融合，并存入 PostgreSQL 数据库。

表 6.1 用于 EP WXT 暂现源判别的 X 射线源表。

名称	能段	覆盖区域
ROSAT All-Sky Bright Source Catalogue(1RXS)	0.2-2.4	全天
ROSAT All-Sky Survey Faint Source Catalog	0.2-2.4	全天
Second ROSAT PSPC Catalog	0.2-12	5982 平方度（全天 14.5%）
The XMM-Newton 2nd Incremental Source Catalogue	0.2-12	40 平方度
Chandra Source Catalogue	0.2-6.0	300 平方度
Swift BAT 105-Month Hard X-ray Survey	14-195	全天

由于已知 X 射线源表的数据条目并不多，不需要对其进行分表存储，仅需构建空间索引即可实现高效检索。将观测源的位置和 WXT 观测误差范围（约 1 角分）作为输入检索已知源表，若检索结果为空，或 P_{any} 值（即存在对应体概率）小于 30%，则认为该观测源为暂现源。将其信息作为输入进入下一步多波段证认步骤。

6.3　暂现源多波段参考数据库

为了实现 EP 暂现源多波段交叉证认在线运行，构建了 EP 多波段参考数据库。该数据库由多波段星表融合而成，采用 PostgreSQL 构建，并以微服务结构提供对外服务。该数据库提供了 API 可供 EP 暂现源处理 Pipeline 直接调用，也提供了 Web 界面供科学用户直接查询。查询结果将返回该源的各个波段属性数据并生成能谱。在可视化界面中，可以查看该暂现源的多波段图像，并可叠加对比分析。多波段参考数据库的运行流程如图 6.3。

EP 暂现源多波段参考数据库的主要数据来源包括如下数据集：

- γ 射线：Fermi、Swift/BAT
- X 射线：ROSAT、XMM、Chandra、Swift

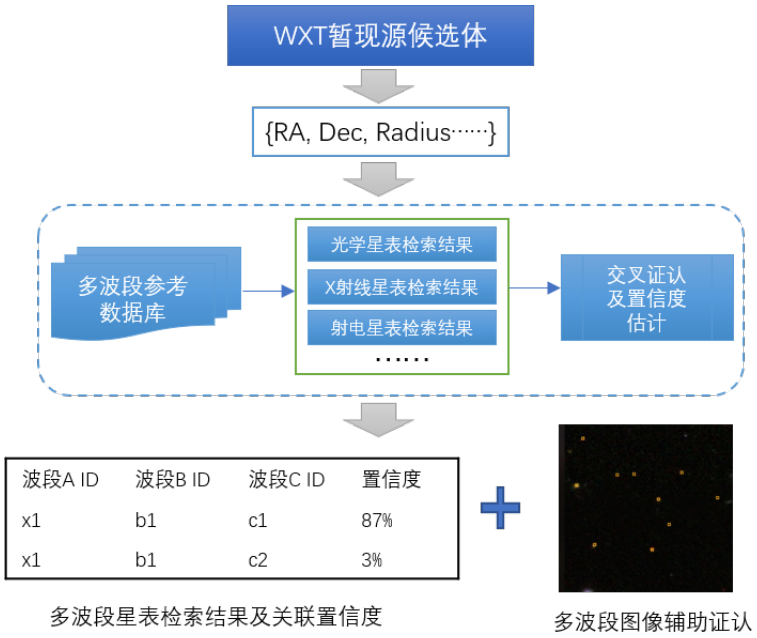


图 6.3 多波段参考数据库运行流程

- 光学：SDSS、LAMOST、GAIA
- 紫外：GALEX
- 红外：2MASS、WISE
- 射电：FIRST
- 引力波：Virgo/LIGO

得益于分表策略和高效的空间索引，完成一组暂现源（少于 50 个）在各个波段对应体搜索时间小于 500ms，完成交叉证认置信度计算时间小于 30s。针对每个暂现源的每组多波段对应体匹配，都会生成一组能谱分布，并展示在可视化界面供科学用户参考。同时基于 HEALPix 索引检索得到的多波段图像也会在可视化界面中进行展示。如图 6.4所示

各个波段的图像数据采用本文第四章中的多波段图像统一组织框架进行管理，将原始 FITS 文件转换为 HEALPix 分块图片文件，并按照 HEALPix 编号分层次存储。原始 FITS 文件的文件名与其生成的分块文件之间的对应关系记录在数据库中，并对 HEALPix 编号构建 B+ 树索引。当请求某个位置的图像时，针对该位置坐标计算 HEALPix 索引编号，并根据该编号调取对应的分块图片。文件通过承载了 HTTP 微服务的 Docker 镜像进行发布，用户可通过 Web 浏览器查看。若需要下载原始图像文件，则会通过数据库检索该分块图片 HEALPix 编号对应的原始文件，通过微服务返回即可下载。

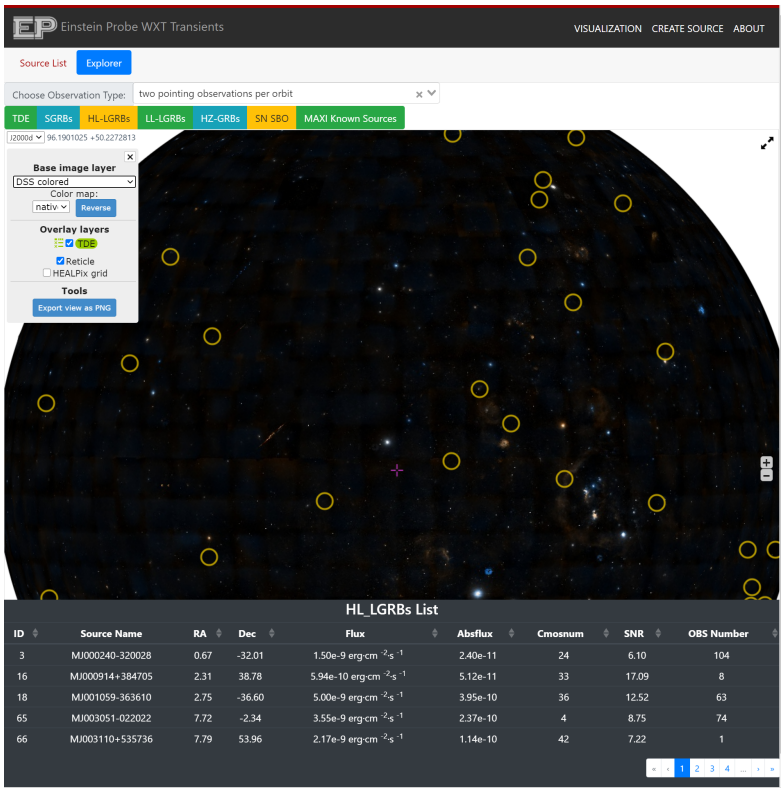


图 6.4 多波段参考数据库界面

多波段参考数据库还包含了参与数据融合的多波段星表的元数据表，记录了这些星表的元数据信息。同时该数据库具备良好的可扩展性，可兼容未来发布的新的多波段数据星表。

6.4　本章小结

满足时域天文观测计划的高时效性数据应用需求是本文的主要研究目的。针对 EP 的暂现源判别及多波段证认的需求，我们运用本文提出的一系列关键技术构建了 EP 暂现源判别工具与多波段证认参考数据库，实现了从 EP 观测源提取、暂现源判定、暂现源多波段交叉证认等一系列功能。基于 EP 的模拟数据对该工具和参考数据库进行了测试，结果显示能够有效满足 EP 暂现源证认的时效性需求。由于 EP 卫星尚未发射，当前的工作仍属科学应用的预研阶段，尤其是模拟数据仍在不断的更新和完善之中，这也对我们的方法提出了数据兼容性和功能可扩展性的新需求。

第 7 章　引力波电磁对应体高效搜寻

引力波事件是当前最主要的多信使天文观测目标。地面引力波源按照数据特点可以分为 4 种类型：连续引力波源（Continuous GW sources）：如快速旋转的中子星，在远比探测器寿命的时期内发射准正弦的引力波；随机引力波（Stochastic GWs）：随机引力波可以采用宇宙学背景的形式，类似于宇宙微波背景，或者可以来自较近距离的引力波源的杂音；引力波爆（Burst GWs）：引力波爆是尚未较好建模或未知波形的暂现信号，其爆发源包括超新星、致密双星的并合及并合后阶段；致密双星绕转（Compact Binary Coalescence, CBC）是相互旋进的二元系统，其中双方或其中一方为黑洞或中子星。CBC 是当前引力波探测仪器最有希望观测到的引力波源，在其设计灵敏度下能够探测到 20 次/年的类似事件。

作为天体物理和宇宙学研究的重要工具，多信使联合观测的一个重要的应用是引力波宇宙学。引力波源 (如双中子星并合) 可以作为标准铃声来测量宇宙的几何和膨胀 (Schutz, 1986)，并可由此检验宇宙学模型、测量宇宙学参数、研究暗能量和暗物质 (Zhu 等, 2001; Zhao 等, 2011)。天文学发展至今，电磁波段是发展最完善、理论研究最透彻的观测窗口，也是现有探测手段与探测仪器最丰富的窗口。无论是要寻找引力波源的天体物理起源，还是对其物理性质开展进一步的研究，都需要通过对引力波源伴随产生的电磁信号的联合探测来完成。因此，从引力波天文学的角度上讲，引力波事件电磁对应体观测研究的意义可相比于引力波信号的直接探测 (高鹤 等, 2001)。

但是地面引力波观测网络对引力波源的定位能力较差，只能得到引力波源的定位误差天区。对于最典型的中子星双星系统来说，单个探测器的源定位的灵敏度非常差，几乎全天响应，很难实现定位。两个探测器对引力波源的定位只能限制在一个环带上。三个探测器或者更多探测器联网，则可以对引力波波源的天空位置进行重构 (Wen 等, 2010; Fairhurst, 2011; Klimenko 等, 2011)。一般情况下引力波源定位的椭圆误差面积在 100 平方度左右 (蔡永志 等, 2017)。根据测算，2020 年后升级的 aLIGO A+ 探测器网络，对引力波事件定位天区范围中值为 110-180 平方度（90% 置信度）(Aasi 等, 2016)。由于引力波探测器的灵敏度及最大探测距离不同（LIGO 设计灵敏度下 190Mpc，Virgo 125Mpc，KAGRA

140Mpc），即使未来所有已规划的地面引力波探测器全部建成使用，对于较远的引力波事件，能够捕捉到其信号的探测器数目减少，其定位天区也会很大。与此同时，探测网络的灵敏度不断提高，将能够探测到平均距离 190Mpc 以内的双中子星并合所产生的引力波事件，预计每年将探测到约 4-80 次双中子星并合事件产生的引力波信号 (Aasi 等, 2016; Schutz, 2011; Zhao 等, 2018; Pan 等, 2019)。事件的高探测率和定位的低精度给引力波电磁对应体的随动观测带来了巨大的挑战，这就需要探索新的方法实现对引力波源的精确定位。

7.1 基于宿主星系筛选的引力波随动观测规划

Gehrels 等 (2016) 最先提出了基于星系观测搜寻引力波电磁对应体的策略。该策略假定引力波源均位于或邻近于某个星系，该星系即称为其宿主星系。利用该寄主星系的信息，可以缩小引力波事件的源天区位置估计。利用星系星表，以引力波定位天区进行查询，可以得到定位天区中的宿主星系候选体，并作进一步筛选，以开展及时的随动观测。基于对国内外相关研究 (Arcavi 等, 2017; Ducoin 等, 2020; Chan 等, 2017) 的总结，我们利用本文的提出的关键技术，结合虚拟天文台领域的一系列相关技术，构建了一个高效且自动化的引力波随动观测规划系统 (GW follow-up Observation Planning System, GWOPS)，该系统主要解决了小视场望远镜搜寻引力波电磁对应体的三个主要问题：

1. 如何在引力波事件的定位天区中高效搜寻宿主星系。
2. 如何排序宿主星系的观测优先级。
3. 如何从观测数据中高效识别电磁对应体。

针对以上问题，本文的主要研究成果给予了解决方案。基于 HEALPix 天球划分的分表分区策略以及基于多层级覆盖天区的高效的空间索引被应用于星表数据库中，并采用基于贝叶斯推断的方法对宿主星系候选体进行筛选和排序。在完成随动观测后，图像数据会被存储在观测数据库中，通过 Pipeline 自动提取观测数据中的观测源，并通过机器学习方法对这些源进行打分。用户可以在 GWOPS 的 web 界面中查看所有的观测数据和自动提取的观测源，通过多波段参考数据来判断它是否是暂现源，并对其进行初步的分类。

GWOPS 由三部分组成，分别是引力波宿主星系筛选组件，暂现源证认组件和数据可视化组件。如图 7.1所示，引力波宿主星系筛选组监听由 LVC 通过

NASA 伽马射线协调网络（GCN）发布的引力波事件警报。收到警报后，它将自动从 GraceDB 中（gravitational-wave candidate event database）下载引力波定位数据。之后根据定位数据查询星系数据库以获取定位区域中的所有宿主星系候选体。由于这些候选者的数量将多达数千个，因此还需要进一步过滤。

GWOPS 应用贝叶斯推断来计算每个候选者是引力波事件宿主星系的概率，并按该概率进行排序。排名的前 150 个候选体被发送到望远镜以进行后续观测。所有观测数据将存储在数据库中。暂现源证认组件用于从观测数据中识别可能的暂现源，即引力波事件的电磁对应体的候选体。这个步骤是通过将观测数据和模板数据相减以获得残差图像，然后通过图像识别方法提取观测暂现源（Observation Transient, OT）。用户可以在 GWOPS 的基于 Web 的可视化界面中对 OT 进行人工验证和分类。数据可视化组件主要通过 HiPS 展示已观测的天区、现有观测数据及已证认的暂现源，且可以展示已观测数据与其它波段图像数据间的对比。

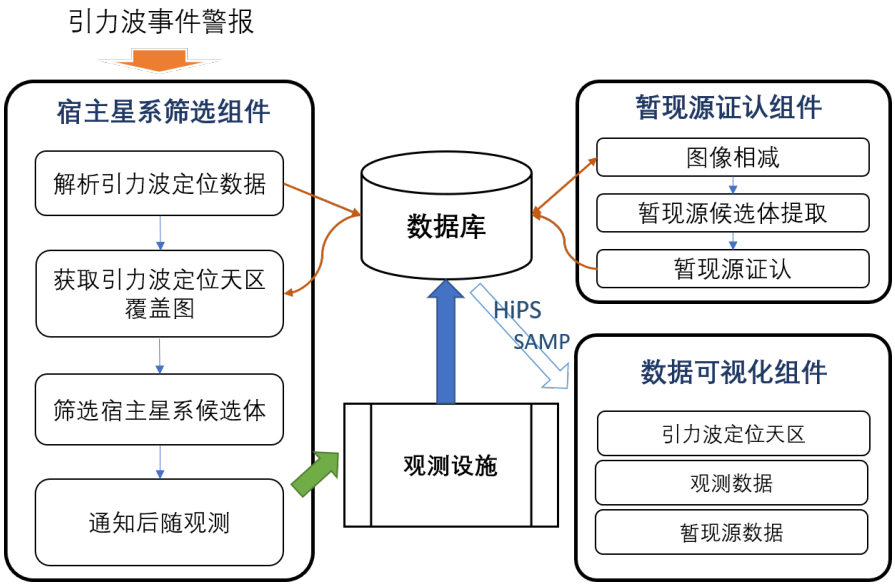


图 7.1 GWOPS 的体系结构

7.2　GWOPS 数据库设计

GWOPS 数据库是基于 PostgreSQL 构建的，其中的核心数据表是引力波事件表、星系表、观测数据表、模板数据表、残差数据表和暂现源候选体表。数据库的结构如图 7.2所示。

1. 引力波事件表：记录每个引力波事件的 ID，发布时间，估计距离和定位天区。

2. 星系表：存储了 200Mpc 范围内的星系信息，包括每个星系的位置，距离，红移以及 B 波段光度。

3. 观测数据表：该表记录了每个引力波事件的观测数据。

4. 模板数据表：包含之前观测的图像数据，作为获取残差图像以提取暂现源的模板。

5. 残差数据表：存储从观测图像和模板图像的图像相减得出的辅助数据。

6. 暂现源候选体表：该表记录从残差图像中提取的暂现源候选体的位置及人工证认的结果。

其中星系表采用的是 GLADE(Dálya 等, 2018)，是通过多个星表融合得到的星系星表，专门用于引力波宿主星系搜索和电磁对应体的筛选以进行随动观测。用于融合得到 GLADE 的五个星表分别是 GWGC (White 等, 2011)，2MPZ (Bilicki 等, 2013)，2MASS XSC (Jarrett 等, 2000)，HyperLEDA (Prugniel, 2005) 和 SDSS DR12Q (Kozlowski, 2017)，它们在波长和星系类型上互为补充。GLADE 总共包含 360 万个星系，其在 37Mpc 以内的完整度为 100%，在 aLIGO 单探测器针对 BNS 探测范围的最大距离（基于 O2 期间的灵敏度，约 100Mpc）上具有约 61% 的完整性。在 O3 期间的 aLIGO 单探测器针对 BNS 的探测范围（约 173Mpc）上的完整性约为 48%(Dálya 等, 2018)。

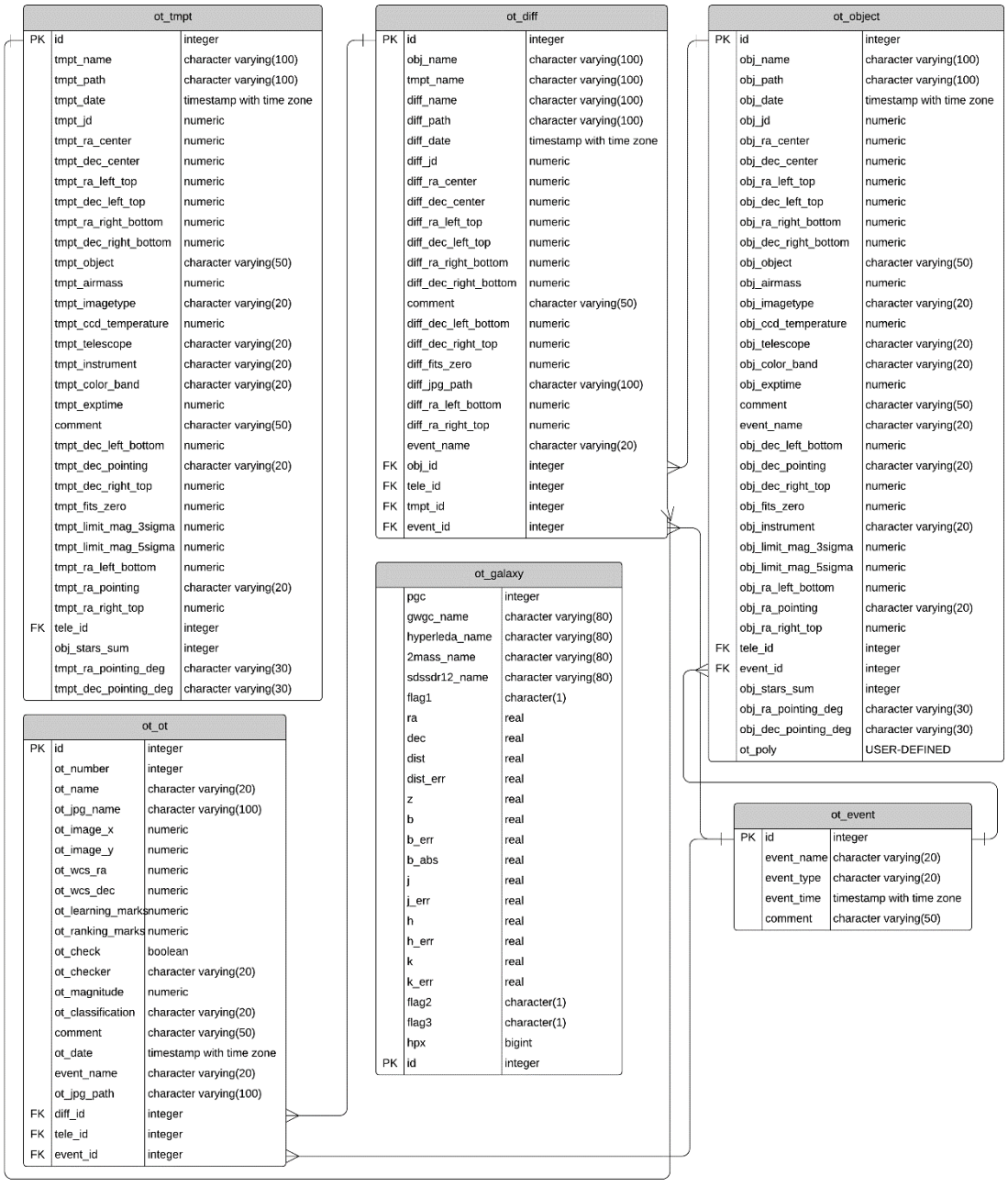


图 7.2 GWOPS 的数据库结构

7.3 引力波宿主星系星系筛选及排序

当探测到引力波事件时，LVC 将通过 GCN 发送 VO-Event 警报。警报包含了具有距离限制的引力波事件定位天区图。通常该定位区域是不规则条带状，而常规空间搜索是圆锥查询或多边形查询。因此，我们利用了第三章介绍的 MOC-Tree 索引的原理开发了一种基于引力波定位区域数据的专用数据库检索方法。引力波定位天区数据是一个 FITS 文件，该文件记录每个 HEALPix 网格上的该事件的发生概率。

要将引力波定位区域用作进行星系检索的条件，需要首先将其转换为 MOC-Tree，并在数据表上构建相应的多层级覆盖天区空间索引。MOC-Tree 能够指定任意天球区域，其机制基于 HEALPix 天球划分算法，是一种将天区映射到按层次分组的预定义单元格中的方法。图 7.3展示了将引力波事件 GW170817 的定位天区转化为 MOC-Tree 作为星系检索条件的天区覆盖情况。

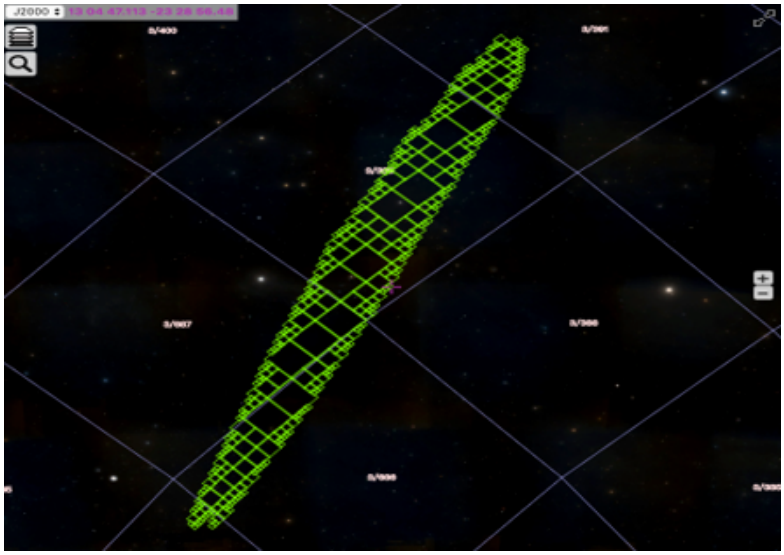


图 7.3 引力波事件 GW170817 的定位天区转化为 MOC-Tree 作为星系检索条件

MOC-Tree 以不同层次的 HEALPix 网格的集合来表示天区的范围。由于基于多层级覆盖天区的空间索引在索引号上能够表征每个天体在不同层级下的 HEALPix 天球划分的相对位置，因此仅通过少量的索引扫描即可完成星表的检索。对于典型的引力波定位区域（100 平方度至 200 平方度），其检索时间在 500ms 以内。

基于引力波定位区域检索到的星系数量约为 10,000 至 50,000 个。在有限的时间内不可能完全观测到如此多的星系，因此还需对它们作进一步筛选并进行

优先级的排序。在 GWOPS 中，我们采用基于贝叶斯推理的方法 (Fan 等, 2014) 来计算该星系是引力波事件宿主的概率，然后根据该概率对星系进行排序。该方法将星系的物理特征（B 波段亮度，距离）和在星系所处位置发生引力波事件的概率作为先验信息，计算出每个星系作为引力波事件宿主的后验概率。公式 7.1用于计算后验概率。

$$p(\gamma|D,S,M)=\frac{p(\gamma|S,M)p(D|S,\gamma,M)}{p(D|S,M)} \tag{7.1}$$

其中参数意义为：

- γ: 引力波事件和宿主星系两组观测的公共参数，具体指覆盖天区中星系的位置参数，α (RA), β (Dec), d (距离)
- D: 引力波事件观测数据相关信息
- S: 电磁波段观测数据集
- M: 宿主星系候选体信息
- p(D | S, γ, M) 是给定 S、γ 时数据集 D 的似然概率, 即引力波定位数据中给出的星系对应位置的概率
- p(γ | S, M) 是先验概率公式

其中先验概率公式为7.2:

$$p(\gamma|S,M)\propto(\frac{D_{gc}}{D_{GW}})^3\frac{1}{N}\sum_{j=1}^{N}\delta(\alpha-\alpha_j,\beta-\beta_j,d-d_j)L_{B_j}+\frac{3L_{Bmean}}{4\pi D_{GW}^3}H(d-D_{gc})d^2 \tag{7.2}$$

其中 N 是星系星表中的星系总数；Dgc 为星系星表的范围；Dgw 为引力波探测器的探测范围；LBj 为宿主星系 B 波段的光度，H 为阶跃函数。范锡龙等人基于致密双星并合的模拟数据对该方法进行过验证，结果显示，在 200Mpc 的范围内，在 8.5% 的模拟结果中，排序中的前 10 位星系是引力波源宿主星系的概率为 50%。在 10% 的模拟中，宿主星系确实位于本方法的前 10 位星系中。表 7.1为 GW170817 的宿主星系候选体排序结果，其宿主星系 NGC4993 排在第二位。此外我们还采用 GLADE 星表中的 K 波段数据作为先验计算因子，得到了一组对比结果。在该结果中，NGC4993 排在第 45 位。

表 7.1 引力波事件 GW170817 宿主星系候选体的排序结果

Rank	gwgc_name	hyperleda_name	2mass_name	sdssdr12_name	RA	Dec	Dist
1	Null	Null	13104593-2351566	Null	197.691	-23.8657	38.7294
2	***NGC4993***	***NGC4993***	***13094770-2323017***	***Null***	***197.449***	***-23.3838***	***39.3549***
3	ESO508-019	ESO508-019	Null	Null	197.466	-24.2394	39.4723
4	IC4197	IC4197	13080432-2347486	Null	197.018	-23.7968	41.0586
5	NGC4968	NGC4968	13070597-2340373	Null	196.775	-23.677	40.4935
6	PGC803966	PGC803966	Null	Null	196.879	-23.1705	38.0388
7	IC4180	IC4180	13065651-2355014	Null	196.735	-23.9171	40.6674
8	PGC772879	PGC772879	Null	Null	198.195	-25.9866	39.1942
9	PGC797164	PGC797164	Null	Null	197.177	-23.7757	36.554
10	ESO576-003	ESO576-003	13103572-2144536	Null	197.649	-21.7482	36.6755

7.4 暂现源证认

根据上一节产生的观测列表所进行的随动观测结果将存入数据库的 Object 表中。系统会用图像相减的方式挑选候选体，并用机器学习方法给出每个候选体为真实暂现源的概率。相减采用的模板图像一部分来自之前的观测，当历史观测图像缺失时，则调用 Pan-STARRS 提供的拼接图像服务获取对应位置的图像做为模板。之后通过 SExtractor(Bertin 等, 1996) 对图像相减得到的残差图像进行 OT 的提取，并采用预训练的随机森林算法为 OT 进行打分，当 OT 得分高于 0.2（满分 1.0）时，才将起列入至人工证认队列，该分数也将作为人工证认时的参考。GWOPS 暂现源证认的流程如图 7.4。

用户可以在系统的交互界面中查看置信度较高的候选体，并对其进行人工证认。在证认过程中，可以查看观测图像、对应的模板图像，以及两者相减后的残差图像。人工判读这些图像，结合候选体的亮度信息并对比其他巡天数据等，可以判断该候选体是否可能为真的候选体，并对其进行分类。分类的类别包括 Supernova、GRB、AGN、Moving Object、Bright Star、Variable Star、Bad Subtraction 等。用户还可以查看多波段参考数据库中该 OT 所在位置的其他波段数据，并生成对应的能谱。也可查看如 SDSS、TNS、NED 等巡天项目或数据库在该位置的对应数据。用于暂现源人工证认参考数据示例如图 7.5。如果该 OT 为真的暂现源，则交给其他随动跟踪望远镜作进一步的光谱证认。

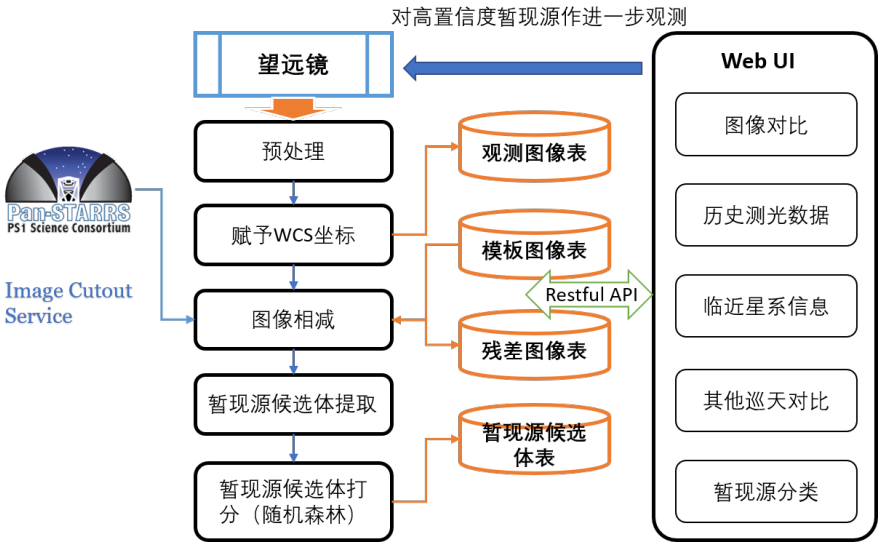


图 7.4 GWOPS 暂现源证认流程

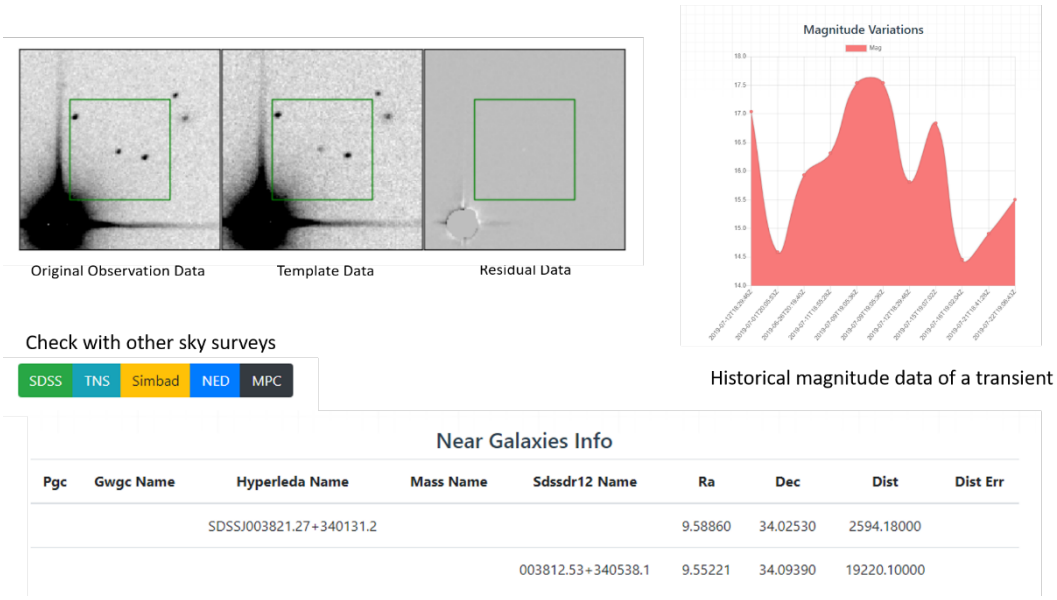


图 7.5 暂现源人工证认参考数据示例，左上是该 OT 的观测数据、模板数据和残差数据；右上为该 OT 的历史光变信息；点击中间的一排按钮可以查看对应的其他巡天在该位置的数据情况；下方是该 OT 位置附近的星系的相关信息。

7.5 GWOPS 可视化组件

可视化组件使得用户直观地查看引力波定位区域中的观测数据。我们采用了本文第五章介绍的方法实现了观测数据的统一组织和可视化功能。将观测数据转换为 HiPS 标准数据集，可以使专用客户端或浏览器工具渐进式地访问和显示数据，其特点是随视野放大的层次越高，所显示的数据细节越多。GWOPS 使用 Aladin Lite (Boch 等, 2014) 作为网络上数据可视化的工具。Aladin Lite 是基于 WebGL 的 Javascript 库，它可以在三维天球上显示加载了 WCS 坐标的 FITS 图像，并基于 JSON 文件在天球上绘制轮廓，这些轮廓来自引力波定位天区数据。使用基于像素并行写入的方法在 Spark 集群上对观测数据进行网格划分，将数据从原始投影转换为 HEALPix 投影，然后根据 HEALPix 的层次划分将数据递归划分为不同分辨率的图像金字塔，其流程如图 7.6。

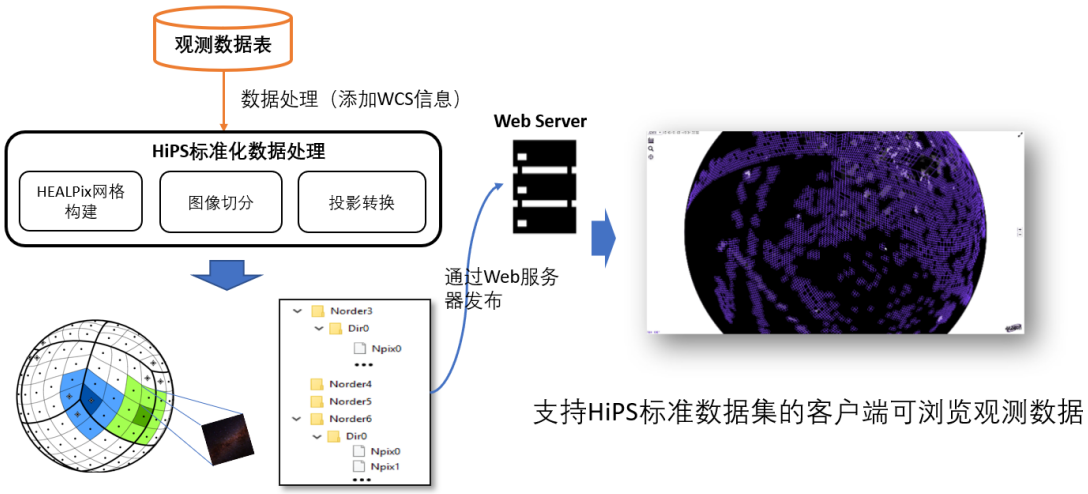


图 7.6 GWOPS 观测数据可视化及流程

GWOPS 中的数据也可以在万维望远镜中可视化。用户需要在计算机上另外安装 China-VO 版本的万维望远镜[1]并预先启动。前文介绍了我们已在万维望远镜中成功实现了对 HiPS 标准数据集的渲染支持 (Xu 等, 2020)，因此 GWOPS 的数据均可在万维望远镜上自由浏览。对于引力波定位数据的可视化，天区文件将被转换为 HEALPix 网格与定位概率一一对应的键值对文件。将这些文件封装为 VOTable，用户可以通过单击数据浏览器页面上的“发送到 WWT”按钮，通过 IVOA SAMP 协议 (Taylor 等, 2015) 将文件传输到万维望远镜。

[1]China-VO WWT 可以从 http://wwt.china-vo.org 下载

7.6　本章小结

本文提出海量星表高效检索方法、天体匹配置信度计算方法以及海量图像数据高效检索与可视化框架是 GWOPS 的核心驱动技术。基于海量星表高效检索方法，实现了以引力波定位天区为检索条件，在星表数据库中高效筛选引力波电磁对应体宿主星系候选体。采用基于贝叶斯的天体匹配置信度计算方法，对宿主星系候选体作进一步筛选，并将它们按照作为宿主星系的概率进行随动观测。对于引力波电磁对应体的随动观测图像数据，采用本文提出的基于像素并行写入的方法转换为 HiPS 标准数据集存储管理及可视化，并通过机器学习方法对图像中的 OT 进行提取及暂现源判别。第六章构建的暂现源多波段参考数据库也能够为引力波电磁对应体的证认提供有力支撑。

在 LVC 的 O3 运行期间，GWOPS 为中国的 6 台光学望远镜提供了引力波随动观测的支持。在截止到 2020 年 3 月的 53 个引力波事件中，GWOPS 都对事件警报做出了实时的响应，平均在 25s 之内完成引力波宿主星系候选体的筛选并提交给望远镜进行观测。在此期间，共观测了 6133 平方度的天区，获取了 21,172 幅观测数据。但是由于 O3 运行期间的引力波事件中只有 8 例是可能的 BNS 事件，且距离太远或不在望远镜的可视范围内，未能够找到引力波事件的电磁对应体。

我们下一步将会对 GWOPS 做进一步的提升，包括提供随动观测列表订阅功能，接入更多波段的观测设施的数据，并基于深度学习方法提升暂现源识别的准确率。

第 8 章　总结与展望

海量数据环境下的暂现源实时交叉证认、多信使事件电磁对应体的高效搜寻，以及暂现源的随动多信使观测证认是当前时域天文学和多信使天文学在观测数据的处理分析领域面临的主要挑战。多波段、多信使数据融合可以得到暂现源在不同波段的对应体资料，并给出相应的多波段图像和能谱分布，从而为研究其物理本质和周围环境提供关键信息。

当前多波段、多信使数据的高效融合仍存在一系列问题，主要表现在海量多波段星表的交叉证认及置信度计算效率较低，不能满足暂现源证认的时效性和准确性要求。且多波段图像的格式、组织异构化，难以实现统一获取和可视化。本文针对以上问题开展技术攻关，分别对海量多波段星表高效检索方法、多波段星表高效交叉证认及置信度计算、以及异构多波段图像组织、检索和可视化框架等开展了研究。

在海量多波段星表高效检索方面，本文提出了多层级覆盖天区空间索引方法及基于天球划分的星表分表分区策略。该方法与其它主流星表检索方法间相比，在大范围天区检索效率上有较大幅度提升，综合检索效率也优于同类型方法。

将多层级覆盖天区空间索引方法和星表分表分区策略应用至海量多波段星表间的位置匹配，有效提升了多波段星表位置匹配的效率。针对多波段星表由于空间定位精度不同带来的多对象匹配问题，本文提出了并行化的贝叶斯推断星表交叉证认置信度计算方法。通过与 XMM-COSMOS 多波段星表的对比测试，结果显示本方法在多波段星表交叉证认的效率及准确率上均有提升。

在异构多波段图像组织、检索和可视化方面，本文设计并实现了基于层次渐进模式标准（HiPS）的多波段图像组织、检索和可视化框架。所提出的基于像素并行写入方法，结合 Spark 内存计算集群，大幅度降低了图像数据转换至 HiPS 标准数据集的处理时间。在 HiPS 标准数据集高效处理的基础上，进而实现了图像数据的高效检索和获取。此外还在万维望远镜软件平台中实现了三维场景下 HiPS 标准数据集的沉浸式可视化，拓宽了万维望远镜软件的数据源。通过与万维望远镜原有数据组织方式下的可视化效率进行对比，本文的实现方法在内存

占用及加载时间上均有优势。

针对 EP 卫星暂现源多波段交叉证认的实际需求，本文应用上述研发的关键技术，构建了 EP 暂现源判别工具和多波段参考数据库。此外还针对引力波电磁对应体高效搜索的需求，构建了引力波事件随动观测规划系统 GWOPS。这两项工作使本文提出的方法能够落地，并得到了很好的检验，也为 EP 和多信使观测未来的科学产出提供了支撑。

在下一阶段的工作中，将继续对本文提出的方法进行改进及完善。其中较为具体的可改进方向包括对海量星表空间索引的物理组织方式可进一步优化，减少检索的对比次数。多波段星表交叉证认的置信度估计可结合实际应用引入更多参量，进一步提高交叉证认的准确率。此外，异构多波段图像数据转化至 HiPS 数据集会出现全天图像亮度不一致问题，还需要进一步解决。随着 EP 的发射和引力波探测的深入，也将对论文研究内容在真实科学应用中进行检验和完善。

在 2019 年由中国虚拟天文台（China-VO）发起的虚拟天文台核心功能需求调查结果中，数据检索与获取、数据可视化和交叉证认服务是 VO 用户最为关心的三个核心需求 (许允飞 等, 2020)。本文的工作一定程度上实现了这些核心需求的关键技术，也完成了一些典型应用，但尚未形成能够供天文研究工作者随时使用的在线服务，且在易用性、便捷性方面仍需进一步的努力。IVOA 针对天文数据的可获取性制定了一系列的技术标准，对服务的发现、注册、接口方式、数据模型等均有详细的定义。这些标准为实现多波段多信使数据融合的在线服务提供了指导，也是我们下一阶段工作的主要依据。我们将在现有工作的基础上，进一步完善 IVOA 协议接口类型，以国家天文科学数据中心为基础平台提供面向天文科研工作者的多波段多信使数据融合在线服务。

此外，本文的工作作为 China-VO 核心功能的一部分，也将基于 VO 平台化的设计路线开放相应的第三方开发接口，吸引感兴趣的开发者基于数据融合关键技术以及 VO 的数据资源做出更多面向具体科学目标的应用，以实现资源与技术向服务的快速转换。

参考文献

蔡永志, 李小龙, 李木子, 等. 引力波探测器联网的定位能力 [J]. 中国科学: 物理学力学天文学, 2017(1): 8.

袁为民, 张臣, 陈勇, 等. 爱因斯坦探针: 探索变幻多姿的 X 射线宇宙 [J]. 中国科学: 物理学, 力学, 天文学, 2018, 48: 039502.

许允飞, 樊东卫, 崔辰州, 等. 中国虚拟天文台的核心功能需求调查分析 [J]. 2020.

高鹤, 范锡龙, 吴雪峰, 等. 引力波爆发事件的电磁对应体的探测 [J]. Science in China Series A-Mathematics, 2001, 44: 249.

Aasi J, Abadie J, Abbott B, et al. Prospects for observing and localizing gravitational-wave transients with advanced ligo and advanced virgo [J]. Living Reviews in Relativity, 2016, 19.

Allen G, Anderson W, Blaufuss E, et al. Multi-messenger astrophysics: harnessing the data revolution [J]. arXiv preprint arXiv:1807.04780, 2018.

Anon. An overview of Hierarchical Progressive Surveys (HiPS) and the HEALPix framework [Z].

Anon. Gw events published during o3 [Z].

Arcavi I, McCully C, Hosseinzadeh G, et al. Optical follow-up of gravitational-wave events with Las Cumbres observatory [J]. The Astrophysical Journal Letters, 2017, 848(2): L33.

Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (surf) [J]. Computer vision and image understanding, 2008, 110(3): 346-359.

Beaver D, Kumar S, Li H C, et al. Finding a needle in haystack: Facebook's photo storage. [C]// OSDI: volume 10. 2010: 1-8.

Becla J, Lim K T, Monkewitz S, et al. Organizing the extremely large lsst database forreal-time astronomical processing [R]. Stanford Linear Accelerator Center (SLAC), 2007.

Berriman G B, Deelman E, Good J C, et al. Montage: a grid-enabled engine for delivering custom science-grade mosaics on demand [C]//Optimizing Scientific Return for Astronomy through Information Technologies: volume 5493. International Society for Optics and Photonics, 2004: 221-232.

Bertin E. Swarp: Resampling and co-adding fits images together [J]. ascl, 2010: ascl-1010.

Bertin E, Arnouts S. SExtractor: Software for source extraction [J]. Astronomy and Astrophysics Supplement Series, 1996, 117(2): 393-404.

Bilicki M, Jarrett T H, Peacock J A, et al. Two micron all sky survey photometric redshift catalog: A comprehensive three-dimensional census of the whole sky [J]. The Astrophysical Journal Supplement Series, 2013, 210(1): 9.

Boch T, Fernique P. Aladin Lite: Embed your Sky in the browser [J]. ASPC, 2014, 485: 277.

Boch T, Pineau F, Derriere S. The cds cross-match service [J]. ASPC, 2012, 461: 291.

Boller T, Freyberg M, Trümper J, et al. Second rosat all-sky survey (2rxs) source catalogue [J]. Astronomy & Astrophysics, 2016, 588: A103.

Bonnarel F, Fernique P, Bienaymé O, et al. The aladin interactive sky atlas-a reference tool for identification of astronomical sources [J]. Astronomy and Astrophysics Supplement Series, 2000, 143(1): 33-40.

Breitenberger E. Analogues of the normal distribution on the circle and the sphere [J]. Biometrika, 1963, 50(1/2): 81-88.

Brusa M, Zamorani G, Comastri A, et al. The xmm-newton wide-field survey in the cosmos field. iii. optical identification and multiwavelength properties of a large sample of x-ray-selected sources [J]. The Astrophysical Journal Supplement Series, 2007, 172(1): 353.

Brusa M, Civano F, Comastri A, et al. The xmm-newton wide-field survey in the cosmos field (xmm-cosmos): demography and multiwavelength properties of obscured and unobscured luminous active galactic nuclei [J]. The Astrophysical Journal, 2010, 716(1): 348.

Budavari T, Lee M A. Xmatch: Gpu enhanced astronomic catalog cross-matching [J]. ascl, 2013: ascl-1303.

Budavári T, Szalay A S. Probabilistic cross-identification of astronomical sources [J]. The Astrophysical Journal, 2008, 679(1): 301.

Cappelluti N, Hasinger G, Brusa M, et al. The xmm-newton wide-field survey in the cosmos field. ii. x-ray data and the log n-log s relations [J]. The Astrophysical Journal Supplement Series, 2007, 172(1): 341.

Chan M L, Hu Y M, Messenger C, et al. Maximizing the detection probability of kilonovae associated with gravitational wave observations [J]. The Astrophysical Journal, 2017, 834(1): 84.

Cutri R, Wright E, Conrow T, et al. Allwise data release [J]. IPAC/Caltech, 2013.

Dálya G, Galgóczi G, Dobos L, et al. GLADE: A galaxy catalogue for multimessenger searches in the advanced gravitational-wave detector era [J]. Monthly Notices of the Royal Astronomical Society, 2018, 479(2): 2374-2381.

Dowler P, Bonnarel F, Tody D, et al. Ivoa simple image access version 2.0 [J]. IVOA Recommendation, 2015, 23.

Du P, Ren J, Pan J, et al. New cross-matching algorithm in large-scale catalogs with threadpool technique [J]. Science China Physics, Mechanics and Astronomy, 2014, 57(3): 577-583.

Ducoin J, Corre D, Leroy N, et al. Optimizing gravitational waves follow-up using galaxies stellar mass [J]. Monthly Notices of the Royal Astronomical Society, 2020, 492(4): 4768-4779.

Fairhurst S. Source localization with an advanced gravitational wave detector network [J]. Classical and Quantum Gravity, 2011, 28(10): 105021.

Fan D, Budavári T, Szalay A S, et al. Efficient catalog matching with dropout detection [J]. Publications of the Astronomical Society of the Pacific, 2013, 125(924): 218.

Fan D, Budavári T, Norris R P, et al. Matching radio catalogues with realistic geometry: application to swire and atlas [J]. Monthly Notices of the Royal Astronomical Society, 2015, 451(2): 1299-1305.

Fan X, Messenger C, Heng I S. A Bayesian approach to multi-messenger astronomy: Identification of gravitational-wave host galaxies [J]. The Astrophysical Journal, 2014, 795(1): 43.

Fernique P, Allen M, Boch T, et al. Hierarchical progressive surveys-multi-resolution healpix data structures for astronomical images, catalogues, and 3-dimensional data cubes [J]. Astronomy & Astrophysics, 2015, 578: A114.

Flewelling H. Pan-starrs data release 2 [J]. AAS, 2018, 231: 436-01.

Fruchter A S. A new method for band-limited imaging with undersampled detectors [J]. Publications of the Astronomical Society of the Pacific, 2011, 123(902): 497.

Fruchter A, Hook R. Drizzle: a method for the linear reconstruction of undersampled images [J]. Publications of the Astronomical Society of the Pacific, 2002, 114(792): 144.

Fruscione A, McDowell J C, Allen G E, et al. Ciao: Chandra's data analysis system [C]//Observatory Operations: Strategies, Processes, and Systems: volume 6270. International Society for Optics and Photonics, 2006: 62701V.

Fu S, He L, Huang C, et al. Performance optimization for managing massive numbers of small files in distributed file systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 26 (12): 3433-3448.

Fukugita M, Shimasaku K, Ichikawa T, et al. The sloan digital sky survey photometric system [R]. SCAN-9601313, 1996.

Gehrels N, Cannizzo J K, Kanner J, et al. Galaxy strategy for LIGO-Virgo gravitational wave counterpart searches [J]. The Astrophysical Journal, 2016, 820(2): 136.

Georgakakis A, Salvato M, Liu Z, et al. X-ray constraints on the fraction of obscured active galactic nuclei at high accretion luminosities [J]. Monthly Notices of the Royal Astronomical Society, 2017, 469(3): 3232-3251.

Gorski K M, Hivon E, Banday A J, et al. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere [J]. The Astrophysical Journal, 2005, 622(2): 759.

Gray J, Nieto-Santisteban M A, Szalay A S. The zones algorithm for finding points-near-a-point or cross-matching spatial datasets [J]. arXiv preprint cs/0701171, 2007.

Gwyn S D. Megapipe: The megacam image stacking pipeline at the canadian astronomical data centre [J]. Publications of the Astronomical Society of the Pacific, 2008, 120(864): 212.

He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

HEASARC N. Heasoft: Unified release of ftools and xanadu [J]. Astrophysics Source Code Library ascl, 2014, 1408.

Ilbert O, Salvato M, Le Floc'h E, et al. Galaxy stellar mass assembly between 0.2< z< 2 from the s-cosmos survey [J]. The Astrophysical Journal, 2010, 709(2): 644.

Jacob J C, Brunner R J, Curkendall D W, et al. yoursky: rapid desktop access to custom astronomical image mosaics [C]//Virtual Observatories: volume 4846. International Society for Optics and Photonics, 2002: 53-64.

Jacob J C, Collier J B, Craymer L G, et al. yourskyg: Large-scale astronomical image mosaicking on the information power grid [J]. Scalable Computing: Practice and Experience, 2006, 7(1).

Jansen F, Lumb D, Altieri B, et al. Xmm-newton observatory-i. the spacecraft and operations [J]. Astronomy & Astrophysics, 2001, 365(1): L1-L6.

Jarrett T, Chester T, Cutri R, et al. 2MASS extended source catalog: overview and algorithms [J]. The Astronomical Journal, 2000, 119(5): 2498.

Jia X, Luo Q. Multi-assignment single joins for parallel cross-match of astronomic catalogs on heterogeneous clusters [C]//Proceedings of the 28th International Conference on Scientific and Statistical Database Management. 2016: 1-12.

Jia X, Luo Q, Fan D. Cross-matching large astronomical catalogs on heterogeneous clusters [C]// 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 2015: 617-624.

Kalpakis K, Riggs M, Pasad M, et al. A system for low-cost access to very large catalogs [C]// Astronomical Data Analysis Software and Systems X: volume 238. 2001: 133.

Klimenko S, Vedovato G, Drago M, et al. Localization of gravitational wave sources with networks of advanced detectors [J]. Physical Review D, 2011, 83(10): 102001.

Koposov S, Bartunov O. Q3c, quad tree cube–the new sky-indexing concept for huge astronomical catalogues and its realization for main astronomical queries (cone search and xmatch) in open source database postgresql [C]//Astronomical Data Analysis Software and Systems XV: volume 351. 2006: 735.

Kozlowski S. VizieR Online Data Catalog: Physical parameters of~ 300000 SDSS-DR12 QSOs (Kozlowski, 2017) [J]. VizieR Online Data Catalog, 2017, 222.

Kunszt P Z, Szalay A S, Thakar A R. The hierarchical triangular mesh [M]//Mining the sky. Springer, 2001: 631-637.

Laity A C, Anagnostou N, Berriman G B, et al. Montage: an astronomical image mosaic service for the nvo [J]. 2005.

Landais G, Ochsenbein F, Simon A. Tapvizier: A new way to access the vizier database [J]. ASPC, 2013, 475: 227.

Lindegren L, Hernández J, Bombrun A, et al. Gaia data release 2-the astrometric solution [J]. Astronomy & Astrophysics, 2018, 616: A2.

Lowe G. Sift-the scale invariant feature transform [J]. Int. J, 2004, 2: 91-110.

Luo B, Brandt W, Xue Y, et al. The chandra deep field-south survey: 7 ms source catalogs [J]. The Astrophysical Journal Supplement Series, 2016, 228(1): 2.

Ma X, Du Z, Sun Y, et al. E-zone: A faster neighbor point query algorithm for matching spacial objects [C]//International Conference on Computational Science. Springer, 2018: 473-479.

Makovoz D, Khan I. Mosaicking with mopex [J]. 2005.

Marchesi S, Civano F, Elvis M, et al. The chandra cosmos legacy survey: optical/ir identifications [J]. The Astrophysical Journal, 2016, 817(1): 34.

McCracken H, Peacock J, Guzzo L, et al. The angular correlations of galaxies in the cosmos field [J]. The Astrophysical Journal Supplement Series, 2007, 172(1): 314.

McCracken H, Capak P, Salvato M, et al. The cosmos-wircam near-infrared imaging survey. i. bzk-selected passive and star-forming galaxy candidates at z$\gtrsim$ 1.4 [J]. The Astrophysical Journal, 2009, 708(1): 202.

McGlynn T, Scollick K, White N. Skyview: The multi-wavelength sky on the interne [C]// Symposium-International Astronomical Union: volume 179. Cambridge University Press, 1998: 465-466.

McMullin J P, Waters B, Schiebel D, et al. Casa architecture and applications [C]//Astronomical data analysis software and systems XVI: volume 376. 2007: 127.

Mi C, Chen Q, Liu T. An efficient cross-match implementation based on directed join algorithm in mapreduce [C]//2011 Fourth IEEE International Conference on Utility and Cloud Computing. IEEE, 2011: 41-48.

Motch C, Carrera F, Genova F, et al. The arches project [J]. arXiv preprint arXiv:1609.00809, 2016.

Nieto-Santisteban M A, Thakar A R, Szalay A S, et al. Large-scale query and xmatch, entering the parallel zone [J]. arXiv preprint cs/0701167, 2007.

Ochsenbein F, Bauer P, Marcout J. The vizier database of astronomical catalogues [J]. Astronomy and Astrophysics Supplement Series, 2000, 143(1): 23-32.

Oh K, Koss M, Markwardt C B, et al. The 105-month swift-bat all-sky hard x-ray survey [J]. The Astrophysical Journal Supplement Series, 2018, 235(1): 4.

Pan H P, Lin C Y, Cao Z, et al. Accuracy of source localization for eccentric inspiraling binary mergers using a ground-based detector network [J]. Physical Review D, 2019, 100(12): 124003.

Pineau F X, Boch T, Derriere S. Efficient and scalable cross-matching of (very) large catalogs [C]// Astronomical Data Analysis Software and Systems XX: volume 442. 2011: 85.

Pineau F X, Derriere S, Motch C, et al. Probabilistic multi-catalogue positional cross-match [J]. Astronomy & Astrophysics, 2017, 597: A89.

PostgreSQL B. Postgresql [Z].

Prugniel P. The Hyperleda Catalogue [Z]. 2005.

Riccio G, Brescia M, Cavuoti S, et al. C 3, a command-line catalog cross-match tool for large astrophysical catalogs [J]. Publications of the Astronomical Society of the Pacific, 2016, 129 (972): 024005.

Roseboom I G, Oliver S, Parkinson D, et al. A new approach to multiwavelength associations of astronomical sources [J]. Monthly Notices of the Royal Astronomical Society, 2009, 400(2): 1062-1074.

Rosenfield P. TOAST Projection [J/OL]. WorldWide Telescope Projection Reference. https://worldwidetelescope.gitbook.io/projection-reference/toastprojection.

Rosenfield P, Fay J, Gilchrist R K, et al. AAS WorldWide telescope: a seamless, cross-platform data visualization engine for astronomy research, education, and democratizing data [J]. The Astrophysical Journal Supplement Series, 2018, 236(1): 22.

Salvato M, Buchner J, Budavári T, et al. Finding counterparts for all-sky x-ray surveys with nway: a bayesian algorithm for cross-matching multiple catalogues [J]. Monthly Notices of the Royal Astronomical Society, 2018, 473(4): 4937-4955.

Schutz B F. Determining the hubble constant from gravitational wave observations [J]. Nature, 1986, 323(6086): 310-311.

Schutz B F. Networks of gravitational wave detectors and three figures of merit [J]. Classical and Quantum Gravity, 2011, 28(12): 125023.

Scoville N, Abraham R, Aussel H, et al. Cosmos: Hubble space telescope observations [J]. The Astrophysical Journal Supplement Series, 2007, 172(1): 38.

Shvachko K, Kuang H, Radia S, et al. The hadoop distributed file system [C]//2010 IEEE 26th symposium on mass storage systems and technologies (MSST). Ieee, 2010: 1-10.

Soumagnac M T, Ofek E O. catshtm: A tool for fast accessing and cross-matching large astronomical catalogs [J]. Publications of the Astronomical Society of the Pacific, 2018, 130(989): 075002.

Sutherland W, Saunders W. On the likelihood ratio for source identification [J]. Monthly Notices of the Royal Astronomical Society, 1992, 259(3): 413-420.

Szalay A S, Gray J, Thakar A R, et al. The sdss skyserver: public access to the sloan digital sky server data [C]//Proceedings of the 2002 ACM SIGMOD international conference on Management of data. 2002: 570-581.

Szalay A S, Gray J, Fekete G, et al. Indexing the sphere with the hierarchical triangular mesh [J]. arXiv preprint cs/0701164, 2007.

Takeda H, Farsiu S, Christou J, et al. Super-drizzle: Applications of adaptive kernel regression in astronomical imaging [R]. CALIFORNIA UNIV SANTA CRUZ ELECTRICAL ENGINEERING DEPT, 2006.

Takeda H, Farsiu S, Milanfar P. Kernel regression for image processing and reconstruction [J]. IEEE Transactions on image processing, 2007, 16(2): 349-366.

Taylor M. Topcat: tool for operations on catalogues and tables [J]. Astrophysics Source Code Library, 2011, 1(S 01010).

Taylor M B. Stilts-a package for command-line processing of tabular data [C]//Astronomical Data Analysis Software and Systems XV: volume 351. 2006: 666.

Taylor M, Boch T, Taylor J. SAMP, the Simple Application Messaging Protocol: Letting applications talk to each other [J]. Astronomy and Computing, 2015, 11: 81-90.

Tody D. The iraf data reduction and analysis system [C]//Instrumentation in astronomy VI: volume 627. International Society for Optics and Photonics, 1986: 733-748.

Vaccarella A, Preston T, Czezowski A, et al. Implementation of the software systems for the skymapper automated survey telescope [C]//Advanced Software and Control for Astronomy II: volume 7019. International Society for Optics and Photonics, 2008: 70192R.

Wang L, Li G L. How to co-add images? i. a new iterative method for image reconstruction of dithered observations [J]. Research in Astronomy and Astrophysics, 2017, 17(10): 100.

Wang S, Zhao Y, Luo Q, et al. Accelerating in-memory cross match of astronomical catalogs [C]// 2013 IEEE 9th International Conference on e-Science. IEEE, 2013: 326-333.

Weisskopf M C, Tananbaum H D, Van Speybroeck L P, et al. Chandra x-ray observatory (cxo): overview [C]//X-Ray Optics, Instruments, and Missions III: volume 4012. International Society for Optics and Photonics, 2000: 2-16.

Wen L, Chen Y. Geometrical expression for the angular resolution of a network of gravitational-wave detectors [J]. Physical Review D, 2010, 81(8): 082001.

Wenger M, Ochsenbein F, Egret D, et al. The simbad astronomical database-the cds reference database for astronomical objects [J]. Astronomy and Astrophysics Supplement Series, 2000, 143(1): 9-22.

White D J, Daw E, Dhillon V. A list of galaxies for gravitational wave searches [J]. Classical and Quantum Gravity, 2011, 28(8): 085016.

Xu Y, Cui C, Fan D, et al. IVOA HiPS implementation in the framework of WorldWide Telescope [J]. Astronomy and Computing, 2020: 100380.

Zhao Q, Sun J, Yu C, et al. A paralleled large-scale astronomical cross-matching function [C]//

International Conference on Algorithms and Architectures for Parallel Processing. Springer, 2009: 604-614.

Zhao W, Van Den Broeck C, Baskaran D, et al. Determination of dark energy by the einstein telescope: Comparing with cmb, bao, and snia observations [J]. Physical Review D, 2011, 83(2): 023005.

Zhao W, Wen L. Localization accuracy of compact binary coalescences detected by the third-generation gravitational-wave detectors and implication for cosmology [J]. Physical Review D, 2018, 97(6): 064031.

Zhu Z H, Fujimoto M K, Tatsumi D. Determining the cosmic equation of state using future gravitational wave detectors [J]. Astronomy & Astrophysics, 2001, 372(2): 377-380.

致　谢

时光如梭，转眼间已加入中国虚拟天文台团队近五年，回想起这五年期间的点点滴滴，不禁感慨万千。在此毕业论文完成之际，谨向所有给予我指导和帮助的人们致以衷心的感谢。

首先，衷心感谢我的导师崔辰州研究员，感谢他给予我的关心、指导和帮助。崔老师治学严谨，视野广阔，为人随和而又不失原则，细致而又不失大局，是青年人的好榜样。他为我提供了优秀的学习和科研条件，并提供了很多交流拓展的机会；鼓励我开拓创新、勤于思考，又不时在方向上给予指导。从论文的选题、课题研究再到论文的撰写，崔老师均给予我巨大的帮助。更重要的是，崔老师作为中国虚拟天文台团队的首席科学家，为团队打造了健康、团结、上进的氛围。这种氛围对于团队和个人的成长都至关重要，我非常荣幸能加入这样的团队，也愿意在这个团队中发光发热，共同发展。

感谢我的另外一位导师樊东卫副研究员。樊博士极其认真负责，技术精湛，且为人热心和善，在工作上给予我很多帮助。在科研和学习上，樊博士都给予了我精心的指导，并总能在我看不清方向时给予指点，让我少走了很多弯路。作为中国虚拟天文台团队的研发负责人，樊博士担负着繁重的研发、管理和沟通工作，但他仍投入了很大精力来指导我的论文。在此，我要对樊博士表示最真挚的谢意。

感谢爱因斯坦探针（EP）卫星首席科学家袁为民研究员。EP 卫星项目是具有开创性技术革新的科学工程，非常荣幸能够参与到 EP 科学应用系统的研发中。在项目的工作中，袁老师给予了我许多关怀和指导，我受益良多。在此衷心祝愿 EP 团队在袁老师的带领下取得更加丰硕的成果。

感谢 EP 团队刘元研究员、金驰川研究员。在参与 EP 项目的工作中，他们给予了我非常多具体的指导和帮助，每次和他们的沟通和讨论都使我受益匪浅。希望在不远的未来，能够和他们一起将 EP 科学应用系统打造成精品。

感谢国家天文台空间部徐栋研究员。在引力波电磁对应体观测的工作中，徐老师给予了非常多准确和实用的建议，并频频迸发出新的思路和想法，促使我对工作不断改进。

感谢哈佛大学史密松天体物理中心 Peter Williams 博士和微软研究院 Jonathan Fay，他们在 China-VO 版万维望远镜软件的研发工作中给予了我非常多建议和支持，并推动了将 China-VO 版万维望远镜中 HiPS 标准数据集的支持功能集成至美国天文学会版本中，使我们的工作成果能惠及全球的万维望远镜用户。

感谢国家天文台赵永恒研究员、黄茂海研究员、张彦霞研究员、韩金林研究员、姜晓军研究员、刘超研究员、李楠副研究员、中国科学院高能物理研究所宋黎明研究员、武汉大学范锡龙教授、华南师范大学李乡儒教授、云南天文台季凯帆研究员、广州大学王锋教授。他们对我的研究工作给予了许多帮助，提出了很好的建议。与他们的讨论经常使我深受启发。

感谢中国虚拟天文台团队的同事李长华、李珊珊、米琳莹、于策、肖健、韩军、陶一寒、杨丝丝、李正、韩叙、何勃亮、杨涵溪。在团队的工作中，大家相互合作、探讨学习，营造了良好的工作环境；在日常生活中，大家志同道合、相互关爱，建立了深厚的友谊；在我的科研工作中，经常向大家请教各自研究方向相关的内容，大家都给予了热心的帮助，并提供了很多素材。与他们的相处令我倍感温暖。

感谢国家天文台人教处杜红荣老师、艾华老师、马怀宇老师、李响老师对我在博士学习期间的关心和支持，我今天的成果离不开各位的亲切关怀和无私帮助。

感谢国家天文台机关许多帮助过我的领导和同事，他们帮我解决了很多工作上的难题，给予我巨大的支持。他们是薛艳杰处长、梁艳春处长、李伟主任、王凤飞、张东坡、谌悦、陈晓艳、云小珊、李会贤、吕品、王博、赵佳、于东升。

感谢国家天文台共同战斗过的同事和同学们，他们是潘海武、朱子佩、曹子皇、胡海波、张墨、包聪颖、刘柱、余邦耀、罗锋、张耀、刘丽佳、邱鹏、崔顺、王川中、宋文明、张震、张磊。

特别感谢我的父母、爱人。你们的付出让我能够安心学业，无后顾之忧。感谢我的孩子，每天回到家，看到你的笑脸我都会疲惫尽消。是你们让这一切更加值得，你们的笑颜是我终身奋斗的意义。

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

许允飞，男，河南省信阳人。

2005 年 9 月至 2009 年 7 月，北京信息科技大学，本科，专业：计算机科学与技术

2009 年 7 月至 2012 年 1 月，北京航空航天大学，硕士，专业：计算机科学与技术

2012 年 4 月至 2016 年 4 月，中国科学院遥感与数字地球研究所，研究实习员、助理研究员

2016 年 5 月至今，中国科学院国家天文台，助理研究员

2017 年 9 月至今，中国科学院大学，博士，专业：天文技术与方法

已发表 (或正式接受) 的学术论文:

1. **Xu, Y.**, Xu, D., Cui, C., et al. (2020). GWOPS: A VO-technology Driven Tool to Search for the Electromagnetic Counterpart of Gravitational Wave Event. Publications of the Astronomical Society of the Pacific, 132(1016), 104501.
2. **Xu, Y.**, Cui, C., Fan, D., et al. (2020). IVOA HiPS implementation in the framework of WorldWide Telescope. Astronomy and Computing, 100380.
3. **许允飞**, 樊东卫, 崔辰州, 何勃亮, 李长华, 于策, 肖健, 李珊珊, 米琳莹, 韩军, 陶一寒 (2020), 中国虚拟天文台的核心功能需求调查分析. 天文研究与技术, 17(1), pp.111-120.
4. 崔顺, **许允飞**, 苏丽颖, 崔辰州, 樊东卫, 韩军, 王川中, 张磊，张洁, (2019). 基于卷积神经网络的全天空地基云图分类研究. 天文研究与技术, (2), p.12.
5. 陶一寒, 崔辰州, 张彦霞, **许允飞**, 樊东卫, 韩叙, ... & 李珊珊. (2020). 深度学习在天文学中的应用与改进. 天文学进展, 38(2), 168-188.
6. 张磊, 樊东卫, 崔辰州, 何勃亮, **许允飞**, 崔顺, & 王川中. (2019). 海量巡天数据在线可视化技术综述. 天文学进展, (2019 年 02), 158-177.
7. 樊东卫, 何勃亮, 李长华, 韩军, **许允飞**, 崔辰州. (2019). 球面距离计算方法

及精度比较. 天文研究与技术, 16(1):69-76.

8. Han, J., Wang, C., Fan, D., Cui, C., Li, S., Mi, L., Li, Z., **Xu, Y.**, He, B., Li, C. and Yang, S., (2018). Amateur public observatory I: The observatory and hardware integration system. Astronomy and computing, 25, pp.89-93.

申请或已获得的专利:

发明专利名称：获取全天相机的自动曝光参数的方法，状态：已公开，发明人：**许允飞**; 崔辰州; 樊东卫; 韩军; 崔顺; 李长华; 李正，申请人：中国科学院国家天文台

参加的研究项目及获奖情况:

1. 国家自然科学基金，天文联合基金项目，U1931132，面向大视场时域巡天观测的大数据检索与融合方法研究，2020.01-2022.12，在研，参与
2. 国家自然科学基金，青年科学基金项目，11803055，基于深度学习等机器学习算法的星系光谱自动分类方法研究，2019.01-2021.12，在研，参与
3. 国家自然科学基金，天文联合基金重点项目，U1731243，面向时域天文学的虚拟天文台核心能力建设与科学应用，2017.01-2021.12，在研，参与
4. 北京市科委项目，Z181100008818076，“FAST 在眼前”互动展品开发，2018.01-2018.12，结题，参与