分类号 _____ 密级 _____

UDC _____ 编号 _____

中国科学院研究生院
博士学位论文

基于虚拟天文台的数据挖掘技术及其在银河系晕结构研究中的应用

刘超

指导教师 赵永恒 研究员 胡景耀 研究员

中国科学院国家天文台

申请学位级别 博士 学科专业名称 天文技术与方法

论文提交日期 2007年12月 论文答辩日期 2008年1月

培养单位 中国科学院国家天文台

学位授予单位 中国科学院研究生院

答辩委员会主席 徐志伟研究员

Typeset by LaTeX 2$_\varepsilon$ at January 17, 2008

With package `CASthesis` v0.1h of CTEX.ORG

Virtual observatory based data mining tools and their application in the study of the structure of the halo of the Galaxy

Chao Liu

Supervisor:

Prof. Yong-heng Zhao, Prof. Jing-yao Hu

National Astronomical Observatories
Chinese Academy of Sciences

January, 2008

Submitted in total fulfilment of the requirements for the degree of Ph.D. in Astronomical techniques and method

摘　　要

技术的不断进步为天文学带来了极大丰富的观测数据，研究手段和方法论正在随着数据的暴涨而悄然变化。信息时代的天文学发展需要和信息科学相结合。虚拟天文台概念的提出顺应了时代的要求，尝试利用计算机网络整合全球的天文资源，让分散在世界各处闲散的数据汇聚到天文学家的指尖。本文详尽描述了如何发展基于虚拟天文台的数据访问和数据挖掘技术、并运用它们完成天文学研究。我们的工作分成三个阶段。首先我们设计并开发了中国虚拟天文台数据访问服务（VO-DAS）来完成对异地异构的海量数据的统一访问。这是一个建立在网格环境中的VO服务，支持多项VO协议，在网格上实现同步和异步数据库访问。然后我们研究和试验了多种可能的天文数据挖掘解决方案：直接采用现在常用的编程工具；完全开发一套网格平台数据挖掘编程环境；还是在现有数据挖掘工具之上进行二次开发，导入VO-DAS的接口并实现和多个虚拟天文台工具的互操作。我们经过评价认为最经济、最快捷、最易实现的是第三套方案。最后，我们把第三套数据挖掘方案应用到银河系晕结构的研究上，从数千万条SDSS DR5的测光数据中，挖掘出了5个银河系矮星系 / 球状星团的候选体。我们还运用这个方案中的数值计算工具对这些候选体的特征进行了估算，对它们可能的类型、同现有的子结构的关系进行了讨论。通过从技术到科学的一系列工作，促生了第一项基于中国虚拟天文台的科学成果，成为了中国虚拟天文台的一个阶段性标志。

关键词： 中科院，学位论文，虚拟天文台，数据挖掘，银河系，银河系结构

Abstract

The progress of technology brings an avalanche of observed data to astronomers. The exponentially increasing data eventually changes the methodology. Astronomy in the age of information needs cooperation with information sciences. Virtual observatory(VO) appeared at the time while the astronomical community were trying to intergrate global astronomical resources so that isolated data can be assembled at the fingertips of astronomers. The goal of this thesis is to develop a means of astronomical data access and data mining based on virtual observatory and push it to practice. Three steps were acheived for the goal. First, a VO based data access service was developed. And an unified interface for query of high volume data on distributed and heterogeneous data resources is available. VO-DAS, which provides asynchronous query as well as synchronous one, is running in a grid environment with multiple VO protocols. Then, a number of solutions of astronomical data mining were studied by investigation and prototyping: to develop a data mining tool with existed programming language, or to invent a new language that calls subroutines in the grid, or to develop plug-in programs on specific data mining tool so that VO-DAS and other VO tools can be interlinked smoothly. The last solution was selected and applied in a study of the structure of the halo of the Milky Way, as the final step of the three. Several ten millions of data of SDSS DR5 were queried and processed. As a result, 5 candidates of satellites of the Galaxy were found. Properties of the 5 candidates were estimated by the numerical tools contained in our data mining solution. And their possible types and relationship to nearby substructure were also discussed. The series of work from technology to science leads China-VO to its first science result, which is a milestone of the development of China-VO.

Keywords: Chinese Academy of Sciences (CAS), Thesis, Virtual Observatory, Milky Way, Galactic Structure

目　　录

表　　格

插　　图

第一章　引言

1.1　现代天文学的新挑战

最近十年以来，天文望远镜和终端设备的技术得到了长足增长。光学望远镜的口径从5米级提高到了10米级，观测的天体数目呈现指数增长。空间天文卫星出现和发展，不仅在光学波段获得了优于地面的成像质量和观测深度，而且在中远红外、紫外、X射线和γ射线等波段的观测上也取得了突破性进展，这使得天文学观测进入了全电磁波观测的时代。大型CCD在天文观测上的广泛使用成倍地提高了望远镜终端设备的信息收集能力。更为重要的是，由于CCD输出的结果本身就是数字化了的，因此可以非常方便在计算机中进行存储、处理和传输，并可以使用数据库管理和发布数据。这激发了越来越多的在线天文数据库的出现。光纤光谱设备的使用使得望远镜一次可以获得一百到数千个天体的光谱数据，观测效率提高了几个数量级。从观测设备到数据收集处理软件都得到发展以后，天文观测数据就如雪崩一样高速生产出来。截至2002年全世界的天文数据库总容量就已经达到了数百TB[1]。最近5年以来数据产出更是以指数增长。从单个设备来看，Hubble空间望远镜（HST）每天的数据产出是5GB，Sloan数字巡天(SDSS)项目从2000年到2005年已经产生了超过16TB的数据，相当于每天平均产生超过20GB[2]。LAMOST正式运行以后，每天也将产生大于30GB的数据。未来的LSST项目每天产生的数据将达到10TB。图1.1显示了层层重叠覆盖的天文观测数据在天空中的分布[3]。颜色越亮的区域被观测的次数也就越多，收集到的数据量也就越大。这张图形象地反映了最近二十年来天文观测获得的巨大的数据集。

天文学的研究对象是天体，近距离的如太阳系有大大小小的天体数千个，中距离的如银河系有恒星$\sim 10^{11}$颗，远距离的在已经探知的宇宙中可能有几千亿个象银河系一样的星系。无论怎么看，天文学研究都需要面对海量的数据。因此，天文学研究对于海量数据有内在的需求。

对于银河系的研究典型地反映了天文学对于海量数据的需求。1785年，William

[1] http://www.astro.caltech.edu/~george/sdt/sdt-final.pdf
[2] http://www.sdss.org/dr6
[3] http://vizier.u-strasbg.fr/

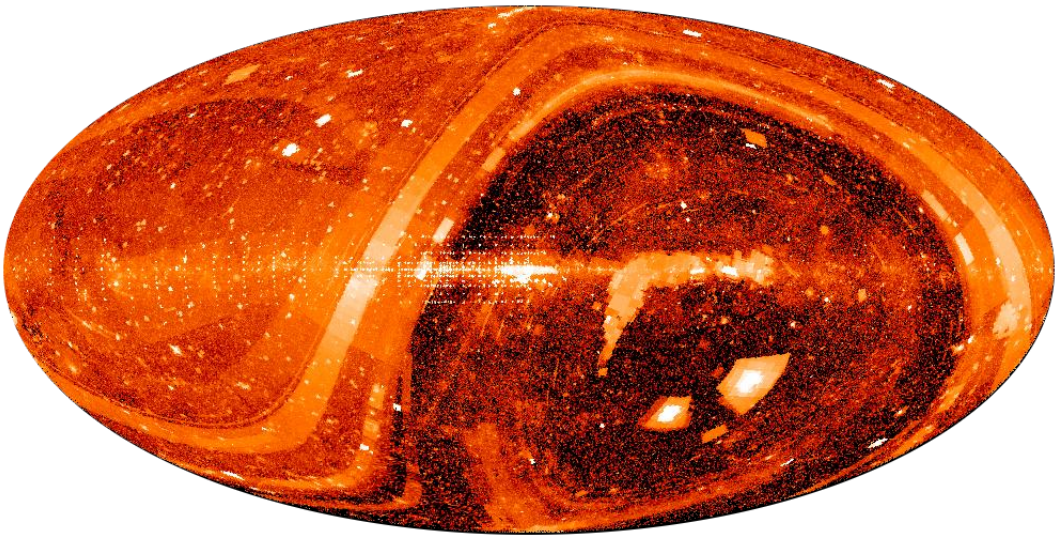

图 1.1: VizieR的天文数据库在天空中的覆盖

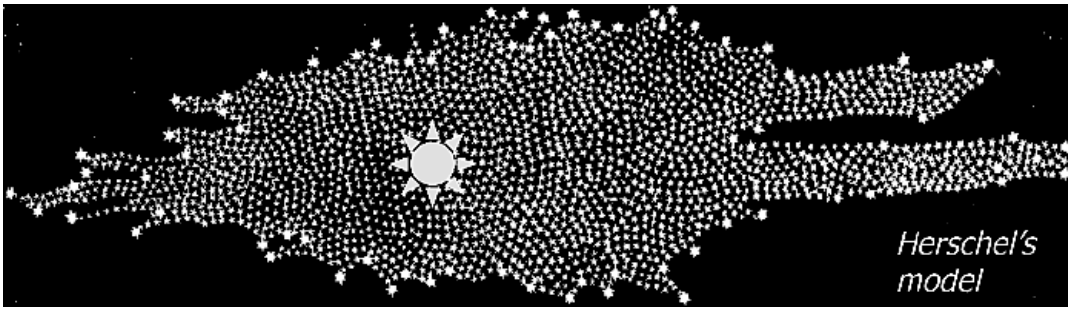

图 1.2: John Herschel描绘的银河系形状

& Caroline Herschel对银河系的683个视线方向进行了首次计数研究，第一次定量地揭示了银河系的形状。John Herschel后来绘制了银河系的形状（见图1.2）。这个形状在现在看来是明显错误的，一方面他们的很多假设是错误的，另一方面，他们所进行的巡天是手工模式的，效率低、数据少。1983年，Gilmore & Reid[1]采用了更多的恒星研究银河系的形状，他们应用约12500颗恒星的观测结果。由于恒星数目有限，他们的结果仅仅描述了银盘和银晕沿着银心距的平均的轮廓特征。这个时期的银河系形状是光滑的，平缓变化的。到2002年，Newberg et al. 使用SDSS的巡天数据对赤道上一条2.5度宽的带状区域进行了计数[2]，距离Gilmore和Reid的工作近20年以后，天文学家终于有了400万颗恒星的数据来刻画银河系的形状，尽管仍然是一条窄窄的区域。这一次，海量的数据让人们对

于银河系的认识又有了新的提升，原来银河系的晕并非是平滑对称的球体，而是充满了子结构的复杂不对称形状（图1.3，这是Newberg et al.(2007)[3]得到的更清晰的结果）。从Heschel的观测到SDSS的观测220年的时间，对于银河系的形状的认识有了质的变化，而刻画银河系形状所使用的数据也增长了好几个数量级。这样的一个历史发展充分说明了天文学研究对于海量数据的强烈需求。到了2003年，2MASS数据释放以后，我们终于获得了银盘的一张完整、清晰的侧面照片（图1.4）。我们看到，只有在观测的数据量大了以后，研究对象的更多细节才可以被揭示出来。

以上的例子说明，天文学研究需要大量的观测数据，无论现在还是将来，只要有更多的数据，天文学家们就会更加深刻地认识宇宙。获取海量数据是天文学研究的内在需求。

天文学的不断深入研究提供了获取海量数据的动力，新技术武装的设备提供了产生海量数据的能力，信息技术的最新发展提供了存储和访问海量数据的能力，天文学因此逐渐进入了新的时代。新时代的标志是伴随着数据雪崩而产生的新的研究手段。

统计学以及从统计学和人工智能等学科发展出来的模式识别等技术大量引入到了天文学研究领域，在星系形态分析[4]、多波段天体证认[5]、星系大尺度结构[6]、银河系结构和演化研究[7]等课题中发挥出越来越大的作用。

天文数据的计算机自动处理也变得越来越重要，以前这些工作可能都是通过人工来完成的，而现在无论对计算机程序抱有多大的成见，你都得接受这样的现实：单靠人工，我们不可能对现有的观测数据作任何大样本的分析和统计。而只有对大样本数据作分析和统计，我们才能减少偶然性对我们的判断的影响。对于星系、宇宙这样一个宏观的研究对象，也只有依据大样本数据才能够对其整体结构和性质做出接近于真实的估计。

新的研究手段又引发了方法论的变革。信息科学的研究方法和逻辑模式被引入到了天文学研究中。例如，一幅被图像处理算法经过滤波处理的图像[8]在天文学家们看来可能是唯象的，缺乏物理依据的，甚至是“虚假的”，不能拿来作为论断的证据。但是图像处理算法不是骗术，它有坚实的数学理论基础作后盾，它相当于设备能力的倍增器，通过它的放大作用，确实揭示了很多用传统天文学方法不能够得到的观测证据。在这种新鲜的，来自其他学科的气氛中，天文学的方法论必然会发生巨大的变化，并带来新的发展。天文学的研究团队中

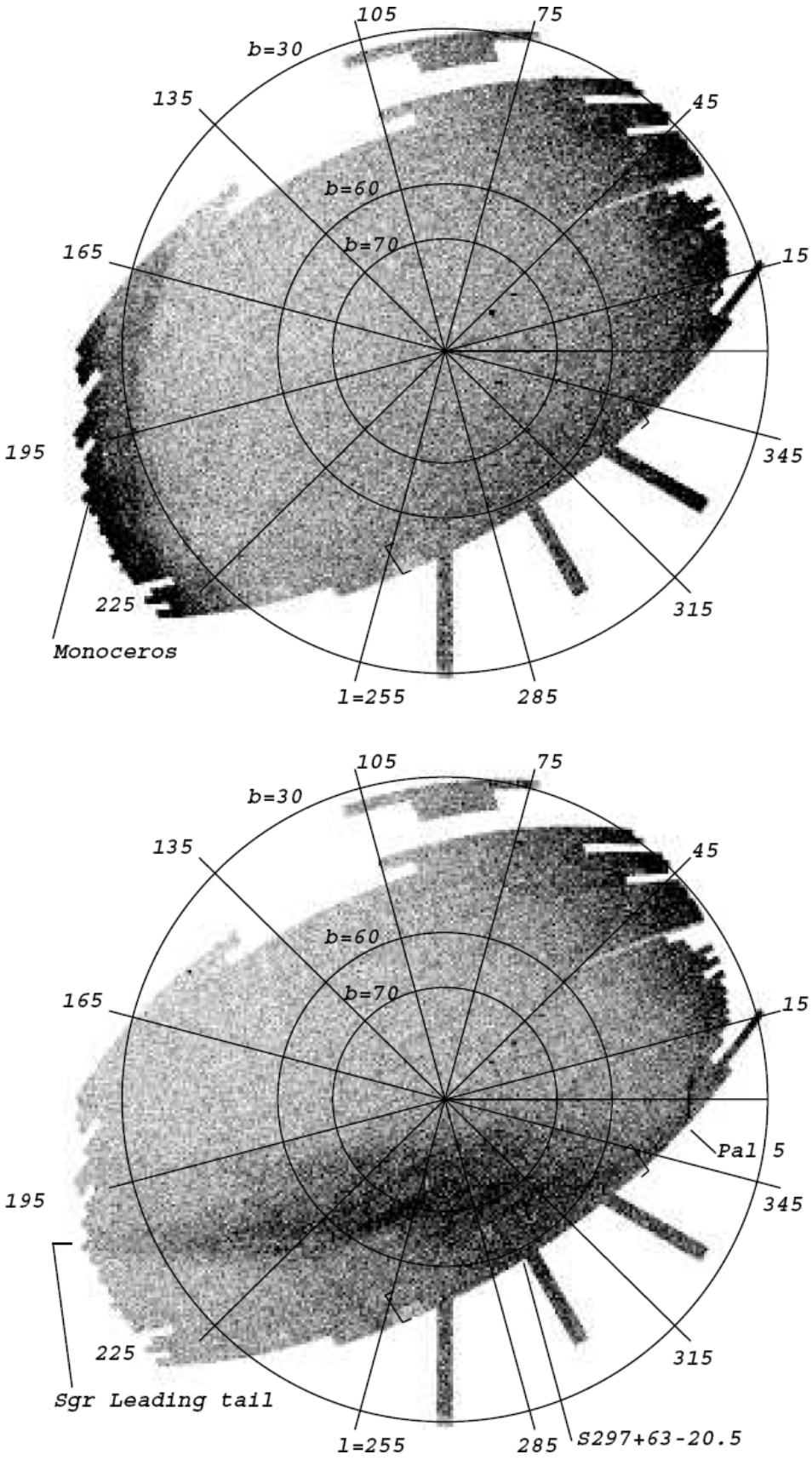

图 1.3: Newberg et al.(2007)得到的银晕的复杂结构

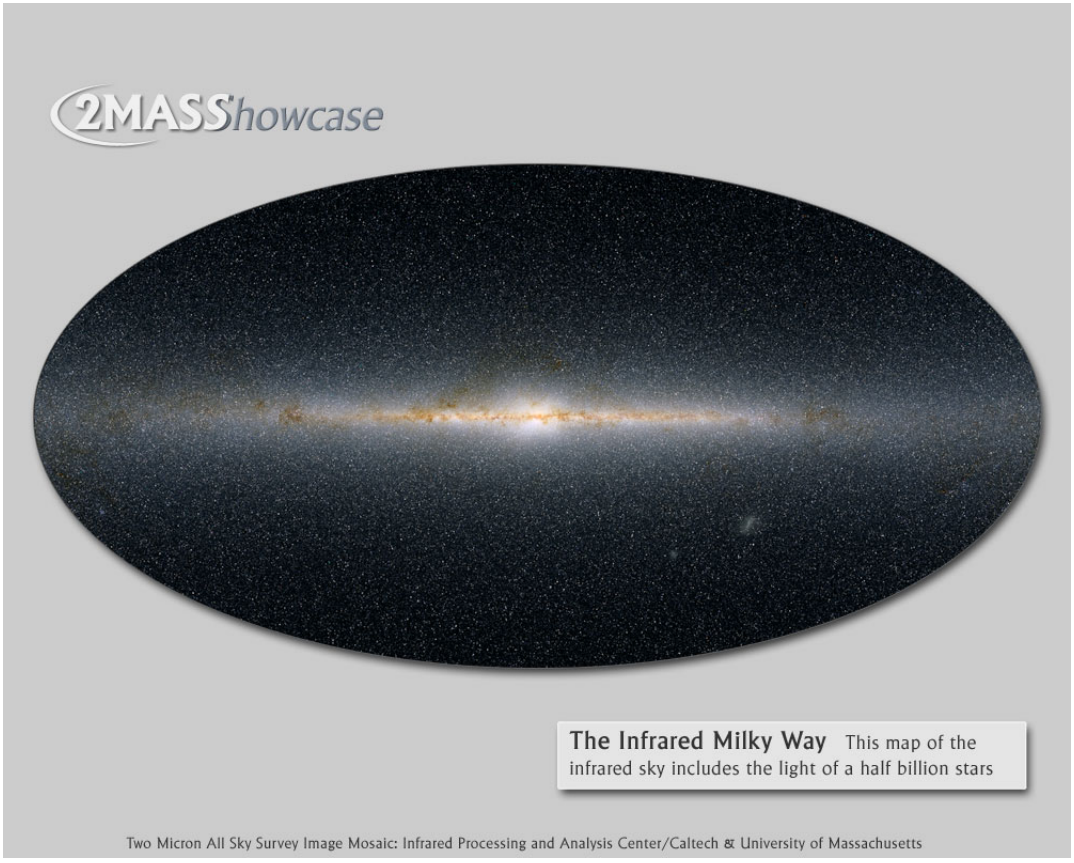

图 1.4: 2MASS观测的银河系照片

不仅有了物理学家、数学家、仪器设计专家，也开始出现越来越多的计算机软件专家、数据处理专家的身影。正如Szalay所言[9]，非天文学元素的引入将会给天文学带来深刻的变革。十年之内，天文学家在他们的每日的研究中使用集成的数据档案、数据挖掘技术就像今天我们使用WWW和email一样普遍。

1.2　虚拟天文台

1.2.1　虚拟天文台的产生和发展

面对数据雪崩，首先要解决的是数据的访问问题。巨大的星表不能再使用一个或几个文件来存放，数据库系统是最好的星表存储和管理工具。但是，不同的星表存放在不同的地方，要无缝地访问这些异地存放的数据，仅仅使用数据库已经不行了。此外，数据量巨大，动辄GB甚至TB计，网络传输的问题变得

至关重要，必须有专门的方法来解决。获得数据以后还需要对数据进行处理和分析。这么多的数据不可能用人工逐个进行处理，因此还需要有自动的数据处理和分析机制。所有这些问题都不能由天文学家独立来解决了，只有引入信息技术，和IT专家一起协作，才能找到解决的途径。虚拟天文台的概念伴随着天文学的信息化困境应运而生了[9]。

虚拟天文台利用最先进的计算机和网络技术将各种天文研究资源以某种统一的服务模式无缝地汇集在系统中[10]。我们将虚拟天文台的主要特征总结为以下三个方面。

首先它表现为对天文服务资源的整合性，即虚拟天文台本身并不产生天文数据、计算和文献资源，它仅仅是将分散在不同地方不同组织中的这些资源整合在一起，以统一的面目出现在使用者的面前。其次虚拟天文台实现了对不同天文数据的联合（federation）。通常的关系型数据库系统的联合查询是通过两个或多个表中的外部键字段实现的。而由于天文数据并不仅仅是表，还包括跨越整个电磁波段的图像数据以及光谱数据。不同组织形式的数据的联合并不是依赖某个键字段，而是通过天体的位置和物理特征来实现的。虚拟天文台需要支持将收集到的天文数据资源按照他们的位置、物理特征等信息进行联合，联合以后得到的一个天体的信息在一定概率情况下是正确的，但也存在一定不正确的概率。第三，虚拟天文台汇集的不仅仅是天文数据资源，还包括天文服务。所谓服务，包括天文计算资源、数据挖掘工具、数据可视化工具、数据存储和发布平台、天文文献等各种类型，五花八门。这么多种类的天文服务要能够在一个统一的环境为了一项共同的研究而相互协同合作，需要它们之间具有良好的互操作性。通俗的讲，互操作性就是指两个来源不同、功能不同的天文应用服务能够互相“理解”对方的数据和操作。

因此虚拟天文台实际上并不是一个简单的软件，它看上去更像是一个操作系统。所不同的是，通常的操作系统是运行在一台计算机上的，而虚拟天文台是运行在整个互联网上的。虚拟天文台的表现形式有很多种，下面介绍几种常见的形式。

虚拟天文台可以表现为一个网站，登录到这个网站以后，天文学家可以访问到各种天文数据资源，包括天体图像、光谱、星表、文献等。法国斯特拉斯堡天文数据中心（CDS）[4] 和Nasa/IPAC ExtragalacticDatabase（NED）[5]属于第一

[4]http://cdsweb.u-strasbg.fr/

[5]http://nedwww.ipac.caltech.edu/

种情况。虽然它们并没有宣称是虚拟天文台，但是从形式上它们已经具备了虚拟天文台的基本特征：对分散在异地数据的整合、多波段数据的联合访问以及交叉证认等。

它还可以提供一个客户端软件，用户借助这个客户端软件，而不是网络浏览器完成天文数据资源的访问、数据挖掘等工作。英国虚拟天文台AstroGrid[6]属于这种情况。它们提供了一个叫做Workbench的客户端软件（图1.5）帮助使用者访问和操作基于网格的各种虚拟天文台服务。使用者不仅仅是简单的访问数据，还可以编辑自己的工作流程，使AstroGrid按照指定的流程自动完成一系列工作。

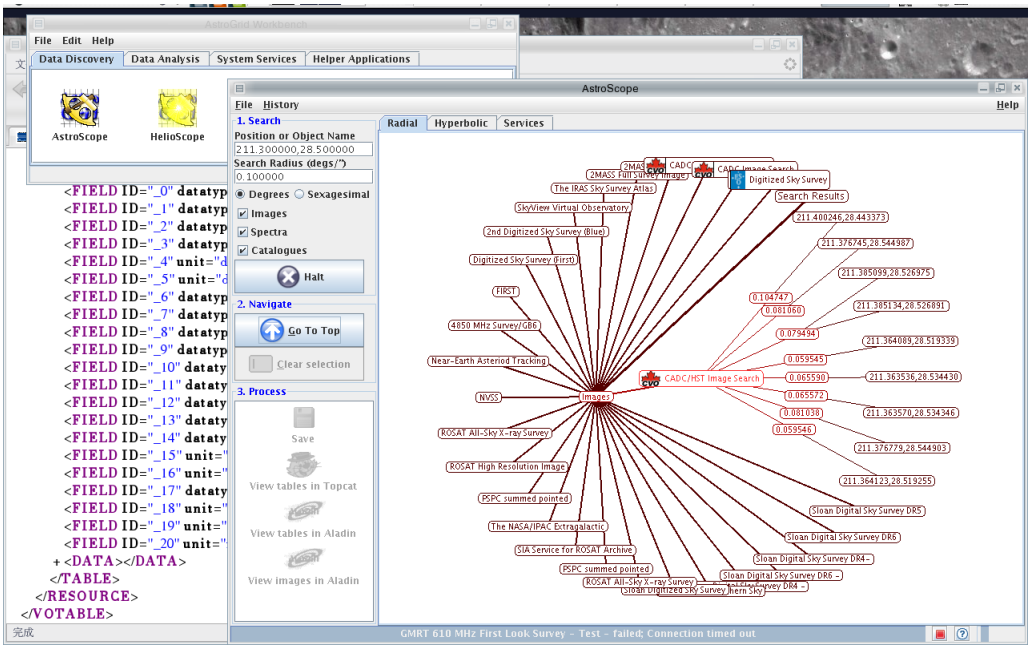

图 1.5: Astrogrid的Workbanch工作画面

它也可以是一组桌面应用程序，通过这些应用程序，天文学家可以完成各种天文数据的访问、交叉证认、统计分析和可视化等工作。Aladin[11, 12]、VOPlot[13]、Topcat[14]等应用软件则属于这种情况。Aladin是由CDS开发的可视化天文数据联合查询工具（图1.6），它可以通过网络获得多个波段的天体图片，并分层显示在用户界面上。这些图片可以按照它们的坐标系统对齐，调整图片的灰度，增加等高线图，用假彩色进行叠加，对图片上的流量进行简单

[6]http://www.astrogrid.org/

分析等。同时，Aladin还可以访问CDS获得数千个星表数据。它可以将星表数据叠加在图片之上，使得用户可以直接将图像上面的天体和星表记录对应起来得到更加详细的信息，如位置、星等、自行等。用户还可以把自己的FITS图像文件和TXT格式的星表文件（如自己观测和分析的测光数据）导入Aladin，用来和网络数据库中的信息进行比较和验证，或者检验自己的星表数据是否正确合理。Aladin因为其功能实用、操作简单、后台数据库丰富等优点得到了比较广泛的应用。VOPlot是一个天文星表可视化工具。它支持VO标准的VOTable[15]格式的星表文件，并可以从星表中选择两列绘制散点图。Topcat工具读取多种格式的星表文件，不仅可以显示散点图，而且可以进行交叉证认（图1.7）。

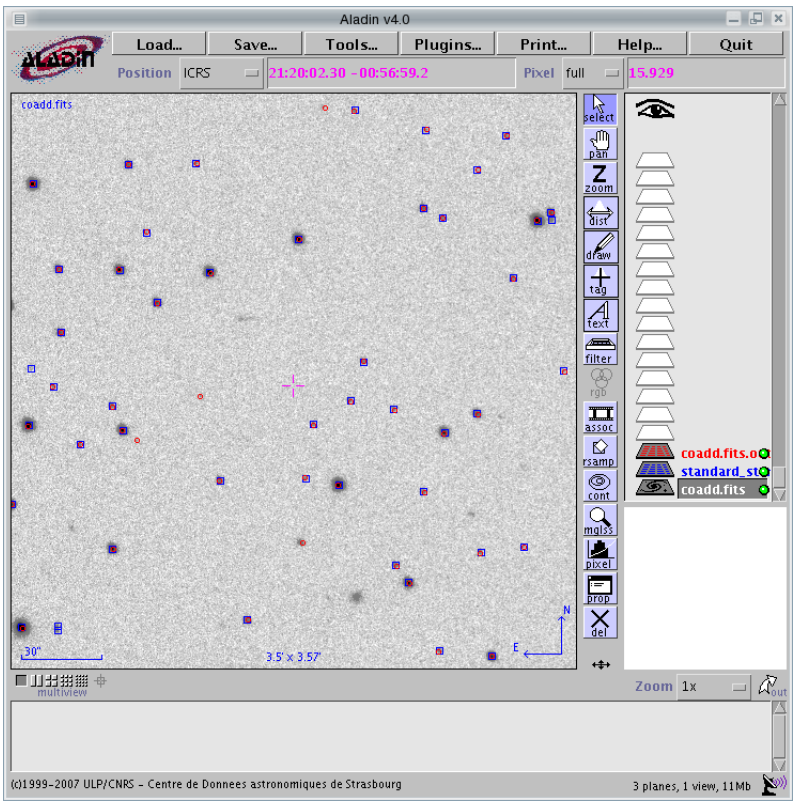

图 1.6: 天文数据联合查询工具Aladin的工作界面

通常，虚拟天文台还会提供一个门户网站，通过这个门户网站，天文学家不仅可以直接访问天文资源，而且还可以找到各种桌面应用程序，下载到本地计算机使用。美国国家虚拟天文台（National Virtual Observatory;NVO）[7]就是一

[7]http://www.us-vo.org/

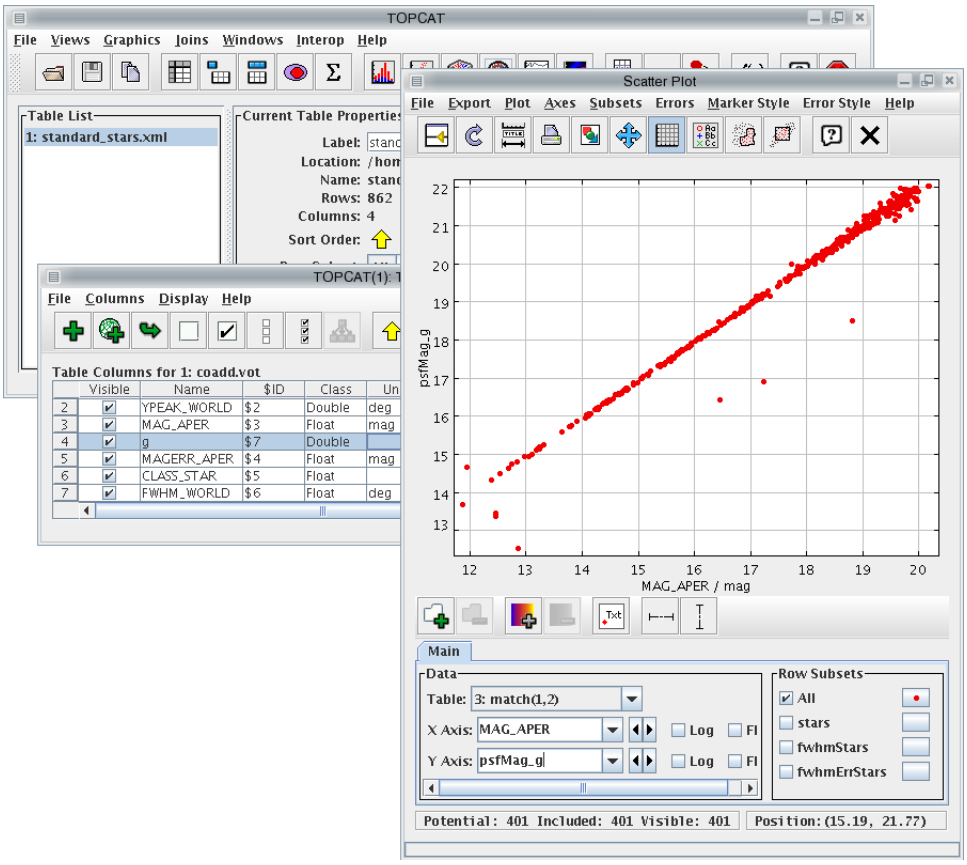

图 1.7: 星表分析工具Topcat的工作界面

个典型的例子。NVO的门户网站看起来很像是Yahoo，访问者不仅可以从那里找到各种VO的新闻，访问和VO有关的社区，了解VO的很多常识，还可以直接查询天文数据资源，搜索各种发布到网络上的天文服务资源，同时它也提供了各种常用VO工具软件和软件库等的下载链接。

虚拟天文台的发展是一种探索性和实验性的。它试图找到天文学和信息科学最有效的融合方式，虚拟天文台的这些多样化的表现形式正是反映了这种探索。没有人能够令人信服的用一种固定模式来描述和开发虚拟天文台，各种尝试都是具有建设性的。但是，虚拟天文台的探索应该有一些最基本的前提：分散数据的集成、数据的联合访问和互操作性。这是虚拟天文台建设的初衷，也是最终目标。因此，建立支持分散数据的集成、数据的联合访问和互操作性的各种协议和规范就是一件必要的事情了。2002年，在德

国Garching召开的“迈向国际虚拟天文台”国际会议[8]上提出了建立国际虚拟天文台联盟（IVOA）的建议。IVOA[9]的重要使命就是推进国际合作与协作，为建设一个能综合利用国际天文数据的、完整的、能协同工作的虚拟天文台，开发、配置必要的工具、系统和组织结构[16]。五年以来，IVOA致力于提高天文信息资源的互操作性的标准和规范，先后确立了数据文件规范VOTable、数据查询语言规范ADQL[17]、虚拟天文台注册(Registry)服务标准[18]、数据模型标准[19, 20]、锥形检索（ConeSearch）[10]，简单图像访问协议SIA[11]、简单光谱访问协议SSA[12]、SkyNode数据访问服务[21]等。这些标准和规范有些已经成为推荐标准，有些还在讨论中。它们为虚拟天文台的开发建立了互操作的基础。理论上，只要各个虚拟天文台项目在这些规范基础上进行他们的开发，完成的各个软件就可以实现某种无缝的链接，可以为了某项科学研究的课题实现协同操作。到2007年，IVOA的成员已经从最初的8家增加到16家。这也表明，几乎所有的虚拟天文台项目都希望参与讨论和制订这些互操作标准，遵循这些规范来开发自己的产品并使之成为未来国际虚拟天文台的一个组成部分。

总体而言，虚拟天文台的研究目前还处于一个比较初级的探索阶段，虽然已经有很多VO工具软件出现，并逐渐被接受和使用，虽然有越来越多的数据能够个很方便地通过互联网访问到并产生了很多科学成果，但是虚拟天文台还没有形成一个产品化的系统服务,很多标准尚在制订过程中。

1.2.2 虚拟天文台的两条发展道路

虚拟天文台的发展正在沿着两个方向进行，一个方向是面向应用，即利用现有成熟技术和标准开发适用于某项天文学研究特点的应用产品；另一个方向是全面制订互操作协议标准，从根本上建立虚拟天文台的规范和框架。两个框架并不矛盾，而是相辅相成地发展的。

基于虚拟天文台的应用集中解决天文学研究中遇到的一些具体的技术瓶颈，试图通过引入虚拟天文台的概念，通过互联网络实现天文数据资源的高效、无缝地访问，同时实现数据分析以及数据可视化。由于虚拟天文台的网络协议尚未健全，因此现有的很多虚拟天文台应用产品仅仅是部分地支持那

[8]http://www.eso.org/gen-fac/meetings/vo2002/
[9]http://www.ivoa.net
[10]http://www.ivoa.net/Documents/latest/ConeSearch.html
[11]http://www.ivoa.net/Documents/latest/SIA.html
[12]http://www.ivoa.net/Documents/latest/SSA.html

些已经比较稳定、成熟的虚拟天文台协议，也主要使用比较成熟、商用化的网络技术。例如，Aladin虽然使用了IVOA的VOTable作为文件格式的标准，使用了IVOA的PLASTIC协议[13]作为与其它应用程序协作的协议，而对于天文网络服务访问，由于没有成熟标准可依，它使用了GLU协议作为服务资源搜索协议（相当于IVOA的Registry服务）和SOAP协议[14]作为网络远程访问协议。随着IVOA的标准不断成熟，新版本的Aladin中也越来越多应用新的IVOA标准。重要的是，不管Aladin用了多少IVOA的协议，是否符合典型的虚拟天文台应用的标准，它确实完成了对多个天文图像、光谱和星表数据库的无缝访问，一定程度上实现了互操作这个目标，也因此得到了广泛地使用。Aladin代表了虚拟天文台研究和开发的一种风格：以科学为目标，以技术为手段，以应用的需求带动标准的确立和成熟，稳步发展，务实而具体。

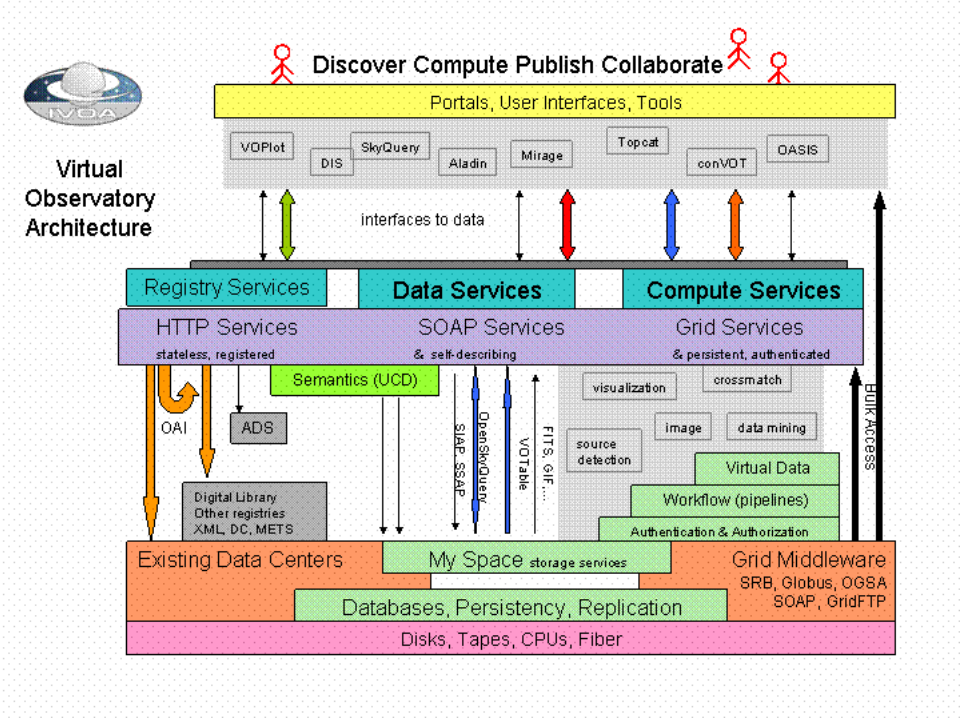

图 1.8: IVOA的VO体系结构

另一个方向是以IVOA组织为代表，致力于标准的制订，规范的确立。IVOA对

[13]http://www.ivoa.net/Documents/latest/PlasticDesktopInterop.html
[14]http://www.w3.org/TR/soap/

虚拟天文台各项标准规范的制订基于图1.8这样的体系结构[15]。IVOA成立了专项工作组，包括应用（Application）、数据访问层（Data Access Layer）、数据建模（Data Modelling）、网格和网络服务（Grid & Web Services）、资源注册（Resource Registry）、语义（Semantics）、标准和过程（Standards & Processes）、VO事件（VO Event）、VO查询语言（VO Query Language）和VOTable。这些工作组的工作完整覆盖了虚拟天文台的体系结构，从数据组织到网络通讯，从查询语言到数据访问模式。如同移动通信领域经历的变革：在一组复杂的面面俱到的协议族的支持下，世界上绝大多数国家的手机不仅可以相互通话，而且可以漫游到对方的国家移动网络中。而做到今天这个成绩，是全世界的移动通信技术厂商以及各个国际通信标准组织二十多年努力的结果。所有的网络协议都采用层次化的结构。在这样的结构中，最底层是物理设备之间的信息交换法则，越向上，协议越接近用户。通过这样的层次化协议（或者称为协议栈），一层一层屏蔽复杂的操作，留给上层越来越简单的对话接口。而且，当硬件或某层网络协议需要升级更新的时候，因为这种层次化的结构而只影响相邻层次的协议，并不影响最上层的用户层或应用层。相比之下，IVOA在其展开工作仅5年的时间里，不可能将虚拟天文台的整个体系结构的关键协议和标准全部建立起来。因此，虚拟天文台的标准和协议的建设仍是一个长期性的工作。

虚拟天文台的发展战略也应该用两个阶段来看待，第一个阶段是近期目标：尽快建立涉及天文学研究领域各个方面的应用产品，并推动这些产品进入研究活动，成为天文学家得力的工作助手。第二个阶段目标是远景目标：建立完善的虚拟天文台体系架构，定义每个环节的协议和标准，确保任何依据虚拟天文台协议族建立的应用软件和数据产品都能够实现互通互联互操作。

不同国家的虚拟天文台依据自身的特点来决定他们的发展路线。英国虚拟天文台依托雄厚的经济实力和技术优势建立一个趋于完备的虚拟天文台系统AstroGrid。这个系统里不仅融合了现有IVOA的很多协议，而且独立制订了很多新的接口协议，例如通用执行架构（Common Execution Architecture; CEA），工作流（Workflow）格式等。它不仅完成了整个虚拟天文台体系结构的各个层次的开发，而且考虑了几种科学应用作为其范例。从数据的访问接口（DataSetAccess），到对已有应用程序进行封装的CEA，从客户端专用程序WorkBench（参见图1.5）到存放用户数据的VOSpace服务，AstroGrid都进行

[15]http://www.ivoa.net/Documents/latest/IVOArch.html

了精心的设计和实现，经过最近一年多两个版本的发展，已经接近实用。可以认为，AstroGrid为虚拟天文台的构架设计确立了一个比较成功的模板，它涵盖了虚拟天文台概念的几乎全部精髓，并使用目前可以采用的比较现实的方案将之实现。反过来，AstroGrid的开发也推动了IVOA各项标准的制订，例如，AstroGrid的VOSpace协议最终成为IVOA的推荐标准。

当然，不可能每个国家的虚拟天文台都有实力象英国虚拟天文台那样将应用的开发和标准的制订有力结合起来。印度虚拟天文台显然走了另外一条道路，在开发能力并不是很强的情况下，他们并没有野心勃勃地展开一个虚拟天文台的体系结构，而是从一个点着手，即星表的可视化。他们的产品VOPlot实现了最简单的VOTable形式的星表文件的二维散点图，直方图的显示（图1.9，摘自Ajit Kembhavi于2007年在印度Pune举办的VO workshop的讲稿）。最新的VOPlot里面允许并列显示多张散点图，并且可以将多张图的数据点关联起来，在一张图中选择数据点以后，相关联的点也会在其它图中被选出来。VOPlot的衍生产品VOMegaPlot实现了海量数据的可视化。这样的工具功能单一，设计简单但是非常实用。在支持了PLASTIC桌面应用连接协议以后，VOPlot可以同Aladin、Topcat等应用产品协同工作：Aladin从网络上查询星表数据，通过PLASTIC协议传递给VOPlot，并由后者显示散点图。虽然整个工作都只是在一台桌面计算机上进行，但是它却体现了虚拟天文台的精髓：互操作性。

综上所述，虚拟天文台的建设并不都是以大而全为目标，而是以实现互操作性为手段，以实现信息化的天文学研究为最终目标。各国虚拟天文台项目只要充分体现各自的特长，发挥自己的优势，就可以在最短时间里设计出满足实际需要的虚拟天文台产品。只要这样的产品体现互操作性的精神，利用现有的可以利用的虚拟天文台标准和成果以及成熟的商用化网络技术，为天文学的科研活动服务，就是一个好的虚拟天文台产品，就会为虚拟天文台的发展作出贡献。

1.2.3　中国虚拟天文台

中国虚拟天文台（China-VO）项目发端于2001年第一次虚拟天文台研讨会。自2004年开始China-VO获得了中国自然科学基金会的资助，开始加速发展。中国虚拟天文台和其他国家的虚拟天文台项目一样有着自己的背景和特点，优势

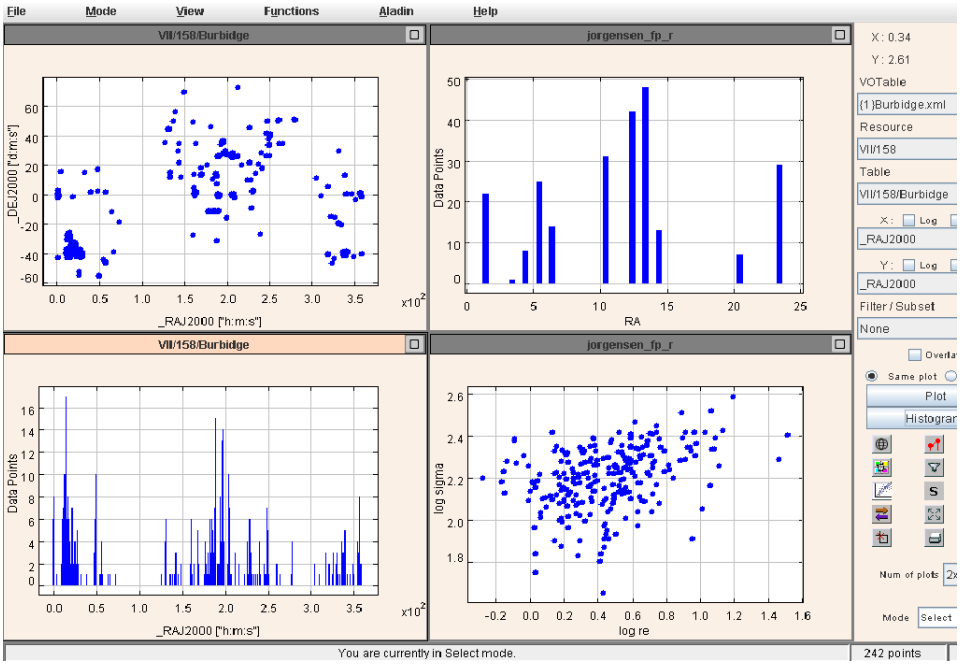

图 1.9: VOPlot的用户界面

和劣势。China-VO最大的特点是：它是伴随LAMOST[22]项目成长起来的，换言之，LAMOST的现实需求刺激了China-VO的创立和发展。此外，China-VO的发展也同时处于中国科学信息化（e-Science）的大潮之中。这样，China-VO就获得了两个外力的带动，一方面它的服务目标比较明确、清晰，那就是紧密联系LAMOST以及相关的科学目标，为中国天文学家建立一个便捷高效的天文资源访问平台；另一方面它的建设基础也将得到e-Science相关项目的支持。

2003年，China-VO的系统设计得到阐述[16]。2004年，崔辰州和赵永恒又提出了China-VO的体系结构[10]。至此，China-VO的概念框架已经形成。在这个概念框架中，China-VO类比于网格的层次结构设计了自己的层次结构（见图1.10）。在这个框架中，China-VO被定义成了四层结构：构造层，资源层，汇集层和用户层。构造层包括了天文数据资源，计算资源，网络资源和存储资源。资源层以开放网格服务框架为基础建立统一的资源访问，计算访问和系统管理功能。汇集层提供具有天文研究色彩的各种VO服务，如数据访问、数据挖掘、统计分析和可视化等。用户层提供客户端程序和VO门户。

China-VO的体系结构是一个庞大的虚拟天文台范例，建立在当时最先进的网格技术之上。但是完全实现这样一个大型系统有很多客观的困难，主要有以

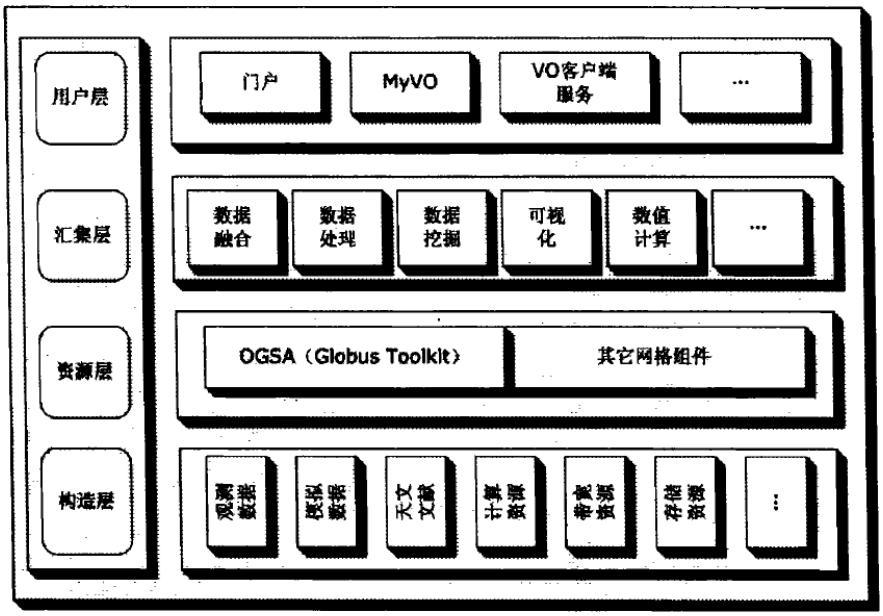

图 1.10: 中国虚拟天文台体系结构图

下几个方面。首先，China-VO的架构得以实现的前提是它所依托的网格技术的成熟和稳定。这一点在3年以前是几乎不可能的。那个时候网格本身的技术尚在不断探索和实验中。不断有前后并不兼容的网格架构涌现出来。即便是一个开发组织，也在不断变换自己的协议规范，寻找最完美的网格解决方案。即便是今天，网格技术仍然没有成熟，难以全面进入商用。建立这样在一个不稳定基础之上的系统，其成功风险是非常巨大的。其次，China-VO的构想非常庞大，牵涉面广，几乎是IVOA虚拟天文台概念架构的翻版。这样，它必然涉及很多并不成熟甚至尚未出现的互操作协议，这就需要中国虚拟天文台的开发者自己独立或半独立完成这些不成熟接口的构思和设计。这是一个巨大的挑战。例如，数据挖掘是一项非常专门的研究方向，将其算法工具化，并无缝地接入虚拟天文台系统的过程中会面临诸如如何完成这么多算法的编程工作，如何设计同这些算法交换数据的参数的接口，如何将天文数据集导入到这些数据挖掘程序中等问题。这些都没有现成的答案可以遵循。以项目管理的观点看，China-VO的架构是一个大型软件项目，需要上百人年的艰巨工作，才有可能完成。而且，由于没有先例可循，中间件技术又不成熟，系统开发者多数对于网格技术并不熟悉，天文背景较少，因而注定如果完全实施这样的一个架构那将是一个极高风险的项目。

因此，中国虚拟天文台项目不能一口气完成一个大系统的研究和开发，而

在这样一个“完美”框架的指导下进行小项目开发才具有切实可行性。在实际实施中，China-VO也确实是在通过不断设计实现小型工具软件和中型应用程序而逐渐向一个天文应用平台靠近。

2005年，China-VO发布了OpenOffice用VOTable转换插件VOFilter[23]。这是一个非常小的转换工具，没有一行执行代码，而是用XSLT实现了VOTable和OpenOffice表格文档格式的相互转换。这个转换工具可以让OpenOffice也能够浏览和编辑VOTable格式的天文星表数据。在VOTable浏览工具还比较缺乏的时候，这样一个小小的插件有助于天文学家从天文数据网站下载了VOTable格式的数据以后迅速浏览和编辑它们，还可以借助OpenOffice将VOTable格式的数据文件转换成自己需要的文本格式，以便于自己编写的程序的读取。

同一年，China-VO发布了类似于Aladin 的VOImpat[24] ，VOImpat 相比Aladin 更加小型化，适合进行快速的星空位置查找和主要星表的联合查询。它支持DSS的图片检索，只要给出天空的坐标就可以快速导入DSS的剪裁图像。同时，为了能够获得天体的更多信息，VOImpat还可以借助SkyPortal（JDL数据挖掘原型中一个数据访问模块，详见第三章有关JDL的介绍）查询USNO、2MASS 、NVSS等不同观测波段的主要巡天星表，将查询到的星表数据和图像数据结合，提供每个天体的综合信息。此外，在要求不高的情况下，VOImpat还可以浏览本地FITS格式的图像文件以及VOTable格式的星表文件，并具有一定的图像处理能力。

2006年，China-VO又发布了SkyMouse，一个桌面天文信息搜索工具（参见图1.11），它可以通过鼠标选取屏幕上的词组，从多个天文信息数据库中查找对应名称的天体数据、图像和相关文献[25]。该工具可以让天文学家在阅读文献的时候快速浏览某个感兴趣的天体的数据信息，简单而实用。它同时还可以成为天文爱好者的小工具，了解更多的天体知识。

2007年又完成了FitHAS和VO-DAS[26]的初步开发。FitHAS是一个批量读取FITS文件头并导入数据库的工具（图1.12）。对于历史观测遗留下来的成千上万的FITS文件，由于以前没有用数据库进行管理，如果需要从中找到有用的内容是非常困难的。FitHAS会自动读取一个文件夹中的所有FITS文件头信息，将它们装载到指定的数据库中。这样，如果需要查找某个特定天体相关的FITS文件，可以直接利用数据库以特定查询条件进行查询，找到相关的FITS文件的相对存储位置。该工具对于整理和管理FITS文件有很大的帮助，从而从数据收集

的角度支持了天文资源共享。VO-DAS则是一个面向异地异构数据库的统一访问服务。本文的第二章将详尽描述VO-DAS的功能、设计与实现细节。

此外，China-VO还在数据访问和数据挖掘服务方面进行了原型研究，包括SkyNode原型的研究[27]和JDL数据挖掘服务原型[28]。后者也将在本文的第三章中做详尽介绍。

在这些VO工具的支持下，China-VO紧密联系实际的天文学研究，把真实的研究课题作为VO应用的科学范例。一年来，在星系测光红移的算法研究、银河系结构研究等方面取得了一定的科研成果。在星系测光红移研究方面，借助China-VO的工具实现了对SDSS DR5约2400万个星系测光红移估计[29, 30, 31, 32]的可视化。在银河系结构的研究方面更是紧密结合虚拟天文台技术，借助多个VO工具发现了五个银河系伴星系/球状星团候选体[33]。关于虚拟天文台环境下银河系结构的研究将是本文论述的重点之一，将在第四章至第五章做详尽描述。

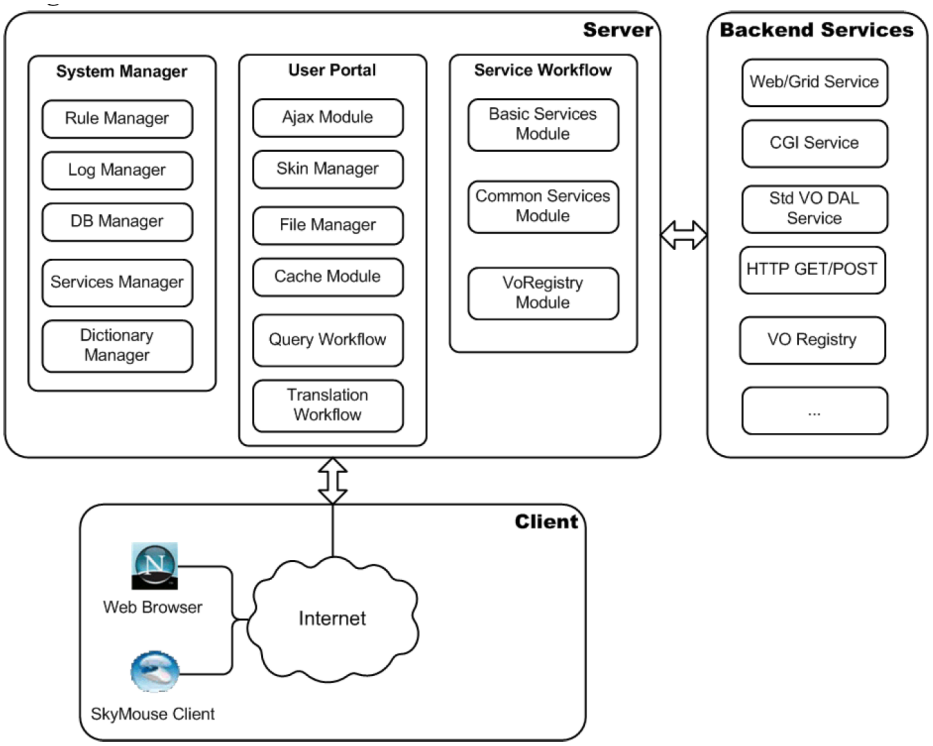

图 1.11: SkyMouse的结构

除了不断开发出独立的VO工具软件满足特定天文研究的具体要求以

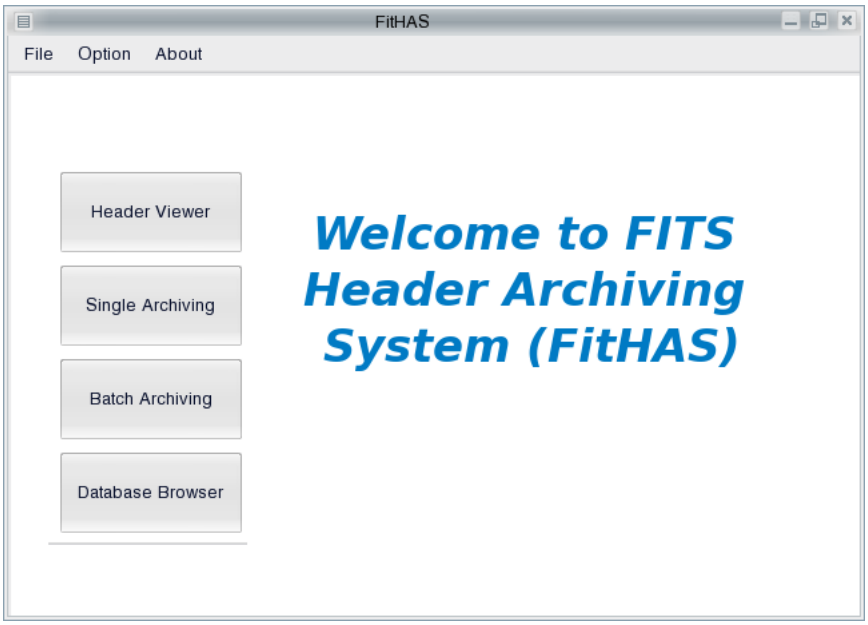

图 1.12: FitHAS的用户界面

外，China-VO还需要满足大科学项目LAMOST的强烈需求。

大天区面积多目标光纤光谱天文望远镜（LAMOST）[22]是一台高效能的光纤光谱巡天望远镜，它被设计成反射式施密特望远镜，有效通光口径4米，焦距20米，具有20平方度的视场。在线尺寸1.75米的焦面上分布有4000根可以局部自由运动的光纤用于对准选定的天体，这些光纤的另一终端连接16台中等分辨率（R=2000）的光谱仪。LAMOST的巡天观测数据将主要用于研究宇宙大尺度结构、银河系的结构与演化以及天体多波段交叉证认。LAMOST在2008年投入运行以后，每个晚上将产生的原始数据有几十GB，而最终得到10^8个天体的光谱数据。无论从LAMOST的科学目标看还是从它的数据产能看，它都和海量数据紧密联系起来了。

如此大规模的观测带来了一系列问题：计划和实施、观测数据的处理、存储、管理和发布等。由于LAMOST望远镜不进行成像观测，只进行光纤光谱巡天观测，因而它必须要有输入星表才能正常工作。研究和产生这份输入星表的工作需要根据特定的科学目标对大量的已经公布的星表数据进行访问、分析，还需要调用庞大的计算资源对观测目标的空间分布、物理特征等情况进行模拟研究以提供更加精确的选源方法。虚拟天文台是输入星表研究的最佳平台。它可以从分布在世界各地的星表数据库中应用VO技术访问所需要的特定

的数据，将不同巡天的不同波段的数据进行交叉证认，选取满足条件的天体产生LAMOST观测需要的输入星表。在LAMOST积累了一定的观测数据并释放出来以后，全世界的天文学家需要以最高效的手段获得这些数据。假如没有虚拟天文台的工具的帮助，使用这样大量的数据开展科学研究将需要较长的准备周期。最后，LAMOST将最终产出TB级的数据，这些数据的存储、备份和传输也需要应用虚拟天文台的技术。因此，中国虚拟天文台将会同LAMOST项目紧密结合在一起，China-VO将成为LAMOST的重要科研平台，而LAMOST会成为VO使能(VO-enabled LAMOST)的大科学项目[34]。LAMOST提供给China-VO的十分具体的功能需求，总结在表1.1中。

表 1.1: LAMOST对China-VO的功能需求

LAMOST需求	China-VO的功能
输入星表	各种常用星表的复杂条件查询，交叉证认功能，输出星表文件
观测光谱文件	按照不同条件（坐标，观测参数）的查询和下载
观测天体星表	复杂条件查询，和其他星表的交叉证认
	用数据挖掘、统计分析等功能对LAMOST发布的数据进行处理和分析
	数据可视化功能，对数据挖掘、统计分析的结果进行显示

此外，最近几年来，以e-Science为代表的科学研究信息化革命风起云涌，许多最先进的计算机网络技术引入了科学领域。中科院的CNGrid 、Vega以及基于网格的数据库和工作流的研发工作正在逐步推进。这些外部条件为China-VO的发展提供了有力的支撑。虚拟天文台本来就是一项跨学科的科研方法革命，仅仅依靠天文学界的力量是很难独自完成的，NVO和AstroGrid都是天文学家和信息技术专家共同合作的结晶。因此，China-VO应该充分利用当前良好的外部条件，借力于国内如火如荼的e-Science来建设自己的基础层面（构造层和资源层），而China-VO团队的主要精力应集中在解决天文学独特的数据需求上来，充分利用现有的成熟网络技术和数据挖掘软件，尽快开发出天文学家能够实际使用的产品，这些产品应该重点突出，简洁实用，不仅满足LAMOST项目数据的需求，而且适合广大中国天文学家使用。

总之，中国虚拟天文台的建设应该以提高天文资源和计算资源的互操作性为核心，以满足国内天文设备建设项目和天文学家的科研为目标，开展短、平、快式的开发，分步骤地逐渐向预定的宏伟的框架设计蓝图靠拢。

1.3 天文数据挖掘

虚拟天文台完成了一个以数据资源为中心的集成的天文学研究环境的蓝图，它将改变天文学家的研究环境和研究习惯，同时给天文学家带来了对海量数据访问的能力。而天文数据挖掘的兴起顺应了海量数据处理的潮流，为天文学研究带来了方法上的革新。

数据挖掘是指自动或半自动地从海量数据中发现模式、相关性、变化、反常规律性、统计上的重要结构和事件[35]。数据挖掘的产生和兴起是同数据库技术的普遍使用和数据的爆炸式增长紧密相关的。它是对60年代以来发展起来的数据库应用的一个提升和发展。使用数据库，人们可以将数据信息通过显式的条件进行过滤，根据关键字的关联实现联系，但是隐藏在数据背后的隐性关联和知识是不能够仅仅通过数据库发现的。数据挖掘技术正是为了发现和挖掘隐藏在海量数据背后的知识和规律应运而生的。数据挖掘是一门边缘学科，它融合了统计学、人工智能、数据库技术等多门学科的前沿成果，在短短十年时间里，已经在经济、信息等领域中得到了长足的发展。

数据挖掘的对象是数据。数据可以是结构化的数据，例如存储在关系型数据库中的以表形式管理的数据。也可以是半结构化的，例如文本、图像、图形。还可以是这些数据类型的组合体。数据可以集中存放在一处，也可以分散存放在互联网上。数据挖掘的技术手段包括数据库、数理统计、人工智能、模式识别等。数据挖掘涉及的算法种类繁多，包括数据库方法、回归分析、判别分析、聚类分析、方差分析、相关性分析，主成分分析、时间序列分析、决策树、人工神经网络、模糊逻辑、统计学习方法、支持向量机、傅立叶分析、小波分析、特征提取、遗传算法等。数据挖掘一般遵循一定的过程，通常这个过程会分为以下几个阶段：数据准备、数据预处理、数据挖掘和解释与评估。由于数据挖掘的工作是以数据为中心的，因此数据准备和预处理两个阶段对于数据挖掘的成败起着至关重要的作用。

数据挖掘的目标是从数据中发现新的未知的知识，带有一定的不确定性。选择什么样的数据集、对数据做怎样的预处理、采用哪种算法处理数据都会对

最终结果有很大影响，经过挖掘以后获得的结果是否具有实际的意义也是需要仔细讨论的。因此，数据挖掘技术并不是一项“傻瓜”技术，希冀不需要人工干预，只要有了算法工具就可以从数据中有所发现的的“天真”想法是要不得的，甚至会花费大量的时间而毫无收获。数据挖掘应用在不同的领域中，需要靠这个领域里深厚的背景知识才能够发挥出它的作用来。

在天文学的研究中，面对越来越多的积累起来的观测数据，自然而然想到将数据挖掘技术引入到天文学研究中来，尝试从数据的海洋中发现新的金矿。这些金矿可能是新的天体、新的现象、特定天体的新的特征，也可能是不同物理量之间的新的关系，不同天体或现象之间新的关联等。

由于天文学数据主要来源于历史上的各次观测，因此数据的原始处理很难做到完全一致，受观测设备、观测天区和观测条件等限制，数据本身的完备性也不是无条件成立的。这为数据挖掘的天文应用带来了一定的困难。一方面，这需要更加细致的数据预处理过程。在预处理过程中，要充分理解所使用的天文数据的来源、观测设备的特征、处理流程的算法和制约条件等等信息。要恰当地选择数据的维度和数据的范围，在选择的时候既要考虑到背后的物理含义也要考虑到数据的完备性。数据预处理的过程中，还要充分考虑观测误差对数据挖掘算法的影响，减少因为误差而产生误判或挖掘出错误知识的概率。另一方面，在算法选择上也要考虑到所处理数据的物理图景，不能仅仅因为计算的方便而违背数据所代表的物理本质。当真实的物理图景确实复杂而需要在进行数据挖掘的时候采用简化的计算时，也要在最后评价和解释数据挖掘的结果的时候有充分的讨论和说明。再有，往往数据挖掘的性能依赖的是统计学的检验，这种检验在物理上并不一定可靠，还需要使用物理的图景进行间接的核对，才能得出最后的结论。

在充分考虑了上述天文学数据分析的特点以后，我们将关注的焦点集中在算法上。并不是每一种算法都是有效的，不同的算法对于处理不同的天文数据效果是不一样的。因此，在对天文数据进行数据挖掘的时候，还要充分考虑算法的局限性。要根据数据的特点和希望发现的知识的特点（尽管在开始我们并不知道要发现的知识到底是什么，但是我们会知道这些知识大致是那一类，是新的天体还是新的现象或者是新的关联等）来选取算法。有的时候并不能确定算法，这时候需要对不同算法进行比较。算法好坏的标准不仅要用统计学的方法定量给出，还需要用天文学和物理的图景给予验证。

Grillmair & Dionatos在2006年[36]利用匹配滤波算法在SDSS几千万颗恒星中发现了一个跨度为63度的星流。我们以此为例，分析在什么条件下数据挖掘才是天文发现的有效方法。他们采用的匹配滤波算法是一种典型的信号处理方法，最早由Rockosi等人借鉴过来，应用于球状星团Pal 5的潮汐尾的发现[8]。匹配滤波算法需要两个模板，一个是球状星团的颜色－星等图，把这个颜色－星等图转换成一个在颜色－星等平面上的概率分布密度函数作为匹配滤波器的信号出现的概率。另一个是场星的颜色－星等图作为匹配滤波中的噪声出现的概率。在天空中一个小的立体角dΩ内，恒星要么属于星流（那么必然和模板星团有一样的颜色－星等图）要么是场星，因此dΩ内的恒星总数为

$$n_{tot}(color, mag) = \alpha f_{cl}(color, mag) + \langle n_{bg} \rangle (color, mag) \tag{1.1}$$

其中，$\mathrm{n_{tot}}(\mathrm{color}, \mathrm{mag})$，$\alpha$，$\mathrm{f_{cl}}(\mathrm{color}, \mathrm{mag})$和$\langle \mathrm{n_{bg}} \rangle (\mathrm{color}, \mathrm{mag})$分别表示立体角dΩ内在颜色-星等平面上任一点的恒星总数密度，星流中的恒星数密度，星流恒星在颜色－星等平面上的概率密度分布函数和场星的平均数密度。匹配滤波方法根据最小二乘方法得到α的估计为

$$\begin{aligned} \alpha = & \Big\{ \sum_{i,j}^{stars} [\frac{f_{cl}(color, mag)}{n_{bg}(color, mag)} \delta_{stars}(color, mag) \\ & - \int f_{cl} d(color, mag)] \Big\} \Big/ \Big\{ \int (f_{cl}^2 / n_{bg}) d(color, mag) \Big\} \end{aligned} \tag{1.2}$$

其中δ_{stars}是delta函数。现在将天空分成小格子，按照公式1.2估计星流在每个格子中的比例，就可以得到一张α在赤经赤纬平面上的密度图（图1.13）。图中，颜色越深的地方表示星流的成员恒星的比例越大。

我们注意到，在上述的例子中，由于我们知道星流可能是球状星团被银河系引力势瓦解而产生的，因而其成员星应该属于一个星族，同时我们还假设星流的各点和我们的距离大致差不多。只有在这两点先验知识的前提下我们才可以使用一个已知的球状星团的颜色－星等图作为匹配滤波器的信号源。另外，场星的颜色－星等图必须是精心选择的区域的，不能选择已知子结构所在的区域，也要考虑到不同天区场星的分布是不同的。不了解这个分布，就会在场星选择上出现疏漏，也就很难让匹配滤波起作用，最终也发现不了这个星流。第三，即便是Grillmair和Dionatos得到了这个匹配滤波的图像，也不能立即确定这

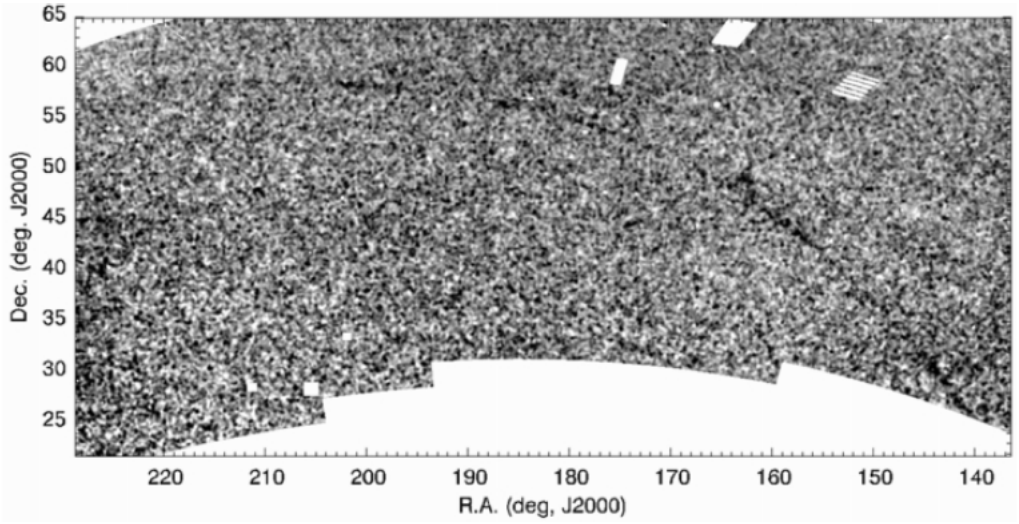

图 1.13: Grillmair和Dionatos在2006年发现的星流

个就是星流，这条长长的痕迹可能是随机涨落，可能是观测的某种选择效应。最终要确定它就是星流，应该还要从它的物理特征进行佐证。例如，观测这个星流上的恒星的光谱获得视向速度与化学丰度信息，并以此加以证认。或者找到这条轨迹上和它具有相同轨道角动量的球状星团主体等。这个例子生动展示了天文数据挖掘的过程和特点。

我们看到，数据挖掘在天文学中的应用并不是仅仅导入其他学科的算法（匹配滤波方法通常用在信号处理领域），而是要将这些算法和天文学的知识充分结合，特别是在数据预处理阶段。本例的滤波模板准备可以认为是数据预处理的一个重要工作。这里需要对研究对象有深刻了解才可以做出适当的预处理。此外，由于数据挖掘方法更多的焦点是放在数据本身上，因此输出的结果还需要根据天文学和物理学的原理进行验证。从统计上讲，这条星流确实是显著不同于背景场星的，但是如果不能从原理上给予验证，仍然不能对这个星流的物理属性、演化特征给出确定性的答案。

天文学家当然可以通过自己写程序完成数据挖掘过程，但是考虑到数据挖掘的算法比较复杂、种类繁多，自己写程序并不是一个高效的做法。因此，提供一个集成的天文数据挖掘工具是一个提高工作效率的想法。天文数据挖掘工具不应该是简单的一个一个算法程序包的集合。为了帮助天文学家方便地获取数据，对数据进行预处理，通过可视化数据而得到数据背后蕴含的物理图景，天

文数据挖掘工具应该是结合了数据访问、数据预处理、多种算法子程序以及数据可视化的一个综合平台。虚拟天文台可以作为天文数据挖掘工具的一个基础，提供对互联网上全部天文数据资源的访问，调用所有可能的数据处理程序，调用计算资源完成数据挖掘，提供多种类型的数据可视化工具。虚拟天文台和数据挖掘结合将是天文学家处理海量数据的有力工具。

1.4 目的、问题和任务

1.4.1 目的

我们把研究目的确定为**通过发展基于虚拟天文台的数据挖掘工具，推动虚拟天文台的应用水平，最终使用这些工具完成实际的天文学研究**。为了达到此目的，我们将采用以下技术路线。我们将自己的研究和开发定位于为实际的天文学研究服务，因此决定了基于虚拟天文台的数据挖掘工具不是要建立一个协议或标准，而是要用最短的时间，花费最少的代价发展出来最有实用价值的应用产品。这样的技术路线也是符合我们在第1.2.3节里论述的China-VO的发展策略的。当然，我们也要注意这样的数据挖掘工具应该是在China-VO的框架结构之下的，同时也应该遵循IVOA已经成熟的各项相关协议和标准。

1.4.2 问题

将虚拟天文台的技术和数据挖掘的技术结合起来应用于天文学研究面临若干技术问题需要解决。首先，如何让天文学家以最简单、直接和便利的方式通过网络获得海量的天文数据。其次，这些数据怎样比较方便地被数据挖掘工具所利用。第三，什么样的数据挖掘平台是适应天文学研究所需要的。具体地讲，我们面临以下技术瓶颈：

1. 数据的分布性：天文数据保存在世界各地的天文数据库中，并没有集中在一起。例如，Sloan数字巡天项目（SDSS）[37]的观测图像、光谱和星表数据保存在Sloan巡天的网站上，2MASS的图像数据和星表数据可以在CDS的数据库中找到。这种分布存放的特征使得对它们的访问变得比较复杂，特别是当需要将不同数据库的数据进行比较或交叉证认的时候，往往花费漫长的数据收集、数据格式调整、数据存放和数据处理的时间。

2. 数据库的异构性：由于天文数据管理没有通用的标准，因此天文数据库无论是元数据还是数据库管理系统都有巨大差异。例如SDSS的数据采用Microsoft的SQL Server进行管理，其元数据采用Schema进行描述，增加了很多和SDSS巡天特点有关的内容。CDS上存放的数据库是一种并不常见的数据库系统，元数据则使用readme文本文件进行描述，要想得到元数据，必须对这些readme文件进行扫描和分析。LAMOST的数据将采用MySQL进行管理，它的元数据则带有鲜明的多光纤光谱观测特点，并使用一个专门的表进行定义。数据的使用者在使用不同来源的数据资源的时候，必须首先学习这些不同系统的操作方法，研究完全不一样的元数据。这种情况将使得虚拟天文台倡导的互操作性荡然无存。

3. 访问海量数据：正如第1.1节中所讨论的，天文学发展到今天产生了数目巨大的观测数据，同时天文学的发展也迫切需要应用这些海量数据完成更加细致、更加精确的发现。首先LAMOST项目及其科学目标就提出了强烈的海量数据访问和数据挖掘的需求（参见表1.1）。这些需求在天文学研究中，特别是大样本天体的研究中还具有普遍性。其次，利用其他大型巡天项目进行研究也需要数据访问和数据挖掘。例如，SDSS观测了超过9000平方度的天空，不仅有5色宽带测光还有中分辨率光谱观测，其中测光天体已经达到2.87×10^8颗，光谱1.27×10^6条。SDSS在宇宙大尺度研究、活动星系核和类星体、星系形成和演化、银河系结构、恒星研究等方面都产生了丰硕的成果。应用SDSS数据进行天文研究必然会面对海量数据，同时很多研究实际上就是在做海量数据的知识发现和数据挖掘的工作，例如从SDSS数据中寻找类星体[38]。但是，海量数据访问的方法尚没有统一解决方案。通常的做法是将数据全部备份到本地服务器上。这种方式对于只使用少数几种巡天数据的情况下问题不突出，但是，当研究需要多个不同的巡天项目的海量数据的时候，就不能采用本地镜像这样高成本方法，而现有的任何虚拟天文台数据访问系统又没有相应的解决方案。举例来说，AstroGrid提供的数据访问功能仅仅能够进行锥形检索；NVO的Open SkyQuery虽然实现了ADQL的兼容，但只能提供最多5000条记录。访问海量数据的能力是完成天文数据挖掘的前提条件，无法绕过。

4. 数据访问和数据挖掘的互操作性：当海量的天文数据能够从网络环境中被查询出来以后，需要和天文数据挖掘工具进行互操作。这需要建立海量数

据和数据挖掘工具之间的传输协议和适当的接口规范。目前，这样的衔接只能依靠研究者自己的努力。

5. 上述技术在天文学研究中的可用性：同科学上对海量数据访问和数据挖掘的迫切需求形成鲜明对比的是目前国际上所有的虚拟天文台实验系统都不能提供科学研究所需要的相应网络数据服务。很多的虚拟天文台研究尚处于独立研究状态，缺少同实际的天文研究进行互动，因而它们的可用性总是受到质疑，除了少数应用软件，大多数没有在天文学家中受到广泛重视和得到频繁使用。

1.4.3 难点分析和任务描述

首先为了解决数据访问瓶颈，我们需要在China-VO的框架下面研究开发一套自己的数据访问服务，这个服务实现一个新的网络服务模式，支持对分布式存放的异构数据库的海量数据访问，这就是我们前面提到的VO-DAS系统。在本文的第二章中我们详细介绍China-VO数据访问服务（VO-DAS）的设计、开发与试验。

然后需要考虑数据预处理和数据挖掘这两个操作。由于数据预处理采用的方法也多是成熟的信号处理、图像处理、多维数据分析等技术，因此可以将这个过程和数据挖掘过程一并考虑。我们面临的问题是，数据挖掘工具种类非常多，但是绝大多数都不能和VO系统实现无缝的连接。很多情况下，天文学家还会自己写算法程序实现特定功能的数据挖掘过程。在虚拟天文台的诸多协议中除了数据格式和数据传输相关的协议外，还没有成熟的数值计算或网络计算相关的协议。AstroGrid采用Common Execution Architecture（CEA）作为数据计算程序的设计框架，NVO的一些大型计算则是采用TeraGrid[16]完成的，上述方案都难以在China-VO的环境下很快应用。因此，我们必须探索各种数据挖掘工具的设计方案，尝试找到针对China-VO基础环境的最适应中国天文学研究现状的数据挖掘工具方案。数据挖掘方案有以下几种可能的选择：

- 利用现在天文学研究中常用的通用编程环境（如Fortran，IDL，python），继承和延伸已有的天文计算包，独自实现数据挖掘算法以及和虚拟天文台数据访问的接口。

[16]http://www.teragrid.org

- 在虚拟天文台的框架之上开发一个平台，将大多数已有的数据挖掘程序封装之后挂载到这个平台下，由这个平台提供天文数据，进行数据挖掘。
- 利用工程领域已经成熟的通用数据挖掘和数值计算软件（如MATLAB, SPSS, SAS, R），在其上进行二次开发，实现和虚拟天文台的互操作，同时增加常用的天文计算工具。

本文的第三章里，我们将深入探讨上述各种选择。

最后，为了能够对数据挖掘的结果进行分析和验证，基于虚拟天文台的数据挖掘工具中还应该有数据可视化工具。相对而言，数据可视化工具的选择比较多，在已经发布的VO工具中不乏优秀的数据可视化工具。例如，FITS文件的可视化可以采用Aladin、DS9等工具实现。那么如果将所有需要可视化的数据图生成FITS格式，就可以使用这些工具进行查看。对于表格数据，新版本的Topcat可以实现常见的几种二维或三维图形，例如散点图、直方图、密度轮廓图等。还有一些通用的数据可视化工具可以利用，如IDL 、Origin 、MATLAB等。但是这些通用的可视化工具必须有能力读取天文数据的特殊格式（FITS 、VOTable）。IDL 和MATLAB 都已经支持了FITS格式的文件读写访问。

总之，基于虚拟天文台的数据挖掘工具应当具有以下特点：

1. 该工具符合IVOA的各项成熟协议和规范，保证它和天文数据资源、服务资源以及其它VO工具之间的互联互通互操作。
2. 该工具符合China-VO的体系结构，可以成为建立中国虚拟天文台的一块基石。同时，它自身的结构应该力求简单、有效，用最小的成本实现最有效的开发。
3. 该工具是面向应用的，因此它的最终目标是能够在天文学研究中得到广泛应用。它的需求来自于LAMOST以及其它大型巡天项目支持的科学研究课题，它最终也要为这些研究服务。
4. 该工具应该尽可能多地集成现有的数据挖掘算法，将这些算法无缝地融合到虚拟天文台的框架中去。
5. 该工具要提供丰富易用的数据可视化工具。

为了让上述虚拟天文台应用产品有的放矢、更具有针对性、更加实用，我们决定采用以**科学驱动技术**的思路，即把虚拟天文台基础上的数据挖掘工具的研究放在一个特定的天文学课题中，在科学研究的同时发展应用工具。这样做的好处是明显的。首先，需求来自真实的研究，这将使得应用工具的开发非常具体，不会流于形式，更加注重那些实用有效的功能。其次，增进了科学研究中虚拟天文台的参与程度，实践了天文学信息化的宗旨。最后，新技术的应用将使得天文学的研究得到更多的助力，有助于解决那些以前用传统方法不能解决的问题。

由于LAMOST还没有完成，观测工作没有展开，因此我们采用和LAMOST非常类似的SDSS的数据来设定科学课题。这样的科学选题具有和LAMOST相类似的数据规模和数据组织，便于以后直接应用LAMOST数据于其上。使用SDSS巡天数据寻找银河系晕的子结构和伴星系是最近一段时间以来备受关注的课题。从科学上讲，对银河系晕的研究有助于解开银河系演化的谜团，同时又为暗物质宇宙学模型提供重要证据（有关银河系晕结构的研究的回顾和概述将在第四章专门讨论）。于是，我们把基于虚拟天文台的数据挖掘工具的科学应用确定为研究银河系晕的子结构以及伴星系。

我们明确研究的任务可以分解为以下几个部分：

1. 研制一种能够处理海量数据的VO数据访问服务，该服务可以对分布存储在不同地方的异构的天文数据资源进行统一访问；
2. 在此之上，研究对海量天文数据的最佳的数据挖掘工具，并实现之；
3. 最后，应用上述工具寻找新的银河系伴星系。

本文后面的各章安排如下：第二章中我们将详细介绍China-VO数据访问服务（VO-DAS）的设计、开发与试验。在第三章，我们将讨论三种数据挖掘方案并通过原型开发和实验对它们的功能和性能进行对比，确定最佳的方案。在第四章，我们将简单介绍银河系晕的结构研究的知识背景和研究现状，为后面章节介绍的科学范例提供必要的知识准备。在第五章中我们将介绍我们利用自己开发的虚拟天文台数据挖掘工具发现的5个新的银河系伴星系或球状星团候选体的具体方法和过程。最后在第六章中作出总结和对未来工作的展望。

第二章　异地异构数据资源的统一海量数据访问

本章将描述VO-DAS的功能、设计和实现以及在其上进行的实验。VO-DAS是一个基于网格技术的虚拟天文台数据访问系统，支持对异地异构数据资源的统一、海量数据访问。它是实现基于虚拟天文台的数据挖掘工具的第一个环节——获取数据挖掘所需要的数据。一种能够集中访问遍布世界各地的天文数据资源的数据访问服务是天文数据挖掘工具的基础设施，因此我们必须首先实现这样一个服务，然后在此基础上完成一个完整的数据挖掘工具的研究和开发。本章关于VO-DAS的描述是基于来刘超等(2007)[26]和田海俊(2007)[39]的工作，并在他们的基础上做了更多详细的展开。

2.1　系统的目标与功能

2.1.1　VO-DAS的目标

自从开始对天体进行照相观测以来，世界各国的天文观测资料不断增长，近年来由于先进的CCD技术、计算机处理技术和大型望远镜技术的快速发展，天文观测数据更是在以指数增长[35]。天文信息爆炸式增长不仅给天文学家带来了天文发现的巨大机遇，也同时带来了数据访问和处理方面的巨大挑战。传统的借助人工来对数据进行分析处理的方式已经不能满足海量的天文数据了。

另一方面，上百年的天体观测资料积累造成了这样一种情况：世界各国的天文资料保存和管理方式存在巨大差异，不同历史时期数据保存和管理方式也存在巨大差异。虽然世界各地的天文数据是自由共享的，但是数据结构的混乱和数据存储的多样性就像语言不通一样严重影响着天文学家对这些数据资料的使用效率。China-VO的数据访问服务系统（VO-DAS）就是为了解决这样的困难而设计。该系统的设计目标有两个：统一访问异地异构数据库，支持访问海量数据。

2.1.2　VO-DAS的功能

统一访问异地异构数据库是指数据访问者不必了解数据资源的具体物理位置以及数据资源的具体组织形式，通过统一的访问接口就可以取得他们希望得

到的数据，并且可以在多个不同的数据源之间实现数据的联合查询或交叉证认。为了能够统一地访问异地异构数据资源，要求数据访问服务具有以下两种能力：从遍布世界各地的网络服务器上发现合适的数据资源；能够整合这些数据资源，当联合查询多个数据库的时候，能够提供一种机制把分散在不同地方的数据库从逻辑上联系起来，实现联合查询，同时尽量减少网络带宽的占用。

发现分布的天文数据资源需要两项技术的帮助：资源注册和元数据描述。资源注册在IVOA制订了规范[18]以后比较成熟，并已经有了数个产品在运行。中国虚拟天文台采用AstroGrid的注册服务器软件，在本地建立了一个Registry服务。我们将在第2.2节详细介绍资源注册的内容和方式。元数据描述包括两类内容。一个是数据资源本身的存放位置、访问方式、发布者信息、版本信息等，这一类型的元数据实际上是用来注册到Registry服务器上的内容，我们将在介绍资源注册的时候一并介绍；另一个类型的元数据是指数据的组织形式。

通常我们经常接触的天文数据资源包括四类资源：星表数据、图像数据、光谱数据和文献，VO-DAS将可以访问除文献外的其它几种数据资源。如果数据是星表数据，通常会按照关系型数据库的方式存放。这样，它的元数据描述应该包括这个星表的基本信息，具有多少个表结构，每个表结构的意义是什么，具有多少列（或用数据库的术语称为字段），每个列的名称、单位、精度、物理意义是什么。可以看出星表的元数据描述是一个树形结构。图像数据本身通常使用FITS文件来保存的。为了能够方便地索引到需要的图像，还要有一个对应的数据库的表，记录FITS文件的空间参考位置、波段、观测时间和地点，观测时候的天空视宁度、云量、投影参数、望远镜参数、CCD或底版的参数等。光谱数据（一维光谱）也是以FITS文件形式保存的，通常每个FITS文件只有一个天体的光谱，因此FITS文件的名字应该能够和一个特定星表的一行一一对应。这样，要想找到一个天体的光谱，就首先在一个星表中找到这个天体，然后再根据星表行的关键字和FITS文件的一一映射关系找到光谱文件。

VO-DAS要为以上三种天文数据设计一个统一的元数据描述方式。注意到FIT图像需要一个索引表检索，因此我们可以将这个索引表看成是和星表一样的。而光谱文件的搜索也一定要借助星表。这样，我们就可以设计一种能够描述表结构的元数据描述形式来描述所有的天文数据资源。

有了资源注册机制，VO-DAS就可以为使用者找到需要的天文数据资源。

有了统一的元数据描述方式，就可以让用户和数据访问服务理解访问的数据是如何组织的、类型是怎样的。这样，无论数据资源放在何处，也无论数据资源是怎么组织的，VO-DAS都可以借助这两项功能找到数据资源并理解数据资源。

统一的数据访问的另一层含义是对异地异构数据的联合访问，包括交叉证认。联合查询两个和两个以上的表，在关系型数据库中需要将两个表的相关行进行组合，然后从中挑出符合条件的行。如果两个表不在一个数据库系统中，甚至不在一个地方，那么必须要将一个表的部分信息传递到另一个表的数据库中，在本地数据库中完成联合查询。当数据量很大的时候，这样的数据传输是不经济的。一种可行的解决方法是把联合查询拆成两步查询，首先在一个表中完成查询，然后把查询结果传递到第二个表所在的数据库中完成第二次联合查询。这样的方法虽然不能完全解决数据传输的问题，在某些最坏的情况下甚至还需要传递整张表，但是大多数情况下，特别是当用一个小天区作约束查询的时候是可行的。

多大的数据算是海量数据？在天文学中没有明确的定义，我们只能定义VO-DAS能够一次访问多少条数据。在NVO开发的Open SkyQuery原型[1]中[40]，他们允许一次访问5000条记录。SDSS的星表访问服务casjob[2]中，用户可以一次访问500M字节的星表数据。记录数会根据列的多少有变化，通常查询10列左右的时候500M字节大约是数十万到数百万条记录（有压缩的情况下）。VO-DAS希望在内存和网络带宽允许的情况下支持访问尽可能多的数据。

为了能够访问更多的数据，传统的数据查询方式需要作出改变。Open SkyQuery使用了Web Service[3]作为提供访问数据服务的模式，这直接导致了数据传输效率的低下。这是因为Web Service通过SOAP[4]协议在网络间传递消息，这就要求所有的从数据库返回给查询者的数据都要封装成XML[5]格式的消息，然后通过HTTP[6]协议返回给用户。对于大数据文件XML序列化（Serialize，即生成XML数据）和反序列化（Deserialize，即解读XML数据）过程非常低效；而HTTP协议是面向无连接的协议，传输海量数据流的效率也非常低。Open SkyQuery选择的技术决定了它需要花很长的时间才能把一个大数据集传递回查

[1] http://www.openskyquery.net
[2] http://casjobs.sdss.org
[3] http://www.w3.org/2002/ws
[4] http://www.w3.org/TR/soap
[5] http://www.w3.org/XML
[6] http://www.w3.org/Protocols/rfc2616/rfc2616.txt

询者。由于HTTP连接在一定时间内（通常10－20分钟，可以设置，但是时间越长吞吐量就越少）不响应会自动中断，因此不能用来等待大数据集的传输。这就是为什么Open SkyQuery限定访问上限为5000条记录的原因。

SDSS的casjob服务可以访问500MB的数据，是因为它采用了异步查询的方式，即在提交了数据查询请求以后，casjob允许查询者断开之间的连接。而查询请求此时已经在服务器端生成为一个任务（job），由服务器来管理这个查询任务。当查询者再次登录进来查看任务进展的状态时，如果此时查询任务已经结束，那么用户会在一个称为MyDB的虚拟数据库中看到自己的查询结果，并可以把它下载回本地计算机。我们从SDSS得到启发，要求VO-DAS能够支持异步查询以实现海量数据的访问。但是对于数据量比较小的查询，异步查询效率比较低，因此还需要VO-DAS能够提供简单实用的同步查询。同步查询应该能够在几分钟的时间内直接将数据返回回来。

数据查询的描述语言应该采用IVOA 标准[7]的ADQL[17] 。这是因为首先ADQL支持特定的天文查询方法，例如ConeSearch和交叉证认。其次，ADQL不依赖数据库的物理位置，这便于实现对分布式数据库查询的描述。最后，也是最重要的，ADQL是IVOA的标准查询语言，支持ADQL有助于VO-DAS和其他VO应用之间的互操作。但是到目前为止，ADQL仅支持对星表的查询，并不支持图像和光谱的查询。为此我们需要将它进行扩展，使它在语法不做任何改变的情况下，语义得到扩充，从而支持图像和光谱查询，并且查询能力等同于SIAP[8]和SSAP[9]。关于ADQL的语法和我们将对它进行的扩展将在第2.2.5节中详细介绍。

所有这些功能及简单说明总结在表2.1中。

2.2 VO-DAS应用到的各项技术和标准

VO-DAS是在计算机领域的最新发展成果的基础上设计和实现的，它还应用了尽量多的IVOA标准来提高它的互操作性。在这一节里，我们将介绍VO-DAS中用到的各项技术和标准。

[7]http://www.ivoa.net/Documents/latest/ADQL.html
[8]http://www.ivoa.net/Documents/latest/SIA.html
[9]http://www.ivoa.net/Documents/latest/SSA.html

表 2.1: VO-DAS的功能

编号	功能	说明
1	数据资源的注册	任何希望被VO-DAS使用的数据资源都要把自己的注册元数据加入到一个Registry服务上，注册的时候需要说明这些数据资源是哪种类型：星表、图像还是光谱
2	数据资源的发现	VO-DAS会在适当的时机向Registry服务查询，搜索所有可以使用的天文数据资源
3	数据资源元数据的获得	VO-DAS在从Registry服务获得数据资源的网络地址后向这个数据资源请求数据组织的元数据，这些元数据可以共享给所有VO-DAS的客户端程序
4	两个存放在异地的异构数据库的联合查询	仅仅支持在一个小天区范围内的这种联合查询，这种联合查询包括两个表的联合和交叉证认
5	能够查询海量数据	海量数据的定义是在内存允许的情况下尽量多的返回数据记录，至少在10^5的数量级，在返回字段不多的情况下可以达到10^7级
6	同步查询数据	客户端一直等待VO-DAS将数据查询出来，并直接把数据返回回来
7	异步查询数据	发出查询请求以后，客户端不必等待数据的返回。VO-DAS会在数据出来以后按照客户端早先的定义，将数据保存到存储服务器上
8	数据查询的状态跟踪	VO-DAS允许客户端的程序跟踪异步查询任务的执行状态
9	支持ADQL的查询方式	ADQL是IVOA的天文数据查询标准，支持ADQL有助于提高系统的互操作性
10	支持使用ADQL对图像和光谱的查询	

2.2.1 网格服务

中科院计算所李国杰院士认为:“网格不同于国外正在搞的Internet 2 或下一代Internet (NGI),网格可以称作是第三代Internet,其主要特点是不仅仅包括计算机和网页,而且包括各种信息资源,例如数据库、软件以及各种信息获取设备等,它们都连接成一个整体,整个网络如同一台巨大无比的计算机,向每个用户提供一体化的服务。” [41]

网格的研究目标从资源的角度上来看和虚拟天文台的目标是一致的,那就是通过基于Internet现有架构实现最大限度的网络资源共享。因此,网格技术自然而然地在虚拟天文台的研究中得到了广泛的应用。由于网格本身就是一个比较新的技术,还处于探索阶段,因此其体系结构也在短短几年里发生了很多的变化,从一开始的OGSA[10]架构转向了WSRF[11]架构。这些架构的变迁影响了虚拟天文台的发展,作为一个通用的网络资源共享的架构,如果不能进入稳定发展,那么虚拟天文台就无从着手,不能享受到新技术带来的好处。

作为网格中间件的提供者和网格应用的推动者Globus Alliance[12]在它的网格中间件系统早期版本Globus Toolkit 3.0采用了OGSA架构而在2005年推出的Globus Toolkit 4.0中又全面支持WSRF构架。

我们决定采用WSRF架构封装现有数据资源,同时作为面向终端用户的VO-DAS也是采用WSRF构架。这是因为首先,WSRF提供的是一个带有状态的Web Service系统,这对我们实现比较耗费时间的海量数据的异步查询将会有较大帮助。其次,WSRF和我们已经比较熟悉的Web Service系统非常相似,学习和理解比较容易。最后,相对于其他的网格中间件,WSRF架构的Globus Toolkit 4.0安装和维护相对比较简单,VO-DAS系统应用以后不会给系统分发和使用带来较大困难。

WSRF的全称是Web Service Resource Framework,它通过一个隐含的资源模式为在Web 服务之间创造有状态的资源定义了一个系统。它是OASIS[13]的Web service规范族中的一个,主要制订者包括Globus Alliance和IBM。WSRF定义了使用Web 服务来访问有状态资源的一系列规范。它包括Web 服务资源特性(WS-ResourceProperties)、Web 服务资源生命周期(WS-ResourceLifetime

[10]http://www.gridforum.org/ogsi-wg
[11]www.globus.org/wsrf
[12]http://www.globus.org
[13]http://www.oasis-open.org

）、Web 服务基本故障（WS-BaseFaults ）和Web 服务服务组（WS-ServiceGroup ）规范。这些新规范的动机是，虽然Web Service在交互的过程中并不维护状态信息，但是它们的交互必须经常性地为状态操作考虑，也就是说，数据的值通过Web 服务交互得以持久化，并且作为Web 服务交互的结果而保存[14]。

Web 服务资源特性（WS-Resource Properties）定义了如何使用Web 服务技术来查询和改变与一个有状态资源相关联的数据。这提供了一种标准的方法，可以用于关联数据和允许由客户端访问的资源。Web 服务资源的特性声明代表了Web 服务资源状态的一个投影或一个视图。这种投影又代表了一种隐含的资源类型，可以用来通过Web 服务接口定义访问资源特性的基础。

Web 服务资源生命周期（WS-Resource Lifetime）定义了两种释放Web 服务资源的方法：直接的和预先计划的。这使得设计人员可以灵活地设计他们的Web 服务应用程序来保证清除不再需要的资源。

Web 服务基本故障（WS-BaseFaults）为基本故障定义了一个XML 模式类型以及Web 服务如何使用这种故障类型的规则。Web 服务应用程序的设计人员经常使用别人定义的接口。当每个接口使用不同的约定来表示故障消息中的常见信息时，管理这种应用程序中的故障就会变得更加困难。对问题确定和故障管理的支持可以通过一种通用的方法指定Web 服务故障消息来加以增强。当来自不同接口的可用故障信息都一致时，请求者理解故障就更加容易了。而与此同时，开发一种通用的工具来帮助处理故障也变得更加可能。

Web 服务服务组（WS-ServiceGroup）定义了一种方法，通过这种方法，Web 服务和Web 服务资源可以为了某个领域的特定目的而聚集或组合在一起。为了让请求者能够根据服务组（ServiceGroup）的内容进行有意义的查询，必须以某种方式来限制组中的成员资格。对成员资格的限制是通过使用分类机制以级别表达的，而每个级别的成员必须共享一组共同的信息来表达查询。

2.2.2　OGSA-DAI

在前面一节中，我们提出VO-DAS访问的数据资源应该是使用WSRF封装过的数据资源。这样，这种资源才能够通过WSRF接口被VO-DAS访问到，并且能够支持海量数据查询。我们希望找到一个中间件，它可以直接对数据资源进行WSRF的封装，OGSA-DAI就是这样的一个中间件。 OGSA-DAI[15]的

[14]http://www.ibm.com/developerworks/cn/webservices/ws-resource/

[15]http://www.ogsadai.org.uk/

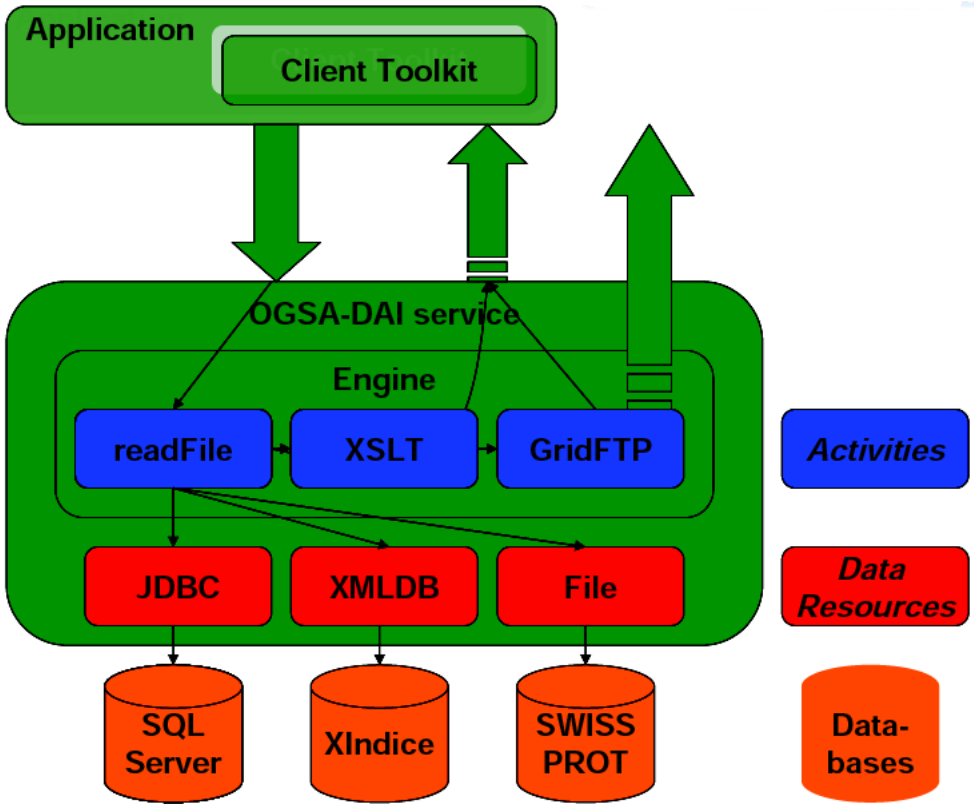

图 2.1: OGSA-DAI的体系结构

全称是Open Grid Architecture Data Access and Integration。该项目是英国e-Science项目中的一个，目标是开发一个中间件，能够对网格中分布存放的数据库进行访问和数据集成[42, 43]。OGSA-DAI中间件在网格上将不同类型的数据库，包括关系型数据库和XML数据库，作为数据资源发布出来。目前OGSA-DAI支持多种常见的数据库，如MySQL, Oracle等。它还支持不同的数据格式转换、数据传输方式等。此外，它还提供了扩展功能接口，便于对功能进行扩展（参见图2.1）。

OGSA-DAI支持关系型数据库，XML数据库以及文件形式存放的资料。我们需要发布的天文数据资源包括星表、图像和光谱，前一种是可以用关系型数据库管理的，而后面两种是以文件形式存放的。OGSA-DAI的支持覆盖了我们所用的全部数据库类型。

OGSA-DAI使用XSLT[16]实现数据格式的转换。在VO-DAS中，我们需要将数据库查询的结果转换成VOTable格式，而VOTable格式本身就是XML描述的，因此，只要定义一个从数据库查询结果XML转换到VOTable格式的XSLT文件，就可以让OGSA-DAI支持VOTable格式了。

OGSA-DAI支持直接将查询结果数据返回给客户端，也可以通过URL、FTP 、GridFTP等传输协议将数据传输出来。这样多种手段的数据传输模式可以根据用户不同的要求和数据的大小进行选择。在VO-DAS中，我们希望支持两种模式的数据访问，一种是同步模式，即数据在查询出来以后直接返回给客户端，在数据查询期间，客户端一直在等待。另一种是异步模式，即客户端在提交了查询任务以后就不再等待了，而由VO-DAS完成数据查询并将数据保存到存储服务器上面去。对于OGSA-DAI，当它支持同步查询模式的时候，可以采用数据直接传回的方式；当它支持异步查询模式的时候，可以采用FTP或GridFTP的方式将数据保存到存储服务器上面。

OGSA-DAI以上这些特征证明它确实很适合VO-DAS作为数据访问和集成的工具软件。我们可以使用OGSA-DAI来封装所有天文数据资源，用它作为访问数据资源的统一接口，在VO-DAS中这样封装的数据资源被称为DataNode。这样，VO-DAS只要访问DataNode就可以访问全部资源了。由于OGSA-DAI也是运行在Globus Toolkit 4之上的，因此使用它封装的数据资源和VO-DAS就运行在一个一致的网格环境中，这为VO-DAS的开发实现和以后的运行和配置代来了不少方便。

2.2.3　天文资源注册

资源注册的概念来自计算网格[16]。在虚拟天文台的概念模型中，天文服务的注册在VO Registry服务上完成的。VO Registry服务接受任何符合IVOA Resource Metadata规范[17]的服务注册。当数据资源完成在一台VO Registry服务器上的注册以后，它的元数据信息会通过Harvest服务传递给其他的VO Registry服务器。很快，全世界的VO Rergistry服务都拥有了这个资源的基本信息。

天文资源在每个VO Registry中存放的内容，也就是它所注册的内容，是由Resource Metadata规范定义的。这是一个XML定义的结构，主要内容包

[16]www.w3.org/TR/xslt

[17]http://www.ivoa.net/Documents/latest/RM.html

括Resource metadata concepts，Data and metadata quality assessment和Service metadata concepts。

Resource metadata concepts包含资源的识别信息、生产者 / 提供者的信息、一般性的信息（如主题、类型、参考资料的URL等）、数据资源的一般信息（如设备、天区覆盖、频率或波段、分辨率等）、空－时坐标元数据，Data and metadata quality assessment包括数据质量、资源有效性分级等，Service metadata concepts包括调用接口元数据、功能描述元数据等。

以上内容视具体资源类型而有剪裁，对于DataNode而言，参考资料的URL就是DataNode的OGSA-DAI的接口地址，而Service metadata concepts的内容不必填写。因为虽然DataNode表现为一个网格服务，但实际上它代表的是数据资源。

2.2.4 数据资源的元数据描述

星表、图像和光谱都需要有一个元数据格式来描述它们是如何组织的，这样的元数据目前还没有统一的规范。但是由于表格数据在虚拟天文台中是用VOTable的格式保存的，因此我们考虑使用VOTable来作为数据资源的元数据描述格式。

VOTable格式[15]是基于XML的一种表格描述格式。它的文档结构分成两个部分，一个部分是元数据的定义，另一个部分是数据。我们可以将一个没有数据的VOTable文件看成是一个星表的元数据描述。在VOTable里面详尽定义了一个表的每一列：它的名称、数据类型、单位、数据精度、UCD[18]等。表2.2描述了我们在VO-DAS中描述元数据所需要的VOTable的元素的具体定义，图2.2给出了一个使用VOTable描述的DataNode的元数据的实例。

2.2.5 ADQL的应用和扩展

ADQL是IVOA提出的天文数据查询语言[17]，虽然已经更新了几个版本，但仍在草稿阶段。为了能够和以后的ADQL正式规范平稳衔接，VO-DAS采用ADQL ver0.9版本作为标准查询语言。ADQL是根据SQL92改进而来的针对天文数据访问特点而设计的数据库查询语言，因而它继承了绝大多数SQL语言的语法特征。图2.3展示了一个典型的ADQL语句的样子，这条ADQL语句

[18]http://www.ivoa.net/Documents/latest/UCD.html

```
−<VOTABLE version="1.1" xsi:noNamespaceSchemaLocation="http://www.ivoa.net/xml/VOTable/VOTable/v1.1">
    <COOSYS ID="J2000" equinox="J2000." epoch="J2000." system="eq_FK5"/>
  −<RESOURCE name="SDSS">
    −<TABLE name="Star2">
        <DESCRIPTION>Velocities and Distance estimations</DESCRIPTION>
        <FIELD name="ra" ID="col1" ucd="pos.eq.ra;meta.main" ref="J2000" datatype="float" width="6" precision="2"
        unit="deg"/>
        <FIELD name="decl" ID="col2" ucd="pos.eq.dec;meta.main" ref="J2000" datatype="float" width="6" precision="2"
        unit="deg"/>
        <FIELD name="u" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="g" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="r" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="i" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="z" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
      </TABLE>
    −<TABLE name="StarDR6">
        <DESCRIPTION>Velocities and Distance estimations</DESCRIPTION>
        <FIELD name="ra" ID="col1" ucd="pos.eq.ra;meta.main" ref="J2000" datatype="float" width="6" precision="2"
        unit="deg"/>
        <FIELD name="decl" ID="col2" ucd="pos.eq.dec;meta.main" ref="J2000" datatype="float" width="6" precision="2"
        unit="deg"/>
        <FIELD name="g" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="ug" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="gr" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="ri" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="iz" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
      </TABLE>
    −<TABLE name="subStar2">
        <DESCRIPTION>Velocities and Distance estimations</DESCRIPTION>
        <FIELD name="ra" ID="col1" ucd="pos.eq.ra;meta.main" ref="J2000" datatype="float" width="6" precision="2"
        unit="deg"/>
        <FIELD name="decl" ID="col2" ucd="pos.eq.dec;meta.main" ref="J2000" datatype="float" width="6" precision="2"
        unit="deg"/>
        <FIELD name="glon" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="glat" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="u" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="g" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="r" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="i" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
        <FIELD name="z" ID="col1" ucd="" ref="J2000" datatype="float" width="6" precision="2" unit="mag"/>
      </TABLE>
    </RESOURCE>
  </VOTABLE>
```

图 2.2: 使用VOTable描述DataNode元数据的实例

```
SELECT p.ra, p.dec,p.r,t.j FROM sdss:photoprimary p, twomass:twomass_psc t
WHERE XMATCH(p,t,1) AND REGION('circle 157.5 55.1 1') AND p.r<22.5 AND t.j<14.5
```

图 2.3: ADQL的例子

表 2.2: VOTable定义元数据的各个元素

Element	说明
COSYS	定义了天体的坐标系统。
RESOURCE	定义了一个或多个资源对象。每个资源对象会有一个DESCRIPTION元素用于文字描述，一个COSYS用于定义资源的参考坐标系。RESOURCE要有一个name或id用于定义一个名字。
TABLE	该元素描述了一个基本的表结构，它包含若干FIELDS元素。它也包含一个DESCRIPTION用于存放描述这个表的文字内容。
DESCRIPTION	纯文字描述。
FIELD	字段的定义。它包括一个name或id属性用于定义名称，datatype属性用于定义数据类型，unit属性定义物理单位，ucd属性定义字段物理含义。

的语法基本上和SQL相同，但有一定扩展。它查询SDSS的photoprimary星表和2MASS的twomass_psc星表，其中SDSS是Sloan数字巡天观测项目[37]的数据资源名称，2MASS是2微米全天巡天项目[44]的数据资源名称。SELECT分句列出查询的列名，分别为photoprimary星表的赤经（ra）、赤纬（dec）、r波段的星等（r）和twomass_psc星表的J波段星等（j）。WHERE分句的第一个条件XMATCH表示这两个星表的天体做交叉证认（参与交叉证认的表为p和t，交叉证认的误差半径为1角秒），这是一个ADQL的扩展函数。第二个条件REGION也是ADQL的扩展函数，表示查询和交叉证认都是在一个小的锥形天区（原点在赤经157.5度、赤纬55.1度、半径1度）内完成的。第三和第四个条件语法同SQL完全一样。通过上述例子我们看到，使用ADQL既可以描述对特定天空区域的查询也可以描述交叉证认，同时它独立于星表的物理存储位置和存储方式。

我们在VO-DAS中对ADQL的语义作了扩展，允许FROM分句中的表的类型是星表（等同于数据库中的表）、图像库（等同于文件集合）和光谱库（等同于文件集合）。当FROM分句列出的是星表的时候，整个ADQL的使用就是按照IVOA的标准进行；当FROM是图像和光谱的时候，SELECT分句的列名实

际是图像和光谱的输出参数名，其中当然包括文件位置URL信息；WHERE分句的条件则是对要查询的图像和光谱文件的输入参数进行的约束。经过这种扩展，当ADQL查询图像或光谱的时候，返回的表中一定有一列是文件URL，这样，再根据URL指示就可以取得图像或光谱的FITS文件了。通过这样的扩展，语法上ADQL没有任何改变，但是语义的扩展便让ADQL支持天文图像数据和光谱数据的查询了。使用ADQL查询图像和光谱所使用的输入输出参数借鉴了SIAP和SSAP两个协议。表2.3和2.4定义了图像查询所使用的参数，表2.5和2.6定义了光谱查询所使用的参数。

表 2.3: ADQL查询图像所使用的输入参数

参数名字	方向	类型	说明
POS	IN	double(2)	希望获得的图像的位置
SIZE	IN	double	以角度表示的角尺寸
INTERSECT	IN	char	图像和目标区域的匹配方式
FORMAT	IN	char	要得到的文件格式
NAXIS	IN	double(2)	输出图像的像素大小
CFRAME	IN	char	坐标系的参考框架
EQUINOX	IN	char	坐标系的历元
CRPIX	IN	double(2)	参考像素的坐标
CRVAL	IN	double(2)	参考像素相对于CFRAME的世界坐标
CDELT	IN	couble(2)	输出图像每个像素以角度计算的大小
ROTANG	IN	double	图像相对于CFRAME的旋转角度
PROJ	IN	char	输出图像的天球投影方式

当ADQL查询语句中涉及到多个DataNode之上的表的时候，需要对异地异构数据库进行联合查询。对ADQL的分析会在VO-DAS之内进行，客户端的用户在编写这样的ADQL语句的时候不必考虑数据库的物理位置和类型。

表 2.4: ADQL查询图像所使用的输出参数

参数名字	方向	类型	说明
Image_Title	OUT	char	图像标题
Inst_ID	OUT	char	设备标识
Image_MJDateObs	OUT	double	观测日期的简化儒略日
pos_eq_ra_main	OUT	double	图像中心的ICRS赤经
pos_eq_dec_main	OUT	double	图像中心的ICRS赤纬
Image_naxes	OUT	int	图像的轴的数目
Image_naxis	OUT	int	每个轴的像素数目
Image_scale	OUT	double	每个轴上像素的以角度计算的大小
Image_format	OUT	char	图像的MIME类型
STC_CoordRefFrame	OUT	char	坐标系的参考框架
STC_CoordEquinox	OUT	char	坐标系的历元
WCS_CoordProjection	OUT	char	FITS WCS定义的天球投影方式
WCS_CoordRefPixel	OUT	double	WCS参考像素的像素坐标
WCS_CoordRefValue	OUT	double	WCS参考像素的世界坐标
WCS_CDMatrix	OUT	double	WCS的CD矩阵
BandPass_ID	OUT	char	波段的ID标识
BandPass_Unit	OUT	char	波段的单位
BandPass_RefValue	OUT	double	波段的特征频率
BandPass_HiLimit	OUT	double	波段的上限
BandPass_LoLimit	OUT	double	波段的下限
Image_AccessRef	OUT	char	图像文件的URL位置
Image_AccessRefTTL	OUT	int	图像URL有效的时间，以秒计算
Image_FileSize	OUT	int	输出的图像文件大小

表 2.5: ADQL查询光谱所使用的输入参数

参数名字	方向	类型	说明
POS	IN	double(2)	位置坐标
SIZE	IN	double	查询的区域范围，以角度计算
FORMAT	IN	char	输出格式
APERTURE	IN	double	孔径
BANDPASS	IN	double	光谱波段
TIME	IN	char	光谱的时间段
REDSHIFT	IN	double(2)	红移范围
SPECRES	IN	double	光谱的分辨率
SNR	IN	double	信噪比

2.3　VO-DAS的系统分析

本节我们根据VO-DAS的功能和我们将要使用到的各项技术和规范分析VO-DAS的组成和工作方式。由于天文数据资源分散在各处，发布和访问的方式也各不相同。为了能够让它们在VO-DAS系统内实现统一的访问接口，我们使用OGSA-DAI来对这些数据资源进行封装。封装之后的OGSA-DAI服务器在VO-DAS系统内被称为DataNode。DataNode应该在一个VO Registry上进行注册，这样在VO-DAS系统中才能够被发现。因此，系统中还应该包括一个VO-Registry。当然，这个组成部分并不是VO-DAS系统所独享的，它可以是同整个虚拟天文台社区共享的一个服务。对于海量数据的访问结果需要一台或多台存储服务器的帮助。因此系统还应该和虚拟天文台社区共享一台支持FTP或GridFTP的存储服务器。最后，作为核心服务器，VO-DAS还应该有一个和客户端直接交互的服务器，它接收客户端传来的ADQL，并对它进行分析，然后决定将这个ADQL描述的数据查询请求送给哪一个DataNode完成。我们把这个服务器称为VO-DAS服务器。它是整个系统的核心服务器。VO-DAS的这四个部分的定义和功能详见表2.7。它们之间的关联参见图2.4。

我们在这里列出VO-DAS的几种常见的工作用例，用例（Use Case）[45]是一种在软件工程中常用的系统操作描述方式，我们在本节将VO-DAS的几个主

表 2.6: ADQL查询光谱所使用的输出参数

参数名字	方向	类型	说明
Dataset_DataModel	OUT	char	输出光谱的数据模型
Dataset_Title	OUT	char	输出光谱的标题
Dataset_SSA_NSamples	OUT	int	数据集的采样数目
Dataset_SSA_Aperture	OUT	double	角度计算的孔径直径
Dataset_CreationType	OUT	char	创建类型
Coverage_Location_Spatial	OUT	double(2)	光谱目标的空间位置
Coverage_Location_Time	OUT	char	观测时间文字描述
Coverage_Location_Spectral	OUT	char	观测波段文字描述
Coverage_Location_Spectral_BandID	OUT	char	波段或滤波器的标识或名字
Coverage_Bounds_Time	OUT	double(2)	曝光的时间范围
Coverage_Bounds_Spectral	OUT	double(2)	频率范围
Coverage_Bounds_Flux	OUT	double(2)	流量范围（Jansky）
Accuracy_Spatial_Calibrated	OUT	boolean	是否定标空间位置
Accuracy_Spatial_Resolution	OUT	double	空间分辨率（PSF的FWHM）
Accuracy_Time_Calibrated	OUT	boolean	是否定标时间
Accuracy_Time_Resolution	OUT	double	时间的分辨率（TSF的FWHM）
Accuracy_Spectral_Calibrated	OUT	boolean	是否定标频率
Accuracy_Spectral_Resolution	OUT	double	频率的分辨率（LSF的FWHM）
Accuracy_Flux_Calibrated	OUT	boolean	是否定标流量
Frame_Time_SIDim	OUT	int	SI因子和维度
Frame_Spectral_SIDim	OUT	int	SI因子和维度
Frame_Flux_SIDim	OUT	int	SI因子和维度
Frame_Flux_UCD	OUT	char	流量的UCD
Access_Reference	OUT	char	输出光谱的URL位置
Access_Format	OUT	char	输出光谱的格式
Access_Size	OUT	int	输出光谱的尺寸

表 2.7: VO-DAS的组成

组件名称	说明
VO-DAS服务器	在一个VO-DAS系统内至少有一个VO-DAS服务器，它直接和客户端程序进行交互，接收客户端的数据查询请求，并替客户端完形成网格环境下的查询。
DataNode	用于封装数据资源成为符合WSRF的网格服务，它会注册到VO Registry上面。
VO Registry	虚拟天文台注册服务器，可以是一个共享的组件。所有可用的DataNode必须在此注册才能够被VO-DAS发现。
存储服务器	支持FTP或GridFTP的存储服务，用来存放查询的结果数据。

要功能用用例图表示出来，并用文字说明来对用例图作出详尽描述。

首先定义数据资源注册的过程。在VO-DAS中，我们把一个数据资源称为Data Resource，而Data Resource是由DataNode的网格服务发布出来的。下面简要记述数据资源注册的步骤：

1. 将数据资源发布到一个DataNode上；
2. 把DataNode注册到VO Registry上。注册的时候需要填写一份完整的注册元数据表单。

然后定义数据资源发现的过程。它的执行步骤如下：

1. VO-DAS向VO Registry询问所有类型是DataNode的服务；
2. VO Registry将自己数据库中的DataNode服务的注册元数据整理成一个XML返回给VO-DAS服务器；
3. VO-DAS根据注册元数据信息找到每个DataNode的URL，然后向它们询问它们所封装的数据资源的元数据。

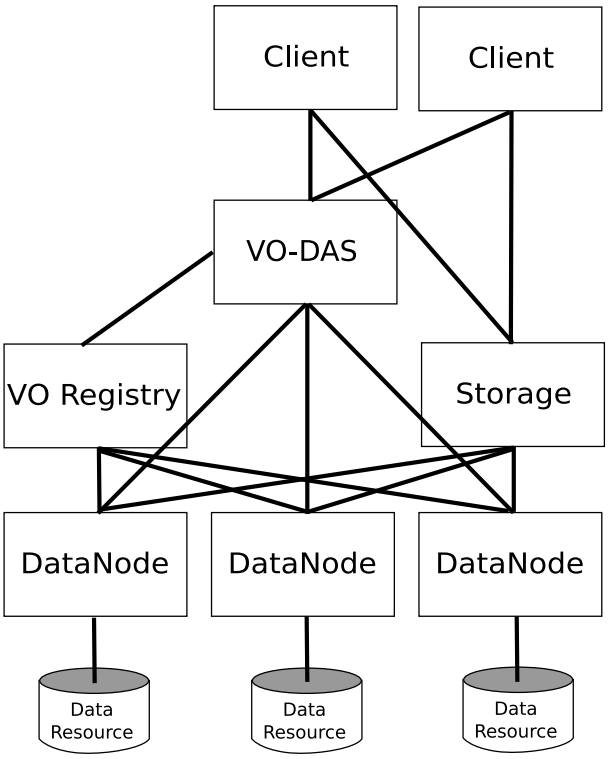

图 2.4: VO-DAS的组成和它们的关联

4. VO-DAS获得所有DataNode的数据资源的元数据VOTable格式的文件。

现在定义VO-DAS的客户端向VO-DAS获取数据资源元数据的过程。这个过程描述如下：

1. 客户端向VO-DAS询问数据资源的列表；
2. VO-DAS根据自己的记录返回给客户端一个数据资源列表，包括名字和描述信息；
3. 客户端向VO-DAS询问某个数据资源的表信息；
4. VO-DAS返回指定数据资源包含的所有的表的名称和描述信息；
5. 客户端向VO-DAS询问一个表的详细信息；
6. VO-DAS返回指定表的各个列的名称、类型、单位、描述等信息。

现在定义一个简单的同步查询的过程，同步查询是指客户端向VO-DAS发出请求后等待查询结果的返回，这是最简单的一个操作（如图2.5所示）。同步查询操作的执行步骤如下：

1. 客户端程序向VO-DAS提交ADQL查询请求；
2. VO-DAS解析ADQL，检查ADQL中涉及的数据资源所在的DataNode是否存在；
3. 如果DataNode存在，则生成DataNode的执行计划对象，并连接DataNode，执行这个计划；
4. 从DataNode获得执行结果，将执行结果返回给客户端。

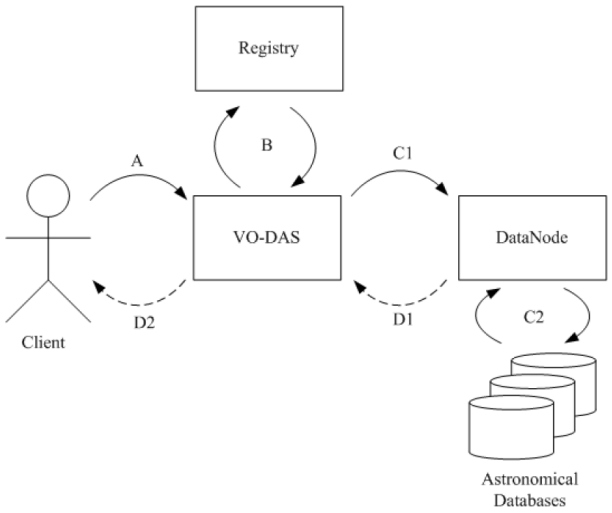

图 2.5: 同步查询的用例

现在定义一个简单的异步查询的过程。异步查询是指客户端在发出查询请求后立即返回，不等待查询结果，结果会以文件形式保存在一个URL处（如图2.6所示）。异步查询操作的执行步骤如下：

1. 客户端将ADQL查询语句和保存结果的URI传递给VO-DAS，然后立即返回，VO-DAS为这次查询建立session；
2. VO-DAS解析ADQL，检查ADQL中涉及的数据资源所在的DataNode是否存在；

3. 如果DataNode存在，则生成DataNode的执行计划对象，并连接DataNode，执行这个计划，查询的结果保存到指定的URI（图2.6中C3虚线部分）；

4. 结果保存完毕，DataNode向VO-DAS通知结束，VO-DAS删除相应的session。

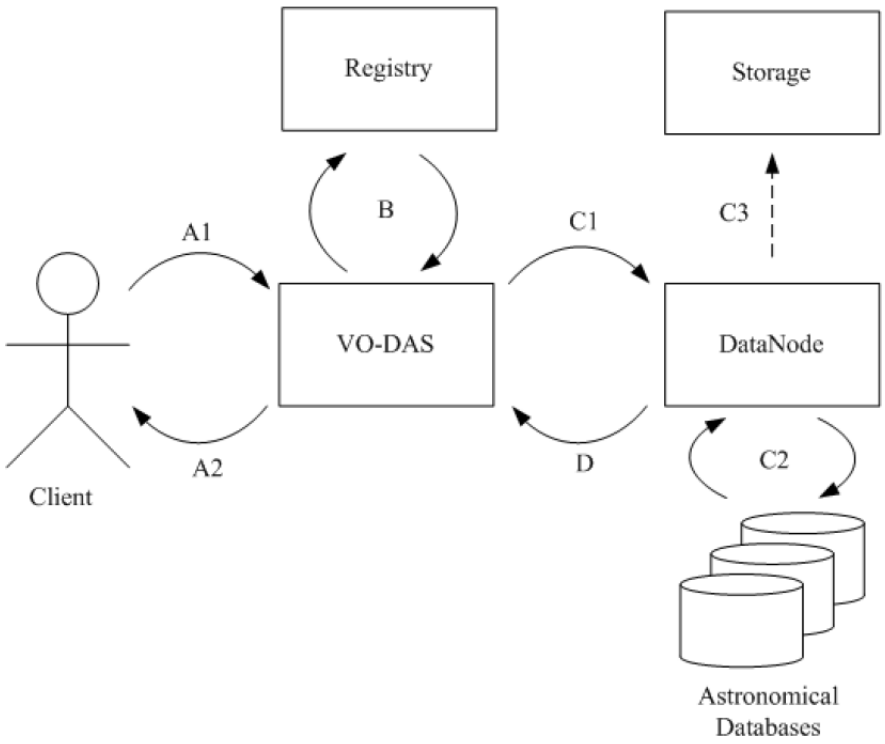

图 2.6: 异步查询的用例

现在定义一个异地数据库联合异步查询的过程。联合查询异地异构的数据库是VO-DAS最复杂的功能，它的操作用例见图2.7，具体步骤如下：

1. 客户端将ADQL查询语句和保存结果的URI传递给VO-DAS，然后立即返回，VO-DAS为这次查询建立session；

2. VO-DAS解析ADQL，检查ADQL中涉及的数据资源的DataNode是否存在，如果存在则按照传输数据最少的原则为所涉及的DataNode排序并生成每个DataNode的执行计划对象（假设这里涉及三个DataNode，排出顺序为DataNode A 、DataNode B 、DataNode C）；

3. 向DataNode A发送执行计划，DataNode A完成查询以后将临时结果直接传递给DataNode B（图2.7中C3虚线表示数据传输方向），数据传输后向VO-DAS报告DataNode A的查询结束；

4. VO-DAS向DataNode B发送执行计划，DataNode B会将来自DataNode A的中间结果保存在数据库临时表中并完成和这个临时表的联合查询，联合查询的结果直接传递给DataNode C（图2.7中D3虚线表示数据传输方向），数据传输后向VO-DAS报告DataNode B的查询结束；

5. 和D的过程相似，只是最后的联合查询结果根据客户端指定的URI存放到一个外部存储空间中（图2.7中E3虚线所示），保存数据后会通知VO-DAS，后者会删除相应的session。

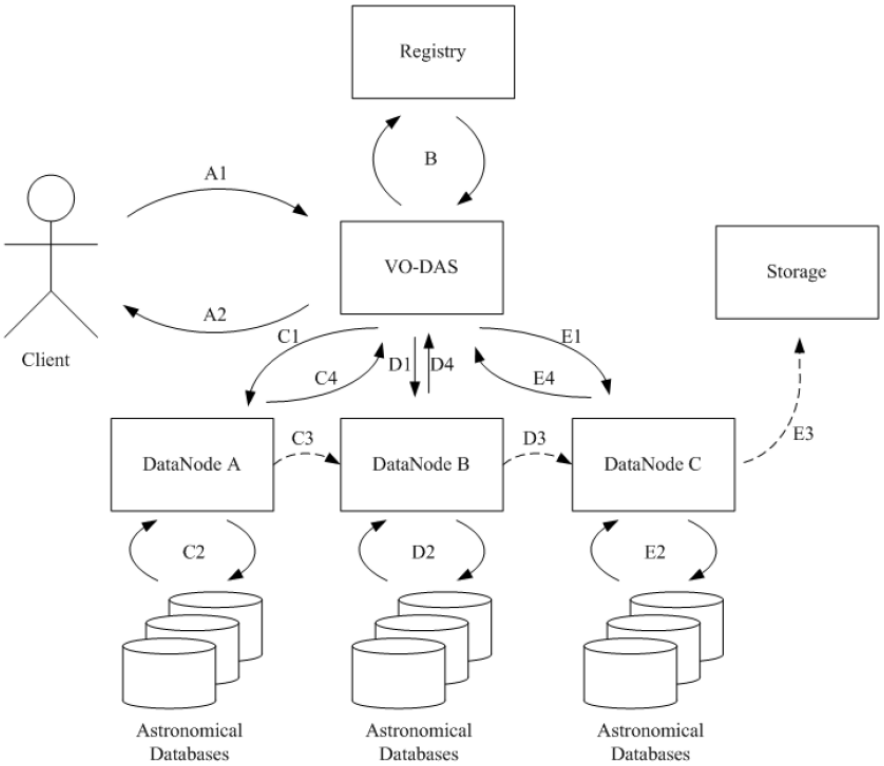

图 2.7: 异地数据联合异步查询

2.4 VO-DAS的总体设计

在对系统进行了分析之后我们对VO-DAS给出一个总体的设计和描述。这个总体设计不再是面向功能的而是面向实现的，也就是说我们会在本节构建一个可以实现的VO-DAS结构。

2.4.1 VO-DAS的框架

图2.8中描述的VO-DAS的总体结构，它可以分成五个模块。图中上半部分最大的模块是VO-DAS服务器的结构，最下面的模块是DataNode的结构。中间的两个模块分别是VO Registry和数据存储服务器（图中以VOSpace作为例子）。此外还有一个没有在此图中明确表示出来的客户端模块。下面我们分别对这些模块给予说明。

VO-DAS服务器是VO-DAS系统的核心模块，它起到承上启下和调度的功能。为了能够支持异步查询功能，我们采用WSRF构架的Web Service方式。因而VO-DAS服务器必建立在支持WSRF的中间件上面。我们采用的中间件是Globus Toolkit 4.0 Java WS Core。这是一个纯Java的服务平台，和其他的网格服务有相对的独立性，既可以独立运行，也可以安装到一个网络服务的容器（如Tomcat）下运行。为了能够提供安全的服务，VO-DAS需要有用户身份认证的机制，因此在GT4 WS Core之上是Authorization模块。认证功能也使用GT4自身的功能模块。由于VO-DAS会与DataNode、Registry以及VOSpace通信，因此需要在其内部有对应这三个部分的客户端模块，分别是OGSA-DAI Client 、Registry Proxy 和VOSpace Client。OGSA-DAI Client是一个标准的OGSA-DAI调用客户端模块，负责将SQL查询和结果数据处理整理成OGSA-DAI支持的Request对象，然后将请求送达指定的OGSA-DAI服务。Registry Proxy（详细的设计类图见图2.9）负责向指定的Registry查询网络中有哪些DataNode，获得这些DataNode的元数据信息，并保存在DataResourceMap对象中。DataResourceMap是一个数据结构，它是一个DataResource对象的数组。每个DataResource对象代表一个数据资源。DataNode可以包含一个或多个Resource，具体数目由DataNode的元数据XML文件中Resource的数目决定。每个Resource元素对应DataReourceMap数组中的Resource对象（详细的设计类图见图2.10）。VOSpace Client目前是一个FTP客户端，负责和存储服务器VOSpace交换数据文件。

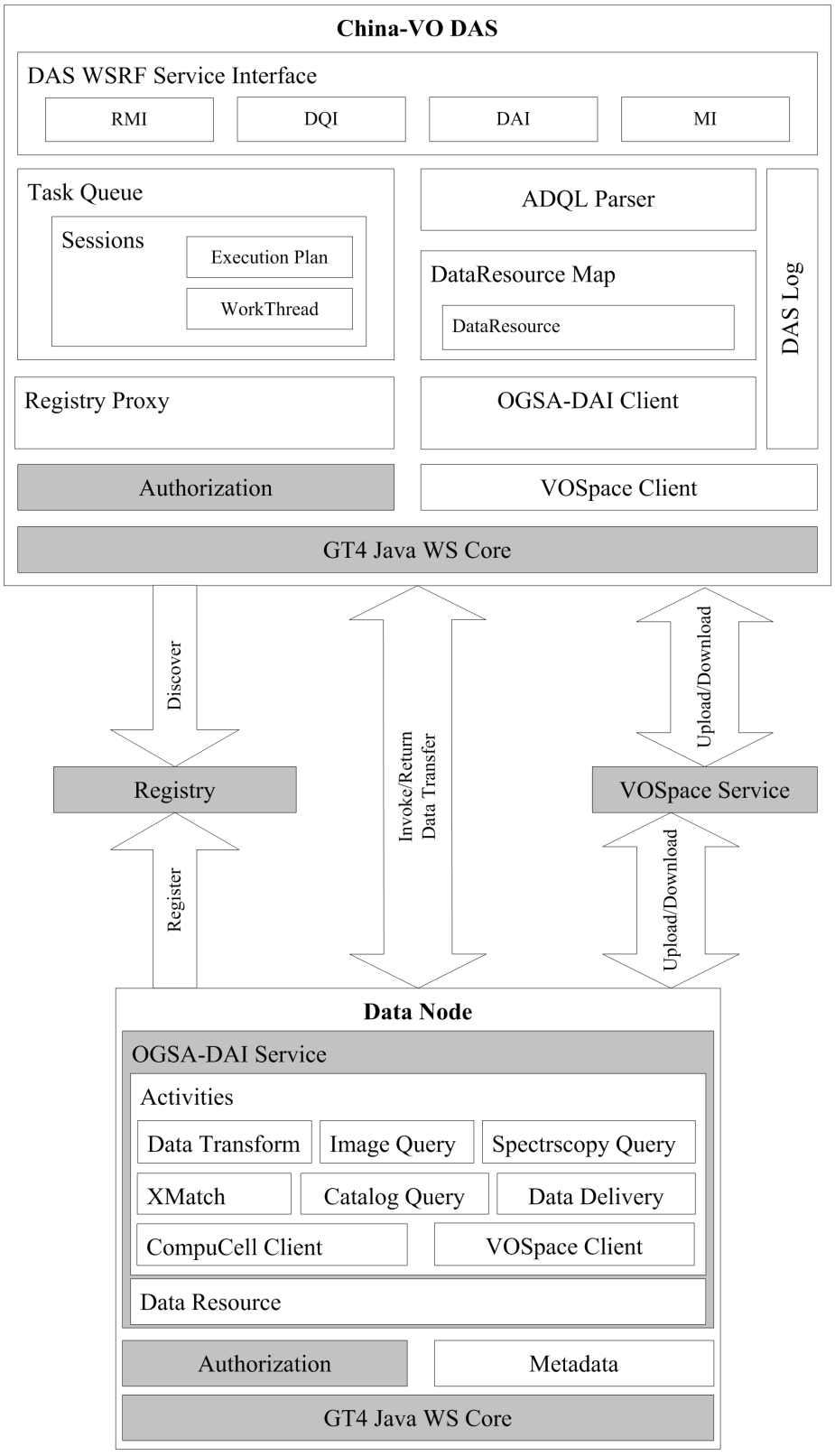

图 2.8: VO-DAS的设计结构

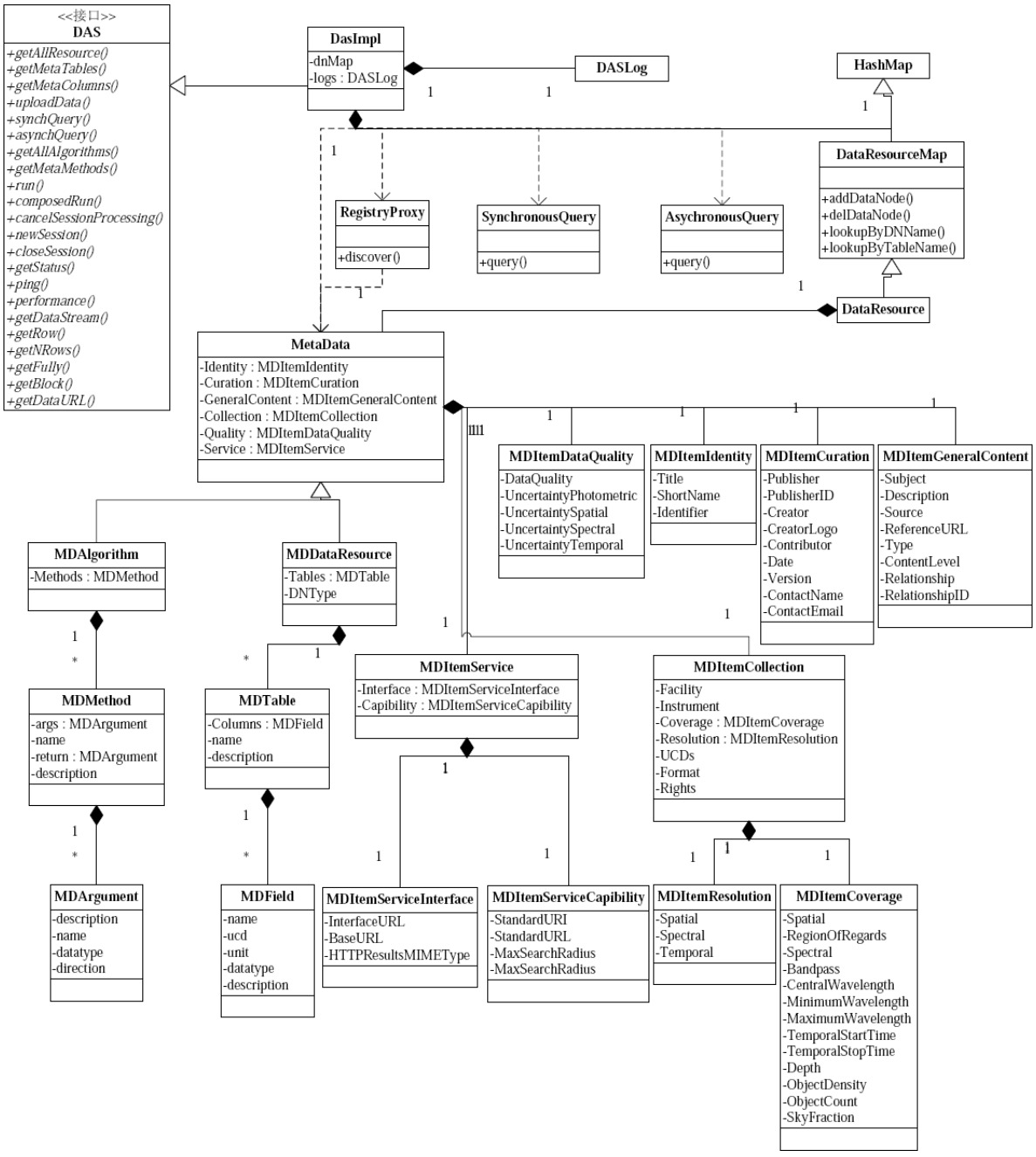

图 2.9: VO-DAS服务器的Registry Proxy模块的设计类图

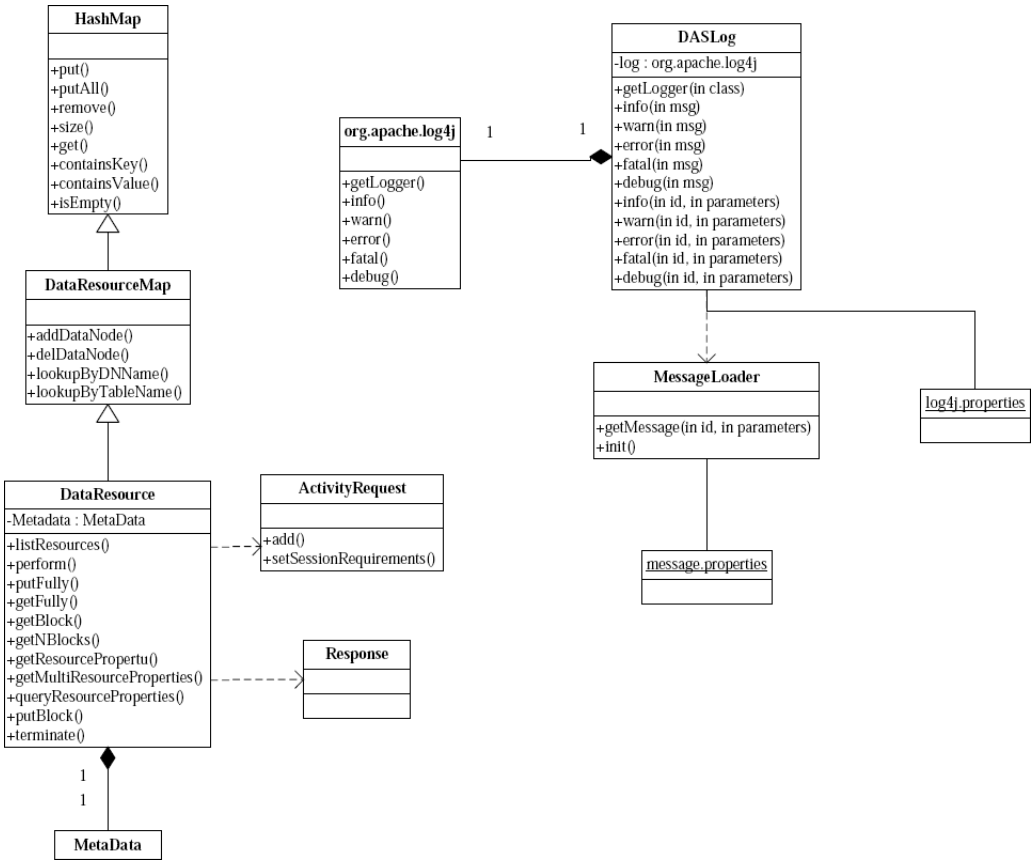

图 2.10: VO-DAS服务器的DataResourceMap和DASLog模块的设计类图

VO-DAS服务器提供给客户端访问的接口有四类：资源元数据接口（RMI）、数据查询接口（DQI）、数据存取接口（DAI）和管理接口（MI），它们的具体定义将在第2.4.5中详细介绍。

在定义VO-DAS功能的时候，我们要求VO-DAS支持ADQL语言的查询。由于DataNode中的OGSA-DAI只支持SQL查询而不支持ADQL，我们需要将客户端传递来的ADQL转换成功能一致的SQL。这个工作要在VO-DAS里面完成，完成这项任务的模块是ADQLParser。ADQLParser支持ADQL ver0.9版本的语法，同时支持我们所做的对ADQL语义的扩展以增加图像和光谱的查询。由于ADQLParser是一个十分复杂的模块，因此对它的设计也进行得十分详尽。在第2.4.3节中我们会描述更多细节。

VO-DAS服务器具有任务调度功能，这项功能的实现是在Task Queue模块中。Task Queue为提交来的每个任务创建一个session对象，所有的执行都是

在session对象内完成的。Task Queue管理将一个session对象组成的任务等待队列和一个任务执行队列，以及一个任务完成队列。任务首先停留在等待队列中，当有空闲资源的时候，它会被调度到正在执行队列，并分配一个工作线程给它。在得到工作线程以后，session会调用ADQL Parser解析ADQL语言并创建一个Execution Plan以向指定DataNode发送查询请求。当查询工作完成以后，session会从正在执行队列中转移到完成队列。

由于各个模块之间的相互联系复杂，系统行为不确定性很大，因此需要有日志管理的功能模块用来记录系统的操作细节。这既便于查询系统历史上所作的操作，方便维护，也有利于开发周期内的调试和测试。这项功能由DASLog完成（详细的设计类图见图2.10）。

DataNode基本沿用了OGSA-DAI的结构，但为了我们的一些特殊功能做了扩充。首先，Catalog Query就使用OGSA-DAI提供的数据库查询功能实现。同时增加XMatch，即交叉证认功能模块。其次，增加Image Query和Spectrum Query两个模块实现图像和光谱查询的功能扩充。Data Transform会支持天文数据文件的特定格式，包括VOTable和常用的ASCII文件。Data Delivery和VOSpace Client联系VOSpace将查询结果文件发送到指定的存储服务器上。CompuCell Client是为数据挖掘功能预留的接口，CompuCell将在第三章中进行描述。

Registry和存储服务器不需要我们实现，而是采用现成的产品。Registry使用AstroGrid的注册服务在本地建立的一个拷贝，而存储服务器目前采用FTP服务器，最终实现IVOA的MySpace协议[19]。

[19]http://www.ivoa.net/Documents/cover/VOSpace-20070626.html

2.4.2　VO-DAS服务器的执行逻辑

WSRF是带状态的Web Service，状态在WSRF中通过Resource来实现。本节中出现的Resource这个词均专指WSRF的资源，而不是VO-DAS中的天文数据资源。VO-DAS的客户端接口是通过WSRF的WS形式发布出来的，它们的状态通过session对象来记录，这样才能够完成异步查询的过程。session的实现正是通过WSRF中的Resource来完成的。图2.11描述了这个实现的过程。这个过程是一个Factory/Instance模式。SessionHome专门用来管理所有session的一个对象，客户端通过工厂服务DASFactoryService间接调用SessionHome对象创建一个新的session对象。DASService是VO-DAS通过WS接口对外提供远程操作，客户端在通过DASService进行远程操作的时候，状态信息全部保存在对应的session对象内。

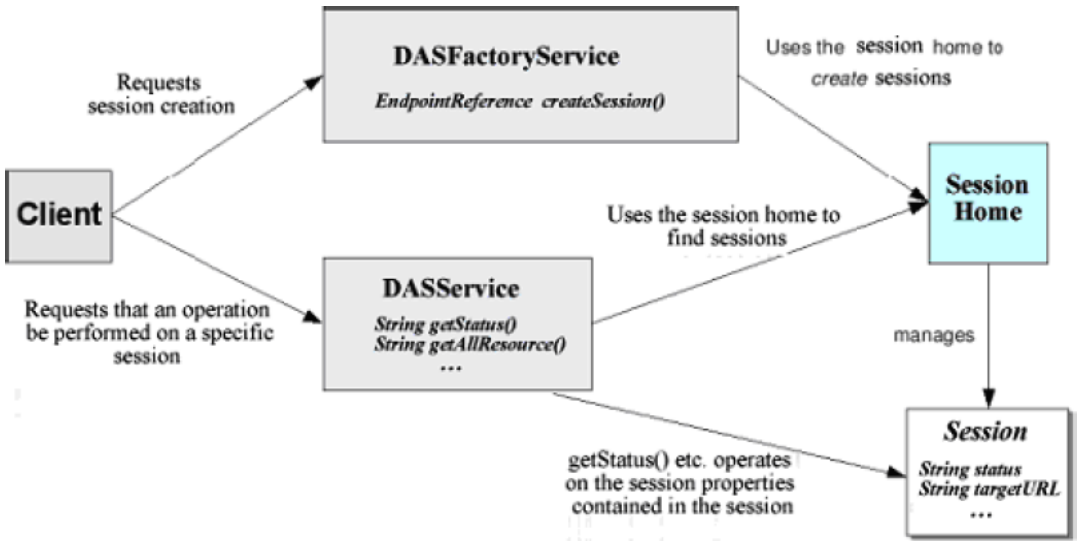

图 2.11: VO-DAS的session实现过程

图2.8中设计的VO-DAS服务器内的对象存在三类生命周期：全局对象的生命周期、非全局对象的生命周期和session的生命周期。全局对象包括维持Task Queue运行的所有对象：各个队列，以及调度任务的监控线程（LoopThread）。此外，DataResourceMap也是一个全局对象。非全局对象主要是指工作线程WorkThread。session生命周期控制Execution Plan对象，每个session的状态对象以及查询结果的URL对象。

全局生命周期的对象在VO-DAS启动伊始就被创建，直到VO-DAS服务器关闭才终结。当LoopThread发现有正在排队的任务和能够处理这个任务的空闲WorkThread的时候，WorkThread会为这个任务重新激活开始一个暂时的生命周期。上述这两种生命周期都相对比较容易控制，需要重点讨论的是session的生命周期。

前面提到的两类生命周期都是由VO-DAS服务器自身管理的。Session的创建却是由客户端发起的。为了统一处理起见，无论是同步查询还是异步查询，VO-DAS都会为这个任务在服务器上创建一个session对象。只不过对于客户端而言，异步查询的session是显式地创建的，而同步查询的session是隐式创建的。Session的结束时机比较复杂，在它创建的时候会有一个预设的最长生命周期，通常在VO-DAS服务器中把这个周期定义为一个星期。当一个session创建了一个星期，用户还没有主动删除释放这个session的时候，系统会强制删除这个session，同时和这个sessin相关的数据存放的URL也消失了；当一个session到了一个星期，还没有执行完毕的时候，系统会再为其延长一段周期，直到它完成或出错。

对于同步查询，查询结果返回给客户端以后，VO-DAS服务器就会主动删除这个对应的session对象。对于异步查询，正常情况下应该由客户端主动删除session对象。当session相关的查询过程中出现错误并发生异常的时候，系统应该主动删除session，以免发生资源耗尽的情况[39]。

有关VO-DAS服务器的执行逻辑模块的设计类图参见图2.12。

2.4.3 ADQL解析器

ADQLParser或称为ADQL解析器是一个ADQL语法解释模块，它的功能是把ADQL字符串转换成一个等效的Java对象树。这个对象树可以被很容易的处理并转换成SQL，或根据涉及的数据资源的不同，分拆成多个发送到不同DataNode的SQL（如图2.7所描述的情况就需要拆分ADQL）。

ADQL解析过程可以分成几个不同的层次。最基本的层次是词法扫描。词法扫描是从字符串中找到所有有意义的符号（token）。第二个层次是根据这些符号的组合找到ADQL的各个分句，如SELECT分句、FROM分句、WHERE分句等。最后，将这些结果全部创建成Java对象树。

词法扫描过程被设计成一个有限状态机，根据输入的字符序列分析得到

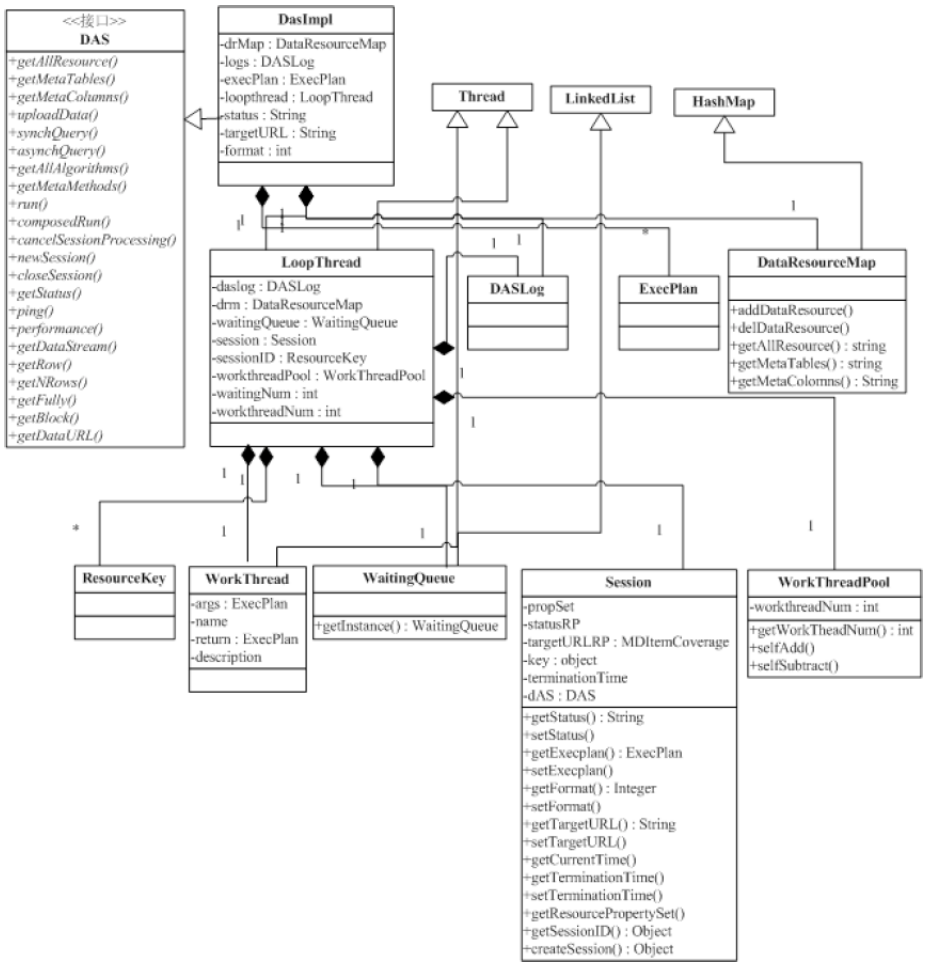

图 2.12: VO-DAS服务器的执行逻辑模块的设计类图

不同的token类型。图2.13中列出了所有token类型，这些类型均根据ADQL的语法定义[20]而设置，包括DBObject（最基本的对象，用于表达表名和列名），ADQLKeyword（ADQL的关键字，由ADQL语法定义决定），ADQLOperator（ADQL规定的操作符），Integer（整型常量），Real（实数常量），Function（函数，包括函数的参数），Variable（变量），Expression（表达式），ArchiveTableItem（表名称，综合了两个DBObject，一个表示资源名称一个表示表名称），TableColumn（列名称，由DBObject组成）等。词法扫描的状态转移图将在附录A.1中详细给出。

在词法扫描进行的同时，对每个解析出来的token进行ADQL的语法解析，

[20]http://www.ivoa.net/Documents/latest/ADQL.html

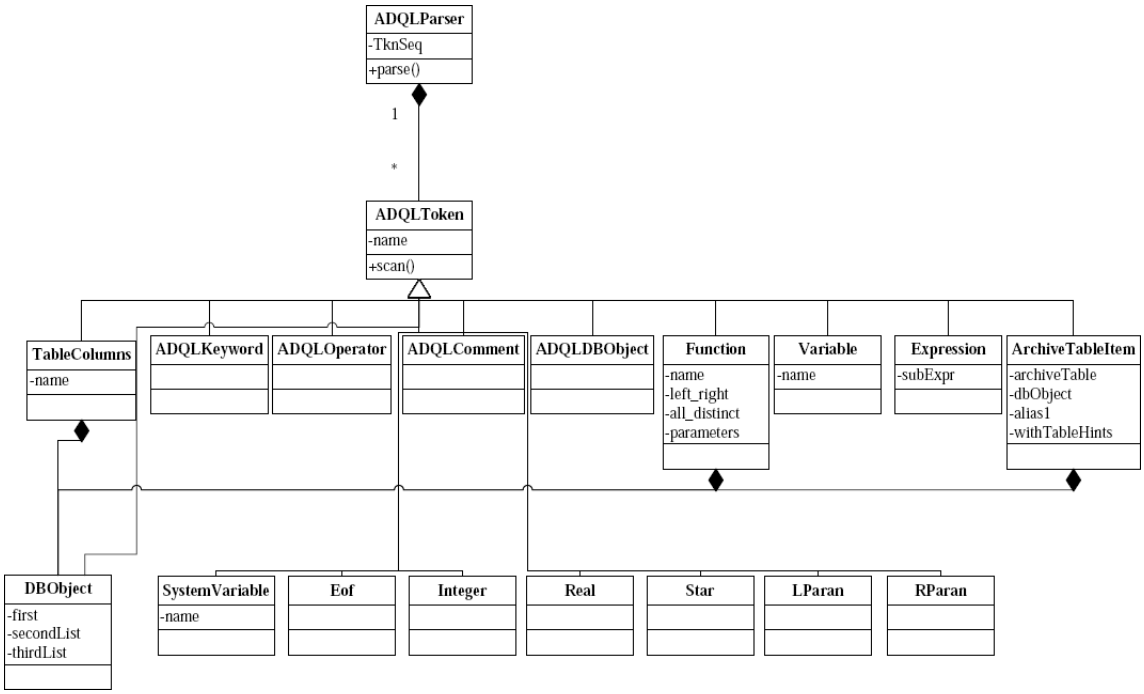

图 2.13: ADQLParser的设计类图之一

也就是按照ADQL定义的各个分句组装各自的Java对象。类SelectStatement是ADQL解析后生成的Java对象的树根类，即所有分句都是这个类的成员类。图2.14 －2.19描述了SelectStatement和它的所有成员类之间的关系。这里最重要的成员是QueryExpression类，它又包括QuerySpecification类，后者包括SelectClause（SELECT分句的对应类）、From类（FROM分句的对应类）、Where类（WHERE分句的对应类）等。SelectStatement还提供了获取各个分句对应对象的方法，包括getFrom 、getSelectClause 、getWhere等。图2.20 定义了语法解析过程，首先从token中寻找是否有SELECT关键字，然后生成SELECT分句对应的Java对象SelectClause。然后寻找token中是否后续的是FROM分句，如果是则开始解析并生成FROM分句对象From。接着检查后续token是否WHERE关键字，如果是则继续解析WHERE分句并生成Where对象。类似地，在WHERE分句解析以后再解析GROUP BY和HAVE两个分句。最后三个分句的解析是可选的，即允许ADQL字符串中不包括这几个分句。每个分句内部的具体解析流程参见附录A.2。

在ADQL语法解析过程中还有一个关键环节就是表达式的解析。ADQL包

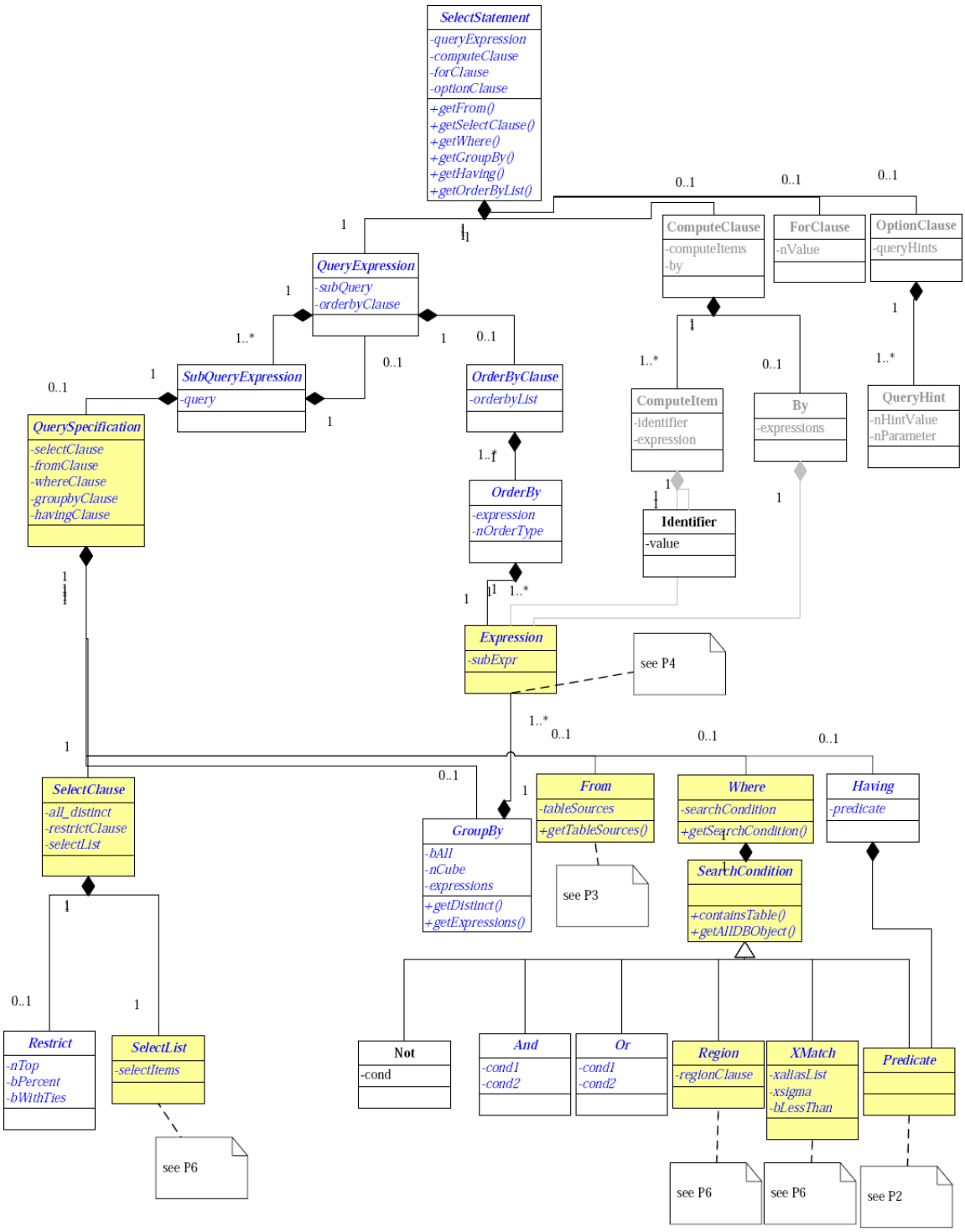

图 2.14: ADQLParser的设计类图之二

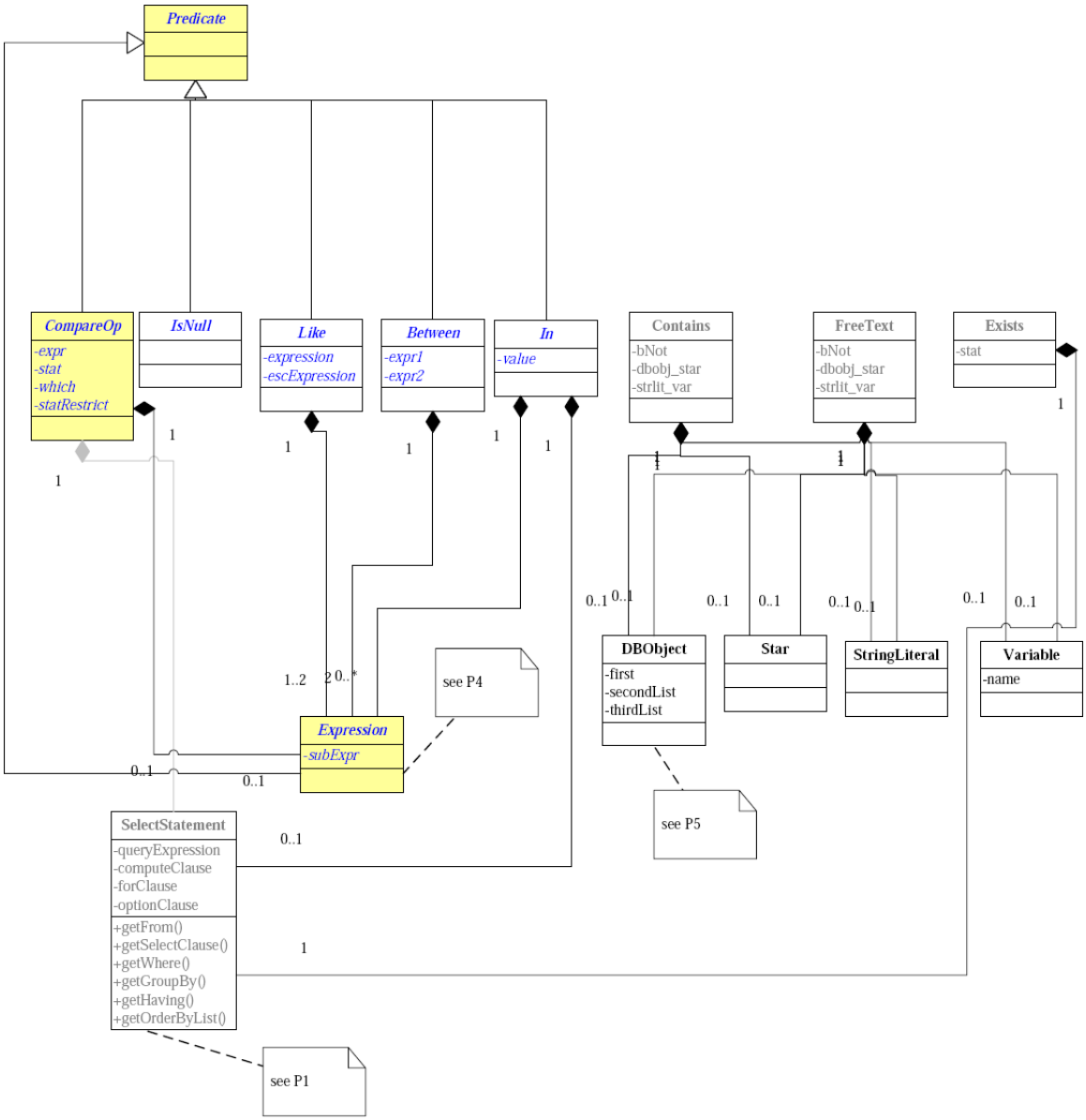

图 2.15: ADQLParser的设计类图之三

含两类表达式，一种是通常的算术表达式，另一种是仅仅出现在WHERE分句中的条件表达式。前者相对简单，是一个典型的操作符中置的二叉树建树过程。后者比较复杂，因为它的操作符除了二元操作符AND和OR以外还有一元操作符NOT，更甚之，还有IS NULL，LIKE等类似于函数的操作符。而最复杂的是NOT和后面两个类型的操作符的组合：IS NOT NULL，NOT LIKE。这种

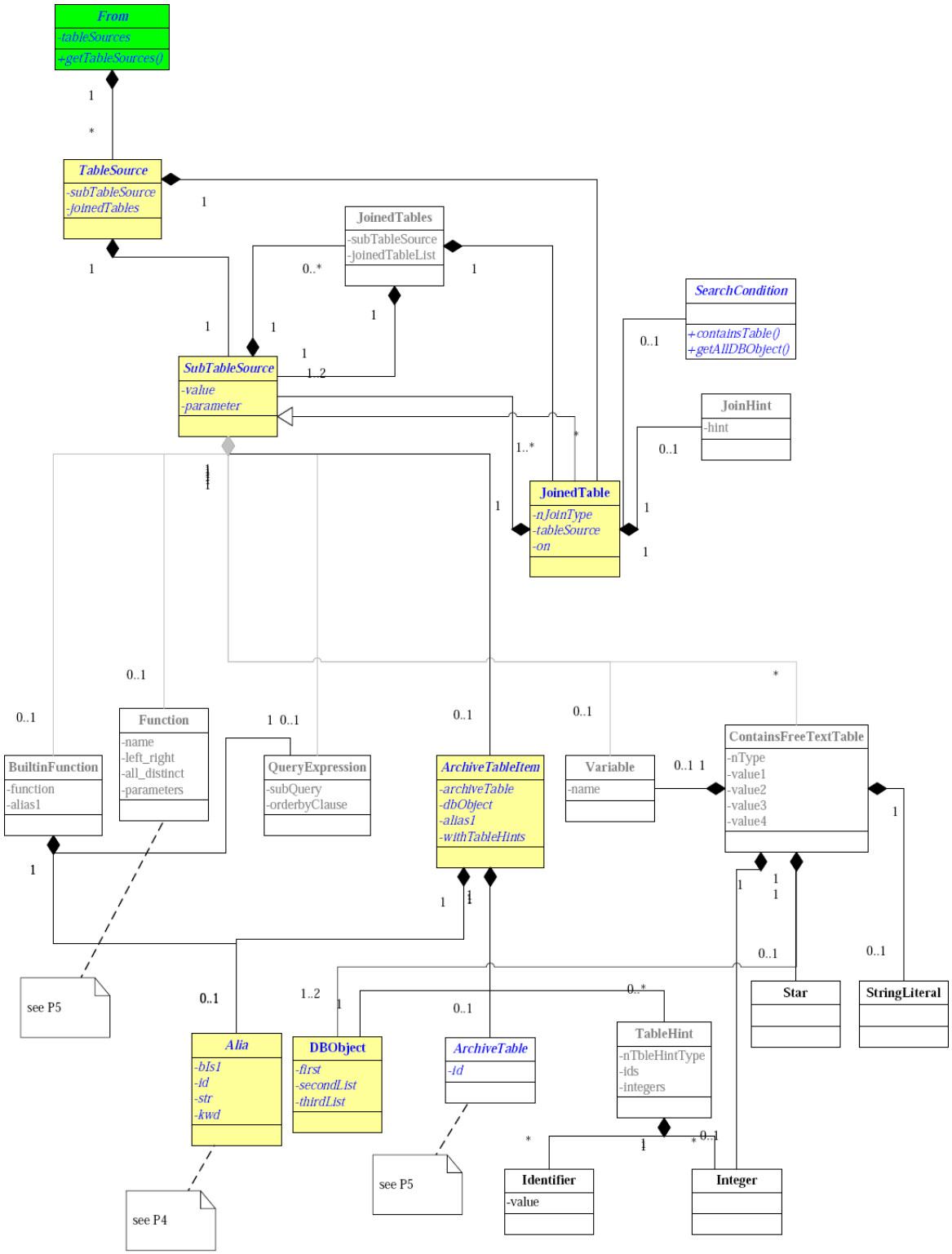

图 2.16: ADQLParser的设计类图之四

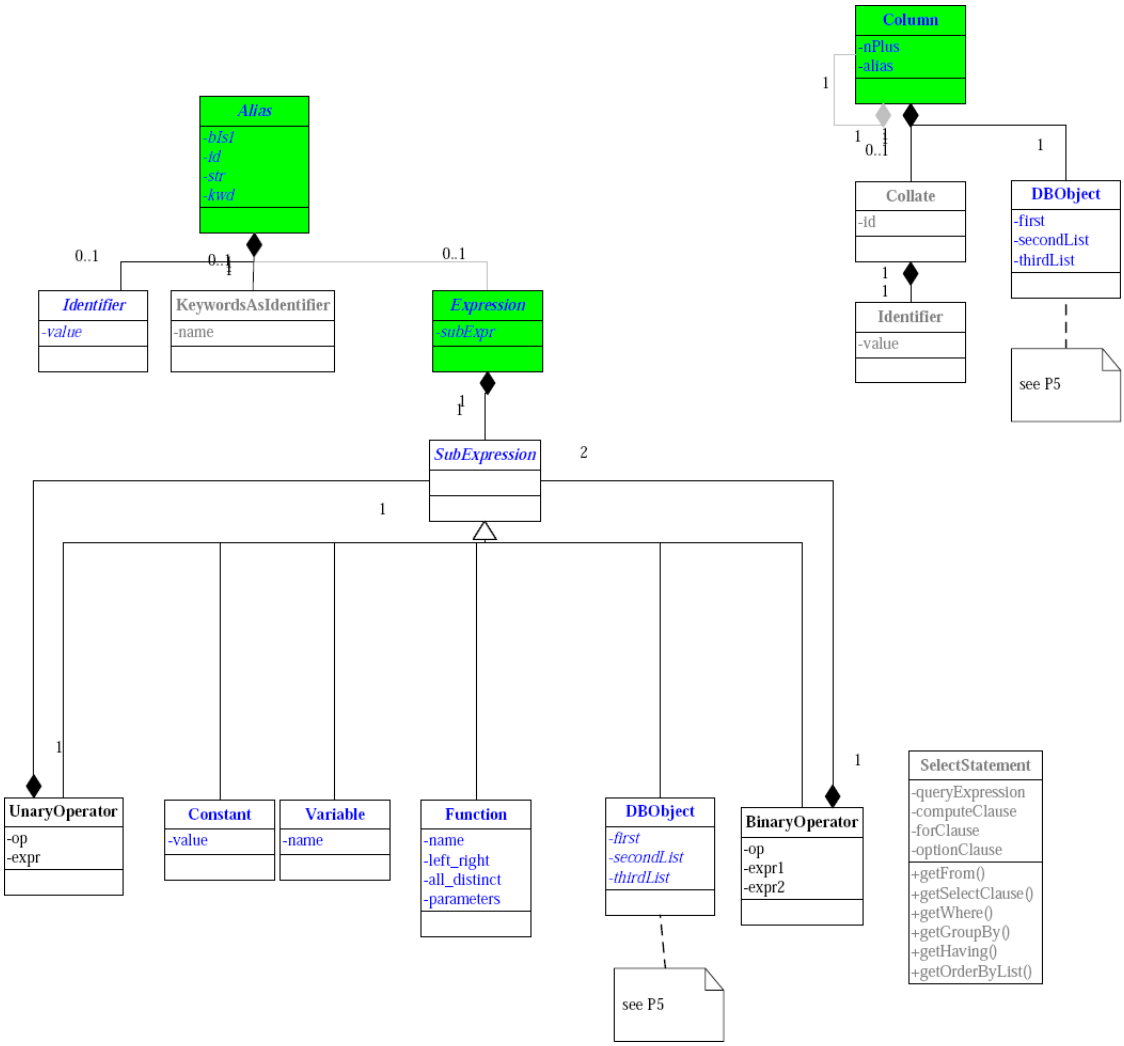

图 2.17: ADQLParser的设计类图之五

接近自然语言的表达式的解析过程比较复杂，需要将各种情况都考虑周到。在附录A.3中我们给出了算术表达式解析的状态转移图，在附录A.4中我们给出了WHERE分句条件表达式的状态迁移图。

2.4.4 ExecPlan执行计划

本节内容基于刘超和高丹在VO-DAS项目中的工作。ExecPlan执行计划模块的目的是把ADQLParser 模块根据ADQL 生成出的SelectStatement对象转换成可以在DataNode 上面执行的数据查询操作。DataNode 的对

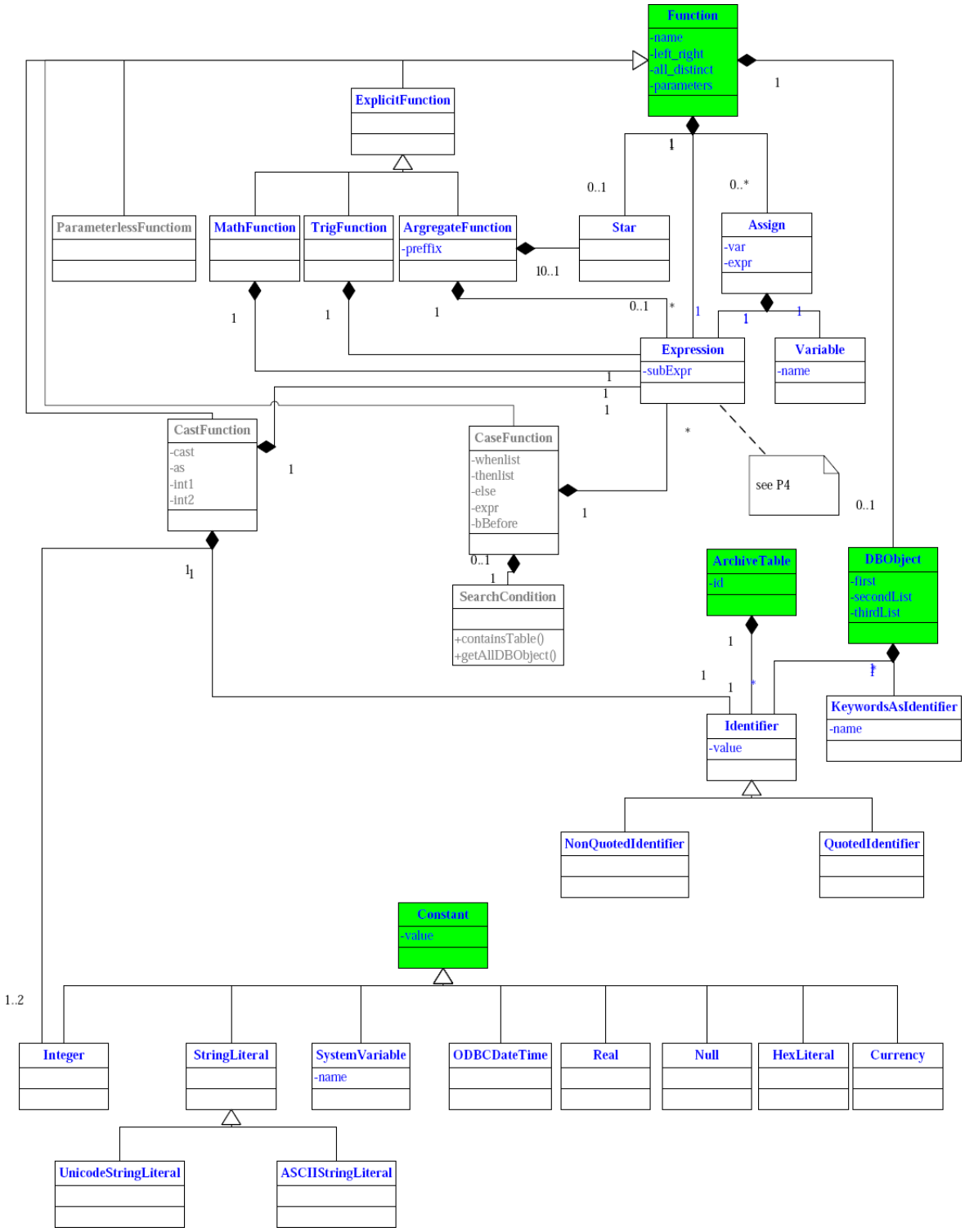

图 2.18: ADQLParser的设计类图之六

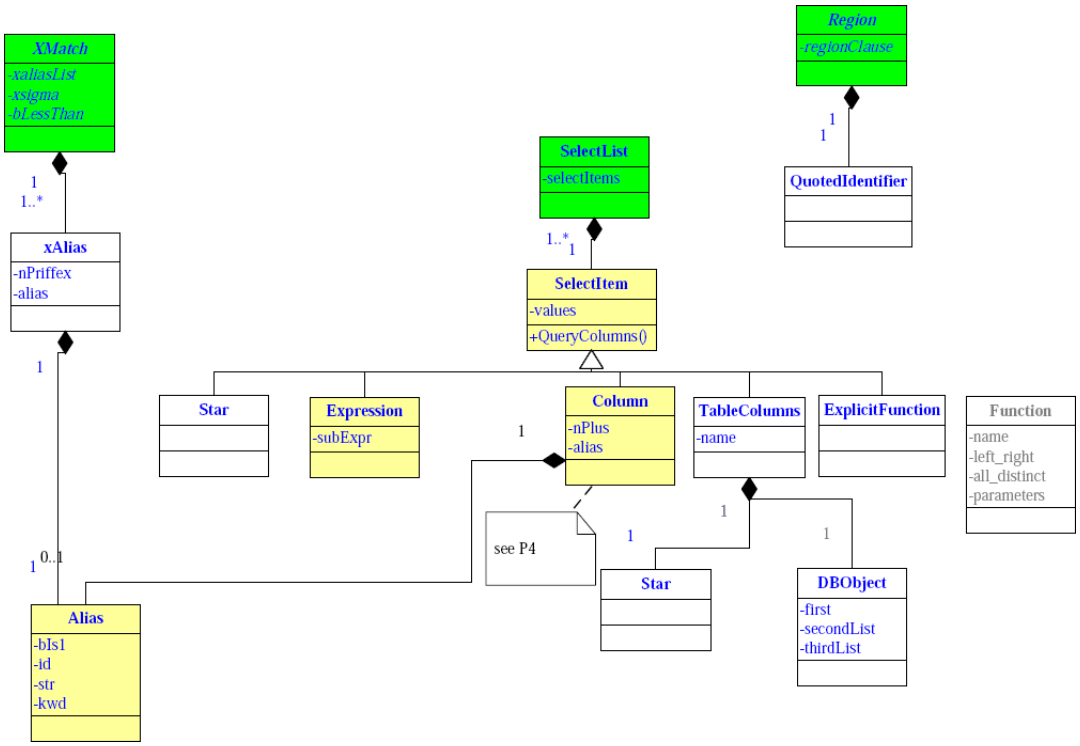

图 2.19: ADQLParser的设计类图之七

外接口就是OGSA-DAI 形式的，因此ExecPlan就是要生成一个OGSA-DAI的ActivityRequest 对象，其中不仅包含SQL查询语句，还包括查询结果文件的转换格式和保存的位置。

这里需要再一次详细描述针对ADQL中涉及多个数据资源的处理方案。假设一个ADQL中涉及了两个不同的数据资源中的表，那么ExecPlan会首先为这两个数据资源的查询次序进行排序，数据量较小的表先查询，数据量较大的表后查询。这样的原则保证了查询间歇两个数据资源中间数据传输量尽可能的小。查询第一个数据资源以后会产生一个中间临时查询结果，这个结果会直接传递给第二个数据资源，后者会使用一张临时表存储这个中间结果。然后在第二个数据资源的DataNode之上完成中间临时表和第二个数据资源的星表之间的联合查询或交叉证认，联合查询的结果作为最终结果输出。中间临时表在查询以后被删除。出现两个以上数据资源在一个ADQL的情况是上述过程的一个简单推广。我们在下面详细描述ExecPlan的动作过程的时候会讨论这个过程是怎样具体实现的。

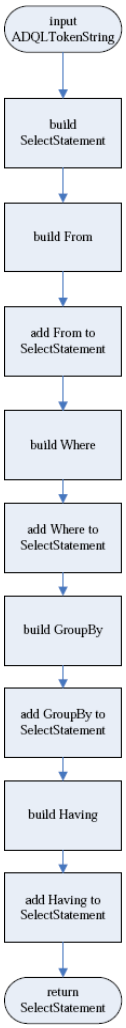

图 2.20: ADQLParser的语法解析流程图

ExecPlan对象为了生成最终的ActivityRequest对象，首先需要将ADQL的SelectStatement 对象转换成一个或多个SQL语句（如果涉及多个DataNode就会拆成多个SQL语句）。在这个过程中，ExecPlan会从SelectStatement对象中列出所涉及的资源有哪些，并检查DataResourceMap列表中是否有这些资源，如果有，则说明这个ADQL可以在VO-DAS上执行，否则说明没有ADQL要求的数据资源，不能继续执行数据查询。如果确认了所有数据资源都存在，则ExecPlan会接着进行SQL生成的工作，其中最核心的是进行名字替换和函数替换。这里指的函数是ADQL所特有的锥形检索函数Region和交叉证认函数XMATCH。它们由于没有SQL中对应的函数，所以需要转换成SQL的一个条件串。另外，一

旦出现XMATCH就意味着至少有两个数据资源参与其中，这个时候一定会对ADQL进行分拆。

ADQL的分拆遵循以下原则。首先，将SELECT中的查询列依据不同的数据资源来源分组。其次将WHERE条件依据不同的数据资源来源进行分组。如果一个条件涉及到了多个数据资源，则在先查询的数据资源对应的SQL中没有这个条件，而后进行查询的数据资源相应的SQL查询语句中会有这个条件。REGION条件则出现在所有拆分的SQL中。在完成分组以后就进行拆分。首先组装传递给第一个数据资源的SQL语句，这里面的SELECT和WHERE分句中应该只含有和当前数据资源相关的项目。在FROM分句中也只有第一个数据资源的表。组装后面的SQL语句的时候假设前面一个的查询已经结束，并自定义一个临时表名，相当于第一个数据资源的查询结果。第二个SQL语句中，SELECT分句中的列来自两个表，一个是第一次查询的结果生成的临时表，另一个是第二个数据资源中的表。在WHERE语句中仅有和第二个数据资源相关的条件以及REGION条件和XMATCH条件。其中XMATCH条件原来的第一个资源的表名字已经替换成了临时表的。同样，在FROM分句中也只有临时表和第二个资源相关的表并列放置在一起。这样的SQL语句能够保证在第二个数据资源所在的DataNode上进行的SQL查询完成了一个带有交叉证认的查询操作。ExecPlan的执行计划生成的具体过程由图2.21－2.23描述。

完成了SQL的生成与拆分以后，为了生成各个相关的DataNode都要执行的ActivityRequest对象，还需要完成数据格式转换和数据传输的设置工作。如果涉及多个DataNode，那么中间结果的数据会直接传递给后面的DataNode，而不是传递到MySpace存储服务器上。只有最后一个DataNode的查询结果才会传输到存储服务器上。如果ExecPlan处理的ADQL最终需要向多个DataNodwe依次查询，那么最终生成的ActivityRequest也有多个，每个DataNode有一个。实际执行的时候，是由Task Queue模块来完成向各个DataNode传递和执行对应的ActivityRequest对象的工作的。

2.4.5 VO-DAS的客户端接口

在上一节中我们提到VO-DAS的客户端接口有四类，下面按照这四类的顺序依次介绍它们的具体定义。

RMI用于获得VO-DAS发现的所有资源的元数据。它提供了如表2.8中所描

表 2.8: RMI接口定义

接口名称	参数	返回值	说明
GetAllResources		String	获得VOTable形式的所有数据资源元数据
GetMetaTables	String	String	获得指定资源名称的表元数据
GetMetaColumns	String(2)	String	获得指定资源指定表名的列元数据

表 2.9: DQI接口定义

接口名称	参数	返回值	说明
SynQuery	String(2)	Dataset	同步查询数据，查询结果直接返回
AsynQuery	String(3)+session		异步查询，查询结果不通过这个方法返回

表 2.10: DAI接口定义

接口名称	参数	返回值	说明
GetTargetURL	session	URL	获得查询结果文件的URL
GetQueryResult	session	Dataset	获得查询结果数据

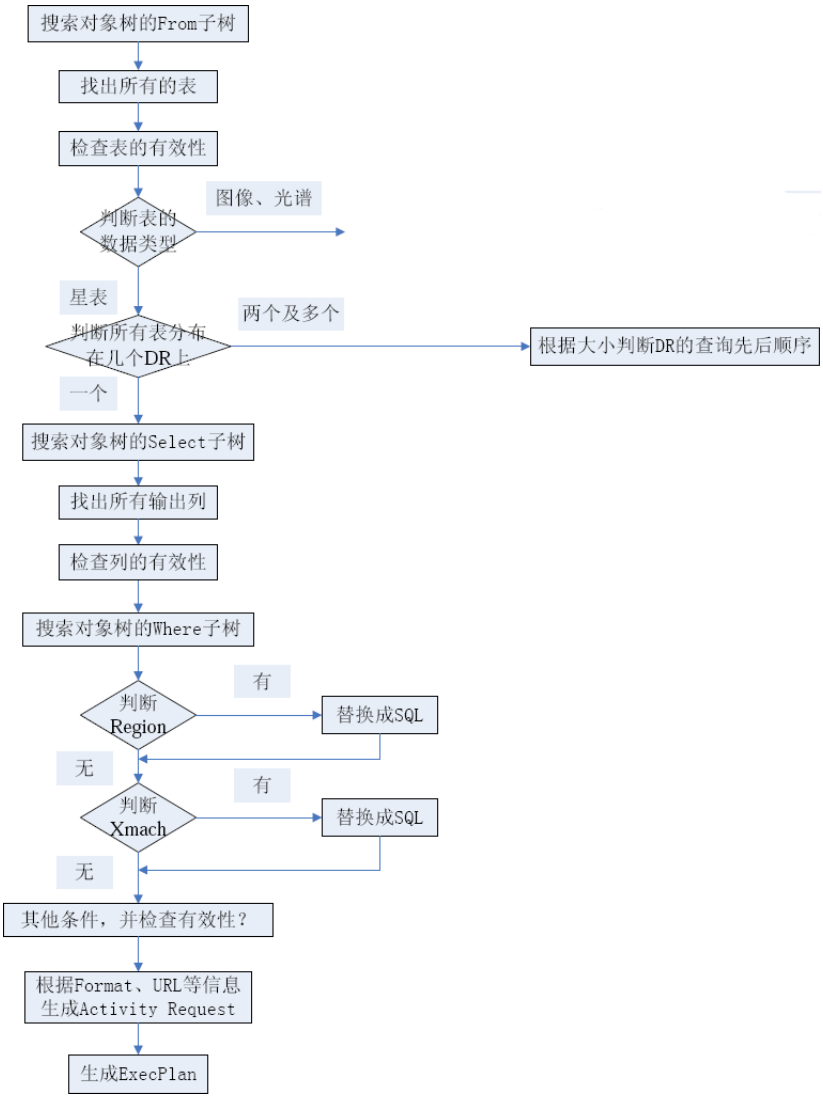

图 2.21: ExecPlan的执行计划流程图之一

述的接口方法。DQI用于发送数据查询请求，其接口方法如表2.9所述。DAI用于访问数据，它的相关方法列在表2.10中。MI是一个管理接口，涉及的方法列在表2.11中。

一般而言，客户端程序首先会使用RMI提供的方法获得资源元数据。这些元数据有助于了解编制ADQL所需要的星表名称，表中各个列的名称和含义。如果客户端程序需要执行同步查询，则调用SynQuery方法。这个方法在调用以后就被挂住，知道查询结果返回回来。因此它虽然使用简单，但是只能访问非常

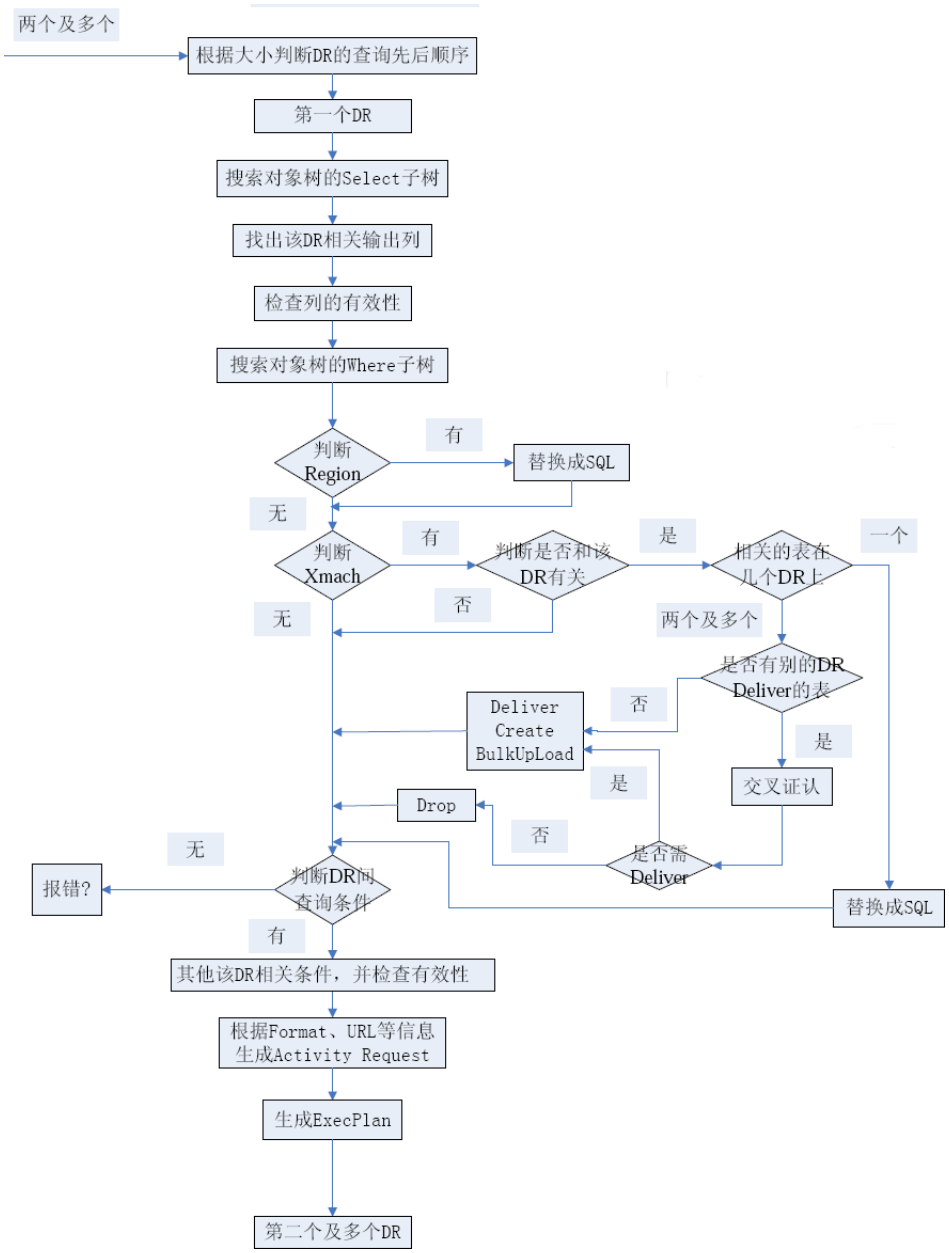

图 2.22: ExecPlan的执行计划流程图之二

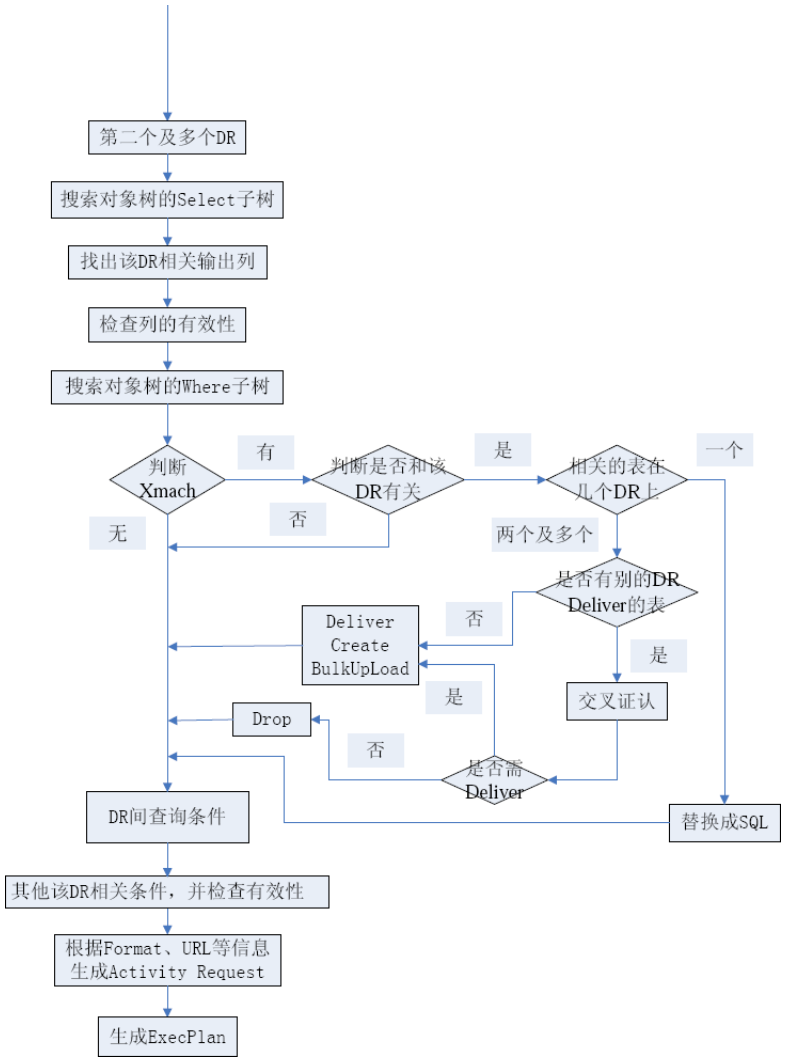

图 2.23: ExecPlan的执行计划流程图之三

少的数据。如果客户端程序要执行异步查询，那么它必须遵循一定的执行步骤：

- 创建session：Session sess=StartSession()
- 执行异步查询：AsynQuery(ADQLstring, dataFormat,sess)
- 查询执行状态：String stat=GetStatus(sess)
- 完成查询以后，取得查询结果URL：URL url=GetTargetURL(sess)
- 完成所有工作，释放session：DestroySession(sess)

表 2.11: MI接口定义

接口名称	参数	返回值	说明
StartSession		session	开始一个新的session
DestroySession	session		释放一个session
GetStatus	session		获得任务当前的状态
GetStartTime	session		查询任务开始时间
GetEndTime	session		查询任务结束时间
GetSubmitime	session		查询任务提交时间

以上介绍的各个接口都是WSRF接口，客户端程序必须要有GT4 Java WS Core的JAR包的支持才可以使用这些方法。因此，这些接口只能应用在Java编写的客户端程序中。在后面介绍VO-DAS客户端的小节里，我们将介绍如何让其它类型的客户端程序也能够连接到VO-DAS。

2.5　VO-DAS的客户端形式

在第2.4.5节中我们定义了Web Service形式的客户端接口。这是最基本的客户端形式，但是需要使用Java语言进行网络编程才能够实现。为了扩大VO-DAS的使用群，最大限度的发挥它的能力，我们定义多种不同形式的客户端。下面分别对这些客户端进行定义和说明，其中GUI客户端的设计来自杨阳和刘超，实现来自杨阳；命令行客户端的设计来自刘超，实现来自杨阳和田海俊。

2.5.1　GUI客户端

GUI客户端是采用Java实现的操作系统无关的图形界面客户端。它提供了简单易用的窗口界面，在连接上一个VO-DAS服务器以后，它会自动从服务器端下载所有数据资源的元数据，并采用树形结构显示出来。它提供了一个简单ADQL编辑环境，编辑好的ADQL只要按下一个按钮就可以将任务送到服务器端。在GUI客户端中，用户可以选择同步查询还是异步查询。如果是同步查询，那么数据会直接返回客户端，需要用户将数据保存到本地文件中。用户也可

以选择通过PLASTIC协议[21]将数据传递到其他的VO工具软件中，例如Aladin, Topcat等。如果是异步查询，那么提交查询请求的同时，用户需要为这次任务起一个容易识别的名字。任务提交以后，在窗口的下方有一个监控窗，监控正在执行的查询任务当前的状态。客户端窗口可以关闭，并不影响任务的执行；重新打开窗口以后，仍然能够继续监控任务的进度。当异步查询的任务结束以后，可以从历史信息中找到异步查询的结果文件URL。用户可以通过FTP客户端将数据下载到本地。

GUI客户端非常适合数据访问频率不是很大，对计算机操作系统不是很熟悉的用户。

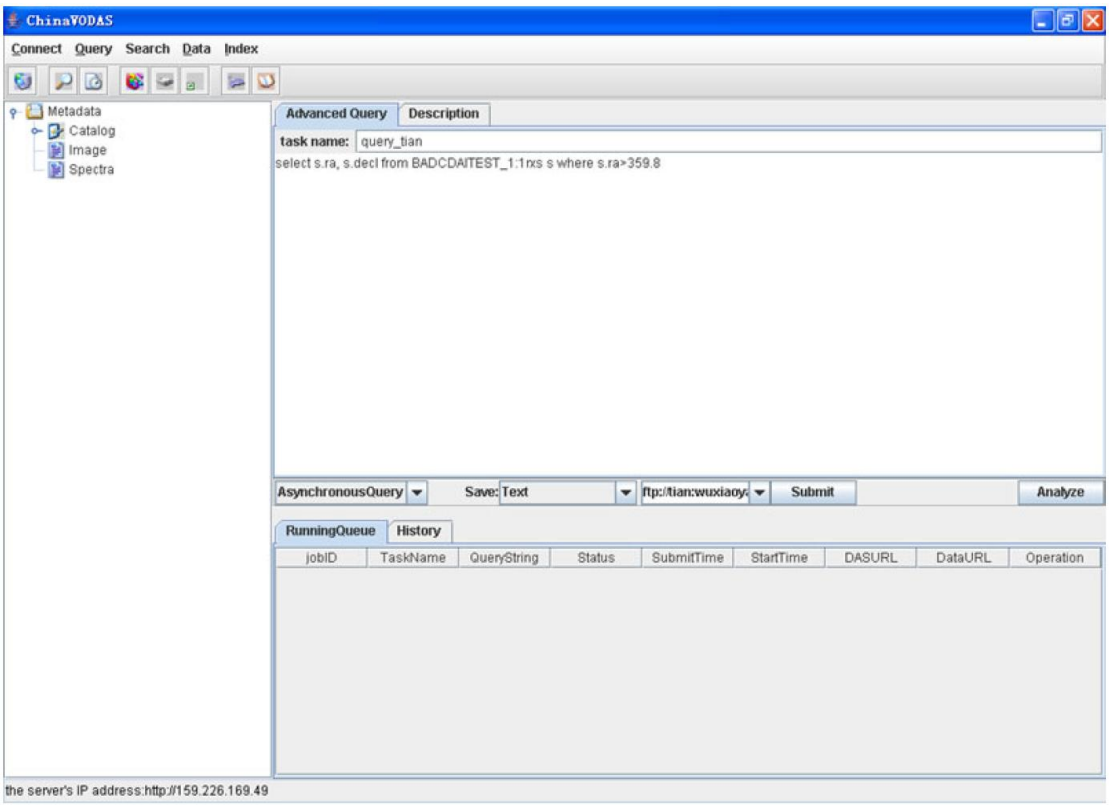

图 2.24: GUI客户端的界面

2.5.2 命令行客户端

命令行客户端提供了一组基本的操作命令，用户不需要图形终端就可以通

[21]http://www.ivoa.net/Documents/Notes/Plastic

过这些命令完成数据的查询。它们是运行在linux环境下的一组shell命令，在后台调用另外一组支持Web Service客户端接口的Java程序完成各个数据查询操作。表2.12中列出了这些命令的细节。诸如VO-DAS服务器地址等配置信息会存放在一个配置文件里而不需要每次执行查询命令的时候都输入。 命令行方式

表 2.12: VO-DAS命令行客户端的命令集

命令	参数	说明
md		从VO-DAS服务器下载数据资源的元数据XML文件
asyn	ADQL文件，文件格式，session文件名	向VO-DAS服务器提交一个异步查询任务，同时这个任务相关的session会保存在参数中指定的session文件中
syn	ADQL文件，文件格式	提交一个同步查询任务，查询结果以字符串形式输出到标准输出上
jobmon	session文件名	查询一个已经提交的异步查询任务的执行状态，返回结果输出到标准输出上
dataurl	session文件名	获得一个已经完成的异步查询的结果数据位置，返回信息输出到标准输出上
close	session文件名	关闭一个session

可以集成到用户自己的程序里面，无论程序是Python还是FORTRAN，都能轻易地通过这组命令完成频繁的数据访问。这种类型的客户端适合访问数据比较频繁或者需要使用自己编写的程序完成数据访问任务的用户。需要说明的是，经过一层封装以后，用户自己编写调用这些命令的程序中并不需要包含Web Service服务支持库。这样，即便用户不了解什么是Web Service也能够在其程序中使用VO-DAS了。

2.5.3 MATLAB客户端

由于MATLAB支持直接调用Java程序包（具体参见第3.4节），因此第2.4.5节所定义的Java程序接口可以直接被MATLAB所支持。在此基础上，通过使用M语言进行一定的封装，我们又得到了MATLAB上面运行的客户端。这个客户端的内容和命令行客户端十分类似。所不同的是，shell命令变成了M语言（参见表2.13）。此外，由于在MATLAB中不便于使用配置文件，因此配置参数也是通过一个M语言实现的。

由于MATLAB还同时支持GUI，因此在提供了这些查询数据的命令函数以后，再将它们集成到一个MATLAB的GUI中。这样，在MATLAB中既支持了命令方式的查询，也支持了GUI。在MATLAB中实现VO-DAS的客户端的最大优点就是它可以将数据查询操作和数据挖掘操作在一个软件之内无缝地连接起来。关于这一点我们将在下面一章里展开来讨论，并介绍我们在MATLAB之上完成的实验开发。

2.5.4 网页客户端

另一类重要的客户端形式就是网页表单形式的客户端。这类客户端的用户群和第一类GUI客户端的很相似。网页客户端使用更加简单，不需要安装任何应用程序，只要用网络浏览器连接到服务器上就可以了。但是由于网络浏览器的独特工作原理，这种类型的浏览器支持异步查询比较困难。

2.6 VO-DAS的实现和测试

VO-DAS的主要实现工作集中在VO-DAS服务器上，只有少量的工作是在DataNode之上，全部开发都是采用Java语言完成的。初期版本的开发完成了ADQL解析的基本语法部分、GUI客户端和命令行客户端、VO-DAS服务器、DataNode的星表查询和交叉证认等功能，其他功能将在后续版本中实现。VO-DAS的运行环境如表2.14所示。

实际测试环境中，VO-DAS环境搭建在6台台式机之上。其中两台VO-DAS服务器，四台DataNode （其中一台和一个VO-DAS服务器共用一台计算机）以及一台存储服务器。客户端分别运行在Windows XP和Linux环境中。测试内容包括ADQL解析器的正确性测试、VO-DAS调度正确性测试、客户端接口测试、运行稳定性测试、大数据两访问测试等。 ADQL解析器经过测试

表 2.13: VO-DAS MATLAB客户端的命令集

命令	参数	说明
activevodas		启动VO-DAS客户端，装载vodas.m文件进入MATLAB环境，这个文件中保存了VO-DAS服务器的地址
md		返回VO-DAS服务器下载数据资源的元数据，是一个MATLAB struct类型的对象
asyn	ADQL字符串，文件格式	向VO-DAS服务器提交一个异步查询任务，同时这个任务相关的session会作为返回值，session返回值是一个struct对象
syn	ADQL字符串，文件格式	提交一个同步查询任务，查询结果以struct的类型返回，struct中包含结果的元数据信息以及一个cell array存放数据。之所以不选择matrix类型返回是因为这种结构既包含返回结果的元数据，又支持返回结果中有非数值类型的列
jobmon	session对象	查询一个已经提交的异步查询任务的执行状态，状态字符串作返回值
dataurl	session对象	获得一个已经完成的异步查询的结果数据位置，并使用FTP将数据下载到MATLAB的workspace，结果数据的数据类型和syn命令得到的一样
close	session对象	关闭一个session

表 2.14: VO-DAS的运行环境

VO-DAS服务器	RedHat Linux或Fedora Core 4以上 JDK 1.5 Tomcat 5.0 GT4 WS Core MySQL 5.0
DataNode	RedHat Linux或Fedora Core 4以上 JDK 1.5 GT4 WS Core OGSA-DAI WSRF 2.2 MySQL 5.0

发现，基本的ADQL语法解释正确，但是复杂的语法，如使用TOP 、GROUP BY、ORDER BY等尚未实现，而WHERE条件非常复杂的时候ADQL解析还不稳定。对VO-DAS调度的测试证实WSRF上的任务调度基本达到要求。客户端的接口测试在不同操作系统下均达到设计要求，但GUI客户端的工作还不是很稳定。DataNode的注册和发现工作尚未能达到设计要求，只能通过静态添加方式实现VO-DAS对DataNode的认知。

运行稳定性测试采用一条能够返回1000行左右的ADQL查询语句通过多个客户端连续不停向服务器发出请求，检查服务器的承受能力。实际测试发现，高负荷的不稳定来源于OGSA-DAI的不稳定性。大数据量访问受限于DataNode所在服务器的内存（2GB)，一般能够达到1M行的数据访问。

我们还测试了系统的访问性能。我们采用程序自动生成查询ADQL，对一个拥有7000多万行数据的数据库进行连续查询，并将查询时间记录下来，同直接使用本地MySQL客户端进行查询的时间比较（图2.25)。查询返回的行数从几万行一直到1百多万行。经过测试发现，除了开始的时候，VO-DAS因一系列初始化工作而花费了很长时间以外，其它访问时间基本上是一个常数。这和本地访问MySQL数据库的花销按查询结果的量成线性增长的趋势不同。这说明在VO-DAS中，即便是一百万行一级的数据访问，主要花销是网格服务带来的，而不是数据库访问产生的。

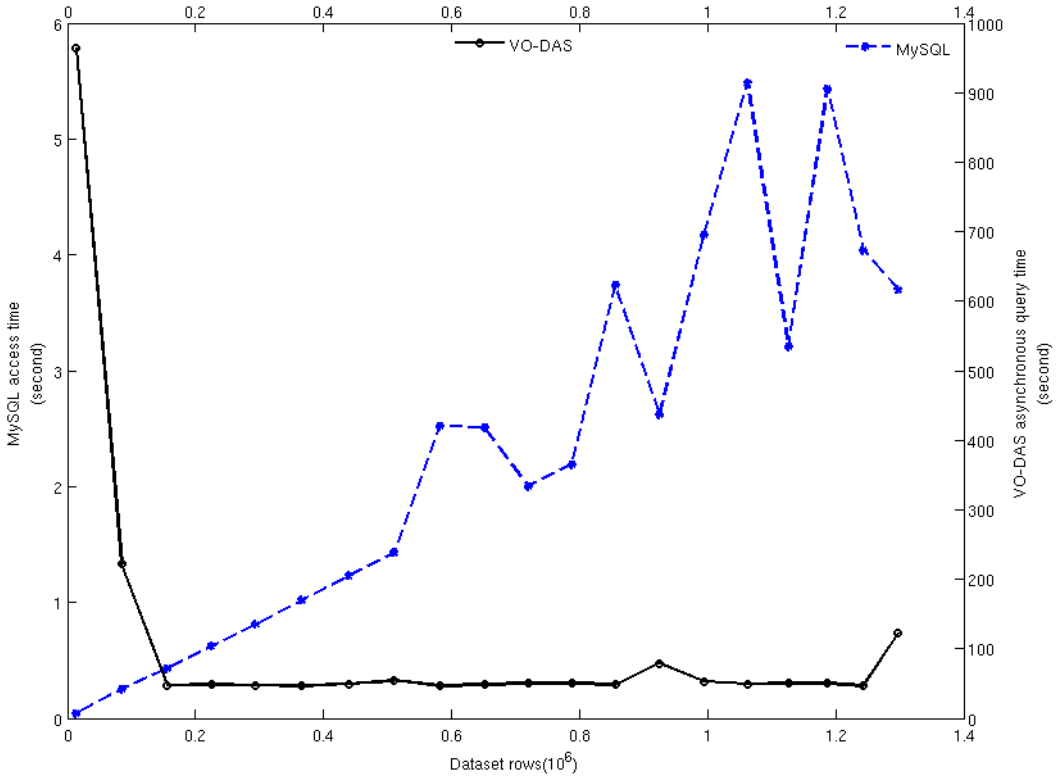

图 2.25: VO-DAS数据访问的性能

测试结果表明，VO-DAS的初步版本可以在内部小范围内使用，还需要在增加更多功能以后，并让系统更加稳定以后再正式发布。

第三章　China-VO数据挖掘工具的设计与实现

我们将在本章讨论各种可能的天文数据挖掘方案。在第1.4节中，我们列出了三种可能的天文数据挖掘工具方案。在本章中我们将依次讨论这些可能的选择，并最终给出我们的意见。对天文数据挖掘工具的研究是一件浩繁的工作，不仅需要对各种数据挖掘算法有大致的了解，还需要对各种现有的数据挖掘工具有较多了解和使用经验。最重要的是，必须使用这些工具进行实际的研究工作，才可能得到最深切的体会和认识。限于时间和经验，我们只能在实际的科研中应用很少几种挖掘工具，积累少量使用的经验。对于不曾使用过的挖掘工具，我们尽可能找到别人的使用经验，尽量作出完整、客观的评价。

本章将分成三个部分，第3.1节介绍现存的几种主要的数据挖掘工具以及它们在天文研究中的应用，并尽量给出我们的评价。在第3.2、3.3节中我们自己定义了一种新的数据挖掘工作语言JDL，并围绕JDL开发出的原型来讨论其可行性。JDL方案可以封装已经存在的各种数据挖掘算法并利用网格环境实现虚拟天文台基础上的数据挖掘。在第3.4节中，我们讨论最后一种可能的选择，在一种已经存在的数据挖掘工具MATLAB之上进行二次开发，增加互操作性，使它可以和现存的虚拟天文台工具协同工作。在第3.5节，通过讨论对上述三种方案给出评价。

3.1　天文数据挖掘相关的编程语言

使用何种数据挖掘工具进行天文学研究是一个开放的问题，答案不是确定的，这决定于研究的性质、团队的知识背景、网络和计算机的基础设施、项目的资金情况等情况。我们在本节对几种常见工具的讨论是忽略了上面的这些前提条件，只是从功能、性能、使用的方便性，与VO的互操作可能性等，从纯粹的软件角度展开的。

3.1.1　FORTRAN

FORTRAN 语言是世界上第一个被正式推广使用的计算机高级语言。FORTRAN 是FORmula TRANslation的缩写，即”公式翻译”的意思。顾名思义，这

种语言适宜科学计算，用它写成的程序中的表达式与数学公式的形式很相似。FORTRAN语言自1954年提出以后，至今已有50多年历史，但仍经久不衰，其间历经多种版本的演变。FORTRAN语言始终是科学计算领域首选的计算机高级语言。

进入70年代后，随着计算机应用的发展，人们已开始感到原有的FORTRAN已不能满足实际使用的要求。为此，美国标准化协会(ANSI)在1976年对ANSI FORTRAN(X3.9－1966)进行了修订，吸收了一些厂商各自扩充的行之有效的功能，同时又增加了一些新的特征，新标准定名为FORTRAN77。1980年，FORTRAN 77被接受为国际标准，它是国内外广泛流行的一个版本。1990年3月，ISO和ANSI双重批准了Fortran语言的最新国际标准，定名为Fortran 90。Fortran 90对FORTRAN 77的主要扩充有:

1. 自由形式的源程序形式，不再受老式面向卡片输入的固定栏目布局限制；
2. 模块化数据与过程定义机制，这样提供了一种数据与过程包装的强有力的而又安全的形式；
3. 六种内部数据类型中派生出用户定义的数据类型；
4. 数组操作机制；
5. 指针机制，允许创建与操作动态数据结构；
6. 数据类型参数化，允许使用多种字符类型，满足各国字符处理的需要；
7. 提供了过程的递归调用机制；
8. 提供了附加的控制结构，如do……enddo、do while等。

由此可以看出，Fortran 90已经是具有强大数值计算能力的现代高级语言，程序的书写更趋结构化[46]。Fortran 95是Fortran90的一个小版本升级，在保持向下兼容的语法的同时增加了并行计算的能力，适合在高性能计算领域的使用。

FORTRAN语言在天文学研究中得到了大范围的普及，几乎每一位研究人员都学习使用过，其高效率的计算和强大的数值处理能力为天文数据分析带来了极大的方便。经过多年的发展，已经有很多适合天文应用的FORTRAN程

序库了，例如基本的数值计算可以使用GSL[1]、IMSL[2]库函数。前者是GNU版权的免费软件，后者是商业化产品，价格根据不同的操作系统从数千美元到一万美元。GSL支持线性代数和向量、矩阵计算、统计、差值、优化算法、FFT、方程求解、数值差分、数值积分、求解微分方程等功能。IMSL支持线性代数、差值计算、积分与差分、微分方程、数学变换优化、统计、时间序列分析等。可以看出两个计算函数库的功能大致相同。还有专门用于数据可视化的FORTRAN库，如PGPlot。很多天文学应用软件都是采用FORTRAN编写的，例如测光用的DAOPHOT，用于N体模拟的各种软件。但是在对虚拟天文台标准的支持方面，除了使用FORTRAN读写FITS文件的库以外还没有见到更多。

简而言之，FORTRAN是一种高级编程语言，可以完成几乎所有可能的功能。但是，由于它对于网络、数据库没有特别方便的支持，没有在虚拟天文台互操作性上提供更多支持，没有集中的高级算法库完成数据挖掘工作，很多算法零散分散在不同的组织（大学、研究机构）中，因此完成一个基于FORTRAN的数据挖掘工具，并能够保证和虚拟天文台有无缝的连接，这是一个非常严俊的任务，需要大量投入人力、财力，作为一项专门的大型项目才可以完成。

3.1.2　Perl

Perl的全称是Practical Extraction and Report Language，它是1987年由Larry Wall 设计的，到今天已经发展到6.0版本。Perl 的设计目标是帮助UNIX 用户完成一些常见的任务，这些任务对于Shell 来说过于沉重或对移植性要求过于严格。Perl 语言中包含了C、C++、shell,script、sed、awk 这几个语言的语法，它最初的目的就是用来取代UNIX 中sed/awk 与脚本语言的组合，用来汇整信息，产生报表。它采用GNU和Artictic两种许可证方式释放，均为自由软件许可证。

Perl语言语法简洁实用，高效而完整。它是一种解释性的语言，但是它又引入了模块的设计思想，这使它的功能可以变得越来越强大。Perl的优势并不在于数值计算，而是在字符串处理方面，因此普遍应用在网站建设方面。

天文学上的Perl应用见诸Jenness et al. (1999)[47]。在夏威夷的Joint Astronomy Center，很多工作都是借助Perl来完成的，其中包括消息系统（Messaging Systems ADAM and DRAMA，主要用于JCMT的望远镜控制和数据处理），文

[1]http://www.gnu.org/software/gsl/
[2]http://www.vni.com/products/imsl/

件I/O库，天文库等系统的接口；它甚至还被用于JCMT望远镜的数据获取和数据处理。

通过使用CFITIO（由Pete Ratzlaff编写的FITS文件读写库），Perl可以对FITS文件进行各种操作。使用PDL（Perl Data Language）模块可以令Perl具备了理解N体模拟的结果的能力，并且可以调用PGPLOT、OpenGL等接口实现数据可视化。和PGPLOT的连接是通过pgperl模块实现的[48]，通过这个模块perl可以调用所有PGPLOT的子程序绘图。

无论和FORTRAN相比还是和后面将要继续介绍的几个工具相比，Perl的数值计算能力是最弱的。只有和其他已有的程序库连接起来，才能够让Perl具有数值计算能力。这已经不是Perl本身提供的功能了，而是把Perl作为封装工具了。

3.1.3 Python

和Perl一样，Python也是一种自由软件式的解释型编程语言。所不同的是，Python采用了面向对象的结构。它是由Guido van Rossum在1989年圣诞节期间开发的，它的语法融合了C++, Java和Shell的特点，很好地弥合了从Shell到常规编程语言之间的鸿沟，十分适合非程序员使用。诞生至今，它被广泛应用于网络开发、图形、文档处理、游戏开发、科学计算、移动通信应用、嵌入式开发和数据库开发等领域。

1999年前后，Python进入天文学家的视野[49]，使用它开发了一些天文计算模块。其后，越来越多的天文项目开始使用Python作为开发工具。通过python-numpy模块，Python获得了数值计算能力。和Perl一样，这些数值计算模块都是被封装的而不是用Python独立开发的。类似的模块还有SciPy 、ScientificPython 、MatPy 、Sparse Py 、stat.py等。特别的，Matfunc是一个纯Python的矩阵计算和曲线拟合的模块。针对FITS文件的处理需求，有pyFITS 、pCFITSIO 、Qfits 、FITS等模块。关于数据可视化，有ppPGPlot和gracePlot.py等。

3.1.4 R

R是一种用于统计计算和绘图的语言及其计算环境，由Bell实验室开发[3]。R也是一个GNU项目，可以免费使用。R具有强大的统计功能，可以完

[3]http://www.r-project.org/

成线性和非线性建模、经典统计检验、时间序列分析、分类、聚类等计算。同时它也具备良好的可扩展性，可以实时调用C 、C++和FORTRAN的代码。在CRAN（Comprehensive R Archive Network）网站上可以下载很多R的扩展模块，这些模块提供了丰富的现代统计学算法。R语言还提供了和数据库连接、SOAP协议等的接口。

R在天文学研究领域的应用，Center for Astrostatistics，Penn State University是最主要的推动者。他们已经提供了很多用于天文学统计的模块[4]。这些模块包括的算法有bootstrap resampling、相关系数、统计分布、经验分布检验、线性代数和方程求解、最大似然估计、方差分析、多元聚类分析、神经网络非线性最小二乘回归、排序、空间分析、统计检验、样条、生存分析、时间序列分析等。

3.1.5　IDL

IDL的全称是Interactive Data Language，它是一个商业化的数值计算和数据可视化语言，由RSI公司[5]开发并投向市场。它在需要数据可视化的领域：航天、军事、医学、地球科学、天文学等领域有广泛应用。RSI还以IDL语言为平台针对特别的应用领域开发了专门的产品，如高级遥感影响处理系统（ENVI ）、数字地形与河流网系分析系统（RiverTools ），科学数据的管理系统（Noesys ）和交互式网上三维数据发布系统（ION ）[50]。

IDL的语法也非常简单，很多方面和FORTRAN很相似。但与FORTRAN不同的是IDL有很强的向量和矩阵的描述能力，这为数值处理提供了很好的条件。它既支持交互式操作也支持过程操作。由于IDL也属于解释语言，因此和前面介绍的Perl、Python一样，执行效率并不高。IDL也支持模块方式的功能扩展，IMSL包以及很多天文应用功能都是这样添加到IDL中的。IDL的特点是数据可视化，可以绘制各种二维、三维的图形，此外它在图像处理领域也有较强优势。IDL不仅有命令行方式的使用界面，用户还可以根据自己的要求定制自己的GUI。IDL由于是面向数据的语言，因此有访问数据库的功能，目前可以通过ODBC接口实现对Oracle、SyBase等数据库系统的访问。IDL可以实现和其它编程语言的集成，它可以和Java程序实现互相调用，也可以通过COM实现和Windows程序的通信。但是除了Java，它不能在UNIX或Linux之下实现和其它编程语言的集成。

[4]http://astrostatistics.psu.edu/statcodes/

[5]http://www.ittvis.com

在天文学领域，NASA在多个太空项目中使用了IDL语言来处理数据，如HST和探索火星的计划。Sloan数字化巡天项目也采用IDL来编写光谱处理的pipeline。IDL的天文的应用模块比较多，仅以NASA提供的模块[6]为例，包括常用工具函数集、测光工具（DAOPHOT ）、数据库模块、磁盘读写模块、FITS头的天体测量和世界坐标系统、图像处理、FITS文件读写、数学和统计模块、绘图模块、健壮性统计模块、星表读写模块、网络套接字模块（web socket procedure），TV播放模块等。可以看到，IDL丰富的天文常用模块已经可以支持正常的天文应用，当然如果需要进行数据挖掘工作，上述模块还仅仅是最基本的，还需要大量算法模块以及高效的针对分布数据库的海量数据访问模块。

3.1.6 MATLAB简介

MATLAB发展自1970年代Cleve Moler 为计算机系的大学生写的小课件，1984年MathWorks 公司[7]成立，MATLAB成为了商业化软件，到1990年代，由于它在控制领域出色的应用使其成为控制工业事实上的行业标准，到2007年MATLAB已经发展到了7.4版本。使用MATLAB的领域有航空、军事、自动化、生物和医药、通讯、计算机、电子、金融、半导体等。

MATLAB的编程语言被称为M语言，是一种面向对象的高级编程语言。它的运算全部采用矩阵运算，使它在描述各种数值算法的时候形式十分简洁，而运行效率也因此高于普通的解释语言。和IDL一样，MATLAB提供了一个集成工作平台，M语言可以是交互式执行的，也可以写成函数，通常函数方式的执行效率更高。MATLAB也支持用户为自己的程序自定义GUI界面。MATLAB具有较强的数据可视化功能，可以绘制各种二维和三维图形，在Mapping ToolBox的帮助下，它可以拥有强大的球面投影绘图功能，满足地球科学的需求同时也令很多大尺度的天文绘图变得非常简单。MATLAB支持直接嵌入Java程序包，也支持调用C和FORTRAN的程序；反过来，MATLAB可以把自己的程序转换成Java，C和FORTRAN。MATLAB支持网络协议，如SOAP协议，也支持数据库访问接口ODBC和JDBC。新版本的MATLAB支持多线程计算和分布式计算。

MATLAB在设计之初就是针对数值计算的，因此诸如线性代数的各种数值算法、常用的数值计算都是它最基本的功能，有些功能只需要非常简单的一个运算符号就可以完成。它的算法扩展通过工具箱(ToolBox)完成。工具箱不仅是

[6]http://idlastro.gsfc.nasa.gov/

[7]http://www.mathworks.com

一种函数调用库，而且包括方便操作的GUI界面。目前MathWorks公司提供的工具箱包括统计、神经网络、分布式计算、地图投影、优化、曲线拟合、模糊逻辑、图像处理、小波分析、遗传算法、符号计算、微分方程数值解等。可以看到，MATLAB不仅有基本的数值计算还提供了丰富的数据挖掘所必须的复杂算法。

在天文应用上，MATLAB并没有成套的天文应用工具箱，象IDL那样。多数工具箱都是由个人开发以后共享出来的。在天文仪器和光学设计领域MATLAB有一些应用。

3.1.7　评价与小结

上面列出的各项可以用来构建天文数据挖掘的编程语言各具特色，并没有哪一种语言有显著的优势，因此不同的天文研究，不同的研究机构所选的编程语言各不相同。下面我们尝试对它们的特性作出一个尽量全面的评价。评价分成如下几个方面：语言特点、功能、性能、易用性、可扩展性、网络连接、数据库连接、天文应用状况、在虚拟天文台框架下的发展潜力以及费用情况。表3.1列出了全部的评价结果。

在语言特点方面，FORTRAN和Perl不是面向对象的语言，其他均采用更容易编程的面向对象语言。FORTRAN是编译型语言，其他全部是解释型语言。IDL有矩阵计算的语言特征，而MATLAB是一个彻底的矩阵计算语言。

功能方面，FORTRAN由于是一个完整的编程语言，因而可实现几乎所有功能，但是这需要付出成本代价。Perl和Python的自身功能非常简单，因为它们本质上都是脚本语言。R，IDL和MATLAB则各有专攻。R擅长统计计算，拥有非常丰富的统计算法库，这是其它语言所没有的。IDL擅长处理图像和数据可视化，虽然也具备比较丰富的计算能力，但是这些计算能力都是基本的计算，借助IMSL和GSL等外挂库，Perl和Python也可以获得这些计算的功能。MATLAB的设计目标就是针对数值计算和数据分析，因此提供了最丰富的数值计算工具箱，这些算法不仅有基础算法还有比较新的数据挖掘算法和数据可视化算法。此外，MATLAB也拥有可以同IDL相比拟的可视化功能。

运行性能方面，FORTRAN因为自身就是编译语言，所以理所当然拥有最高效的运行速度，同时也由于它有半个世纪的历史，因此稳定性非常好。Perl和Python本身因为简单所以并不容易产生不稳定，但是正是这种简单

表 3.1: 天文研究常用编程语言的比较

评价项目	FORTRAN	Perl	Python	R	IDL	MATLAB
语言特点	结构型编译型语言	脚本语言	面向对象的脚本语言	面向对象解释语言	面向对象解释语言	面向对象＋矩阵计算解释语言
功能	全面	简单	简单	专于统计	擅长数据可视化	擅长数值计算
性能	稳定、高效	依赖于模块的实现方式	依赖于模块的实现方式，可以加速	未知	速度低	速度低，但可以加速
易用性	需要较高使用技巧	简单实用	简单实用	简单，集成环境	简单，集成环境	简单，集成环境
扩展性	一般	非常好	非常好	一般	好	非常好
网络	不好	好	好	好	好	好
数据库	不好	好	好	一般	一般	好
天文应用	广泛使用	有一定应用基础	有较多应用	较少应用	很多应用	较少应用
VO潜力	不好	好	好	一般	一般	一般
费用	免费	免费	免费	免费	费用高	费用高

会带来其他问题，例如由于没有一个好的命名机制，会导致同名但功能不同的模块发生混淆，为系统的运行带来隐患。此外，虽然它们可以通过调用外部库——这些库的代码可能是FORTRAN或者是C——而提高运行速度，但是这本质上并没有减轻编程的工作量，只是提供了一种更加容易的系统集成方法。值得一提的是Python已经有并行处理的版本了[8]，这应该能够提高它的运算效率。R语言因为没有实际使用来做大量测试，也没有相应的文献介绍，因此对它的运算性能并不了解。但是由于它是解释语言，因此效率不会比其他解释语言更好。IDL在Sloan在项目中的使用经验表明它用于大量的数据处理效率并不高。MATLAB在使用M语言进行数值计算的时候，也没有较高的效率，但是可以通过生成C和FORTRAN代码，或者通过并行计算提高运行效率。

易用性方面，由于FORTRAN仍然带有上一代编程语言的特色，因此并不适合非程序员来使用，只有对计算机的工作原理有较深刻了解才能够写出高效、稳定的FORTRAN程序。除此以外的其他语言由于都是解释语言，在编写程序的时候，使用者不用考虑内存与线程等问题，因此掌握起来比较容易，代码编写的正确率也较高，比较适合非程序员使用。特别的，R、IDL、MATLAB均提供了集成环境，不仅简化了编程的工作量，而且提供了调试环境。MATLAB甚至提供了Profile功能帮助使用者研究自己的程序各个部分的调用频率和运行效率。

扩展性方面，FORTRAN仍然保留着封闭系统的特点，不能够用比较简单的方式进行扩展。Perl和Python是脚本语言，他们的功能主要依靠扩展性实现，因此它们的扩展性最好。R语言专门用于统计，这制约了它的扩展性，尽管它可以在统计算法领域内具有良好的扩展性。IDL虽然也具备扩展性，但是对和其他变成语言的集成仍然比较有限，特别是对COM的支持限制了IDL的夸平台可移植性。MATLAB的扩展性和IDL近似，但是在同其他变成语言的集成上比较有优势。

对网络的支持，FORTRAN本身并不是为了网络应用而设计的，因此对于网络的支持并不出色。Perl依赖它出色的文字处理能力具有非常好的网络编程能力。Python也提供了Internet数据处理的基本功能。R和MATLAB支持WEb Service的通信协议SOAP以及HTTP和FTP协议。IDL也具备网络数据访问的能力。

[8]http://www.parallelpython.com/

在数据库方面，FORTRAN没有专门的支持，而Perl和Python可以通过模块来支持。R语言提供了ODBC 、Oracle 、PostgreSQL 、MySQL等数据库的驱动。IDL 支持ODBC 接口和Oracle 和SyBase 连接，而MATLAB 用支持ODBC 和JDBC 各种数据库。

在天文学领域的应用方面，FORTRAN是应用最广泛的，IDL因为首先在NASA得到应用后来在Sloan项目也得到应用，因此有较多工具可供使用。Python是最近几年以来逐渐在天文学研究中流行起来的，因此也积累了很多实用的模块。Perl语言在天文学研究中有一定的应用，例如Joint Astronomy Center就使用Perl完成了很多仪器控制和数据处理的工作。MATLAB在天文学研究中较少应用。

所谓VO潜力就是看编程工具是否提供良好的互操作能力，能够同VO工具协同工作。FORTRAN因为系统的封闭性，很难实现良好的互操作性。又由于不能很好地支持网络，所以很难在网络环境下得到使用。但是FORTRAN编写的数据处理和数据挖掘算法程序却可以在适当的封装后成为VO的工具。Perl和Python均具有良好的VO互操作潜力。实际上，Python已经应用于AstroGrid项目中，用来描述AstroGrid的工作流。R和IDL的开放性并不好，需要添加专门的互操作模块才能够支持VO的协同工作环境。MATLAB虽然有较好开放性，但是也需要有更多开发才能够支持VO。

最后从费用上看，除了IDL和MATLAB其他均为免费软件。

综上所述，从虚拟天文台的立场出发，以我们的任务来衡量，没有一种现成的编程语言可以在不经过多少开发的情况下担当数据挖掘工具。比较有潜力进行改造的是Perl 、Python 、R和MATLAB。前面两种语言的优势在于集成性，便于用它们实现对现有数据挖掘工具的集成。后两种语言的优势是算法，它们可以提供丰富的算法给天文数据挖掘工具。

3.2 天文数据挖掘的工作流描述语言——JDL

在本节中我们开始一个新的尝试，综合上述几种编程语言的优势，设计一种专门用于虚拟天文台的网格环境中实现数据挖掘的编程语言。这就是任务描述语言－－JDL[28]。

3.2.1　JDL设计的目标和功能

JDL设计的初衷是用一种建立在网格服务上的非常简单的脚本语言来串联起VO数据访问和各种已经存在的数据挖掘工具，各种不同层次的用户都可以使用它整合VO数据访问和数据挖掘工具。初级用户可以通过图形化的集成环境完成可视化编程，而高级用户可以直接编写JDL脚本。

JDL应该是一种在网格上运行的编程语言，它在被用户编辑好以后应该在服务器上运行，用户不必在本地建立复杂的虚拟天文台工具的运行平台，只需要将JDL提交给服务器，服务器将按照JDL的描述调用网格中的各种资源完成用户的任务。JDL应该既可以完成数据查询的工作也可以完成数据挖掘的工作。使用者不必为从那里找到数据而担心，也不必为从那里找到数据挖掘算法和算法运行环境而担心，一切安排都是由处理JDL的后台网格服务完成的。数据挖掘的过程十分复杂，经常需要反复迭代。因此完全自动的过程并不能完全满足用户的需求，这就需要有一个交互的工作方式作为自动化工作流的补充。我们把JDL的功能总结如下：

1. 高度简单的脚本语言，容易图形化编程
2. JDL应该在网格上运行
3. JDL既可以交互式运行也可以自动化运行
4. 将数据访问和数据挖掘在网格环境中有效结合起来，用户既不必担心访问的数据的输出格式也不必关心数据量的大小
5. 具备可扩展性，可以随时添加数据资源共和数据挖掘算法

3.2.2　JDL定义

为了满足上节中提出的设计目标和希望达到的功能，我们考虑JDL是一种具备两种等价形式的脚本语言，一种形式是适合人工阅读的通常的脚本语言，另一种形式是适合计算机程序自动理解和处理的XML形式。这种“双面”程序的设计可以满足以下要求：普通脚本适合人工编写，XML适合机器生成。前者为高级用户准备，后者给图形化编程准备。XML形式的语言更加适合在网格中传递和被各种服务程序所理解。为了描述方便，我们称脚本形式的JDL为JDL/s而XML形式的脚本语言为JDL/x。

JDL/s的语法采用Extended Backus Normal Form(BNF)的描述规则进行定义，参见附录B。JDL/x的结构采用Schema定义参见图3.2－3.4。图3.1给出了一个JDL的例子，上面的框中是JDL/s下面的是等价的JDL/x。

```
project cc
    job gettable
        function t=main()
            t=query(“select glon,glat, j_m, h_m, k_m from TwoMass where glon>=270 and glon<271 and
glat>-10 and glat<10”);
            t=addcol(t, 5, “h-k”, t(“h_m”)-t(“k_m”));
            t=addcol(t, 6, “j-h”, t(“j_m”)-t(“h_m”));
        end
    end
    job cchist
        function m=main()
            t=jobresult(“gettable”);
            m=hist(t, “h-k”, “j-h”);
        end
    end
end
```

```
<?xml version="1.0" encoding="UTF-8"?>
<PROJECT ID="cc" name="cc">
  <DESCRIPTION/>
  <JOB ID="gettable" name="gettable" type="job">
    <DESCRIPTION/>
    <FUNCTION ID="main" name="main">
      <DESCRIPTION/>
      <STATEMENT>
        <OPERATOR ID="assign" type="assign">
          <VARIABLEREF ref="t"/>
          <FUNCTIONCALL ref="query">
            <PARAMETERS>
              <PRIMITIVEB value="select glon,glat, j_m, h_m, k_m
from TwoMass where glon&amp;gt;=270 and glon&amp;lt;271 and glat&amp;gt;-10 and glat&amp;lt;10"
/>
            </PARAMETERS>
          </FUNCTIONCALL>
        </OPERATOR>
        <OPERATOR ID="assign" type="assign">
          <VARIABLEREF ref="t"/>
          <FUNCTIONCALL ref="addcol">
            <PARAMETERS>
              <VARIABLEREF ref="t"/>
              <PRIMITIVEA value="5"/>
              <PRIMITIVEB value="h-k"/>
              <OPERATOR ID="subtraction" type="subtraction">
                <VARIABLEREF ref="t">
                  <PRIMITIVEB value="h_m"/>
                </VARIABLEREF>
                <VARIABLEREF ref="t">
                  <PRIMITIVEB value="k_m"/>
                </VARIABLEREF>
              </OPERATOR>
            </PARAMETERS>
          </FUNCTIONCALL>
        </OPERATOR>
```

```
        <OPERATOR ID="assign" type="assign">
          <VARIABLEREF ref="t"/>
          <FUNCTIONCALL ref="addcol">
            <PARAMETERS>
              <VARIABLEREF ref="t"/>
              <PRIMITIVEA value="6"/>
              <PRIMITIVEB value="j-h"/>
              <OPERATOR ID="subtraction" type="subtraction">
                <VARIABLEREF ref="t">
                  <PRIMITIVEB value="j_m"/>
                </VARIABLEREF>
                <VARIABLEREF ref="t">
                  <PRIMITIVEB value="h_m"/>
                </VARIABLEREF>
              </OPERATOR>
            </PARAMETERS>
          </FUNCTIONCALL>
        </OPERATOR>
      </STATEMENT>
    </FUNCTION>
  </JOB>
  <JOB ID="cchist" name="cchist" type="job">
    <DESCRIPTION/>
          ... ...
  </JOB>
</PROJECT>
```

图 3.1: JDL实例

JDL的语言结构是一个层次化的结构（图3.2－3.4），最外层是project体，被关键字project和end包裹。project可以有一个名字。一个JDL文件只允许有一个project体。如果存在多个project体，JDL解释器将只处理第一个project而忽略后面的。Project体内包含0个或多个job体。job体被关键字job...end包裹。Job可以有一个名字，每个job体之前还可以有0行或多行有关这个job的注释。当JDL文档中包含多个job的时候，解释器会依次执行每个job。每个job内包含三种对象：整个job生命周期内都有效的变量声明；一个main函数定义；以及若干其他函数定义。每个对象之前都可以有0行或多行注释。Job体内声明的变量是在这个job的每个函数体内都有效的变量。大多数情况下，JDL不要求在使用一个变量之前声明这个变量。但是如果希望这个变量的作用域超出函数体，就需要

在job体的开始部分声明这个变量。Job对象内的main函数可以有返回值，这个返回值可以被后续的所有job对象引用，这样job之间就可以传递信息了。

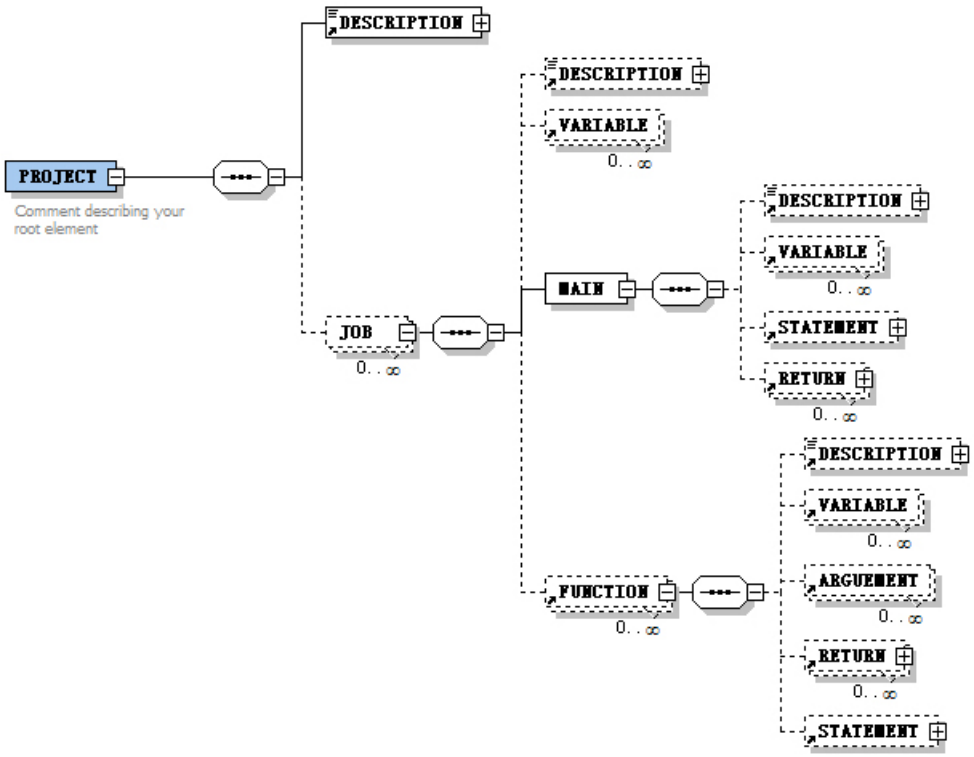

图 3.2: JDL语言的结构之一

JDL的函数体内是一行一行的表达式，表达式的语法非常类似于MATLAB。首先它的运算符是以矩阵计算为缺省的运算方式的，其次程序的控制流语法也同MATLAB类似。

JDL和在前面介绍的所有语言最大的不同并不是在它的语法特点上，而是它的运行方式上。它是一个纯粹的网格语言，不会在用户本地运行，而是在网格环境中运行。JDL所支持的函数扩充是通过Registry上的资源发现完成的，用户可以通过从服务器获得元数据而了解新发现的JDL函数库的使用方法。当这些函数库在JDL中应用的时候，用户不必在程序中注明这些函数存放在哪里，而是由服务器根据文件名字查询Registry上的服务资源，找到到底哪个服务支持这个函数，然后通过调用那个服务完成这个函数的执行。下面我们通过我们开发的一个数据挖掘原型来描述JDL这个独特运行方式。

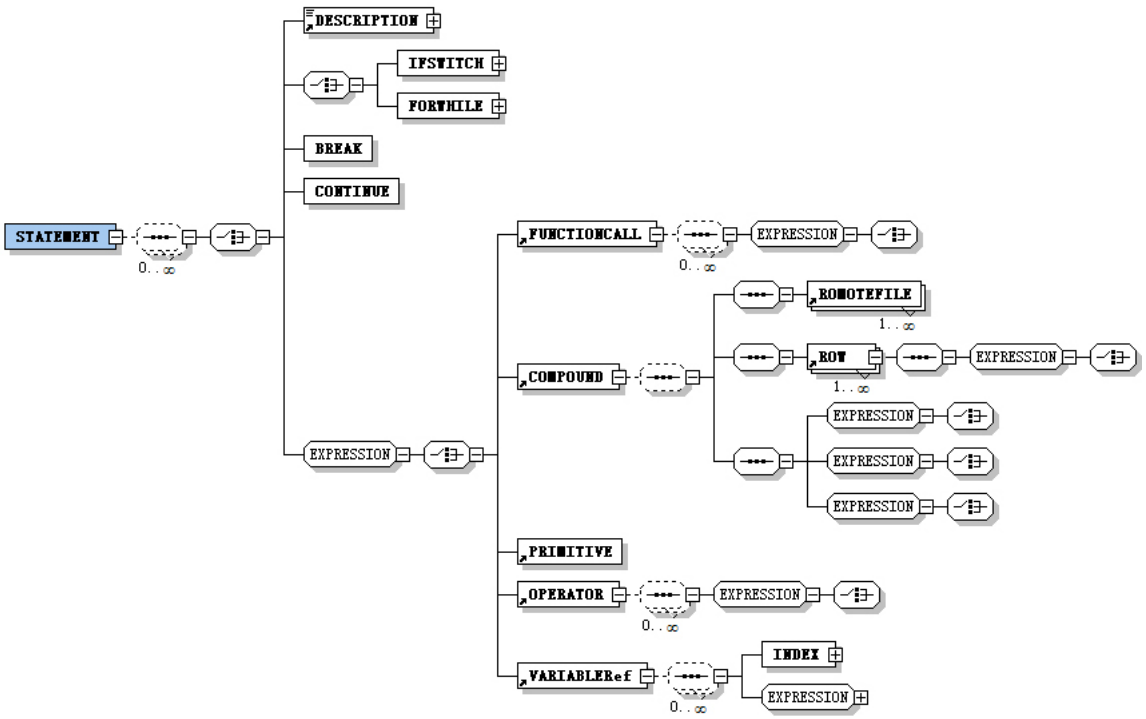

图 3.3: JDL语言的结构之二

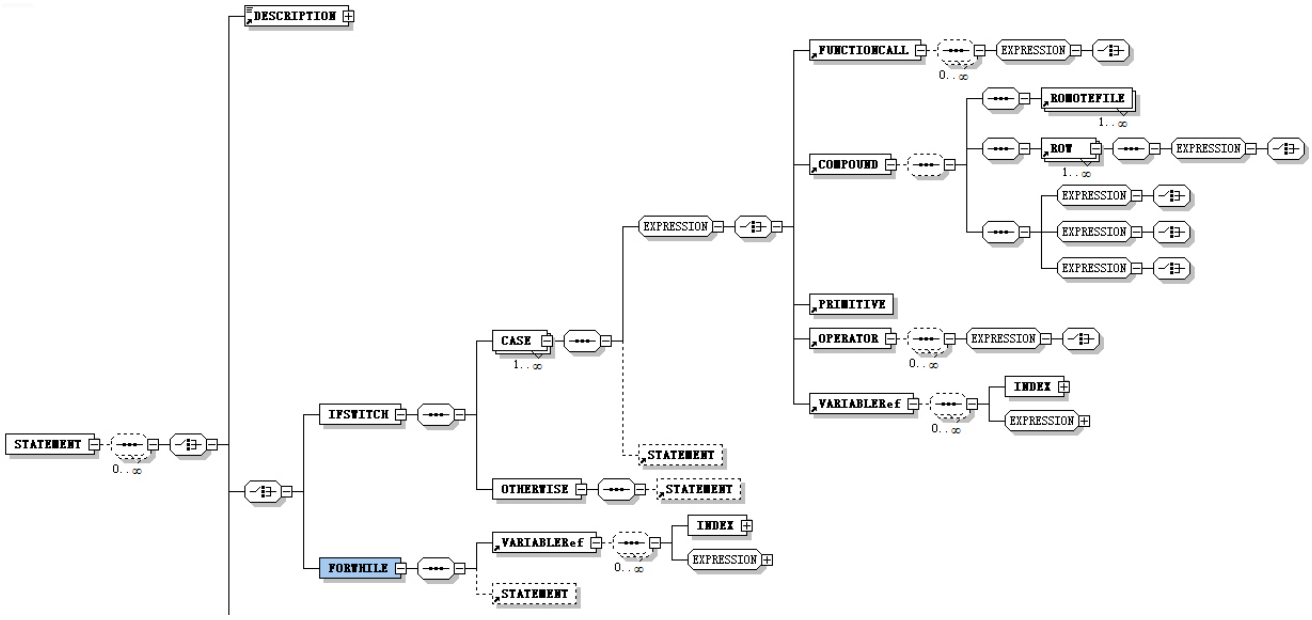

图 3.4: JDL语言的结构之三

3.3　基于JDL的天文数据挖掘工具原型

3.3.1　原型结构

基于JDL的数据挖掘工具模型旨在研究JDL的可行性，展示在JDL的驱动下虚拟天文台的数据访问和数据挖掘是如何结合起来的，其结构参见图3.5。原型的最上层是客户端，被称为Portal。用户在Portal内完成JDL的编辑以及提交到服务器上的JDL的运行状况监控。中间层的服务是JDL Job Engine和JDL Intepreter。JDL Job Engine是一个JDL任务调度引擎，它会根据用户提交的JDL的顺序依次调度他们的执行。由于这是一个原型，因此它的结构非常简单，并没有考虑复杂的调度和稳定性等问题。JDL Interpreter就是JDL的解释器。

在最底层起支撑作用的服务是那些提供JDL中真正功能的服务，它们是SkyPortal和CompuCell。SkyPortal是一个数据访问接口，它采用Web Service封装数据访问，只要发送ADQL到SkyPortal就可以得到数据。相比在上一章中提出的VO-DAS，SkyPortal只能提供同步查询。采用同步查询的原因是因为这仅仅是一个原型，如果使用VO-DAS的接口，将大大增加JDL Job Engine的开发工作量，以处理复杂的异步调度工作。可以把JDL Job Engine和VO-DAS服务器比较一下，我们就会发现无论从结构上还是从代码的工作量上，JDL Job Engine都远远不能和VO-DAS服务器相比。CompuCell是一个封装数据挖掘算法包的接口服务，它可以将C语言、Java语言开发的算法包接入JDL环境中，由JDL调用它们发布出来的函数，然后由JDL Job Engine替JDL发送请求给对应的CompuCell完成实际的计算。除了上面提到的三个层次的服务，还有一个Registry。它为各个服务提供类似于Google这样的服务，便于JDL Job Engine找到合适的JDL Interpreter或者SkyPortal或者CompuCell。

最后说明的是，这个原型没有采用GT4中间件支持的WSRF构架，而是采用了更加简单的Web Service接口，这样的选择主要是出于简化开发的考虑。

3.3.2　原型的设计与实现

原型的具体设计参见图3.6，JDL的执行过程可以根据图3.6简述如下。

首先，用户在Portal中进行JDL编辑操作。在Portal内有一个对象称为File Parser，用来对用户编辑的JDL的语法正确性进行检验。JDL文件编辑好并经过检验以后，通过Web Service接口提交给JDL Job Engine。JDL Job Engine在

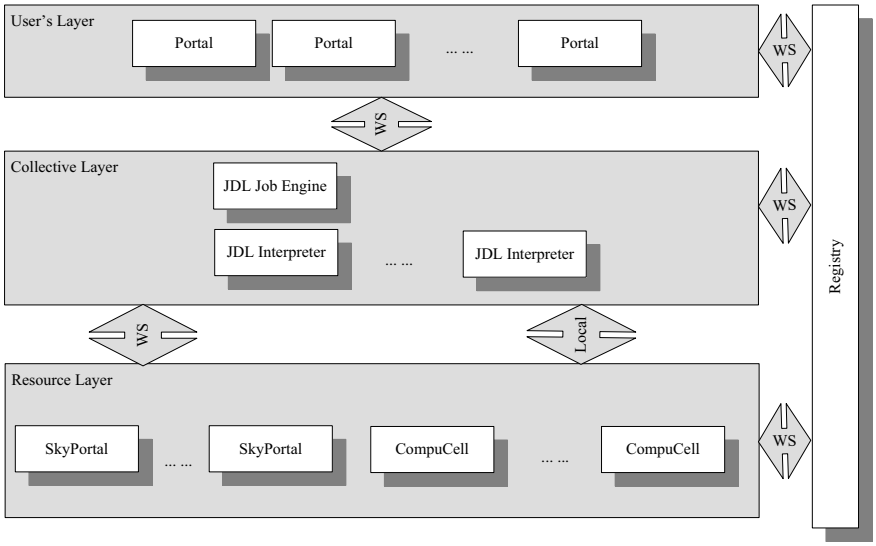

图 3.5: 基于JDL的天文数据挖掘工具原型结构

得到一个任务以后会启动一个工作线程，进行处理。首先它会找到一个JDL Interpreter web service，然后利用它解释JDL。JDL Interpreter在解释JDL的时候总是先处理第一个job对象，执行它的main函数。如果在main函数执行到内建函数query，则根据query参数的SQL查询哪个SkyPortal可以提供查询，然后把SQL发送给SkyPortal等待查询结果。查询结果会被转换成JDL的数据类型保存在JDL Interpreter中。当在执行main函数的过程遇到一个函数调用的时候，JDL Interpreter首先检查这个函数是不是已经在job对象中有了定义，如果没有定义就会向Registry下查询谁提供过同名的函数，找到那个提供这个函数的CompuCell服务。然后把函数的参数转换成JDL/x的形式，通过Web Service传递给CompuCell服务并等待计算结果。CompuCell会调用它所封装的算法程序完成计算，并把结果作为返回值传递回JDL Interpreter。JDL Interpreter得到的是JDL/x形式的数据结果并继续后面的执行。

图3.7显示了原型中JDL的客户端程序Portal的运行界面。Portal在后台服务正在运行的时候，会定时向JDL Job Engine询问已经提交的任务的进度，如果任务完成，就会从JDL Job Engine把结果数据取回，并可以调用VOPlot等VO工具实现简单的数据可视化。

通过这个JDL原型，我们可以得出如下结论：首先，JDL的思路是新颖的，它从功能上类似于工作流描述语言，如BPEL4W，但是语法上更接近计算性

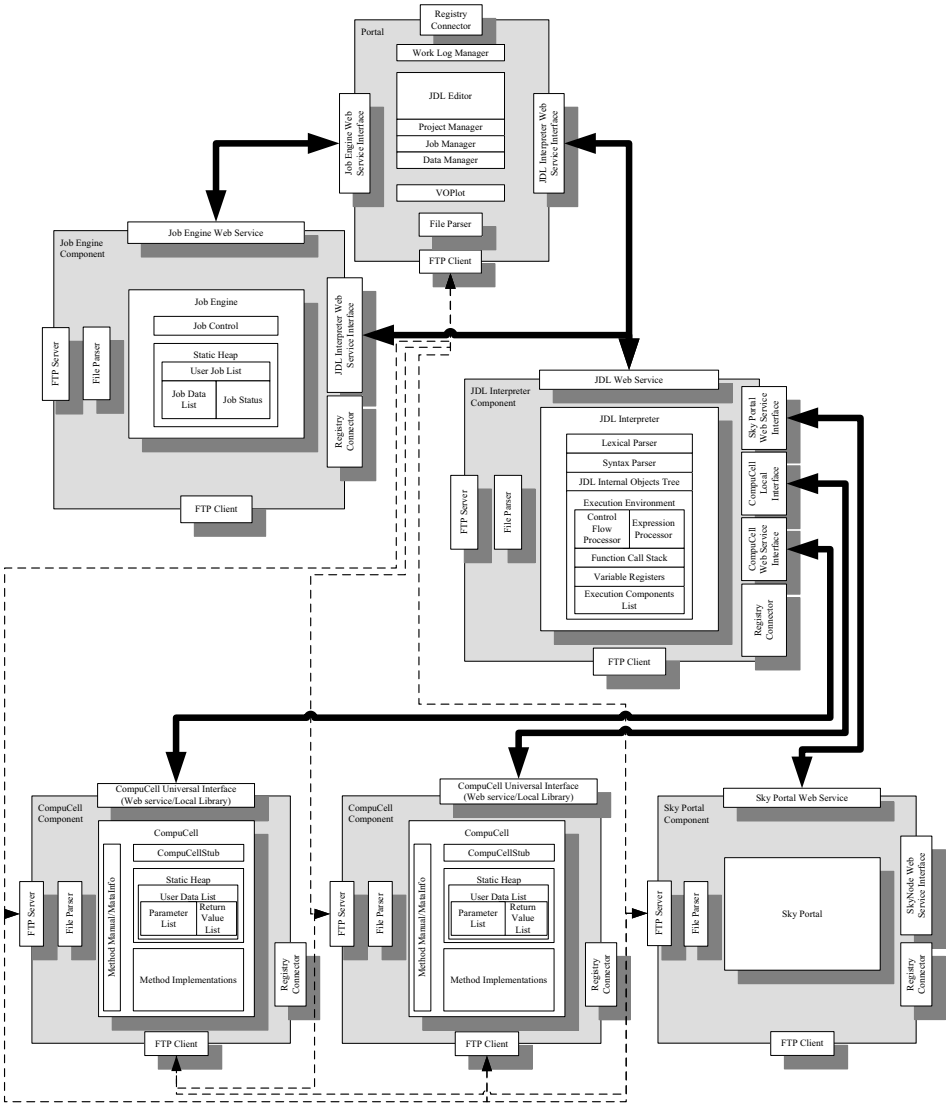

图 3.6: 基于JDL的天文数据挖掘工具原型的设计

的语言。这样有助于将基于网格的工作流式的处理方式同数据挖掘计算相结合。其次，我们只使用了Web Service就成功让这样一个流程运转起来说明如果加上WSRF，那么JDL Job Engine的结构设计会进一步简化，系统的性能会有更大提高。第三，我们也发现这样的一个运行结构执行效率非常低，特别是算法简单的情况下更加突出。此外，数据处理量也因为JDL的运行限制而不能很大。最后，我们认为JDL的思路是可行的，但是如果采纳这样的思路完成一个数据挖掘工具，带来的最大问题是工作量的激增，主要体现在JDL解释器的编写

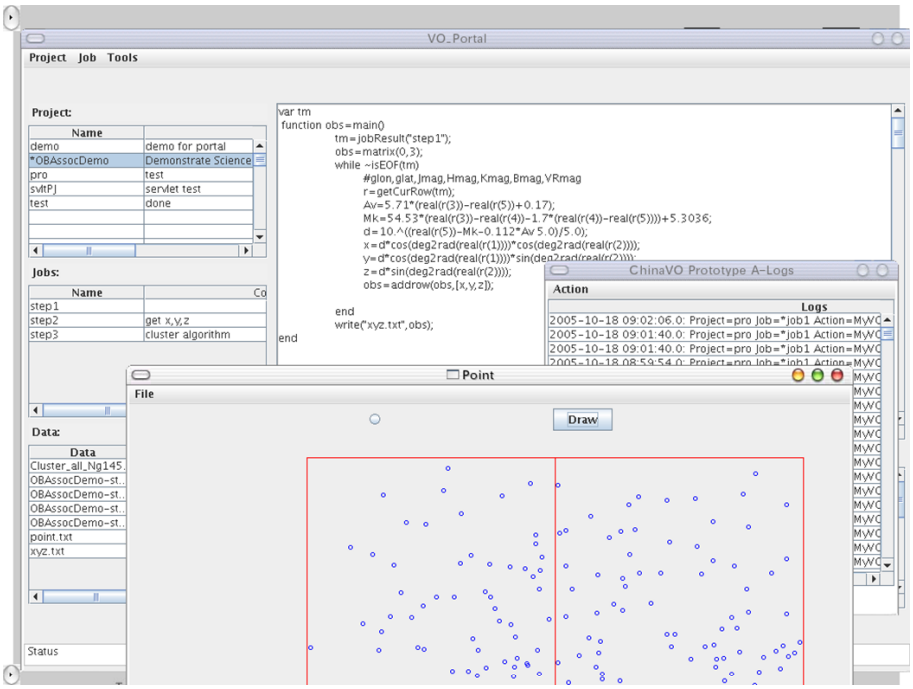

图 3.7: 基于JDL的天文数据挖掘工具原型的客户端外观

和CompuCell对各种算法的封装等。

3.4 用MATLAB实现天文数据挖掘

3.4.1 MATLAB实现数据库查询

本节我们将讨论第三种可能的数据挖掘工具解决方案，即对某个现成的数据挖掘工具进行二次开发，使其能够和各种VO协议兼容，实现VO的互操作性。我们选择MATLAB作为二次开发的对象。这样的选择主要是考虑它有很好的扩展性，在XML和网络应用方面有一定优势，同时又能够提供大量的数值计算工具。

首先介绍如何让MATLAB具有数据库访问的能力。有两种方法可实现这样的功能。第一种是使用MATLAB自身的Database Toolbox实现对数据库的查询；第二种是利用MATLAB支持Java的特点，直接调用VO-DAS提供的WSRF接口函数，实现对VO-DAS的直接连接。

我们利用MATLAB的Database Toolbox通过JDBC接口访问MySQL数据库。首先将JDBC驱动程序的Java包复制到MATLAB指定目录下，并更改静

态装载Java包的配置文件。重新启动MATLAB之后，JDBC驱动就被装载上了。采用MATLAB提供的database和fetch函数就可以很方便地查询数据库了。但是，这种方法对于天文数据库的查询特点没有任何支持，它不支持锥形检索也不支持交叉证认，更不支持分布式的异构数据库。

3.4.2　MATLAB实现天文数据挖掘

为了解决天文数据挖掘所遇到的这些特有问题，我们将MATLAB和VO-DAS联系起来，设计在MATLAB之上的VO-DAS客户端。这样的客户端表现为一系列MATLAB函数（参见表2.13）。这些MATLAB函数可以和数据挖掘计算组合使用，将数据查询和数据分析、数据挖掘无缝地连接起来。而且利用MATLAB的GUI自定义功能，可以将上述命令函数封装到一组美观大方的GUI背后，让最终使用者感觉不到编程的困难而使用到最便捷的数据挖掘工具完成特定的任务。

图3.8中显示的就是我们在MATLAB上完成的一个天文数据挖掘范例。这个数据挖掘范例的目的是估计SDSS巡天的恒星数据中星流的特征参数。首先我们从文献中得到Monoceros Stream出现的位置。在图3.8中所示的GUI界面中输入需要访问的天空的坐标，通过数据查询命令（MATLAB自带的database或者VO-DAS提供的syn命令）向SDSS数据库请求一个锥形检索，锥形检索除了位置要求以外还有星等范围和色指数范围可供选择。得到的结果包含多个波段的星等，按照选择的波段和色指数绘制成颜色－星等图的密度图（Hess图）。接下来这个工具会利用Girardi et al (2004)[51]的星族的理论等年龄线对数据库查询出来的Hess图进行等年龄线拟合，以得到Hess图上显示的这个星流的年龄、金属丰度和距离。

这个范例用一个比较简单的实际研究实例展示了MATLAB如何把数据访问和数据分析有机融合在一起。对于最终使用者而言，他们完全可以将精力全部集中在研究本身上，而不必再被诸如数据在哪里、如何写算法、怎么解析文件格式等等这样的技术细节所困扰。

通过这个实验，我们发现在MATLAB上实现天文数据访问和数据挖掘的结合是一件非常容易的事情。只要设计好一套天文专用的MATLAB ToolBox，就可以让很多用户不需要懂得M语言就可以在MATLAB上完成很多天文研就工作。性能上看，在MATLAB上完成数百万条数据的处理和计算在台式机上也是

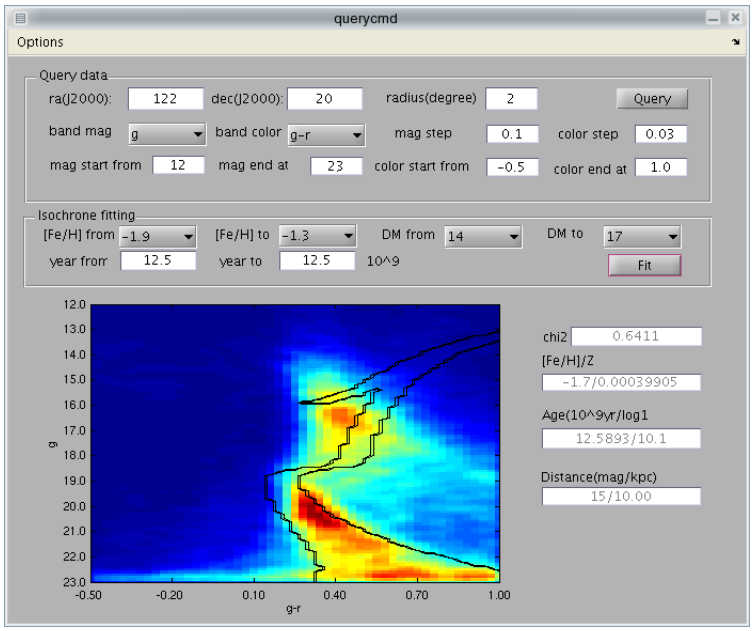

图 3.8: MATLAB二次开发后进行天文数据挖掘的实例

有可能的，如果考虑到它对并行计算的支持，那么对于比较复杂的计算和比较巨大的数据的处理都是可以较好地完成的。

3.4.3 MATLAB和其他VO工具的互操作

一些VO工具具有独特功能，如果MATLAB支持和VO工具之间的互操作，便可以充分利用利用现有的VO工具带来的丰富的功能，结合MATLAB的强劲的数值计算能力就可以组成非常有竞争力的天文数据挖掘工具集。为了让MATLAB和更多的VO工具实现互操作，我们将PLASTIC协议[9]引入MATLAB。

PLASTIC的全称是Platform for Astronomy Tool InterConnection，这是一个连接天文桌面应用程序的通讯协议。PLASTIC的哲学是天文学家使用一群软件——而不是一个全能软件——完成研究工作，每个软件完成一项专门的功能。而PLASTIC则是这些软件的黏合剂，将这些小的工具软件结合在一起，使得它们之间可以进行数据交互，协同完成一项工作。这个设计哲学也是虚拟天文台的基本思想。

PLASTIC的应用场景通常是这样的：用户通过VO-DAS查询星表数据，得

[9]http://www.ivoa.net/Documents/latest/PlasticDesktopInterop.html

到的结果返回到VO-DAS的GUI客户端上，这时用户希望使用Topcat查看星表数据，画出简单的图形，这个时候可以使用PLASTIC协议自动将星表数据从VO-DAS的GUI客户端传送到Topcat上面。当用户希望检验这个星表对应的天区的图像的时候，可以使用Aladin搜索图像数据，然后再利用PLASTIC将星表数据从Topcat送到Aladin。在Aladin显示叠加了星表的图像的的时候，可以在Topcat的散点图（例如颜色－星等图）中选择某个感兴趣的数据点（例如可能的蓝离散星或水平分支星），这个时候可以利用PLASTIC即时在Aladin的图像上点亮对应的数据点。

PLASTIC协议采用两种可选的进程通讯手段，一种是不依赖编程语言的XML-RPC方式，另一种是Java下的Java-RMILite。PLASTIC的主要通讯方式是传递消息（Message）。在桌面上必须首先运行一个PLASTIC Hub的监控进程，它是所有PLASTIC应用程序的消息中转站。每个PLASTIC应用在启动以后都需要或显式或隐式地寻找正在运行的PLASTIC Hub，并把自己注册上去，以便让其他应用程序能够看到自己。当一个PLASTIC应用希望向另一个传递消息的时候，它就把消息连同目的应用的名字一同传递给Hub，由Hub代为转发。如果一个应用需要将数据传递给另一个应用，首先需要把数据保存到一个临时文件中，然后将这个文件的URL作为参数连同表达数据传递的消息一起发送给Hub。接收的应用，从Hub得到消息以后，会从消息的参数中获得这个URL然后打开这个文件。缺省情况下，所有的在PLASTIC应用之间传递的数据都是采用VOTable格式保存的。最后，除了向某一个应用发送消息，PLASTIC还允许向Hub内注册的所有应用广播消息。

目前已经支持了PLASTIC协议的桌面VO工具包括Topcat 、Aladin 、Astro-Grid AstroScope 、VO-DAS GUIClient 、VisiVO 、VOPlot 、VOSpec等。这些工具的功能包括分布数据库的数据查询、星表显示、星表和图像的联合、数据可视化、光谱数据显示等。我们看到还没有一个具备丰富计算能力的桌面工具支持PLASTIC。MATLAB虽然可以作为服务器或计算机集群中使用的并行计算平台，但是大多数情况下用户还是喜欢在桌面应用它完成小规模但非常频繁的计算工作。在天文数据挖掘过程中，也并不是所有算法都需要使用并行计算才可以完成，一般规模的模型拟合，数据处理都可以在普通的台式机平台上完成。因此，有很多情况下，MATLAB是运行在桌面的。通过对MATLAB进行PLASTIC扩展，可以让它成为VO工具组中的一员，完成数值计算的专项功能。这样，天文学家桌面上的VO应用就会更加完整、功能更加强大，VO工具集

也就变得更加实用。

MATLAB下加入PLASTIC协议支持主要需要两个方面的工作，一个是实现PLASTIC的消息收发，另一个是能够读写VOTable文件。我们利用Java程序实现了一个PLASTIC客户端，并将它接入到MATLAB中，同时创建了一个有GUI界面的MATLAB函数plastic。只要在MATLAB中运行plastic就会启动一个PLASTIC客户端GUI，使用这个GUI完成向Hub注册或反注册的工作。注册以后MATLAB就处于一种等待数据状态，这个时候MATLAB仍然可以进行别的工作。当其它PLASTIC应用程序向MATLAB发送消息的时候，MATLAB窗口中就会产生提示信息，如果传送的是数据，那么可以利用PLASTIC客户端将数据保存到当前的工作空间（workspace）中。在保存过程中，数据已经从VOTable格式转换成MATLAB的cell array格式了。这个转换工作是通过调用Starlink提供的STIL软件包[14, 52]实现的，STIL提供了VOTable到各种文件格式的转换方法。图3.9是我们开发的MATLAB的PLASTIC扩展功能的工作画面，图中展示了MATLAB通过PLASTIC协议从Topcat获得VOTable数据并转换成MATLAB数据类型的过程。

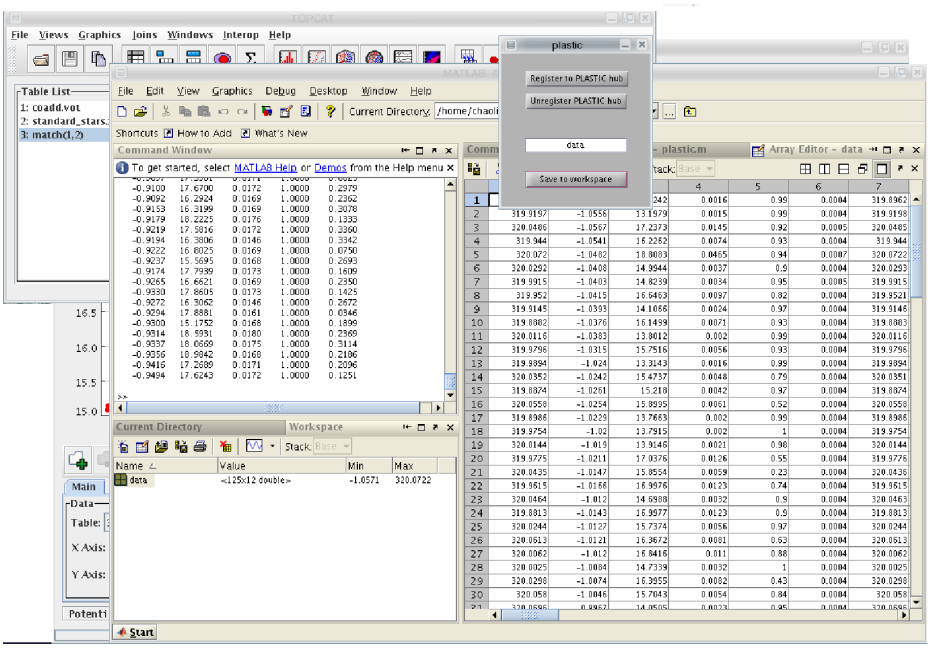

图 3.9: MATLAB的PLASTIC扩展的工作画面

3.5　本章小结

本章我们通过调研和多项原型实验研究了三种可能的基于虚拟天文台的数据挖掘工具的解决方案。我们看到，如果直接使用现有的数据挖掘工具则难以立即遵守VO的各种协议并与VO环境产生互操作。基于网格平台的专门的数据挖掘语言JDL的实验说明完全创立一套新型的针对虚拟天文台环境的数据挖掘语言从技术上是可行的，但是需要较长周期的开发和很大的投入。关键是是否能够获得很多用户的青睐还需要作进一步的调研和分析。由于MATLAB在数值计算方面表现十分突出，而它的可扩展性方面也令人满意，我们在其上完成了第三种方案的实验，即在MATLAB作二次开发，实现和虚拟天文台的数据访问服务VO-DAS的无缝连接以及与其它VO工具的互操作。通过具有实际研究背景的范例进行测试，证实了我们对MATLAB的判断，并说明这是一种有希望的天文数据挖掘方案，它不仅具有开发容易、风险较低、充分发挥现有VO工具优势、算法丰富等优点，还具有使用界面友好、易学习易使用等特点，唯一美中不足的是MATLAB仍然是比较昂贵的软件产品。

回顾我们在第1.4节所确立的目标，到此章为止，我们已经完成了目标的一半，即通过发展基于虚拟天文台的数据挖掘工具，推动虚拟天文台的应用水平。现在我们已经拥有了至少一套可用的数据挖掘工具：VO-DAS+MATLAB+PLASTIC。尽管这还是一套实验性的工具，无论从功能上还是从稳定性上都还不能说是成熟产品，但是它已经能够在实际的天文学研究中初露头角了。在后面的各章中，我们将基于这套实验性的数据挖掘工具展开科学研究，实现我们后一半的工作目标：最终使用这些工具完成实际的天文学研究。

第四章　银河系：结构、形成历史和研究新进展

4.1　银河系的基本结构

由于我们的太阳系身处银河系之中，因此全面观察银河系的形状和结构是一件十分艰巨的事情。自从伽利略首次将望远镜指向天空（1609年）以来，人类对于银河系的认识逐渐深入。

早期对银河系的认识主要是由哲学家们来做的，例如Immanuel Kant就在1755年的时候提出银河应该具有透镜一样的形状，恒星都在围绕着中心旋转。第一个进行定量研究的是William & Caroline Herschel，他们使用48英寸望远镜分析了683个视线方向上的恒星计数。但是那个时候他们不知道星际消光，并假设所有恒星都是一样亮的，而且认为银河系的恒星是均匀分布的。这些错误的认识导致了他们给出的银河系图景是完全错误的。其后的一百多年间，三角视差方法得到发展，近邻恒星的真实距离能够被测量出来，这对建立精确的银河模型提供了重要帮助。1922年Jacobus Kapteyn[53]根据邻近恒星（主要是盘星）的视差和自行首次定量化提出银河系的模型。在他的模型中银河系是一个中间密外围松的扁平盘子，跨越15kpc长，3kpc厚。几乎与此同时，Harlow Shapley（1918）[54]根据天琴座RR变星和室女座W变星的周光关系确定了一批球状星团的距离，从而断定银河系的中心在人马座方向，是一个扁平的盘状，而直径有100kpc之多。太阳则位于距离中心16kpc远的地方。由于两人研究对象的选择效应（一个是盘星一个是球状星团）所以他们描述的银河系有巨大差异。

又经过了几十年的观测，在得到了大量的恒星巡天数据，并且得到了空间观测结果，在深入研究了星际消光，并对它作出修正以后，我们终于初步了解了银河系结构的主要特征。现在一般认为，银河系是由极薄盘（Extreme disk ）、薄盘（Thin disk）、厚盘（Thick disk）、核球（Bulge）和晕（Halo）组成的（图4.1，摘自Eva Grebel在2007年IMPRS Summer School上的讲稿）。最近的研究还表明银河系中心有存在棒(Bar)的迹象[55]。

从统计的角度看，银河系不同结构上的恒星的类型是不同的，这暗示着这些结构可能有不同的起源和演化。采用星族来描述这些处于不同结构的不同类型恒星合适的。通常称薄盘上的恒星为星族I而晕中的恒星称为星族II，这种分

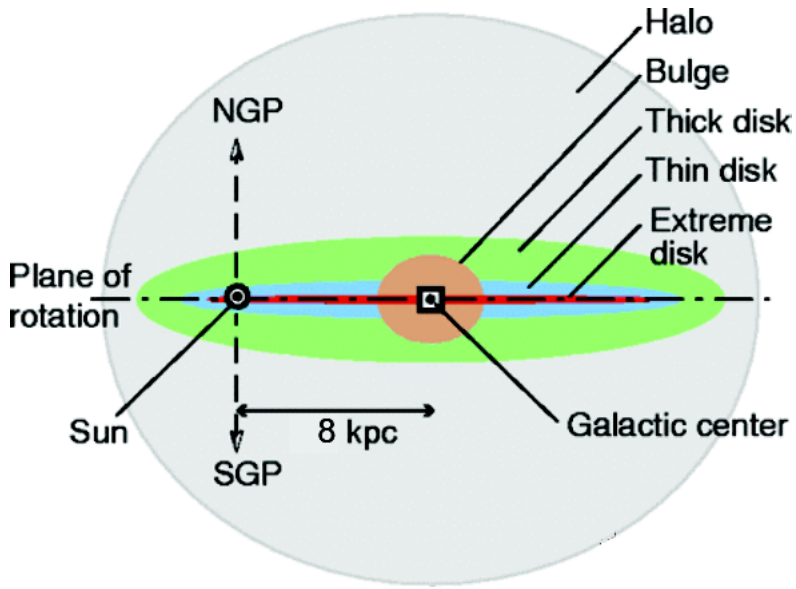

图 4.1: 银河系的结构

类最早来自W. Baade (1944)[56]。星族是一群特征相近的恒星的集合。一个星族最关键的特征是年龄，即一个星族的恒星年龄都是在一个特定范围内的。除此以外，星族中恒星的初始化学组成、初始质量函数（IMF）、运动学特征、距离特征、空间分布、恒星形成历史也都表现出共性。一般认为，薄盘上的星族I是比较年轻的恒星，具有较高的金属丰度，绕银心旋转轨道相近，轨道比较圆，速度也差不多，速度弥散小。晕中的星族II恒星是比较年老的恒星，具有较低的金属丰度，在大偏心率轨道上运动，轨道大多不相同，旋转速度较低且速度弥散大。处于薄盘和晕之间的厚盘恒星的特征介于星族I和星族II之间，具有中等金属丰度和中等速度弥散。

银河系的形状是一个旋涡星系或棒旋星系，在盘上有旋臂并且聚集了丰富的气体和尘埃。HII观测勾画出了旋臂的结构，而OB星协则将太阳邻近的几个旋臂显现了出来[57]。目前对旋臂的动力学的解释最成功的是密度波理论[58]。银晕的情况没有那么清楚，传统认为除了球状星团以外的晕星组成一个椭球状对称的结构。在对银河系恒星进行计数研究的时候，总是假设它只有平滑的成份[59, 60]。在建立银河系恒星计数模型的时候，晕中的场星分布规律会用描述河外星系的de Vaucouleurs律[61]，或者幂律来描述[60]。无论是哪一种分布，都是假设晕是平滑的而且是对称的。最近几年，晕中的子结构被揭示出来，晕再

也不是平滑对称的，而是充满了子结构，并且有迹象表明它也可能不是对称的。在第4.3节中，我们会详细介绍这方面的进展。

4.2　银河系的形成假说

随着对银河系的了解越来越深入，对它的形成历史的探讨也越来越活跃。但是到目前为止，还没有一种假说能够得到一致的赞同。两种著名的假说主宰了银河系形成历史的讨论几十年。1962年，Eggen, Lyden-Bell和Sandage（ELS）[62]发现太阳附近的恒星随着金属丰度的降低，轨道偏心率和垂直方向的振动能量增大，而角动量下降。从这个事实出发，他们认为，要么银河系早期有过剧烈活动，要么贫金属恒星不是在离心力支撑的盘上形成的。他们假设了这样一种银河系的形成景象：开始的时候贫金属的原始星云塌缩，凝结出贫金属的晕星和晕中的球状星团。随着早期恒星中的超新星爆发，使原始星云的金属丰度提高，当星云塌缩到10倍大小的时候，盘开始形成[63]。这样的两个阶段的形成历史造成了现在贫金属恒星和富金属恒星运动学上的差异。特别地，当原始星云塌缩形成晕和球状星团的时候，由于塌缩是自由落体过程，而自由落体时标远远短于增丰的时标，因此球状星团的金属丰度不存在丰度梯度。

Searle & Zinn（SZ）[64]对此提出不同的见解。他们认为ELS模型不是解释球状星团的金属丰度没有梯度的唯一原因。他们更加直观地假设这些球状星团是在一些小的团块中形成的，小团块的质量大约$10^8 M_\odot$。由于超新星的增丰对各个团块各不相同，因此团块初始金属丰度各有不同。另外，他们假设增丰事件是由球状星团的诞生产生，这样每个团块就只有非常少的增丰事件，这就保证了团块最终金属丰度的弥散。

SZ假说后来受到重视的一个重要原因是在宇宙学研究中提出来的CDM模型[65]的N体模拟显示宇宙引力膨胀过程中不断合并小的系统形成大的系统。一个宏观的模型和一个局部的模型竟然能够如此巧合地对应起来，这是非常吸引人的理论前景。HST升空以后，也确实不断观测到宇宙深处正在发生许许多多的星系并合事件，这提示我们回顾我们的银河系，在银河系中是否也发生过，或正在发生这样的并合事件呢[66]？首先值得怀疑的是银河系晕中的球状星团，它们是否曾经是邻近的矮星系的附属，随着矮星系被吸积进银河系，它们也就成为了银河系的附属？其次，银河系邻近的矮星系特别是矮椭球星系是否是并合过后遗留下来的呢？场晕星中也应该有很多是并合进来的吗？银河系如果发生

过并合事件，那么就应该会在晕中和盘中留下些许证据。一种观点认为，厚盘就是由于吸积事件加热薄盘而产生的。在晕中也应该有一些并合产生的子结构存在，这些子结构应该能够从化学丰度的分布中找到，或者从运动学的分布中找到。

4.3 银河系结构研究的新进展

自从1990年代后期以来，大天区的巡天项目陆续展开，2MASS观测了全天几乎全部区域；SDSS也已经完成了接近10000平方度的天区的观测，而且还准备观测更多天区（SDSS III）。Majewski在1993年的时候[66]已经预言了子结构的存在，并且判断它们可能通过对丰度分布的分析或奇异的运动学特征（如逆行轨道）的分析而揭示出来。但是大天区的恒星巡天又让另一种子结构探测技术成为可能，那就是直接看到子结构。能够直接看到子结构当然是找到它们的最佳方法了。

Yanny et al. (2000)[67] 用SDSS 中挑选出来的BHB 星进行计数，发现了一块不能用银河系模型进行解释的过密度区。同年，Ivezić从SDSS中发现的天琴座RR 变星候选体的分布中也发现了这个子结构[68]。Newberg et al. (2002)[2]则利用SDSS 的F型星的计数清晰刻画出了这个过密度区，并根据颜色－星等图的相似性认为它应该是Sagittarius dwarf galaxy被银河系吸积而产生的潮汐流。最重要的进展出现在2003年，Majewski预言子结构10年以后，他自己采用2MASS测光数据将这个Sagittarius dwarf tidal stream在全天的分布[69]描绘了出来。图4.2摘自Majewski et al.(2003)，在通过颜色选择挑选出2MASS中的M巨星后，这些巨星在晕中清楚勾画出Sagittarius dwarf galaxy南北两侧的两个潮汐流，这是迄今发现的最显著的银河系并合证据。Newberg et al. (2003)[70] 发现了位于90kpc远的Sgr stream的潮汐尾。这之后，Belokurov et al. (2006)[71]对SDSS 的数据进行分析后又发现Sgr stream的nothern arm有一个分叉。而Sgr stream扫过的球状星团和它之间的关联也开始被研究。

越来越多的细节被揭示。Fellhauer et al.(2006)[72]尝试应用数值模拟的方法解释Belokurov e al.(2006)[71]发现的Sgr stream的分叉现象。他们认为这两个分支分别为Sgr dwarf galaxy绕银河系旋转两次留下的潮汐流。Martínez-Diagado (2007)[73]也运用数值模拟手段分析了Sgr stream和Virgo overdensity的关联，认为后者是前者在太阳系邻近的一个部分。Monaco et al.[74]对Sgr dwarf

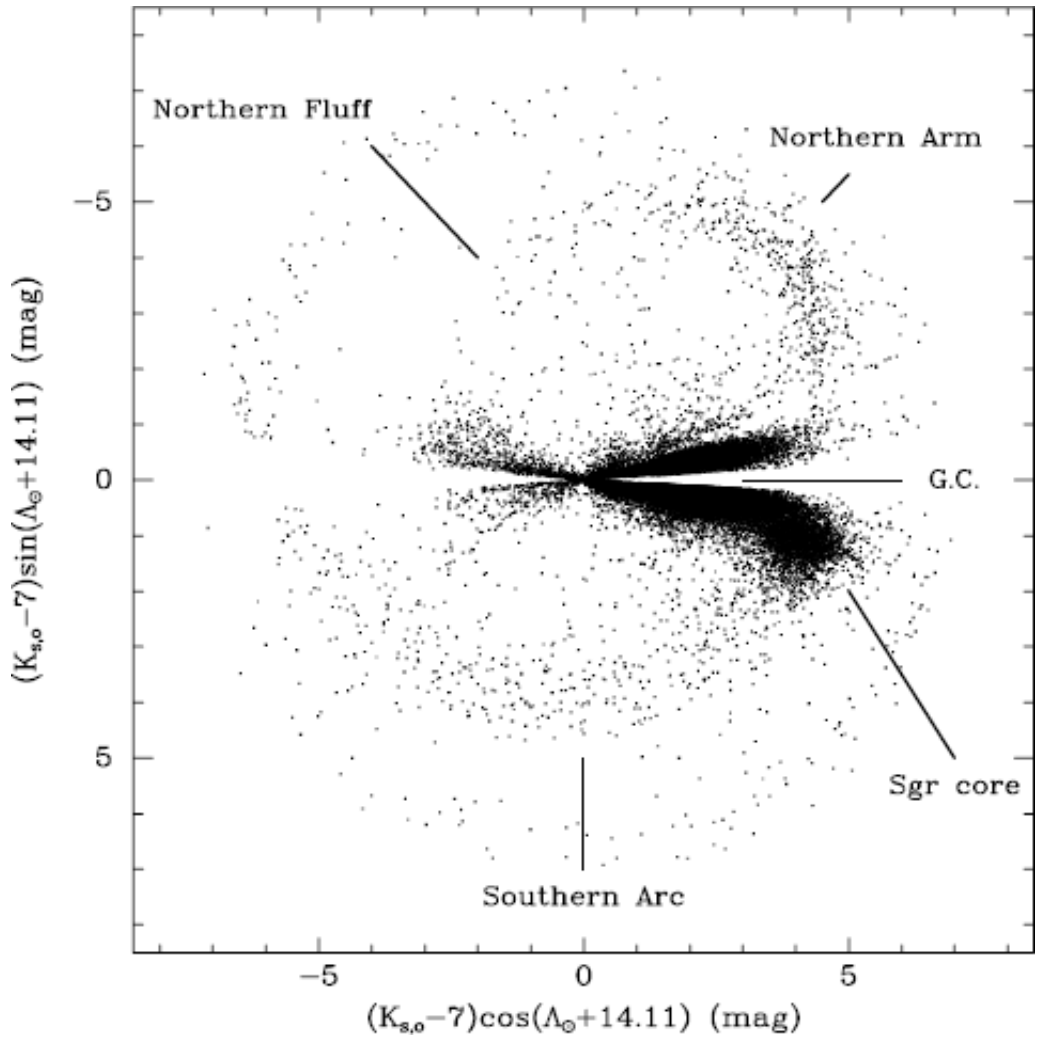

图 4.2: Majewski et al.(2003)图9显示的Sagittarius dwarf tidal stream

galaxy和Sgr dwarf tidal stream内的RGB星进行了高分辨率光谱研究，发现了星流的速度弥散只有8.3km/s，比矮星系本身的速度弥散11.2km/s还要低。而且星流中的金属丰度（$\langle$[Fe/H]$\rangle \sim -0.7$）也比矮星系本身的（$\langle$[Fe/H]$\rangle \sim -0.35$）要贫。而先形成的星流（Northern arm）比晚形成的星流（Southern arm）更贫一些。

Sagittarius dwarf tidal stream是目前比较确定的矮星系并合现象，还有很多并不能很明显确定其起源的子结构存在于晕中。实际上，现在已经发现的最大的子结构不是Sgr stream，而是被称为Virgo overdensity的，它在天空中占据

了约1000deg^{2}的面积。最早揭示出这个子结构的是Newberg et al.(2002)[2]，他们把SDSS数据中位于赤道上的F turnoff类型的恒星挑出并进行统计，发现很多不能用现有银河系模型解释的过密度区（overdensity），其中最强的一个就是Virgo overdensity。（图1.3是Newberg et al.(2007)[3]重新画出的清晰的Virgo Overdensity的空间位置图，图中下半部分标注了S297+63-20.5的高密度区域就是了，它距离太阳20kpc。S297+63-20.5是Newberg et al.(2002)中最早对这个overdensity的叫法。）

Virgo overdensity的起源众说纷纭，Juric et al.(2005)[75]利用SDSS测光的色指数估计恒星的距离，在一个更大的空间范围内发现Virgo Overdensity覆盖了1000deg^{2}的天区，距离在5kpc $\sim$ 15kpc，他们认为这个巨大的密度超出要么是一个星流，要么本身就是一个矮星系。Newberg & Yanny(2005)[76]和徐岩等（2006）[77]分别用三轴椭球体的银河系模型拟合SDSS的观测数据，以解释Virgo overdensity。对视向速度的研究认为[78, 3]Virgo overdensity包含几种中不同的恒星成份：不对称造成的场晕星和不同运动特征的星流。但是怎样从众多恒星中将这两种成份剥离开来，还没有很好的手段。

另一个比较显著的而又很奇怪的星流被称为Monoceros Ring，它也是最早出现在Newberg et al.(2002)[2]的恒星计数图中，位于反银心方向，很靠近银盘的低银纬上。Yanny et al.(2003)[79]使用SDSS对几个反银心方向的低银纬区进行光谱观测发现这个星流的视向速度弥散只有22 $\sim$ 30km/s明显小于厚盘和晕的平均值，虽然这些样本的视向速度方向和银盘的一致，但是圆周速度只有110km/s，比厚盘的平均圆周速度慢很多，但是又比晕星的快。此外他们还获得了这批样本的金属丰度为[Fe/H] $= -1.6 \pm 0.3$。通过数值模拟分析，有人认为[80, 81]这个星流是Martin et al.(2004)[82]发现的Canis Major矮星系的潮汐流。Canis Major被认为是目前发现的距离银河系中心最近的矮星系，只有$\sim$ 13kpc。对于Canis Major矮星系和它的星流对于银盘的动力学影响，特别是它们对于银河系盘的翘曲是否作出了贡献这些问题尚在探讨之中。

Orphan stream最早出现于Belokurov et al.(2006)[71]中（图4.3，即该文的图1，显示出Sgr stream分叉的Northern Arm，蓝色区域是Monoceros Ring，还有位于$(\alpha, \delta) = (165°, 0°) \sim (150°, 40°)$的较暗的Orphan stream）。这是一个不知道起源的星流，Belokurov et al.(2007)[83]对它的特征进行了研究，认为它最近的点距离20kpc。目前虽然已经有人给出了各种可能解释[84, 85, 86]，但是仍然没

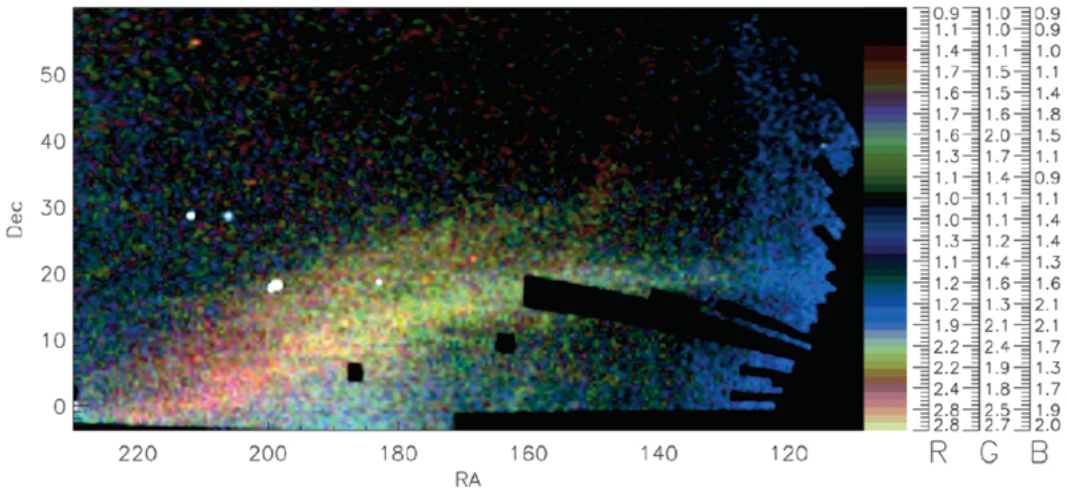

图 4.3: Belokurov et al.画出的几个显著的星流

有最后确定它和哪一个矮星系相关联，但很大可能最近发现的Ursa Major II矮星系是这个孤单的星流的孕育者。

除了上面提到的，还有一些稍小规模的子结构。它们有Anticenter stream[87]、cold stellar stream[36]、Pal 5 tidal tail[88]、NGC5466 tidal tail[89]。这些子结构大多是星团级的，相对前面提到的星流，它们弥散很小，所以轨道被刻画地更加精确，这对于研究银河系的引力势大有帮助。

这些子结构的发现大大改变了银河系晕的外观，原来那个平滑的对称椭球已经被一个“彩色毛线球”所替代。由于几个主要的并合事件涉及的恒星数目都有可能在$10^8 M_\odot$级，而银晕的总恒星质量也不过在$10^9 \sim 10^{10} M_\odot$，因此并合事件对银晕的结构研究的影响已经达到了不能忽视的程度。在对银河系建立计数模型的时候，不得不考虑这些星流在局部对恒星计数的巨大影响。Bell et al. (2007)[90]在尝试用各种银河系模型对SDSS的恒星计数进行拟合的时候发现由于这些巨大的子结构的存在，很难应用原来的平滑的模型进行拟合。最好的拟合也仍然有$\gtrsim 40\%$的误差。根据星系并合的数值模拟结果他们认为银河系的恒星晕有很大的比例是由于吸积而产生的。

CDM宇宙学模型描绘的是一个层级式的宇宙，很小的质量团块并合成大一些的，大一些的并合成更大一些的。按照CDM模型的预言，银河系周围应该有很多矮星系。在SDSS之前，共发现了9个矮椭球星系（dwarf spheroidal，dSph），它们是Draco、Ursa Minor、Fornax、Carina, Sculptor

、Leo I 、Leo II 、Sextans和Sagittarius。其中有七个都是在照相底片上用眼睛找到的。Sextans是Irwin et al. (1990)[91]在进行底片自动扫描的过程中发现的，Sagittarius 是Ibata et al.(1995)[92] 通过在核球的视向速度巡天中用运动学特征确认的。除此之外，还应该加上Martin et al. (2004)[82] 发现的Canis Major。SDSS巡天数据发布以后，一批新的dSph 很快被挖掘出来。它们有Ursa Major I[93] 、Canes Venatici I[94] 、Bootes[95] 、Ursa Major II[96, 97] 、Coma Berenices 、Canes Venatici II 、Leo IV 、Hercules[98]和LeoT[99]。此外，还从SDSS数据中发现了低光度球状星团：Willman 1[100] 、Segue 1[98] 、Koposov 1 和Koposov 2[101]。 最近从SDSS数据中发现的这些新的银河系dSph以及球状

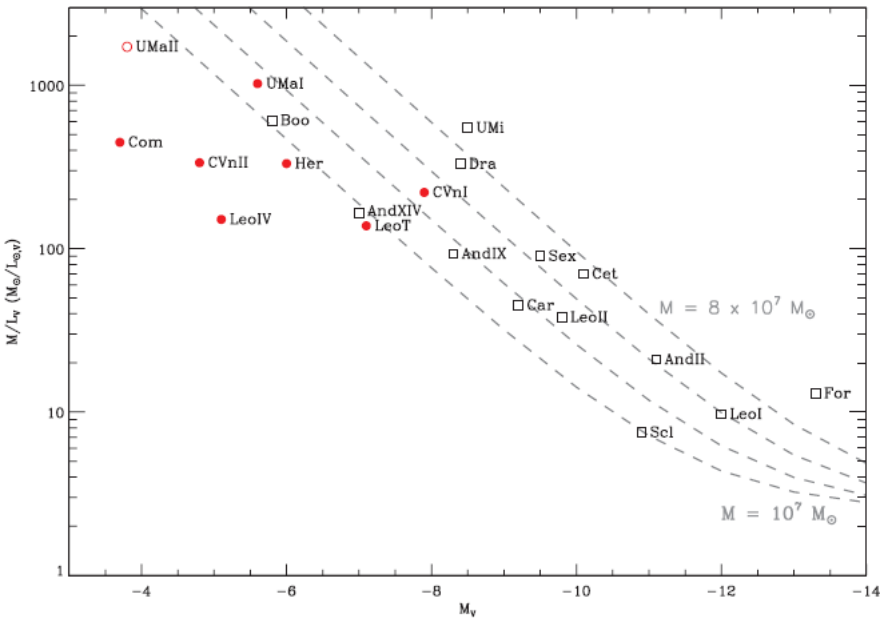

图 4.4: Simon & Geha(2007)中的矮星系绝对星等－质光比图

星团无一例外的属于低光度天体。虽然他们很暗，但是这些发现的意义却十分重大。Simon & Geha(2007)[102]使用Keck望远镜对这些新发现的极暗的矮星系中的8个（除去Bootes）进行了光谱研究。根据成员星的速度弥散，他们估计了这些极暗矮星系的质光比和绝对星等的关系（图4.4，来自Simon & Geha(2007)的图15）。这张图充分说明了矮星系是暗物质主导的星系，这同CDM宇宙模型的预言一致。Simon & Geha还认为，应该还有更小质量的矮星系尚待发现。根据CDM的预测，在SDSS DR5的观测范围内应该观测到的矮星系是现在的4倍左右。此外，他们的观测发现Ursa Major II是一个正在被潮汐力瓦解的矮星系，这

和先前认为的Ursa Major II 是Orphan Stream的主体的论断[84, 85, 86]是不矛盾的。

根据上面的概述，我们认为，首先，SDSS覆盖的天区内应该仍然有数十个尚待发现的矮星系；其次，这些矮星系一方面将对于CDM宇宙模型的结论作出有力的验证，另一方面也对银河系晕的并合历史有重要意义。据此，我们将开展对SDSS巡天结果的数据挖掘，希望从中找到新的矮星系。在第五章中我们将详尽描述我们采用的方法和得到的结果。

第五章　寻找新的矮星系和球状星团

5.1　数据的处理

寻找矮星系的数据资源是SDSS DR5[103]，它提供了u, g, r, i和z五个波段[104]的8000deg^{2}的测光数据。我们选择其中的点源。这些点源包括恒星，类星体和非常暗的分类错误的星系。我们首先要从如此浩繁的天体中找出可能的有矮星系的位置。矮星系所在的区域恒星密度应该比周围大。但是如果矮星系非常暗，恒星密度很低，就不能从前景场星中区分出来。通过特定星等和颜色剪裁，突出矮星系的成员星的特征而压制场星的影响，就可能让矮星系从背景中突出出来。我们选择的星等范围是$19 < \mathrm{i} < 22$，选择的颜色范围是$0 < \mathrm{g} - \mathrm{i} < 1$。这里的g和i星等都已经经过了消光改正，使用的消光值来自Schlegel et al.(1998)[105]。取这样的星等范围就可以有效避免厚盘的恒星的干扰，而颜色选择在这个范围可以避开更红的薄盘晚型星和更蓝的类星体。在这个颜色范围内，晕星和矮星系成员星的主序星，亚巨星和红巨星都分布其中。削弱了薄盘和厚盘场星的影响以后，天空投影的恒星数密度主要贡献来自晕星，而晕星的密度较低，这就有可能让矮星系从背景中突出出来。为了避开Sgr stream和Virgo overdensity的影响，我们将考察的区域设定在$120° < \alpha < 270°$，$25° < \delta < 70°$。为了后面的应用，初始的数据查询并没有进行星等和颜色剪裁，而是把所有考察区域内的$\mathrm{i} \leqslant 23$的点源全部下载下来。这个过程是通过SDSS提供的casjobs平台[1]查询的，大约3.7×10^7个天体被下载到本地的数据库中。该数据库使用DataNode发布数据，可以使用VO-DAS进行数据访问。

我们使用Java编写程序，做成一个VO-DAS客户端，在这个客户端之上实现计算。我们从DataNode中取得星等范围是$19 < \mathrm{i} < 22$，颜色范围是$0 < \mathrm{g} - \mathrm{i} < 1$的样本，在每个尺寸为$0.2° \times 0.2°$的格子里计算恒星计数。之所以使用这样的格子尺寸是因为绝大多数矮星系的的尺寸都在$\sim 10'$一级，同时考虑取整方便，所以将格子的尺寸定成这个数值。

共统计了750×225个格子，计数之后的密度分布如图5.1所示。银心方

[1] http://casjobs.sdss.org/

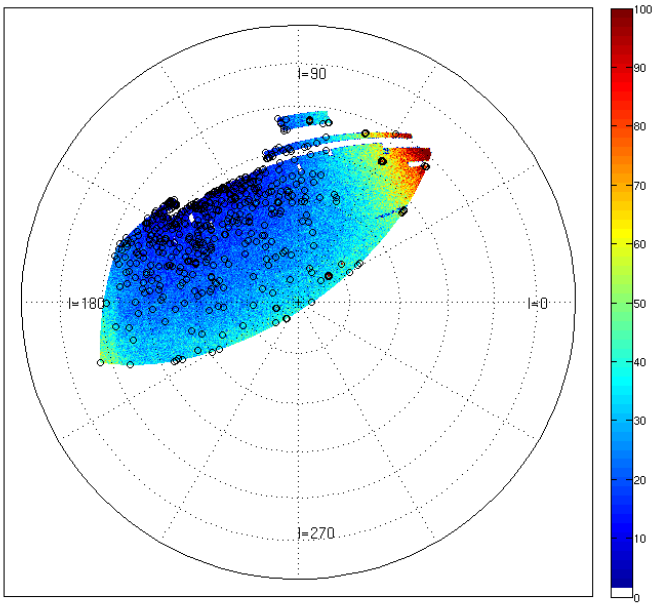

图 5.1: 格子为0.2° × 0.2°的恒星数密度在银道坐标系的投影以及过密度点

向在图的右边。可以从颜色的发布大致看出，除了Anticenter stream在左边缘，Monoceros stream的一小部分在左下角，Sgr stream的一小部分在下边缘有稍许增强以外，恒星计数的主要贡献来自于晕中的场星。从这样的着色图用眼睛很难分辨独立的过密度点，只有象NGC 5466，NGC 6205等这样较近的球状星团可以依稀分辨出来。因而，我们需要使用程序自动搜索那些孤立的过密度点。

考虑到场晕星在空间分布并不均匀，但是局部起伏较小，因此我们对于每个图中的考察点使用一个以考察点为中心的11 × 11大小的窗口，只计算这个窗口内的场晕星。除去中心的考察点以外的所有点（11 × 11 − 1）的平均值作为该局部的场晕星平均密度，记为μ。这些点的随机起伏用方差来衡量，记为σ。对于每个考察点建立一个统计量

$$\tilde{n}_{center} = (n_{center} - \mu)/\sigma \tag{5.1}$$

其中n_{center}是中心考察点的恒星计数。对于我们考察的750 × 225，共168750个点，除去其中有18627个没有观测数据的点，它们对应的统计量$\tilde{n}_{center}$服从$\tilde{\mu}$ =

0.33，$\tilde{\sigma} = 0.56$的正态分布。我们选择那些$\tilde{n}_{center} > \tilde{\mu} + 3\tilde{\sigma} \sim 2$的点为过密度点，共计524个。对这个正态分布积分后，我们估计出正态分布允许全部考察点中有大约0.14%，或者207个过密度点是因为随机起伏而落入我们选择的范围。换句话说，我们期待我们筛选出来的524个过密度点中有317个不是偶发的过密度点，它们是我们要重点检查的。附录C列出了全部524个点的位置坐标、$\tilde{n}_{center}$值和对它们进行的证认。

5.2　候选体甄选方法

从这524个过密度点中找到可能的银河系的伴星系是非常困难的。这里面除了前面分析的随机涨落带来的混入以外，还有巨大的星系的外晕不能区分的星团被SDSS的pipeline当成恒星处理，造成的成团，有遥远的星系团因为没有正确进行恒星 / 星系分类而混入，也有少数亮星附近因为光晕的污染而没有得到测光数据，致使考察邻近区域时场星背景便暗而带来的混入，当然也有在位于探察区域内的球状星团和矮星系的干扰。另外，524个过密度点如果距离边界小于0.5°，则也会被忽略掉。

我们把所有已知天体证认出来，发现所有已知的球状星团和矮星系，包括低光度矮星系都落入我们第一轮筛选的524个过密度点内了。它们是Pal 4, UMa I, UMa II, CVn I, Draco, NGC 6341, NGC 5466, NGC 6205, NGC5272和Willman 1。除了这些已知的源，剩下的高密度点仍然有432个。为了对这432个目标进行证认，我们需要它们局部的颜色－星等图（CMD）的帮助。我们使用VO-DAS的客户端接口函数通过Java编程，根据附录C中所列出的坐标，向数据库依次查询以每个点为中心的$1° \times 1°$范围所有恒星。为了验证MATLAB的可用性，我们还同时使用了MATLAB替代Java编程，完成同样的工作。比较发现两者的效率相差不大。我们以9个已知星团 / 矮星系，Pal 4, UMa I, UMa II, CVn I, Draco, NGC 6341, NGC 5466, NGC 6205和Willman 1，在本系统内输出的CMD作为模板，考察所有未知的点的CMD。这9个模板涵盖了距离模数从14.2mag到21.7mag的范围，既有高密度的球状星团，也有极暗的dSph，具有较广泛的代表性。尝试用Hough变换方法进行匹配，发现由于未知点的CMD通常所包含的恒星数目比较小，因此噪声很大，匹配误差较大。于是采用人工分级方法检出最可能的样本出来。分级的原则是看CMD中是否出现星族的特征：明显的主序、HB形状、RGB形状、蓝离散星（BS）的形状，相似形状越多等级越高。同时还要看在

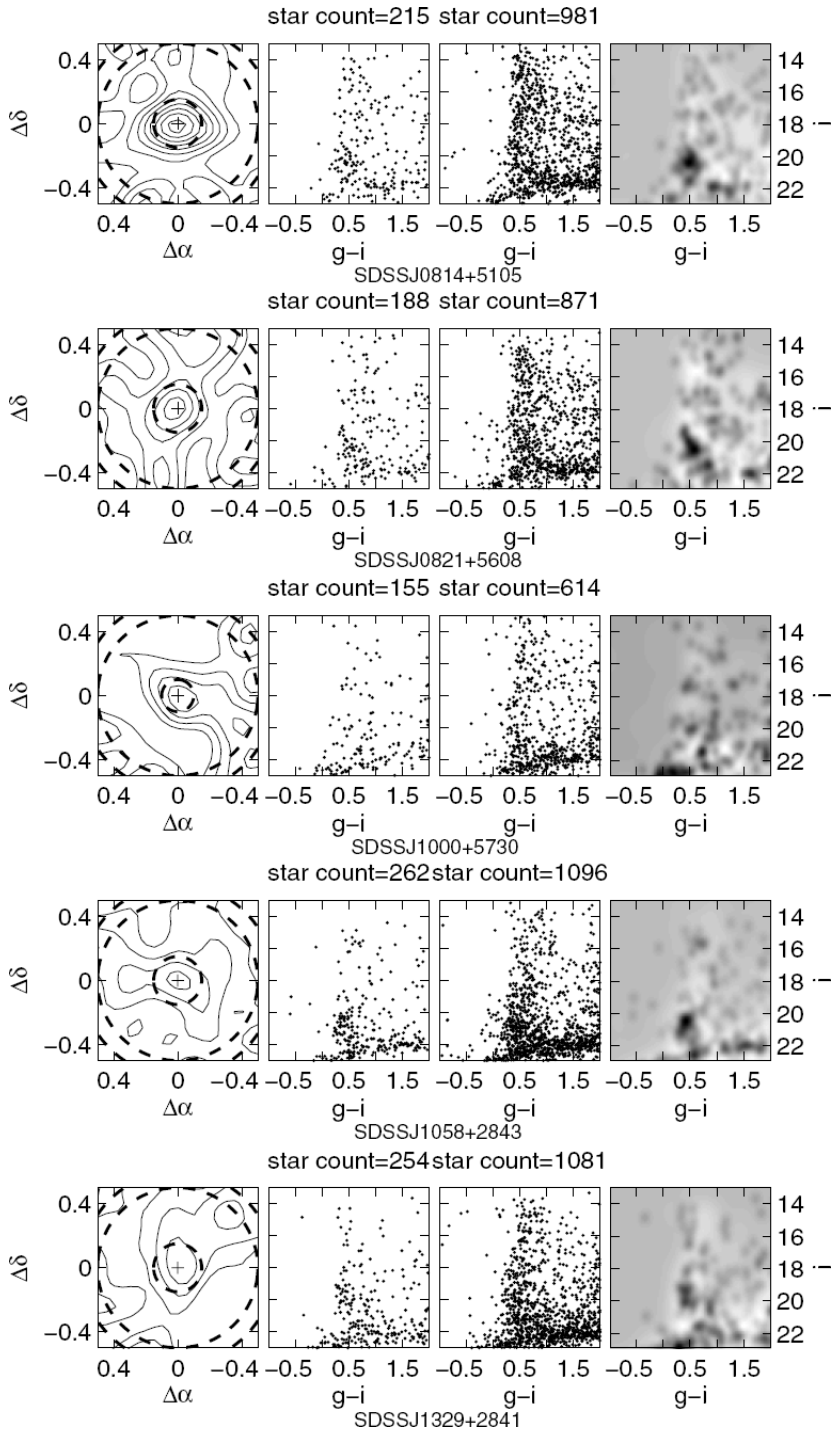

图 5.2: 5个候选体的空间密度轮廓，CMD和减掉场星后的Hess图

空间的密度，凝聚度高的等级高。综合这两方面的等级，挑出等级最高的，再进行详细甄别和讨论。首先挑出了20个候选体，从中又筛选出10个。仔细考虑这10个的CMD，检查除了形状以外，各个分支的位置和相对密度是否合理，以及考虑SDSS两次扫描的接缝问题等其他因素，最终选择了5个最可能的候选体，参见图5.2。图中第一列为在赤道坐标系下各个候选体的恒星数目的分布轮廓。第二列是0.1°或者0.15°半径的中心圆形区域（第一列中的虚线小圆以内）的恒星的CMD。第三列是作为参考的背景场星的CMD图（第一列中两个大一些的虚线圆形所夹的环形区域，内半径为0.5°，外半径0.6°）。第四列是第二列转换成密度图以后减去第三列的密度图之后的Hess图，相减之前已经做了等面积的归一化。第四列反映了去除背景干扰以后的颜色－星等的特征。

5.3　候选体的距离估计

对图5.2的第四列，我们可以通过用不同参数的星族合成等年龄线进行拟合来估计5个候选体的距离。我们选择Girardi et al.(2004)[51]提供的SDSS五色测光系统的等年龄线。为了能够进行匹配，我们把每一条等年龄线转换成带有一定宽度的0－1模板，等年龄线通过的附近为1，其他地方为0。生成模板的时候，等年龄线中的AGB星已经被扣除掉了。模板尺寸为100×200，横轴为色指数$g-i$，从-1mag变到2mag，步长0.03mag；纵轴为星等i，从-10mag变到10mag，步长为0.1mag。

令金属丰度Z的取值为0.0001, 0.0004, 0.001, 0.004，年龄A的取值从9.5到10.25，步长0.05，单位$\log_{10}(\mathrm{yr})$，距离模数DM从13mag变到21.9mag，步长0.1mag。这样我们就获得了5760个的模板。将5个候选体的去除背景的Hess图归一化到横轴为色指数$g-i$，从-1mag变到2mag，步长0.03mag；纵轴为星等i，从13mag变到23mag，步长为0.1mag，成为100×100的灰度图。图像中的灰度值归一化为单位面积恒星数目。我们用5760个模板依次和各个候选体的Hess图进行匹配，优化公式定义为

$$R = \frac{\sum\limits_{g-i,i} (C(g-i,i)(1-T(g-i,i,Z,A,DM)))}{\sum\limits_{g-i,i} C(g-i,i)}. \tag{5.2}$$

其中，C(g−i, i)表示候选体的Hess图，T(g−i, i, Z, A, DM)表示模板的0－1图。R是衡量匹配好坏的标志。R越小，匹配得越好。 最佳匹配的R，$R_{\min}$，表示候选

表 5.1: 等年龄线拟合方法对已知天体的实验

名称	Z	A ($log_{10}(yr)$)	DM (mag)	文献DM (mag)	R_{min}	文献
UMa II	0.0004	10.1	17.9	17.5	0.50	Zucker et al. [96]
UMa I	0.0001	10	19.9	20	0.41	Willman et al. [93]
Pal4	0.001	10.2	20.2	20.02	0.37	Harris [106]
CVn II	0.0004	10.2	20.9	20.9	0.25	Belokurov et al. [98]
CVn I	0.0004	10.15	21.7	21.75	0.28	Zucker et al. [94]
NGC5272	0.001	10.1	15	15.04	0.21	Harris [106]
NGC5466	0.001	10.05	16.1	16.1	0.34	Harris [106]
NGC6205	0.004	9.95	14.7	14.28	0.21	Harris [106]
NGC6341	0.001	10	14.9	14.59	0.28	Harris [106]

Z：金属丰度；
A：年龄，单位为$log_{10}(yr)$；
DM：我们拟合的距离模数；
文献DM：文献查到的距离模数；
R_{min} ：公式5.2的最佳拟合值；
文献：参考的文献。

体的Hess图中有多少比例的恒星没有落在最佳匹配的等年龄线附近。在最极端的情况下，如果候选体的Hess图上的灰度是均匀分布的，考虑到等年龄线很细，这样得到的$R_{min} \sim 1$。也就是说，Hess图中绝大多数的恒星都是和这条等年龄线无关。另一种极端情况是，候选体中所有的恒星都集中在等年龄线附近，此时，$R_{min} \sim 0$，即几乎没有和这条等年龄线无关的恒星。R的计算过程也是在两种环境下进行的，一种是Java编程，另一种是MATLAB编程。相比而言，MATLAB的编程进度很快，运行效率也比较高。这种模板匹配方法比较容易采用矩阵计算的形式描述，而这正是MATLAB的优势。

为了检查R的有效性，我们对附录C中所有的点对应的Hess图都应用公式5.2进行了计算，附录C各个表的第三列即是计算的结果。图5.3中短线组成的

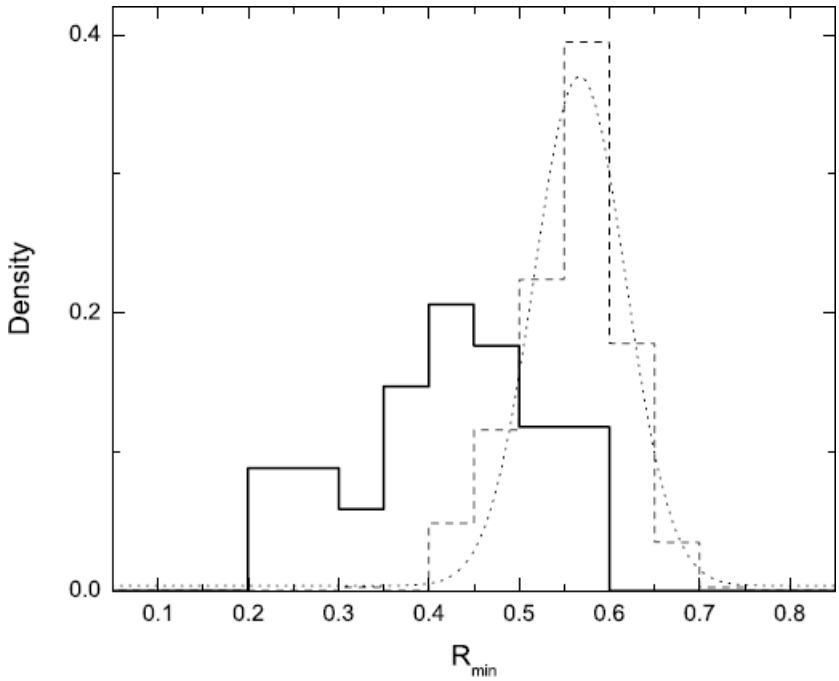

图 5.3: 所有过密度点的R_{min}的分布和已知天体的对比

阶梯线是所有过密度点的R_{min}的密度分布图。点线组成的曲线是对它的高斯拟合，均值为0.57，方差0.1。实线的阶梯线是其中已知的球状星团，矮星系，星系，星系团和5个候选体的R_{min}的密度分布。可以看出，所有已知天体的R_{min}值虽然和显著小于全体的，而且存在双峰：峰值在0.2和0.3之间的是球状星团的R_{min}值，峰值在0.4的主要是矮星系和候选体的贡献。这张图说明球状星团由于其星的密度高，因此CMD图的特征最明显，所以和等年龄线的匹配效果也最好；矮星系虽然很暗，其成员星被观测到的较少，但是同那些可能仅仅是随机起伏的过密度点比较，仍然呈现出可以识别的CMD特征，因此也能得到相对较低的R_{min}。已知天体中UMaII的$R_{min} = 0.5$，是其中最差的。这也和UMaII正处于瓦解状态这个事实相一致。 在确认了等年龄线拟合的方法是有效的以后，我们还需要确定这个方法对于各项参数估计的精度。由于我们只使用了四个金属丰度，因此并不期冀这个方法可以给出精确的金属丰度。年龄和金属丰度是耦合的，因此，尽管年龄的参数步长足够细致，但是仍然不能得到准确的年龄。年龄参数设置成小步长的目的是保证金属丰度和年龄不会对距离估计的精确度有太大影响。用8个已知天体作为测试，拟合的结果保存在表5.1中。使用我们的等年龄线拟合方法得到的结果和文献中给出的结果比较，距离模数估计的标准差为0.23mag。

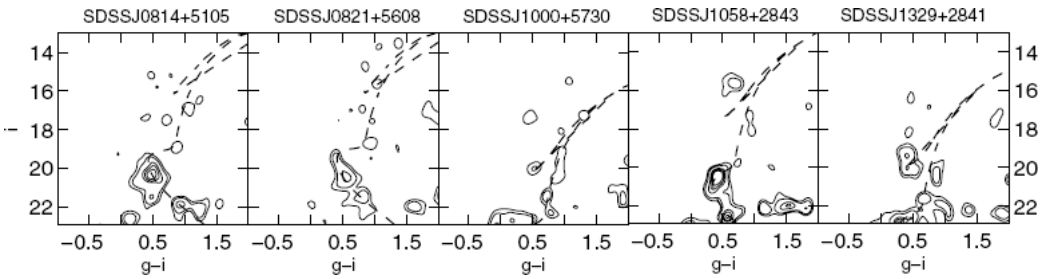

图 5.4: 5个候选体的等年龄线

这样的结果是可以接受的。

使用该方法对5个候选体的距离进行估计，同时得到最佳拟合的等年龄线。表5.2中给出了5个候选体的距离估计结果，图5.4中画出了5个候选体对应的等年龄线。

5.4 候选体的物理和几何特征

表5.2中还列出了每个候选体的基本物理参数：中心位置，几何特征和绝对星等。本节介绍这些特征的估计方法。

为了算出候选体的中心位置，需要首先确定哪些恒星属于它的成员。我们采用两个判别标准进行选择。首先在空间分布中，选择那些落入初选范围$1^\circ \times 1^\circ$范围内的星，然后选择落在图5.2最右列的Hess图中0.05mag/bin或者更密的区域内的星。在Hess图中，星的密度接近0或者是负的就认为是噪声和背景干扰。

将选出的恒星画在经过球面投影改正以后的平面上，通过高斯核的非参数密度估计计算出经验密度分布，并利用投影平面上的位置坐标的一阶矩求出中心位置来。一阶矩的计算公式为

$$\begin{aligned} x_0 &= \frac{\sum_i x_i I(x_i, y_i)}{\sum_i I(x_i, y_i)}, \\ y_0 &= \frac{\sum_i y_i I(x_i, y_i)}{\sum_i I(x_i, y_i)}. \end{aligned} \tag{5.3}$$

其中，x_i, y_i是密度分布中第i个像素的坐标，I(x, y)是(x, y)这个位置的密度。通

表 5.2: 5个候选体的各项特征参数

Parameter	SDSSJ0814+5105	SDSSJ0821+5608	SDSSJ1000+5730	SDSSJ1058+2843	SDSSJ1329+2841
赤经(J2000)	$08^h13^m42^s$	$08^h21^m15^s$	$10^h00^m28^s$	$10^h58^m04^s$	$13^h29^m13^s$
赤纬(J2000)	$+51°05'27''$	$+56°08'16''$	$+57°30'10''$	$+28°42'39''$	$+28°41'27''$
银经(deg)	167.743	161.665	155.506	202.649	45.716
银纬(deg)	33.449	34.615	47.372	64.966	81.513
金属丰度（Z）	0.004	0.004	0.001	0.004	0.0004
年龄($\log_{10}$(yr))	10.05	10	10.15	9.95	10.1
距离模数$(m-M)_0$(mag)	15.7	15.7	19.6	16.9	19.4
R_{min}	0.5	0.45	0.47	0.36	0.58
距离(kpc)	$13.8^{+1.5}_{-1.4}$	$13.8^{+1.5}_{-1.4}$	$83.2^{+9.3}_{-8.4}$	$24^{+2.7}_{-2.4}$	$75.9^{+8.5}_{-7.6}$
指数模型半光半径r_h(arcmin)	6.2 ± 1.0	4.7 ± 1.0	8.1 ± 2.7	4.7 ± 0.7	8.6 ± 2.5
Plummer模型半光半径r_h(arcmin)	5.4 ± 0.8	4.3 ± 0.8	8.3 ± 2.2	4.8 ± 0.5	8.8 ± 1.9
指数模型半光半径线尺寸(pc)	24.9 ± 4.0	18.9 ± 4.0	196.0 ± 63	32.8 ± 4.9	189.8 ± 55
Plummer模型半光半径线尺寸(pc)	21.7 ± 3.2	17.3 ± 3.2	200.8 ± 53	33.5 ± 3.5	194.2 ± 42
背景水平(arcmin^{-2})	0.11	0.1	0.02	0.12	0.12
指数模型绝对星等M_V(mag)	-0.77	-1.63	-4.15	-2.99	-3.91
Plummer模型绝对星等M_V(mag)	-0.81	-1.42	-4.16	-2.98	-3.92

过二阶矩求出主轴的方位角和偏心率，

$$\theta = \frac{1}{2} arctan \frac{2\sigma_{xy}}{\sigma_{yy} - \sigma_{xx}},$$
$$e = \frac{\sqrt{(\sigma_{xx} - \sigma_{yy})^2 + 4\sigma_{xy}^2}}{\sigma_{xx} + \sigma_{yy}}. \tag{5.4}$$

其中σ_{xx}, σ_{yy}和σ_{xy}是密度分布的二阶矩。但是由于这5个候选体的恒星数目太少，其构成的形状非常杂乱，因而噪声主导了我们看到的形状，这样求出的椭率和主轴方位角是没有意义的。因此，我们简单地将各个候选体的投影形状当作圆形轮廓来处理。背景水平采用距离中心位置$30' \sim 60'$内的全部恒星的平均密度来估计，并从候选体的投影密度轮廓中减去（各个候选体对应的背景水平已经在表5.2中列出）。图5.5显示出5个候选体的密度轮廓，等密度线从外至内分别代表背景水平之上2, 3, 4, 5, 6, 7和7.5σ的密度。

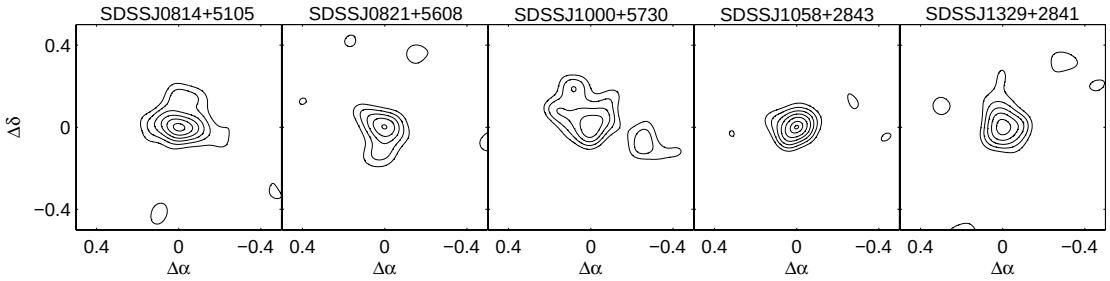

图 5.5: 5个候选体的等密度线

在得到密度轮廓以后，我们可以按照从中心向外算出各个距离半径处的平均恒星密度，得到径向轮廓（Radial profile）。我们希望使用经验的轮廓模型对其进行拟合进而得到候选体的几何特征。由于我们并不清楚它们的类型，是球状星团，dSph还是潮汐力瓦解掉的矮星系碎片，因而我们不清楚它们的恒星密度分布到底怎样的。我们不知道哪轮廓种模型是适合的，于是就选择了最简单的几种。我们应用McConnachie & Irwin(2006)[107]对M31的dSph的结构进行研究时运用的三种矮星系径向轮廓模型同时进行拟合，它们是King模型，Plummer模型和指数模型。King轮廓是King(1962)[108]给出的简单经验模

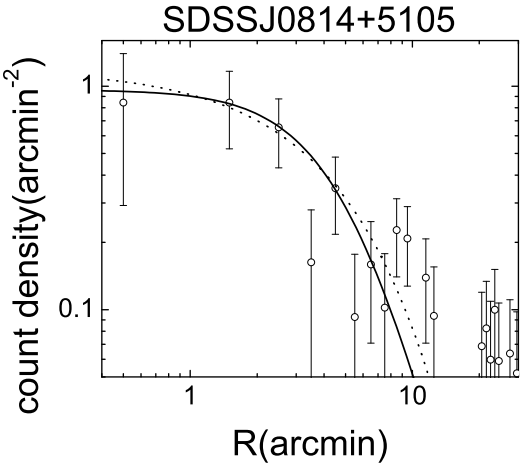

图 5.6: SDSSJ0814+5105的径向轮廓

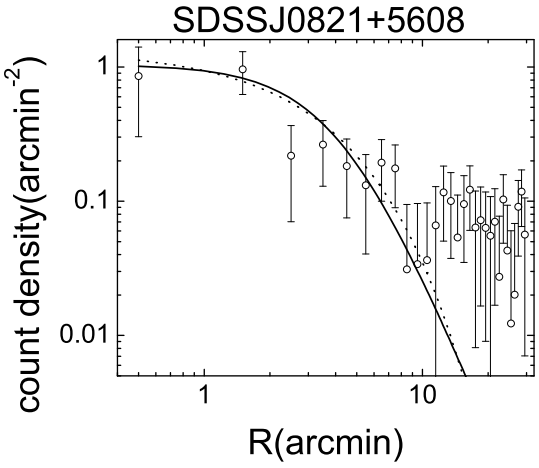

图 5.7: SDSSJ0821+5608的径向轮廓

型，其表达式为

$$f_K = A\left\{\frac{1}{[1+(r/r_c)^2]^{\frac{1}{2}}} - \frac{1}{[1+(r_t/r_c)^2]^{\frac{1}{2}}}\right\}^2. \tag{5.5}$$

其中A是比例参数，r_t称为潮汐半径，r_c称为核半径。Plummer模型是经常在进

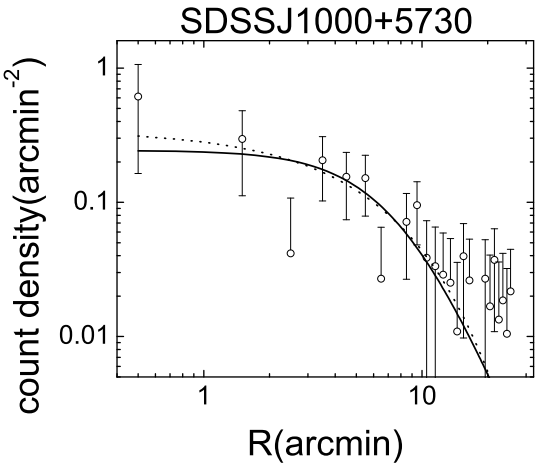

图 5.8: SDSSJ1000+5730的径向轮廓

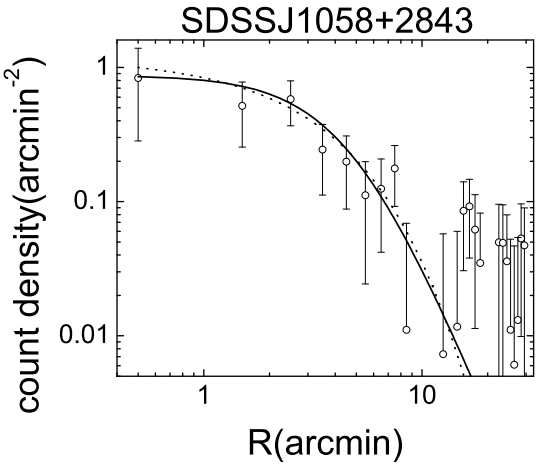

图 5.9: SDSSJ1058+2843的径向轮廓

行N-体模拟的时候使用的模型，其表达式是

$$f_P = B\frac{b^2}{(b^2+r^2)^2} \tag{5.6}$$

其中B是比例参数，b是Plummer核半径。指数模型最早由Faber & Lin(1983)[109]引

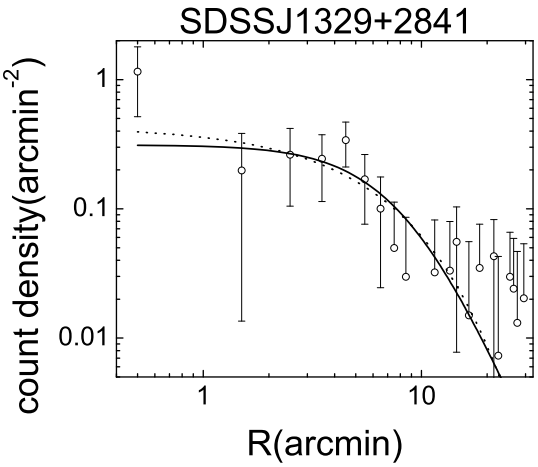

图 5.10: SDSSJ1329+2841的径向轮廓

入，用来描述dSph的投影表面密度分布，其表达式是

$$f_E = Cexp(-\frac{r}{r_e}). \tag{5.7}$$

其中C是比例参数，r_e被称为有效半径。

由于星数太少，经过实验发现拟合三参数的King模型的结果误差很大，所以我们摒弃King模型，而使用参数较少的后两个模型进行轮廓拟合。拟合的结果在图5.6-5.10和表5.2中。图中空心圆是观测数据，实线是Plummer拟合结果，虚线是指数拟合结果。在表中我们直接没有列出Plummer和指数模型的参数，而是列出了由它们拟合的轮廓的半光半径（Half-light radius）的值。这个值是从中心向外积分拟合后的表面密度，积分值等于总积分一半处的半径值。

为了检验上述方法的估计效果，我们对两个已知天体球状星团Pal 5和dSph Bootes进行了检验。对于Pal 5，我们用指数模型得到$r_h = 2.64' \pm 0.07$，用Plummer模型得到$r_h = 2.91' \pm 0.07$。和Harris(1996)[106]得到的$r_h = 2.96$是相近的。对于Bootes，我们用指数模型得到$r_h = 14.4' \pm 1.8$，用Plummer模型得到$r_h = 14.6' \pm 1.5$。这和Belokurov et al.(2006)[95]得到的结果$r_h = 13.0' \pm 0.7$和$r_h = 12.6' \pm 0.7$也是相近的。因此，尽管我们考察的对象的观测数据误差较大，但是，我们采用的方法是可行的，结果是可靠的。

上述拟合计算全部采用MATLAB编程，充分利用MATLAB提供的曲线

拟合工具箱，只要我们自己定义拟合函数的表达式（公式5.5, 5.6和5.7）的MATLAB形式就可以了。

最后，我们估计各个候选体的绝对星等。我们把减掉背景场星影响以后的r_h半径之内的颜色－星等图画成流量单位表示的Hess图（类似图5.2的第四列，但选择的半径有所不同），其中所有取值为正的像素点和一个加宽的最佳拟合的等年龄线做成的0－1蒙版相乘，积分以后的结果就是总流量的一半。用这样的方法计算g波段和r波段的总流量，并根据下面所列公式转换成B和V波段[104]，最终求出绝对星等。

$$(B-V)_{tot} = (g_{tot} - r_{tot} + 0.23)/1.05. \tag{5.8}$$

$$V_{tot} = r_{tot} + 0.49(B-V)_{tot} - 0.11. \tag{5.9}$$

$$M_{V,tot} = V_{tot} + 5 - 5log(d). \tag{5.10}$$

应用不同模型的半光半径r_h算出的5个候选体的绝对星等列在表5.2中。作为对算法的测试，我们应用同样方法对Com, CVn II, Her和Leo IV这些用SDSS发现的低光度dSph的绝对星等进行估算。估算的结果和文献[98]结果进行比较（表5.3），发现我们给出的估计最大和文献值差0.55mag，比文献估计的绝对星等的标准差0.6mag还要小。这个检验证明了我们的方法的可行性。

表 5.3: M_V估计方法的测试

名称	文献M_v	$M_{v,exp}$	$M_{v,plummer}$	$\Delta M_{v,exp}$	$\Delta M_{v,plummer}$
Com	-3.7 ± 0.6	-3.99	-3.71	0.29	0.01
CVn II	-4.8 ± 0.6	-4.9	-4.77	0.1	0.03
Her	-6 ± 0.6	-5.82	-5.81	0.18	0.19
Leo IV	-5.1 ± 0.6	-4.55	-4.56	0.55	0.54

文献M_v：文献中得到的矮星系绝对星等；
$M_{v,exp}$：指数模型的绝对星等；
$M_{v,plummer}$：Plummer模型的绝对星等；
$\Delta M_{v,exp}$：指数模型和文献的绝对星等之差；
$\Delta M_{v,plummer}$：Plummer模型和文献的绝对星等之差。

5.5　讨论

在从524个过密度点中选出这几个候选体以后，我们需要分析它们可能的类型是什么。这些候选体有可能是球状星团，有可能是dSph，也有可能只是被潮汐力瓦解掉的矮星系的碎块（Tidal debris）。简单地看它们的半光半径，SDSSJ0814+5105，SDSSJ0821+5608和SDSSJ1058+2843很象球状星团，SDSSJ1000+5730和SDSSJ1329+2841很象dSph。我们参考Belokurov et al.(2007)[98]中的方法，使用半光半径（r_h）－绝对星等（M_V）图来区分球状星团和矮星系。图5.11中给出了150多个球状星团[106]（图中的空心三角形），SDSS发现的dSph和球状星团[93, 94, 95, 96, 98, 100, 98]（图中星形表示），本星系群中的矮星系[110, 111]（图中空心菱形表示）以及Andromeda附近的dSph[107, 112]（图中空心矩形表示）在r_h－M_V平面上的位置。图中虚线族是等表面亮度线。可以看出，除了Seg1以外，球状星团都位于图的左侧，而矮星系都在右侧，中间有一个窄窄的空间是没有任何天体的（Seg1和Com之间的空隙）。利用这个性质，我们将5个候选体也点在这个平面中后（图中实心圆表示）看到SDSSJ0814+5105和SDSSJ0821+5608落在球状星团一侧，但是比最暗的球状星团AM4还要暗一些。SDSSJ1058+2843虽然绝对星等落在球状星团的范围内，但是半光半径偏大，落在空隙的边缘。而SDSSJ1000+5730和SDSSJ1329+2841完全落在了矮星系的范围内，只是表面亮度非常低。

由于SDSSJ0814+5105，SDSSJ0821+5608和SDSSJ1058+2843的距离较近，在SDSS的数据中就已经能够看到主序，我们可以对它们的潮汐半径作出估计，判断它们是否还处于引力束缚的状态。潮汐半径的估计可以使用Binny & Tremaine（1987）[113]给出的公式:

$$r_{tidal} \sim R_{cand}\left(\frac{\mathcal{M}_{cand}}{3\mathcal{M}_{MW}}\right)^{\frac{1}{3}}. \tag{5.11}$$

式中，R_{cand}是候选体到银心的距离，$\mathcal{M}_{\rm cand}$和$\mathcal{M}_{\rm MW}$分别是候选体和银河系的总质量。我们假设太阳到银河系中心的距离是8kpc，旋转速度$v_c = 220 kms^{-1}$。我们用候选体和Pal 5进行比较来估计它们的潮汐半径。如果$r_{\rm tidal,Pal5} = 110$pc, $R_{\rm Pal5} = 18.6$kpc（来自Harris(1996)[106]）和$\mathcal{M}_{\rm Pal5} = 8318\mathcal{M}_{\odot}$[114]分别是Pal 5的潮汐半径，银心距离和质量，那么

$$r_{tidal,cand} = \frac{r_{tidal,Pal5}R_{cand}}{R_{Pal5}}\left(\frac{\mathcal{M}_{cand}}{\mathcal{M}_{Pal5}}\right)^{\frac{1}{3}}. \tag{5.12}$$

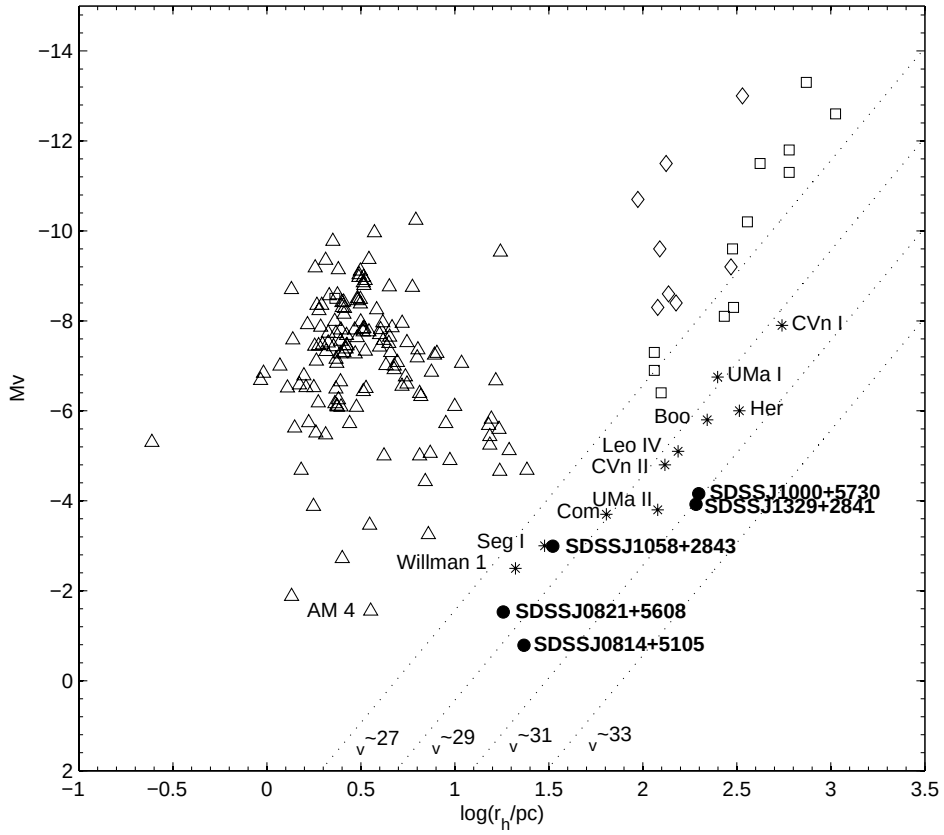

图 5.11: 矮星系和球状星团的分类

为了估计候选体和Pal5的质量比，我们用候选体的光度函数和Pal 5的光度函数进行对比（图5.12）。为了减少误差，我们只使用$M_V = 4 \sim 6$之间的部分进行比较。这部分的恒星主要是主序上的F、G型星，是我们在颜色－星等图上看到的三个候选体分布密度最高的部分。我们最后得到SDSS0814+5105的潮汐半径为41pc，SDSS0821+5608和SDSSJ1028+2843的分别为44pc和60pc。它们的潮汐半径都大于它们的半光半径，这说明它们目前都还处于引力束缚状态。

最后我们在图5.13中给出5个候选体在银道坐标系下的位置。图中实心圆是5个候选体，星形点是我们考察的天区内覆盖到的球状星团，加号点为已知的矮星系。三条平行的曲线标识出Anticenter stream的位置。我们发现SDSSJ0814+5105和SDSSJ0821+5608在投影上刚好和Anticenter stream重合。

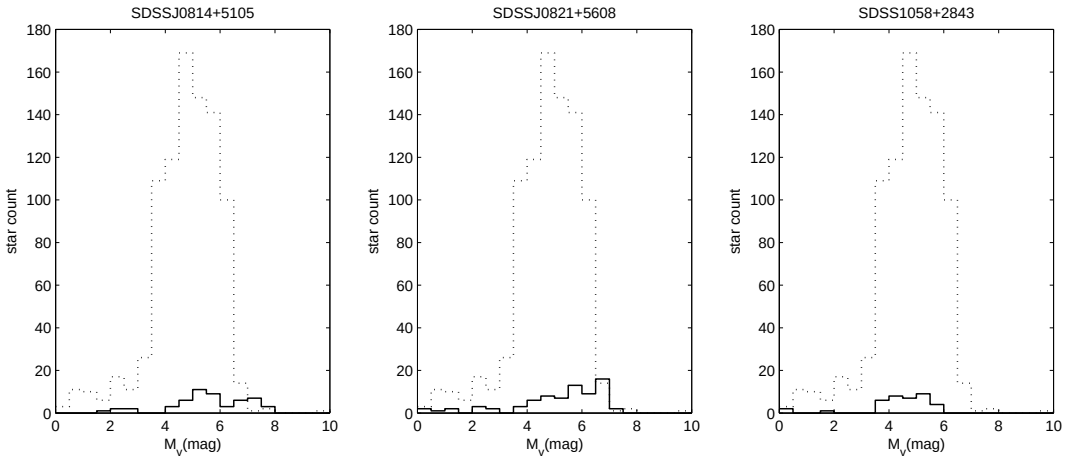

图 5.12: 3个候选体的光度函数

而在附近还有Monoseros stream经过。那么这两个候选体和这两个星流是否有关呢？

在Grillmair(2006)[87]中，Anticenter stream的主序拐点的位置在g = 18.7，我们从图5.2中估算这两个候选体的主序拐点在g = 19.5，这和Anticenter stream的距离并不一致。但是，附近的Monoceros stream的主序拐点为$g = 19.4$[2]，这刚好与两个候选体的距离一致。因此我们猜测，这两个类似于球状星团的候选体虽然投影上和Anticenter stream重合，但是距离并不一致，所以和这个星流的关联不大，反而和Monoceros stream相关的可能性更大，因为它们在位置上相近，距离上也相似。它们很可能是Monoceros stream的球状星团，或者是潮汐力瓦解出的大块碎片。

在上面叙述的这项研究课题中，我们大量使用了VO-DAS于Java和MATLAB程序中。我们还在MATLAB中完成了几项拟合计算：等年龄线拟合，径向轮廓模型的拟合。在研究早期我们还应用MATLAB完成Hough变换的图像识别计算。在后期我们还将数据访问接入到MATLAB之中，实现了数据访问和数据挖掘的无缝融合。我们部分使用了Aladin和MATLAB配合完成对候选体进行证认的工作。

VO-DAS提供的海量数据访问能力正是这项研究所需要的。MATLAB提供的数据分析和数据挖掘能力也帮助本项研究大大提高了效率。很多处理涉及到对数百个数据集做出同样类型的计算，利用数据访问和数据挖掘的集成环

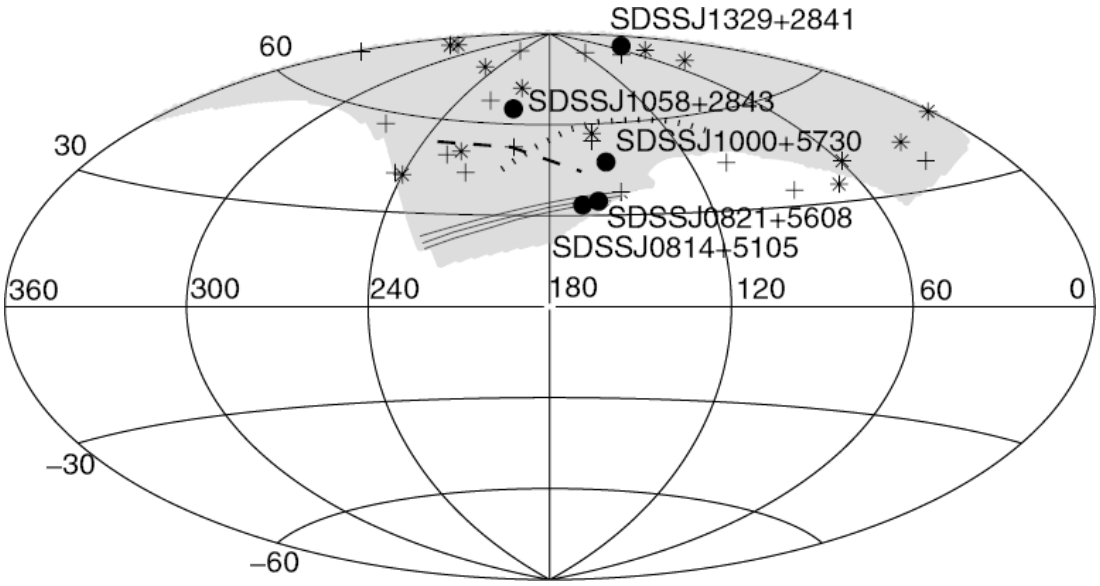

图 5.13: 银道坐标系下的5个候选体和已经知道的天体及子结构

境（VO-DAS+MATLAB），可以自动批量完成数据查询＋数值计算的一系列操作。很多计算，例如曲线拟合算法，是MATLAB已经提供了的，我们只要将需要拟合的公式提供给MATLAB就可以完成拟合，减少了编程的工作量。由于MATLAB具有很强的数据可视化能力，因此计算完成以后的可视化操作也可以很顺利完成。本章中引用的所有图形均使用MATLAB生成。特别值得指出的是图5.1和5.13，一个是Lambert投影图，一个是Aitoff投影图，这两个球面坐标投影都是使用MATLAB Mapping Toolbox完成的，只需要几行代码，就可以用赤经赤纬画出上面的银道坐标图，中间不需要再转换坐标系。

总而言之，本章的工作让我们实现了用现有虚拟天文台数据挖掘工具辅助天文学的研究的目标。虽然用户界面，工具的稳定性和集成度方面还有很多要改进的，但是这至少在功能上虚拟天文台数据挖掘工具已经能够为研究提供足够的服务了。

第六章　总结、讨论与展望

我们利用了网格服务体系中的带状态的网络服务构建了VO-DAS系统，通过建立VO-DAS服务和DataNode之间的动态网格连接，实现了对分布式异构天文数据资源的异步查询数据模式。通过这个模式，海量数据访问得以实现。我们还通过将OGSA-DAI嵌入DataNode中获得了一种对异构数据库的接口封装。由于系统构建在网格之上，借助VO Registry服务，VO-DAS可以发现分布在各处的DataNode，通过一套复杂的session控制和任务调度机制实现了对这些分布数据库的联合查询和交叉证认。这样，一个异地异构数据资源的海量访问服务被建立起来，支持采用IVOA的ADQL查询语言标准访问星表、图像和光谱等各种天文数据资源。它多样灵活的客户端支持不同要求的应用，也便于同数据挖掘平台实现良好的互操作。

我们考察了多种可能作为天文数据挖掘的计算机编程平台和语言，经过对比、原型实验和科学范例研究，我们认为MATLAB具备了成为天文数据挖掘通用工具的潜力，通过在其上的二次开发，我们不仅将VO-DAS的数据访问接口融合到MATLAB之中，实现了数据格式的无缝连接，还实现了MATLAB和其他天文桌面工具的互联。我们还在MATLAB之上针对特定的天文研究开发了专门的算法程序。这些工具在作为科学范例的寻找银河系伴星系的研究中发挥了作用。

我们把寻找银河系伴星系的研究作为驱动技术成熟的催化剂，推动上述虚拟天文台数据挖掘工具的产生和成熟。同时，通过应用这些工具对SDSS恒星数据的研究，我们成功发现了5个新的极暗的伴星系 / 球状星团的候选体，并运用这些工具探讨了这些候选体的性质，对它们的可能类型做了推断。我们认为其中有三个从几何尺寸和光度两个方面看，都很像是非常暗的球状星团，而另外两个则很像是矮星系。

我们发现，完成天文数据挖掘工具箱这样的工作和天文学研究一同相伴进行，是一个双赢的结果。一方面面对天文学课题，技术的发展受到应用的迫切要求，因而发展出来的技术是最实用的。我们在设计VO-DAS的过程中经常面临这样的决择：两个都很有吸引力的功能，应该先开发哪一个？通常，我们的决定是根据实用的需要。哪一个应用更迫切就应该先做哪一个。例如，VO-DAS的客户

端的开发，我们首先开发出来的是一个GUI客户端，这样的客户端看上去感觉更好，在这个客户端上可以做更多的工作，让它的功能更加华丽，例如实现格式化显示数据，提供FTP下载功能，实现ADQL语法的检查，用彩色标识ADQL的不同部分，从资源元数据中通过双击实现对一个资源查询的ADQL语言模板的生成等。这些工作可以让GUI非常强大，用户应用起来更加方便，从界面的可用性角度可以和其他有实力的虚拟天文台的产品相媲美。但是，另一方面，我们还面临一大批用户，他们使用FORTRAN，他们使用IDL，他们使用Python，他们也许还使用Java但是对WSRF不甚了了。面对这些用户，一个漂亮的GUI客户端并不能解决他们的现实问题：对这些类型的用户而言他们的需要是将数据查询和他们的程序无缝结合起来。因此，与其花费很多时间和精力完善GUI界面，还不如开发一个命令行客户端满足这批用户的需求。虽然命令行客户端的用户易用性较差，但是更加实用，应用范围更广，特别适用于将VO-DAS集成到自己的工具箱中的客户。于是我们决定增加一个命令行客户端。此外，在数据挖掘工具应用于银河系晕结构的研究的过程中，我们发现有很多专门应用于天文学的算法需要完善和补充。例如，球面投影的计算是MATLAB的Mapping Toolbox已经提供了的，但是，在把恒星坐标投影到与球面相切的平面上以后（即球心投影），我们需要在这个坐标系下做一系列的统计计算，这些计算都是现成的工具所不具备的。这就需要我们完成这些算法的开发以适应研究的需要。这些计算工具用于数据挖掘中数据预处理的过程，不仅对于银河系矮星系搜寻，对于其他的研究也会有很强的实用价值。经过不断应用数据挖掘工具到科学研究课题中，并且不断从中总结那些实用的算法，最终我们就会积累出一个非常可观的适应天文学研究特点和需求的天文数据挖掘工具箱。

另一方面，在实际的科学研究中，如果没有一套适宜的天文数据挖掘工具的帮助，很多工作就会花费非常长的时间。我们在开始银河系晕结构研究之前，曾经试图用当时的技术条件完成另一个科学范例：寻找银河系中的OB星，并通过它们确定银河系的消光在三维空间的分布。为了寻找隐藏在浓密气体和尘埃之后的OB星，我们需要使用2MASS星表进行筛选。然而由于消光的影响，以及银盘上随着银经变化恒星分布的不均匀性，我们很难确立一个简单的OB星判断标准。在这个研究过程中，我们需要反复实验不同条件的效果。这就需要反复对2MASS星表进行海量数据查询，查询数据高达2×10^8之多。由于当时没有VO-DAS，只有作为原型的SkyPortal作为数据查询工具，我们的数据访问工作非常费时，需要将数据分成很多部分，通常按照银经分成360份，分

别查询以后再进行合并。而在之后我们对银河系晕结构所作的研究中，由于有了VO-DAS的帮助，可以很容易从数据库中取得10^7级的数据，这也使得我们的工作没有在数据获取这第一个阶段就搁浅，很快度过了数据选择和查询这个阶段，进入了实质性的数据分析阶段。在估计候选体的几何特征，特别是径向轮廓的时候，正是因为有了MATLAB这样的工具提供了丰富的数值拟合算法，所以我们可以将精力集中在选择矮星系面密度模型上，而不是如何实现牛顿下山法或者Levenberg-Marquardt算法等繁复的非线性加速优化算法上。总之，辅以实用的基于虚拟天文台的数据挖掘工具，天文学的研究可以得到加速，天文学家可以将更多精力集中在天文学本身的问题上，而数据将在指尖上呼之即来，算法将会成为放在桌面上的工具随时可以派上用场。

在本文所描述的所有工作，还仅仅是探索性的，为了能够真正投入实用，让没有技术背景的天文学家也可以很容易使用，并且很愿意使用这种集成化的工具软件集，我们在未来还有很多工作需要完善。首先，由于实验探索性质，无论是VO-DAS还是基于MATLAB的数据挖掘工具箱还不是一个可以发布的产品。VO-DAS系统的功能还要有进一步完善，系统的稳定性也是一个重要的课题。一个不稳定的系统会严重打击用户的使用信心。所以，在系统没有完全进入稳定状态以前，发布给用户很可能适得其反，延缓虚拟天文台工具的应用推广进程。MATLAB上的天文工具还处于一种零散的状态，需要对它们进行整合，优化程序和使用接口，编写完善的使用文档，同时进行更多测试。下面分别讨论这两个部分未来的发展思路。

VO-DAS面临的主要问题是功能还不完善，性能有待提高。功能方面需要完善的主要有：ADQL解析器需要更加复杂的解析能力，支持ADQL的全部语法，图像查询，光谱查询和更加丰富的客户端。性能方面，需要对ADQL解析器进行彻底测试完善，对不同拓扑结构的应用环境进行测试，提高系统稳定性。这些都将是后续工作中需要完成的。VO-DAS未来的发展基本是确定性的软件开发工作。

MATLAB数据挖掘工具需要完善的主要有：已经开发的各种计算程序需要汇总，统一它们的接口描述，需要增加GUI的增加GUI界面；添加对已有部分MATLAB的算法程序的封装，使得接口更加符合天文学研究的需要；完善和测试VO-DAS接口；完善和测试PLASTIC协议接口；在新的研究范例中继续总结和开发新的工具。

在未来的数据挖掘工具研究中，我们将仍然运用在本文中成功运用的工作方式：**科学驱动技术**。在银河系晕结构的研究中有很多非常需要数据挖掘工具帮助的课题有待于研究。对这个领域的研究将继续深入下去，通过将基于虚拟天文台的数据挖掘工具应用于其中，可能可以更高效的完成这些研究课题。

在上述数据挖掘工具做得更加完善以后，我们可以再对SDSS DR6的天区进行搜索寻找更多的矮星系候选体，因为根据Simon & Geha(2007)[102]依据CDM模型所作的预测，在SDSS的观测天区内还应该有四倍的矮星系尚未被发现。

晕的子结构的演化研究是一个吸引人的领域。通过虚拟天文台工具的帮助，我们将有可能对SDSS DR6已经观测到的晕中所有的恒星的金属丰度分布作出统计，将金属丰度分布函数（MDF）和子结构的空间位置联系起来，就有可能透过MDF的特性发现子结构的原身矮星系的化学演化历史。进一步应用场晕星在不同局部空间的MDF的不同特征，研究是否可以用矮星系并合解释比较均匀分布在晕中的场星的来源。这项研究的结果将可能给各种银河系晕的起源假说带来更加细致的观测证据。研究过程中将会带来巨大的计算工作量，需应用并行计算方法来实现。这将促使天文数据挖掘工具针对高性能计算作出改进和扩展。

晕中的子结构，特别是那些跨越了广大的天区的星流结构，还是银河系引力势的最好的示踪物。如果将它们看成是一个质量一定的天体在银河系引力场中的运动轨迹，那么综合这些星流的不同形态，有可能有效约束银河系引力势的形状。这项研究也将会有大量微分方程数值计算和相空间曲线可视化。这方面的计算目前在数据挖掘工具中虽然有基本算法，但是大量应用于天文学尚未尝试。因此对这个课题的研究有助于完善数据挖掘工具对各种银河系引力模型的支持，所积累下来的算法在以后的研究中还会大有用途。

附录 A　ADQL解析器的详细设计

A.1　词法扫描的状态迁移图

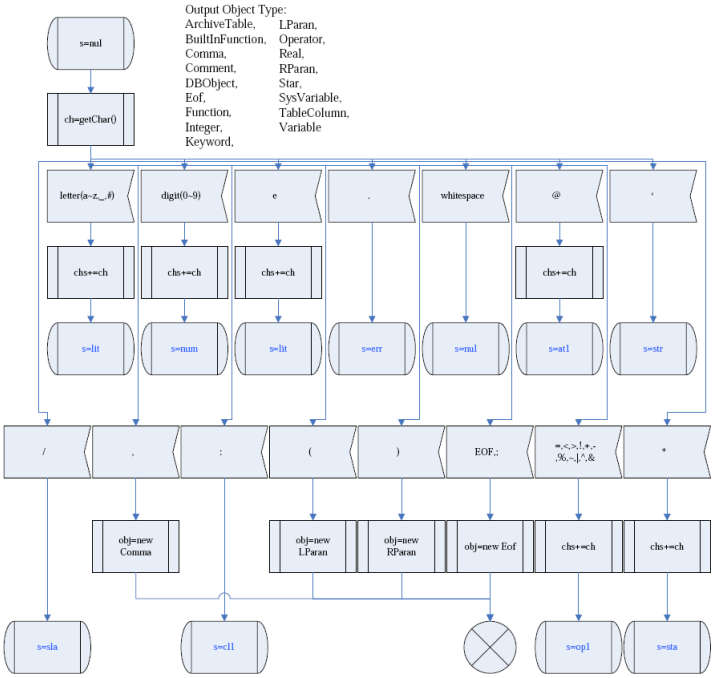

图 A.1: 词法扫描状态迁移图之一

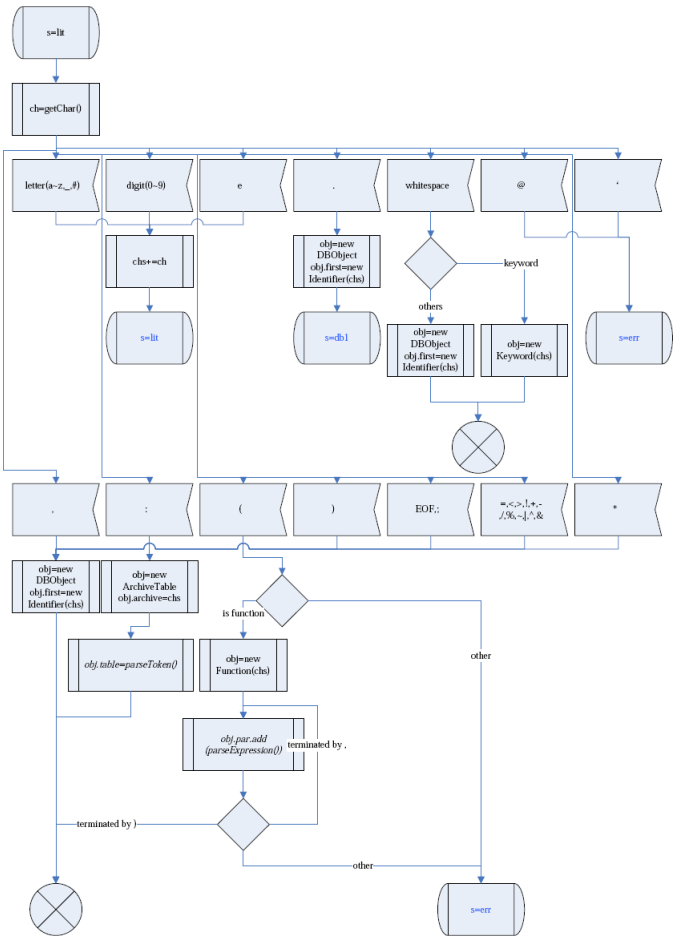

图 A.2: 词法扫描状态迁移图之二

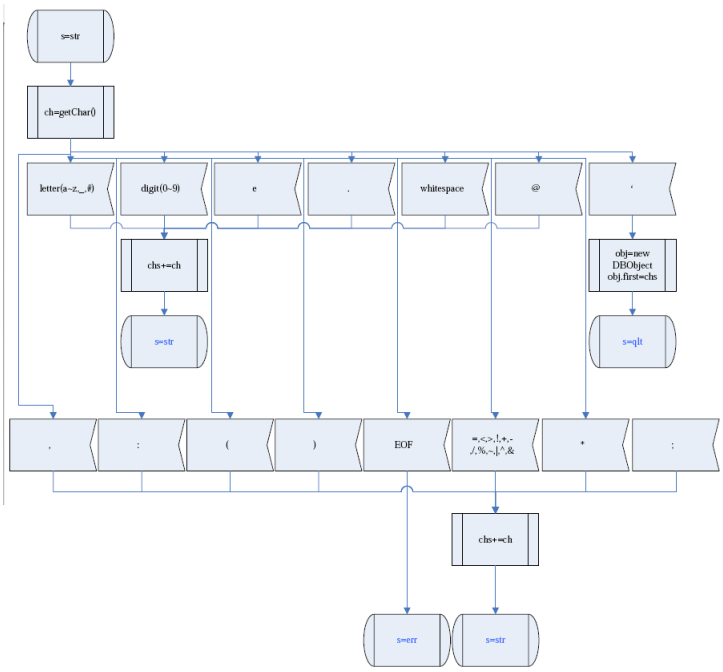

图 A.3: 词法扫描状态迁移图之三

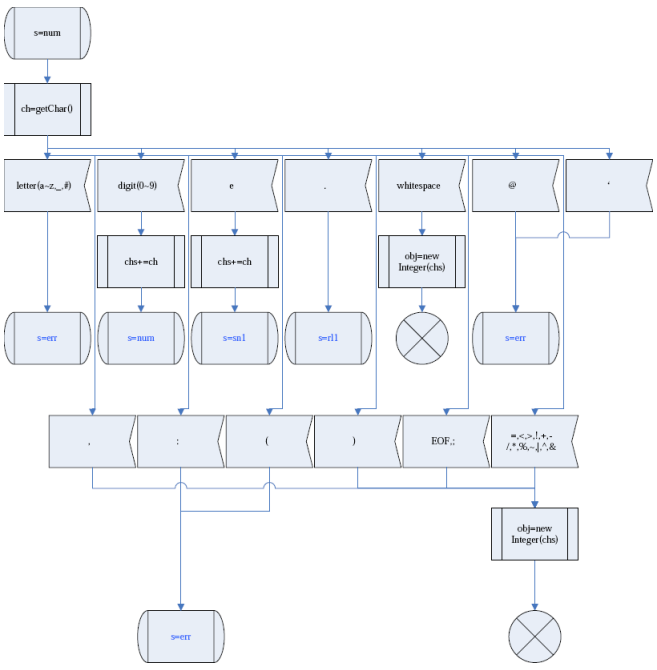

图 A.4: 词法扫描状态迁移图之四

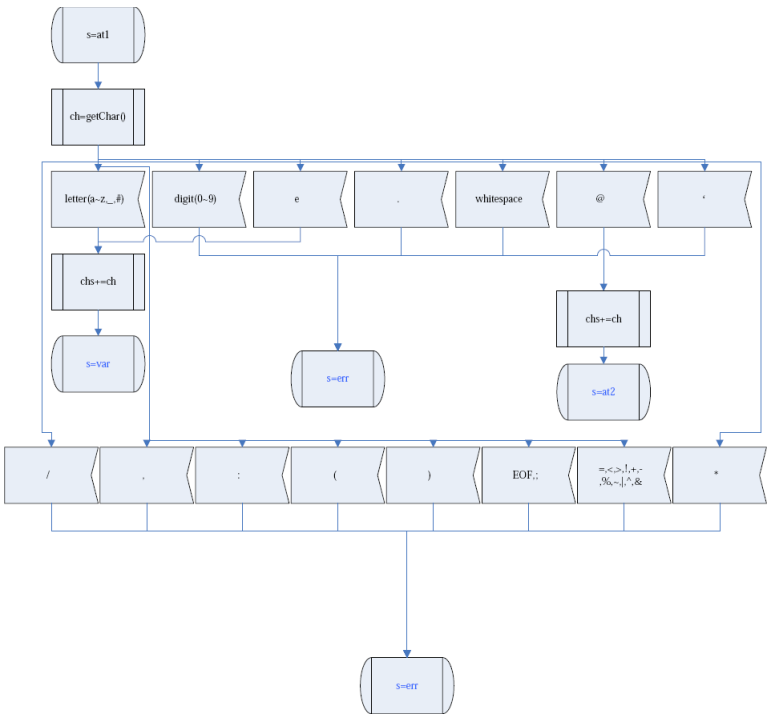

图 A.5: 词法扫描状态迁移图之五

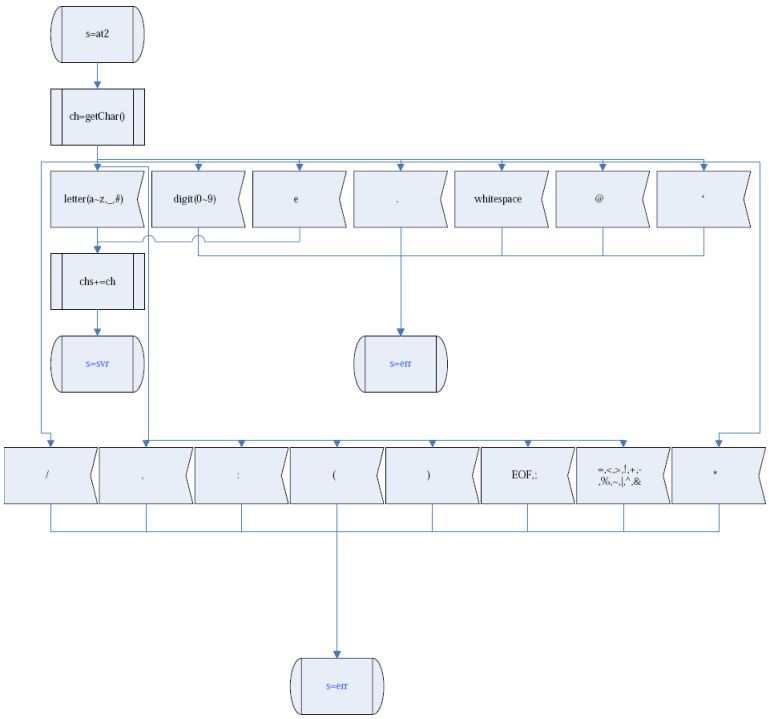

图 A.6: 词法扫描状态迁移图之六

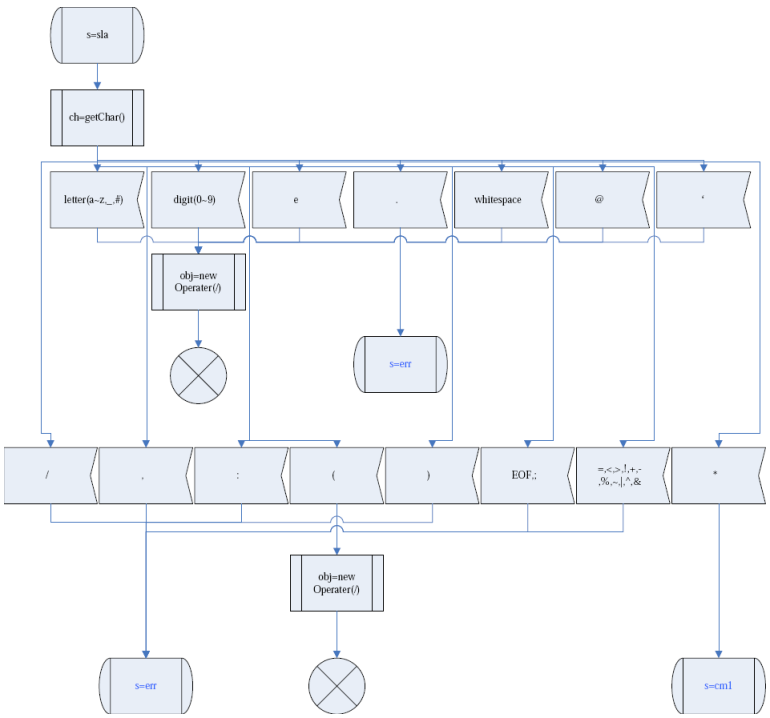

图 A.7: 词法扫描状态迁移图之七

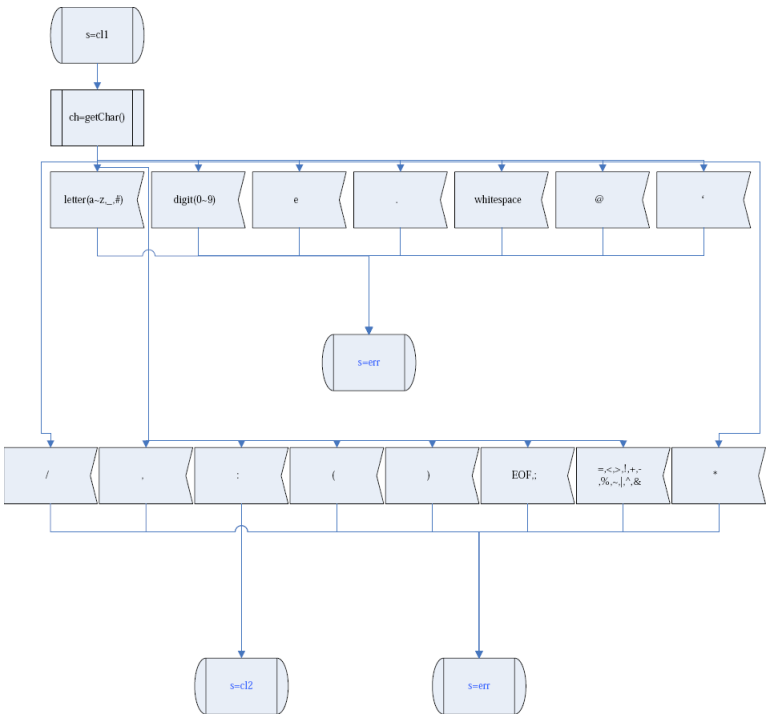

图 A.8: 词法扫描状态迁移图之八

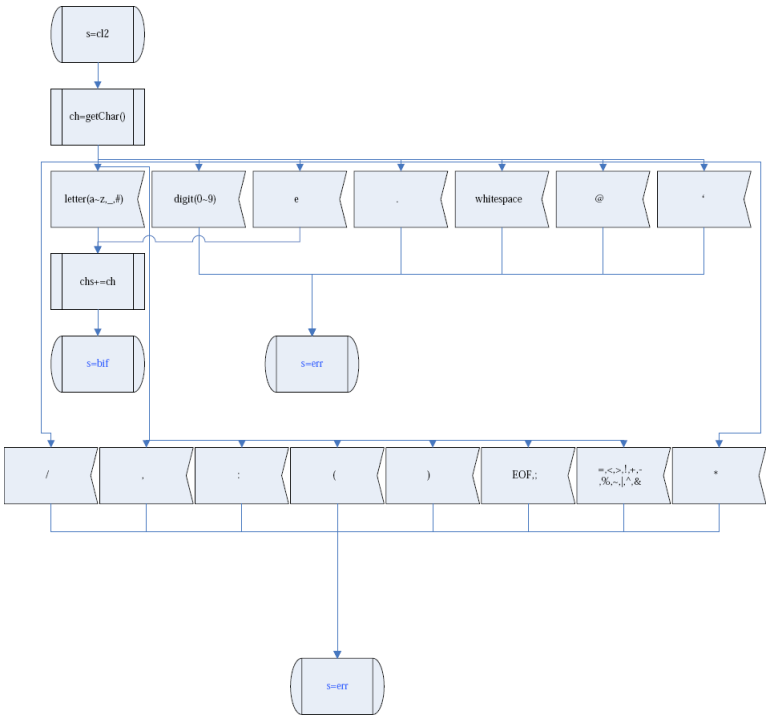

图 A.9: 词法扫描状态迁移图之九

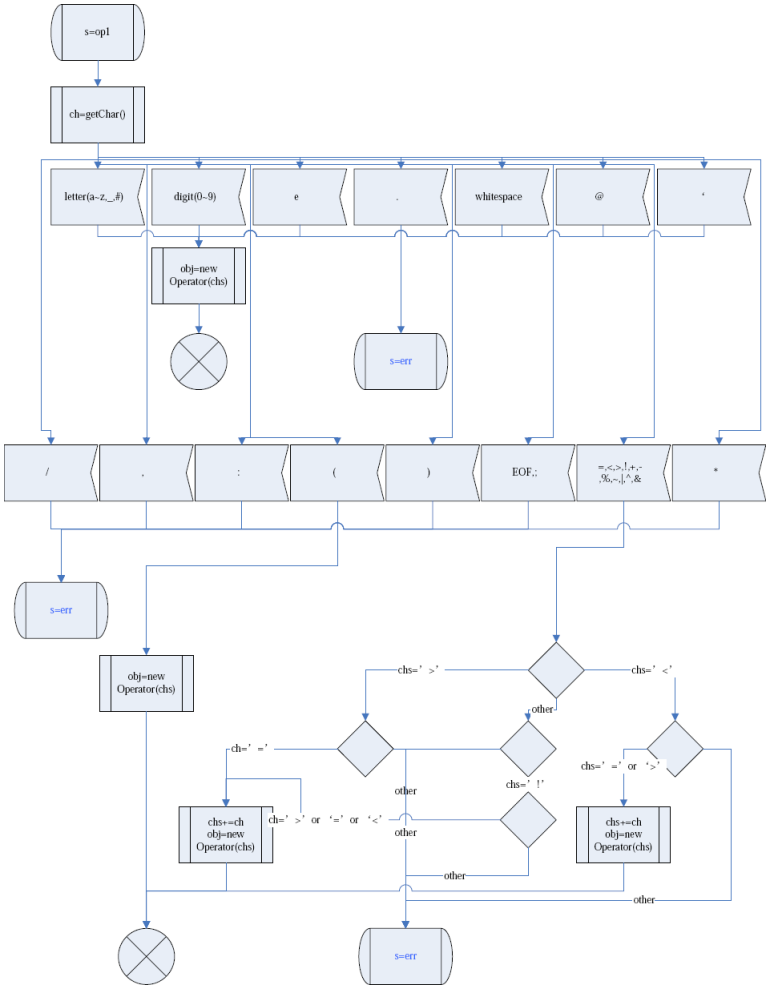

图 A.10: 词法扫描状态迁移图之十

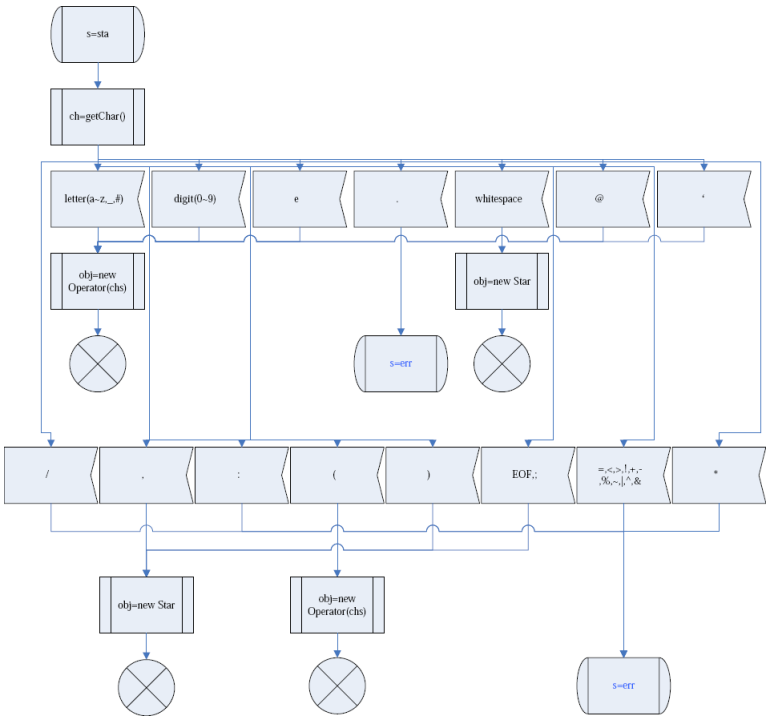

图 A.11: 词法扫描状态迁移图之十一

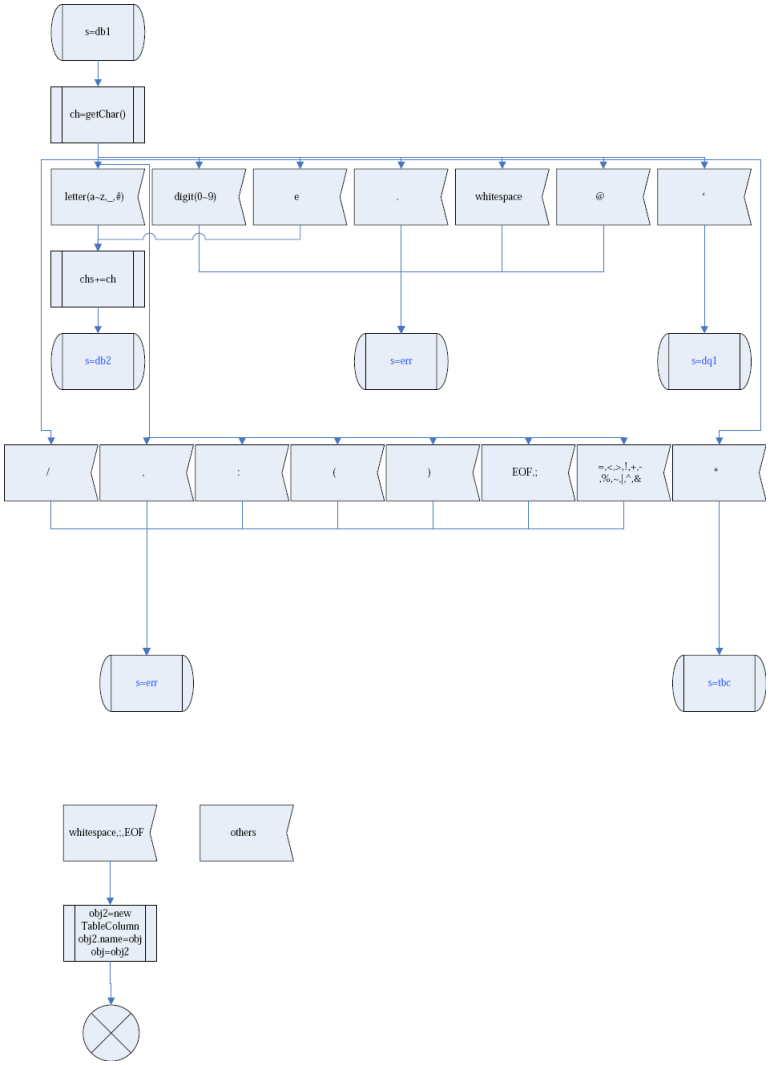

图 A.12: 词法扫描状态迁移图之十二

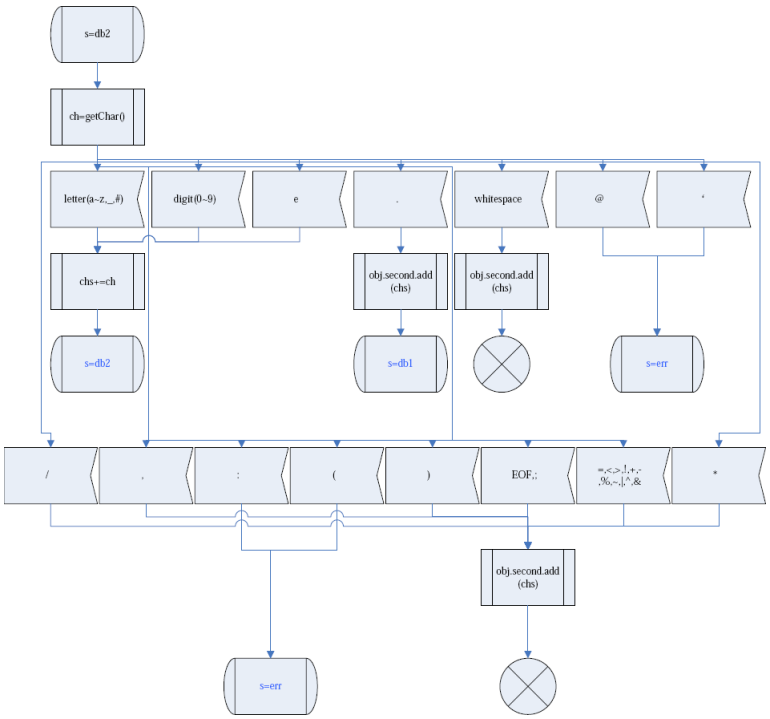

图 A.13: 词法扫描状态迁移图之十三

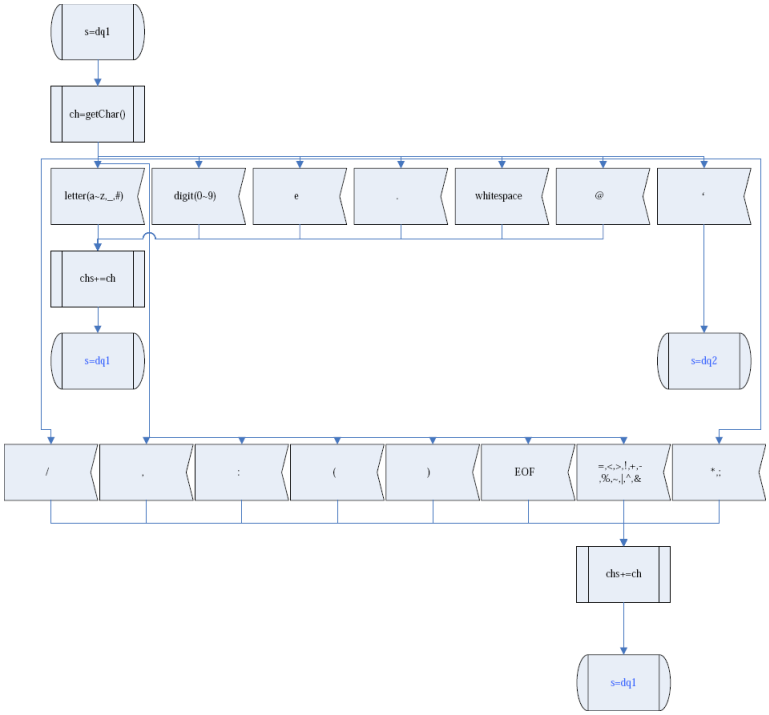

图 A.14: 词法扫描状态迁移图之十四

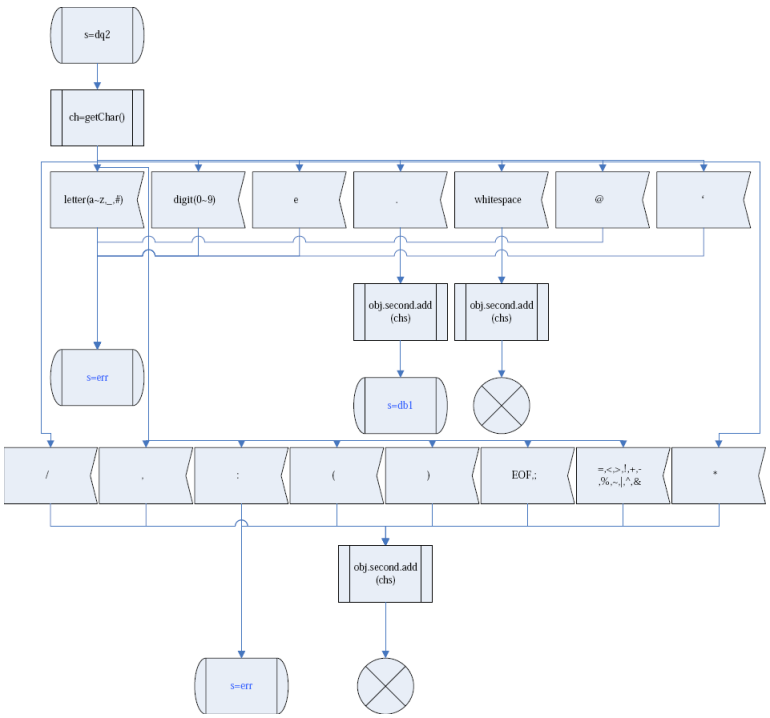

图 A.15: 词法扫描状态迁移图之十五

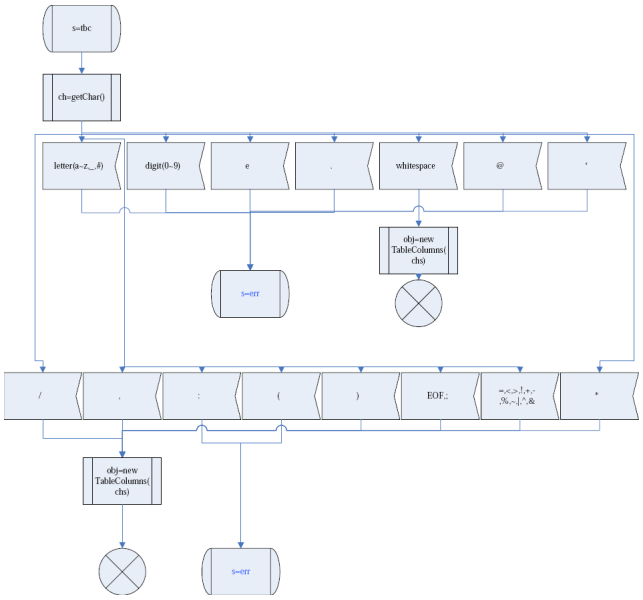

图 A.16: 词法扫描状态迁移图之十六

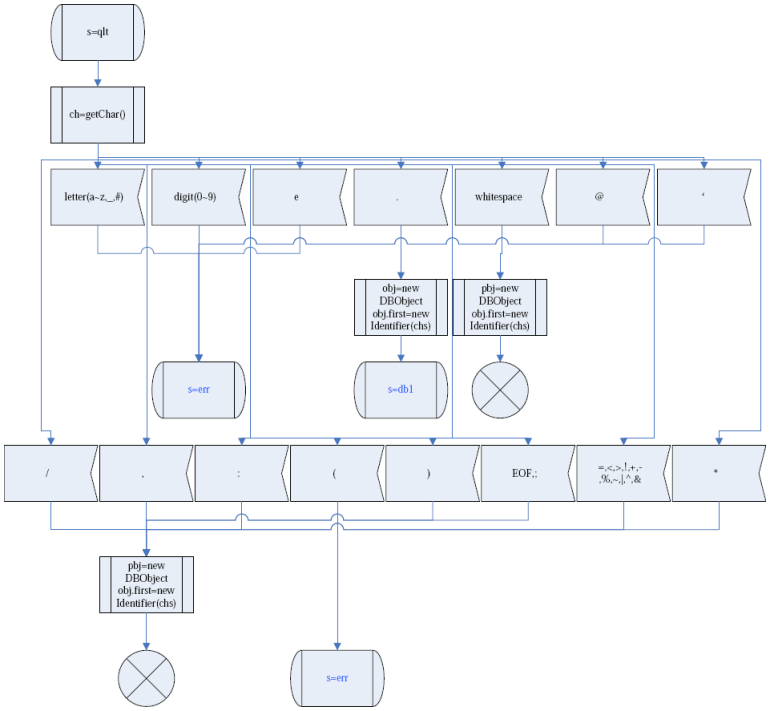

图 A.17: 词法扫描状态迁移图之十七

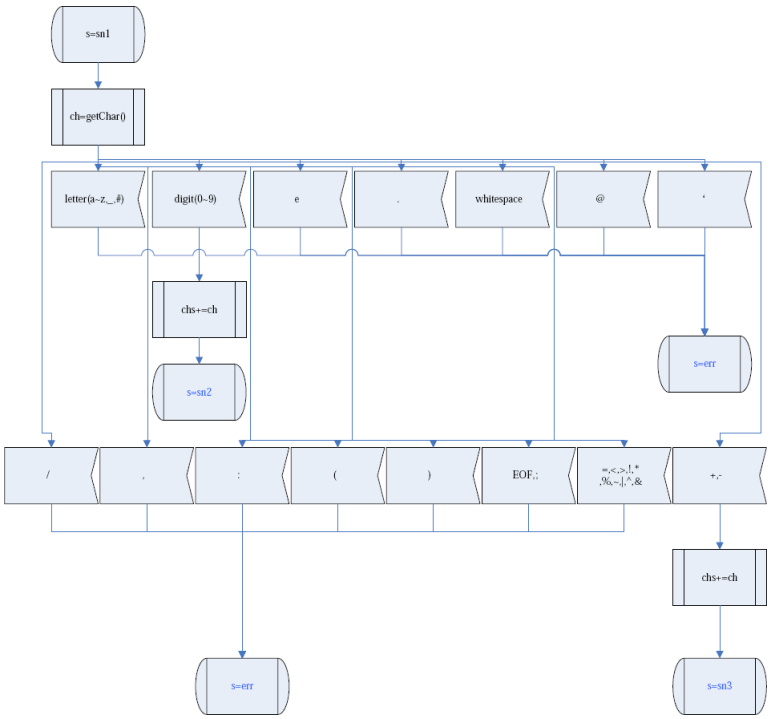

图 A.18: 词法扫描状态迁移图之十八

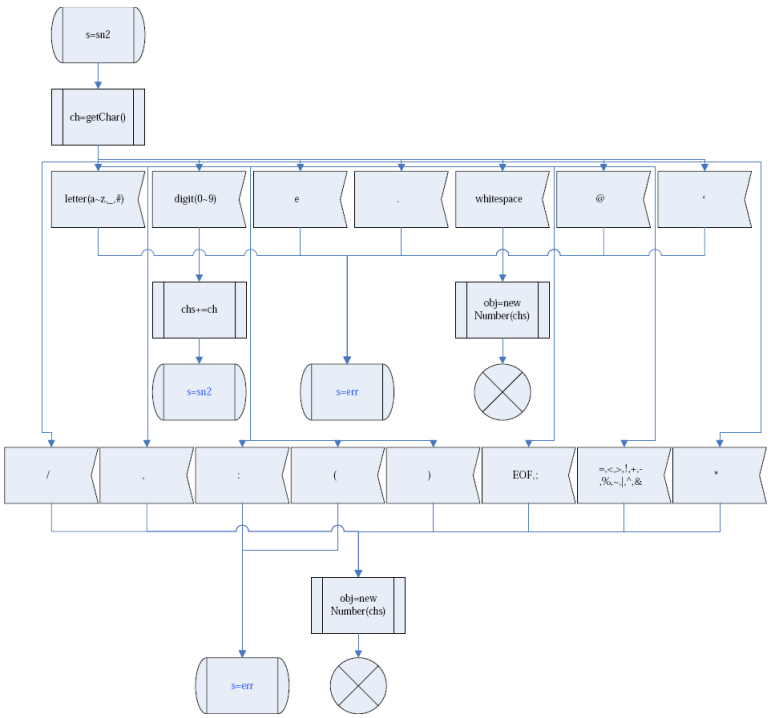

图 A.19: 词法扫描状态迁移图之十九

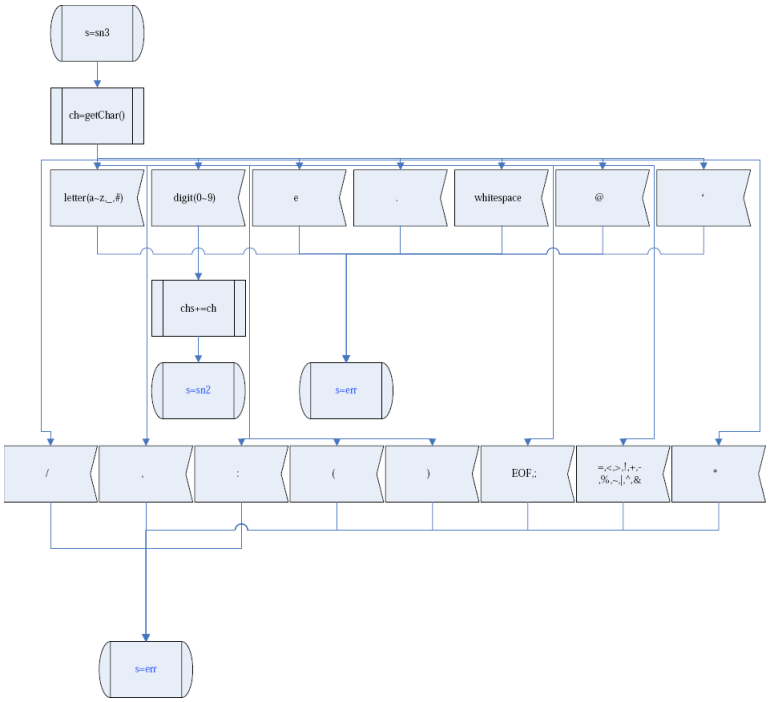

图 A.20: 词法扫描状态迁移图之二十

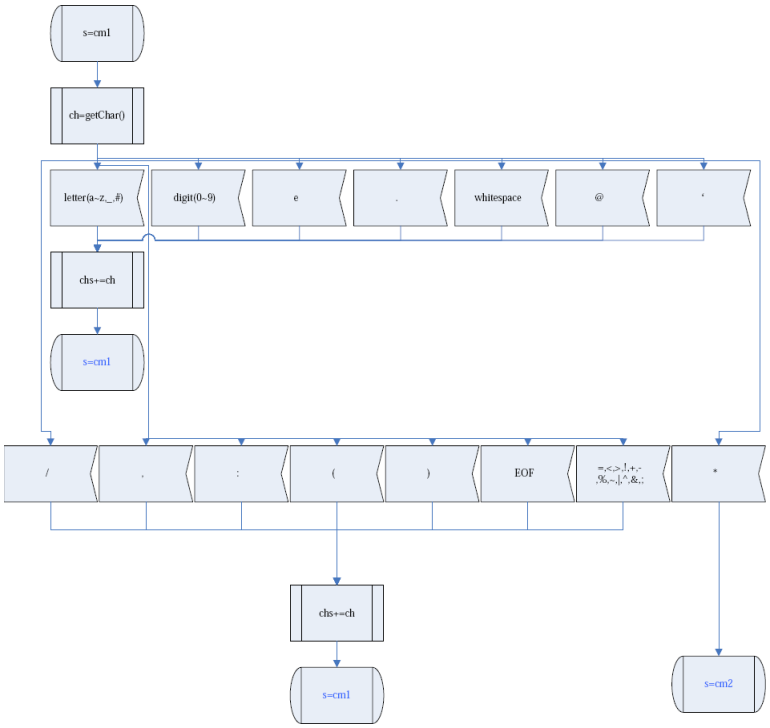

图 A.21: 词法扫描状态迁移图之二十一

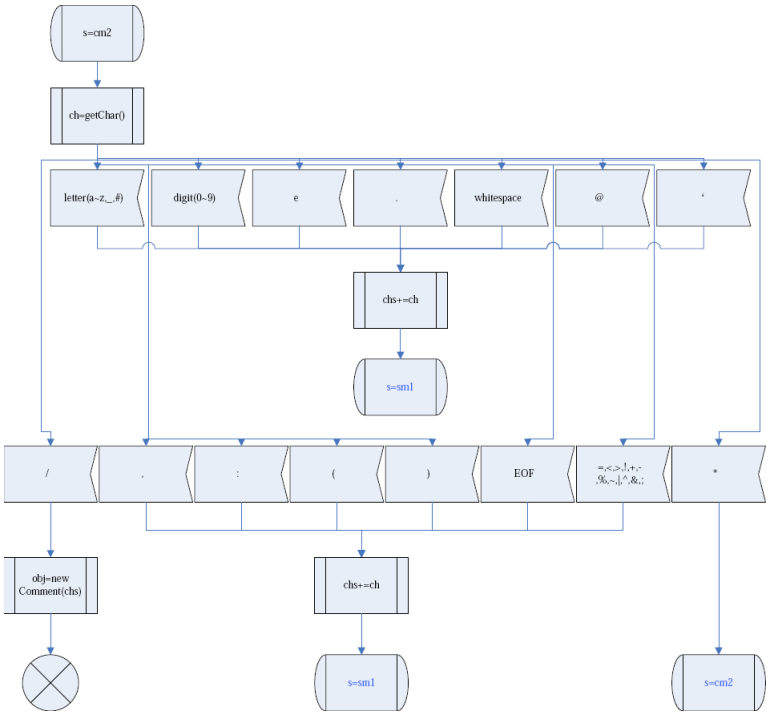

图 A.22: 词法扫描状态迁移图之二十二

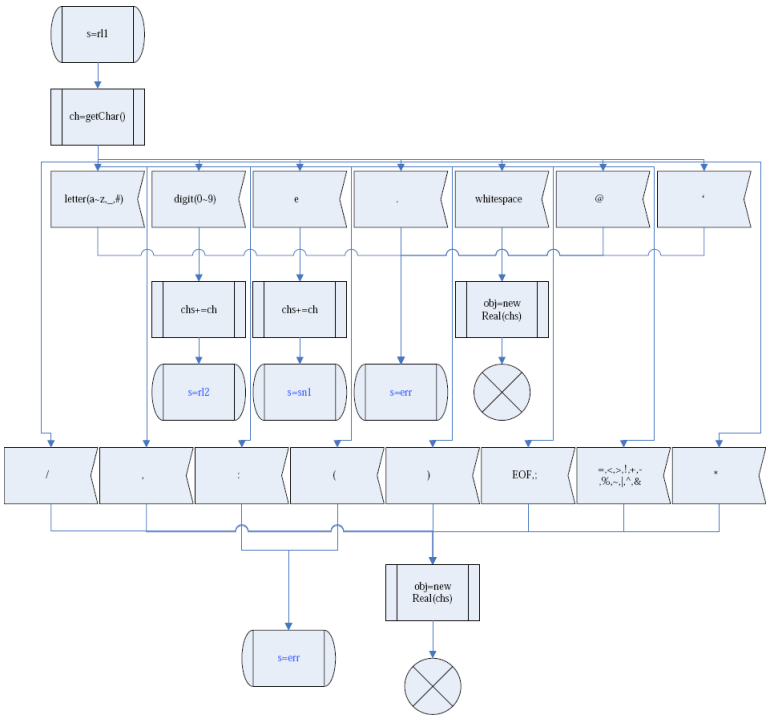

图 A.23: 词法扫描状态迁移图之二十三

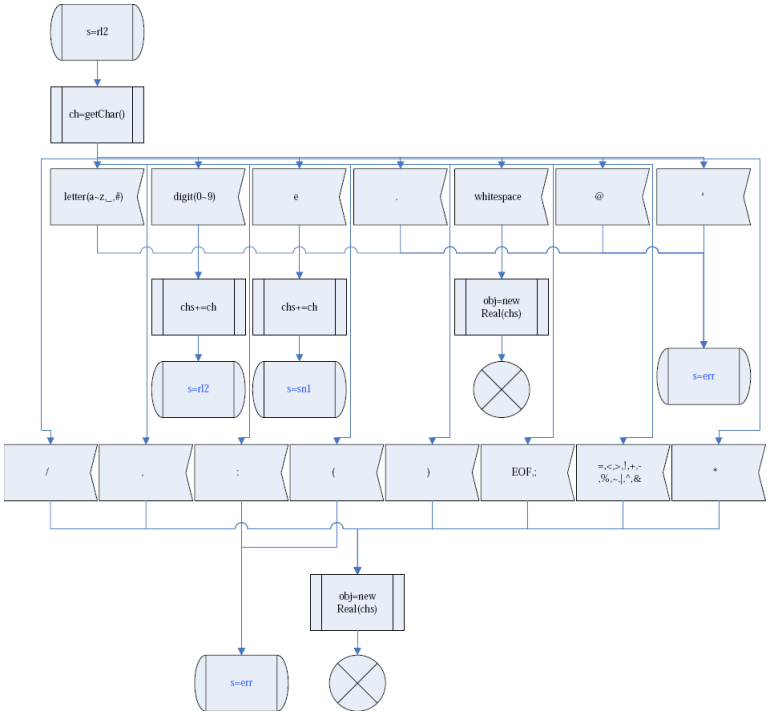

图 A.24: 词法扫描状态迁移图之二十四

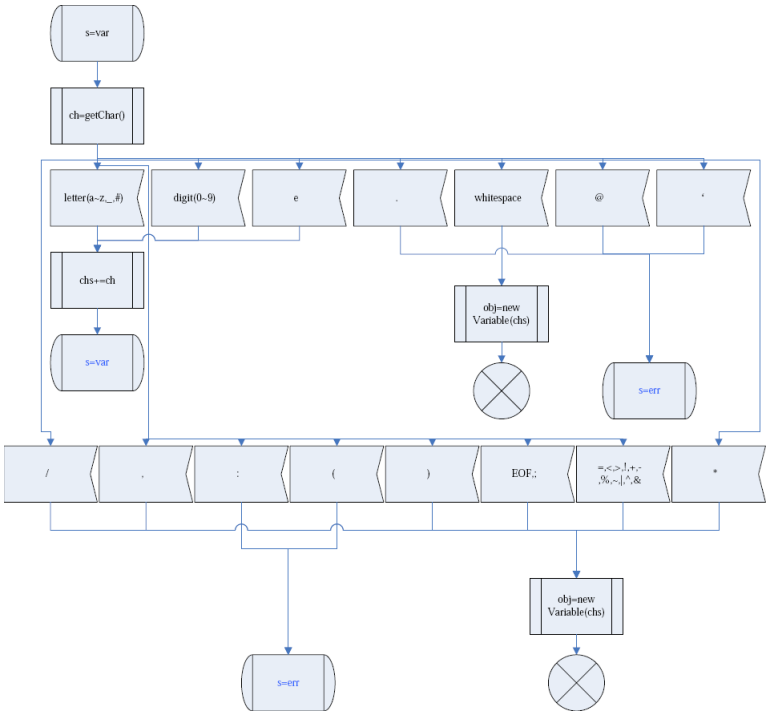

图 A.25: 词法扫描状态迁移图之二十五

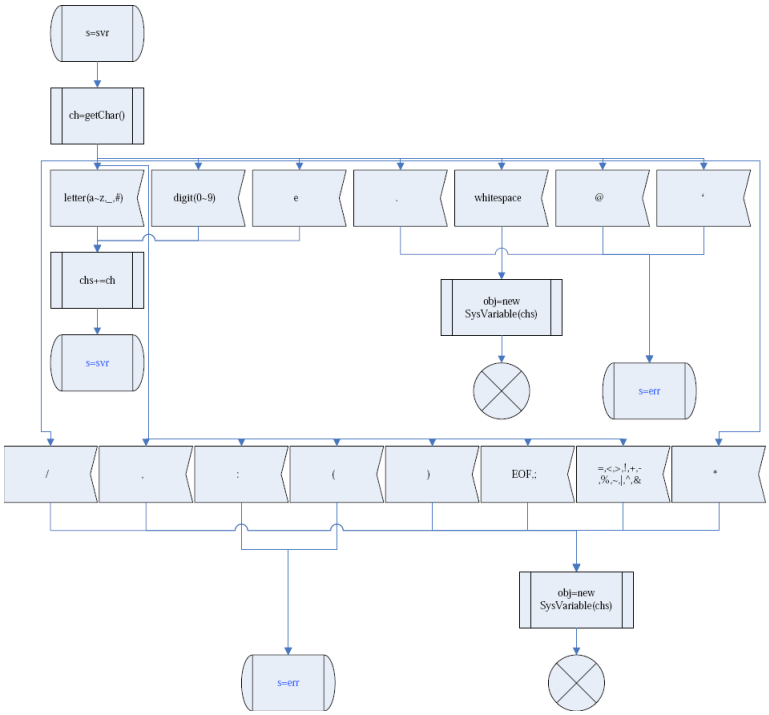

图 A.26: 词法扫描状态迁移图之二十六

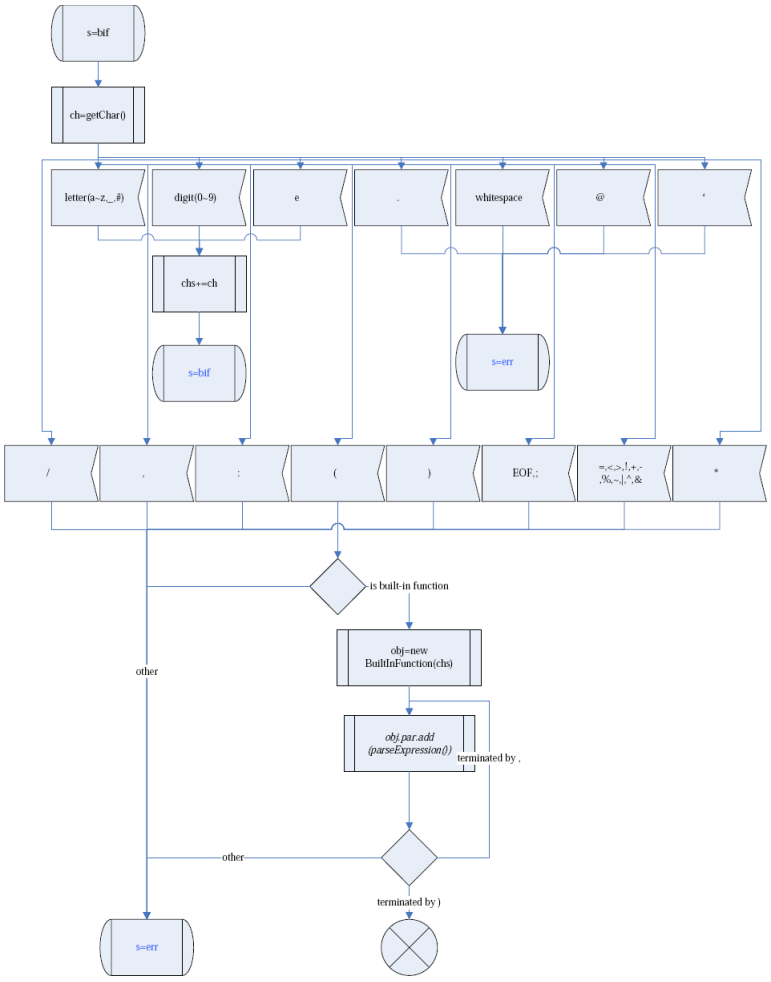

图 A.27: 词法扫描状态迁移图之二十七

A.2　ADQL各个分句解析的流程图

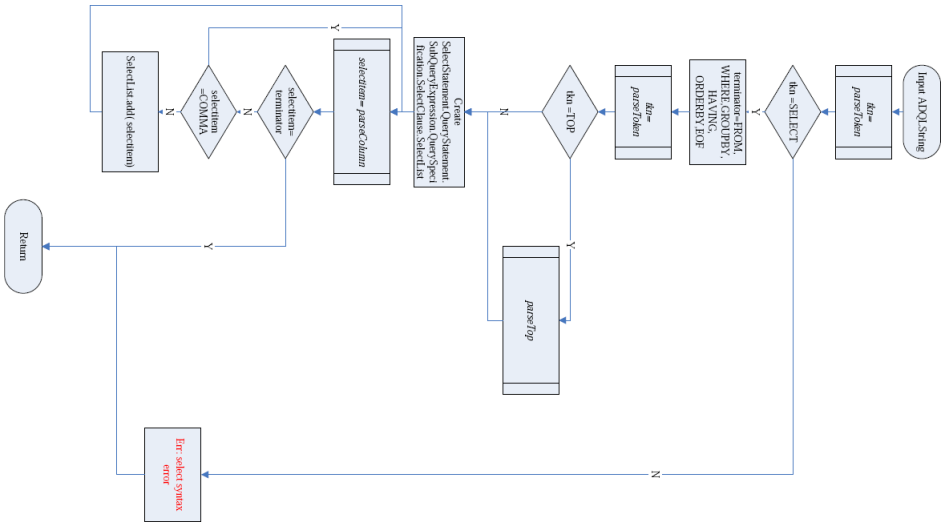

图 A.28: ADQL分句解析流程图之一

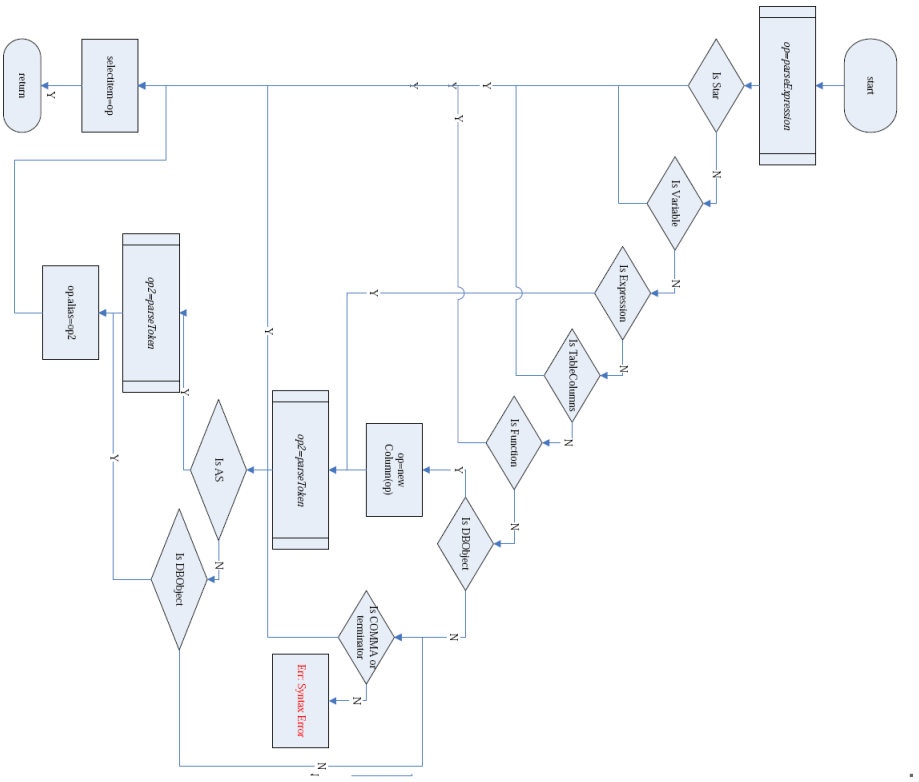

图 A.29: ADQL分句解析流程图之二

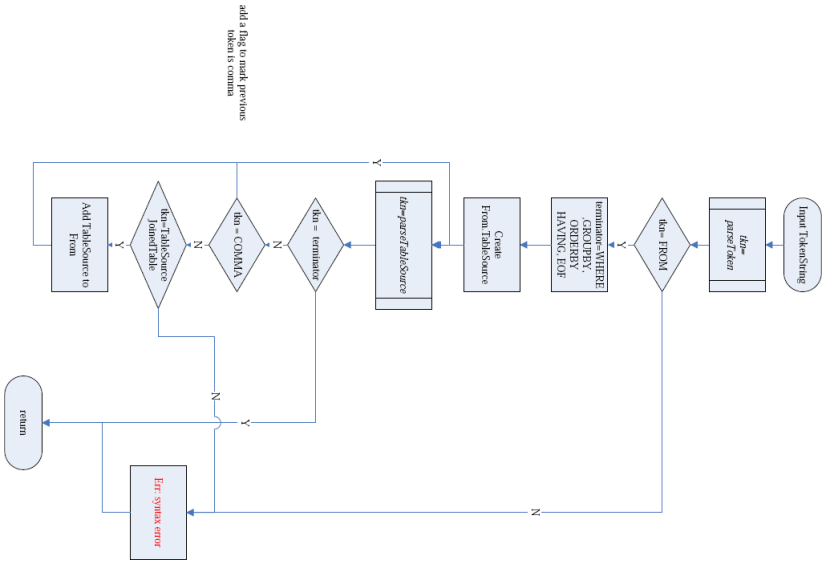

图 A.30: ADQL分句解析流程图之三

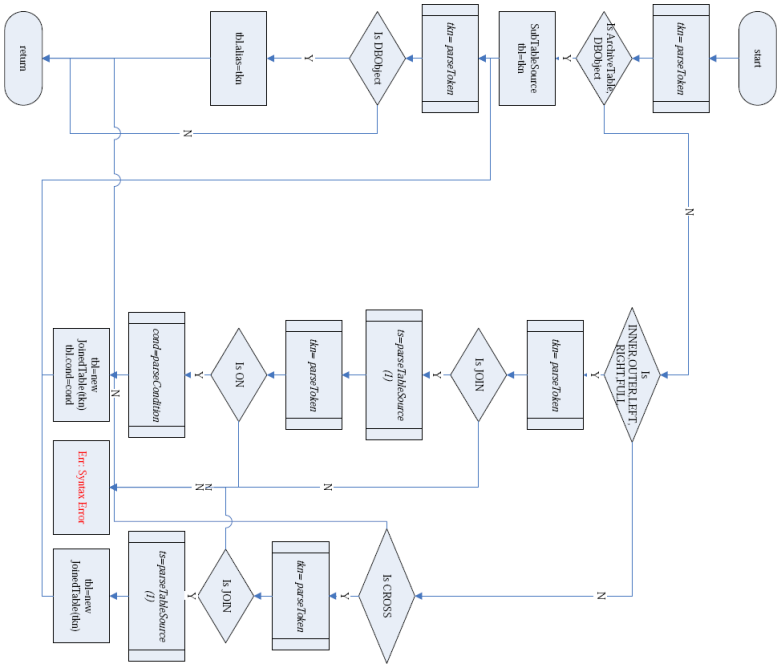

图 A.31: ADQL分句解析流程图之四

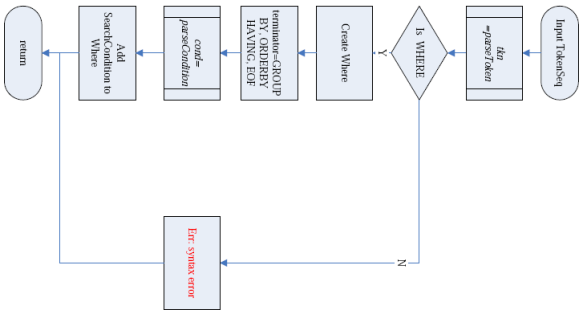

图 A.32: ADQL分句解析流程图之五

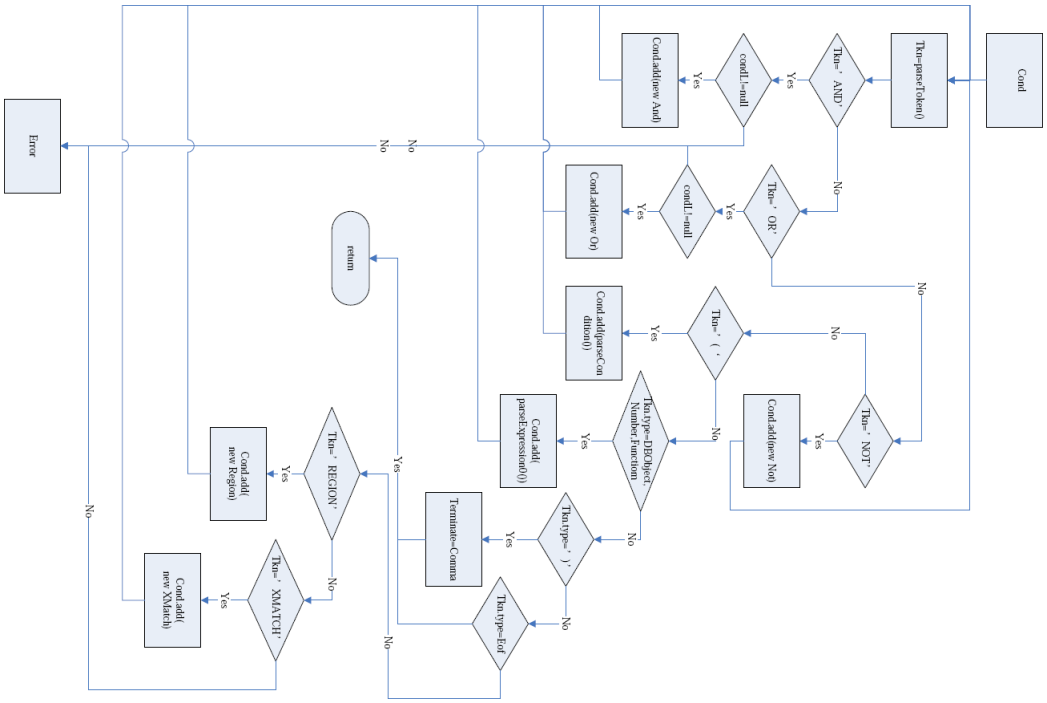

图 A.33: ADQL分句解析流程图之六

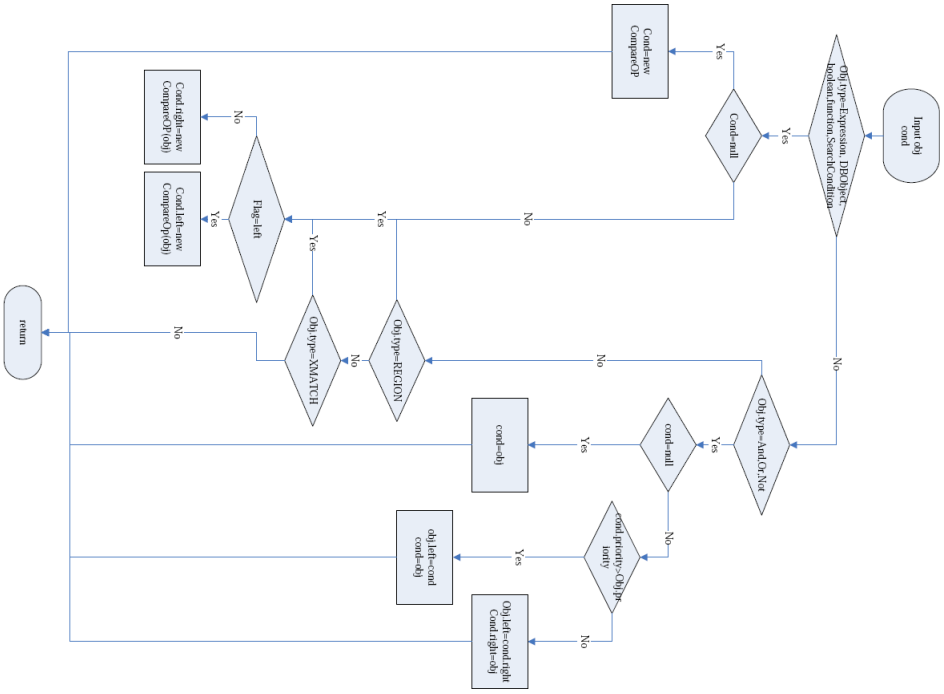

图 A.34: ADQL分句解析流程图之七

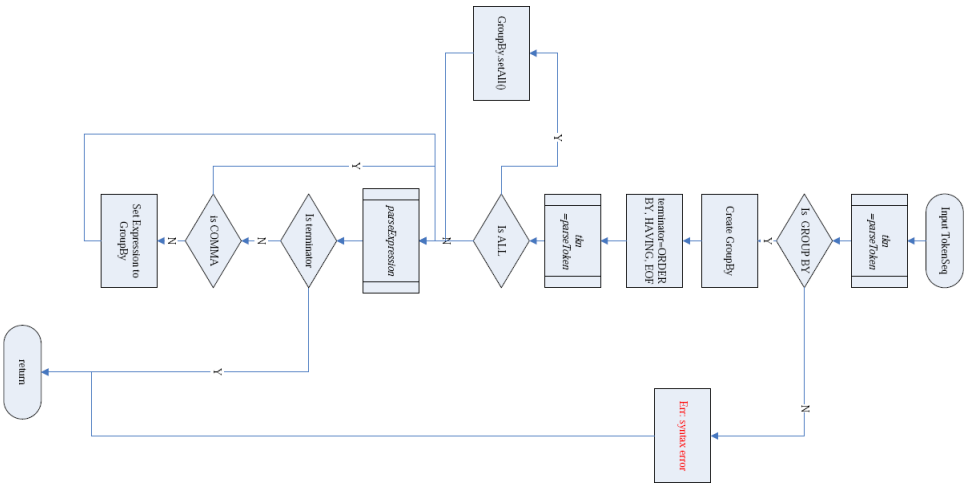

图 A.35: ADQL分句解析流程图之八

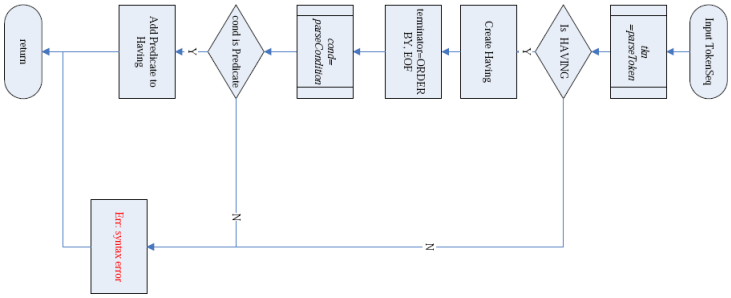

图 A.36: ADQL分句解析流程图之九

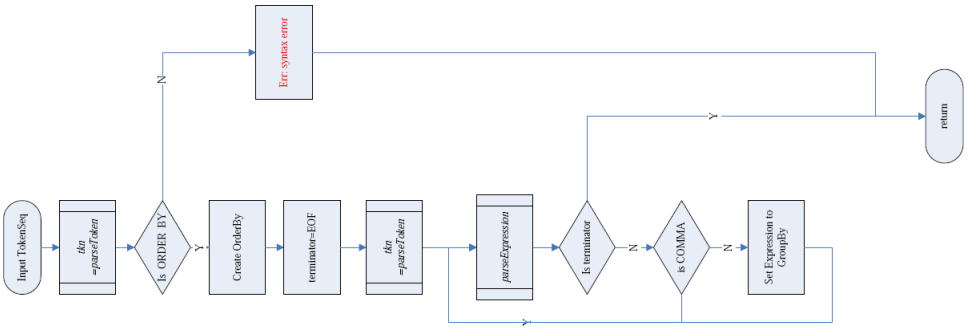

图 A.37: ADQL分句解析流程图之十

A.3　表达式分析的状态迁移图

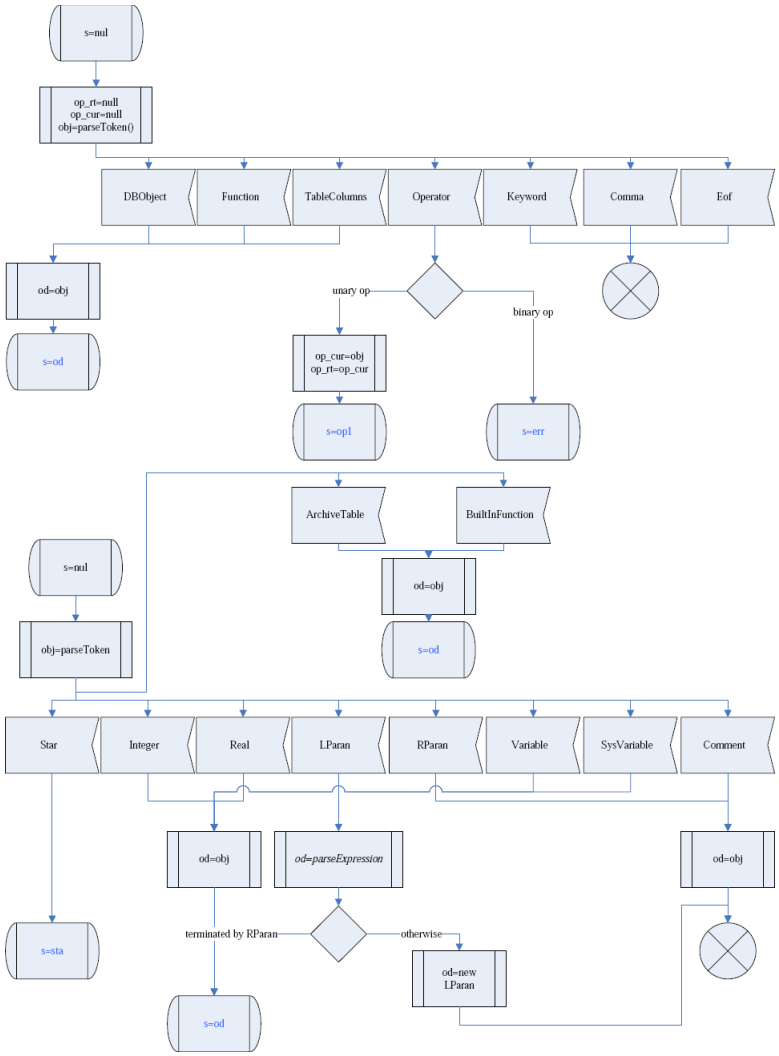

图 A.38: ADQL表达式状态迁移图之一

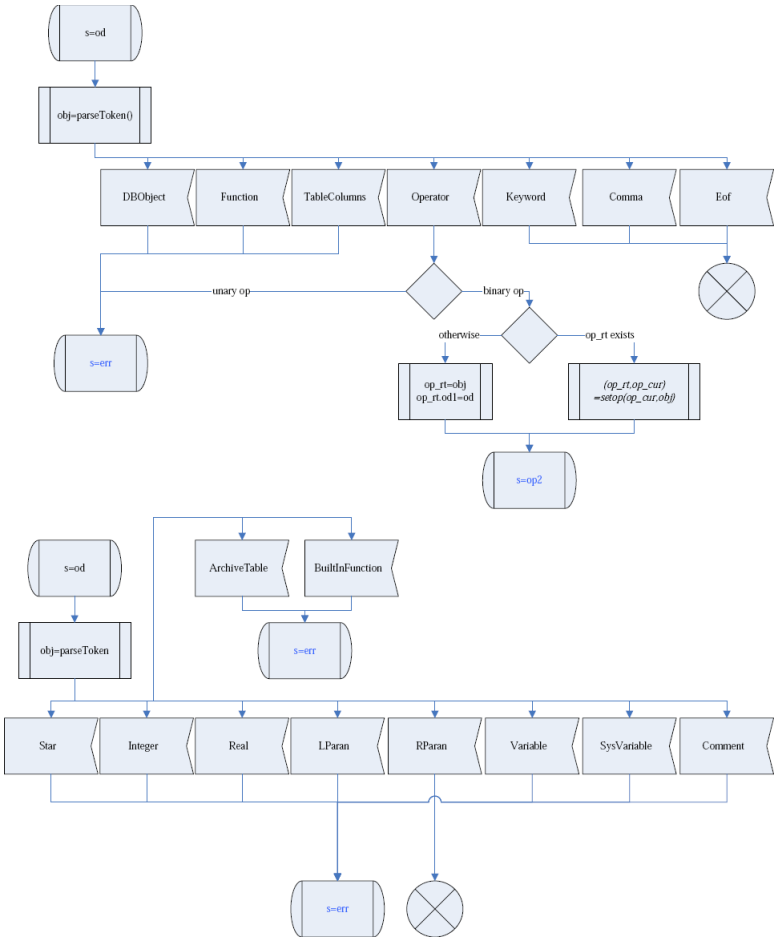

图 A.39: ADQL表达式状态迁移图之二

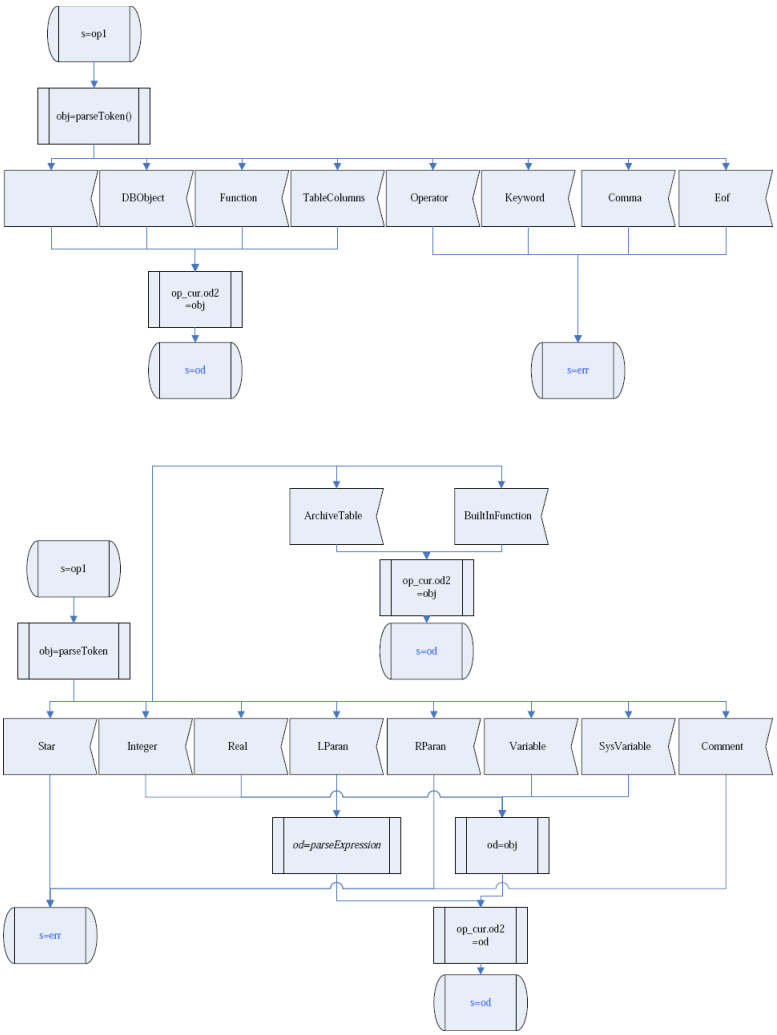

图 A.40: ADQL表达式状态迁移图之三

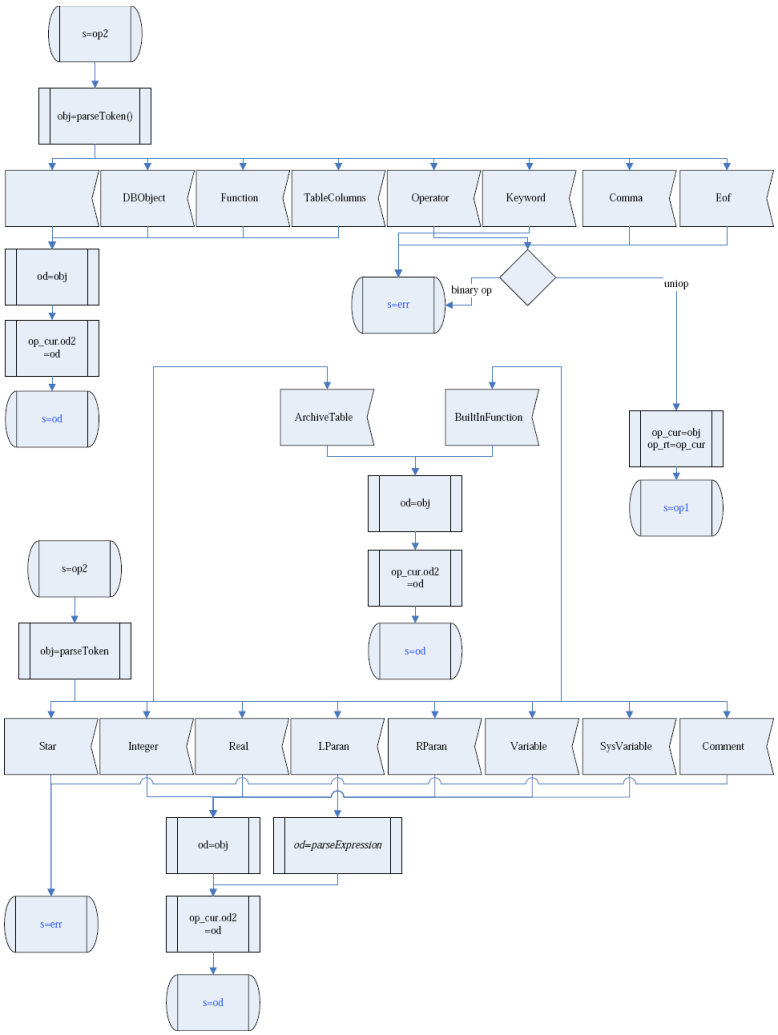

图 A.41: ADQL表达式状态迁移图之四

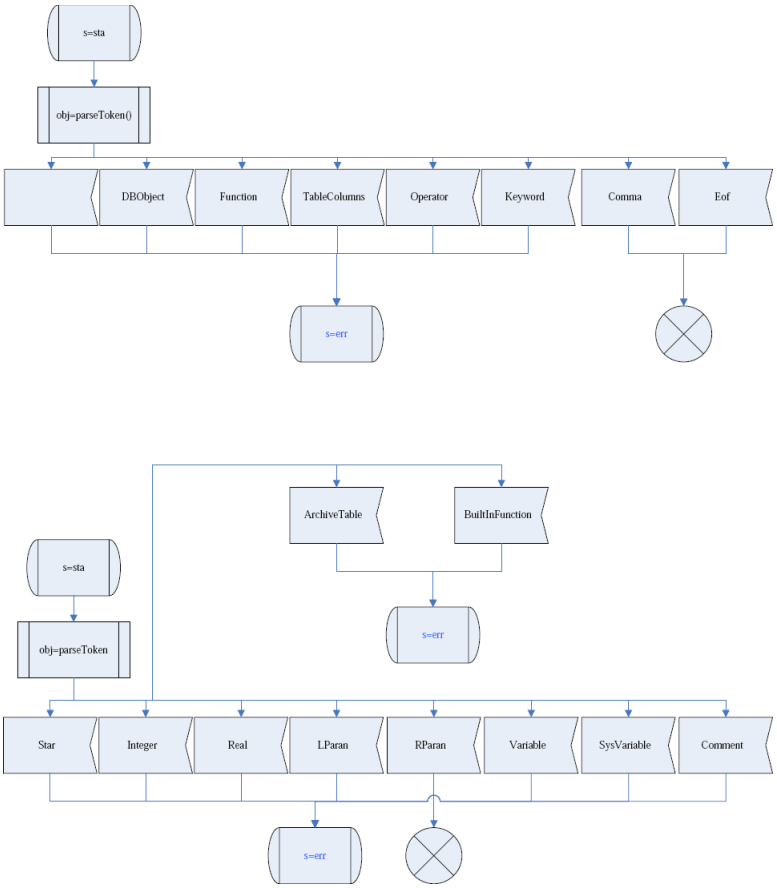

图 A.42: ADQL表达式状态迁移图之五

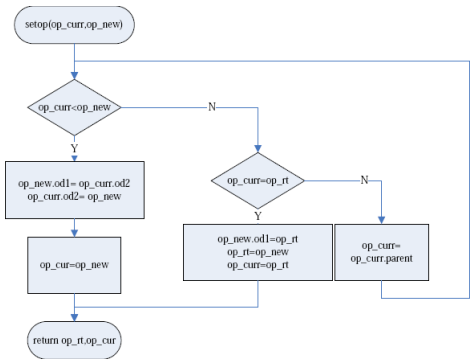

图 A.43: ADQL表达式状态迁移图之六

A.4 WHERE分句条件表达式分析的状态迁移图

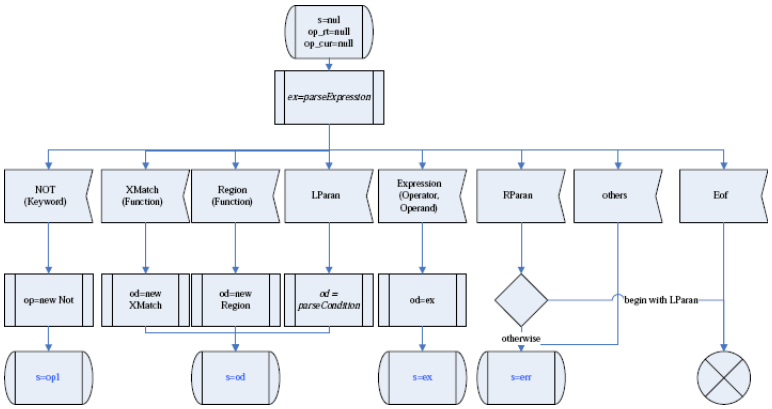

图 A.44: ADQL条件表达式状态迁移图之一

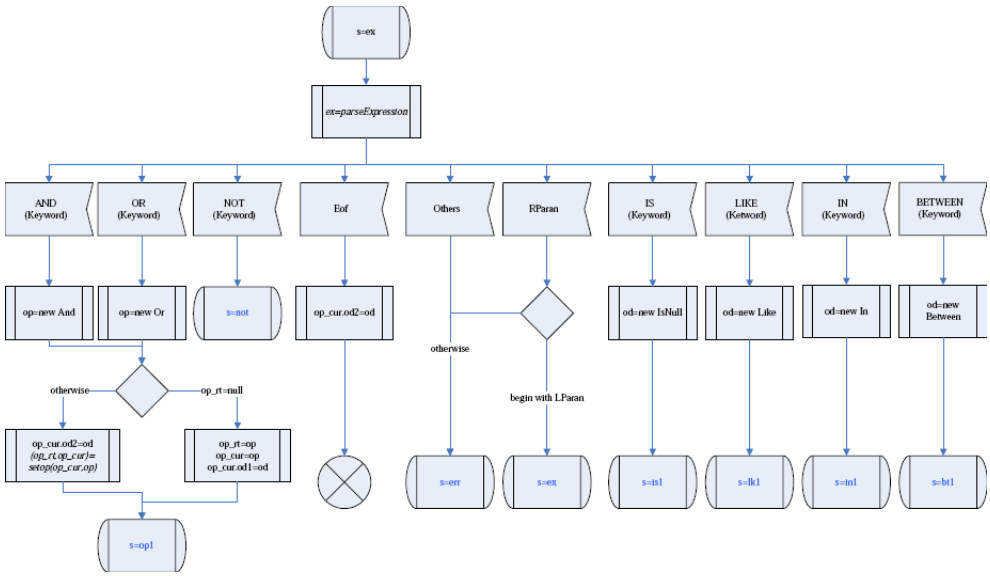

图 A.45: ADQL条件表达式状态迁移图之二

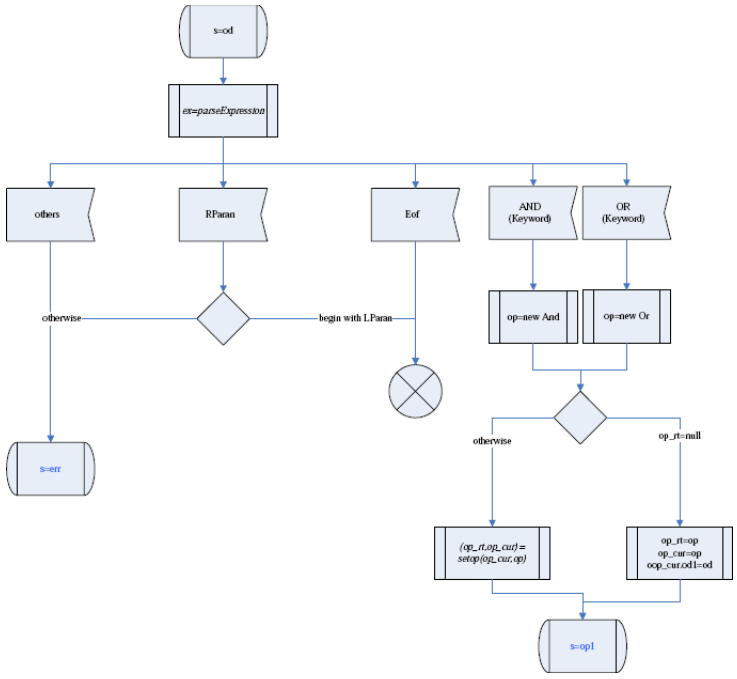

图 A.46: ADQL条件表达式状态迁移图之三

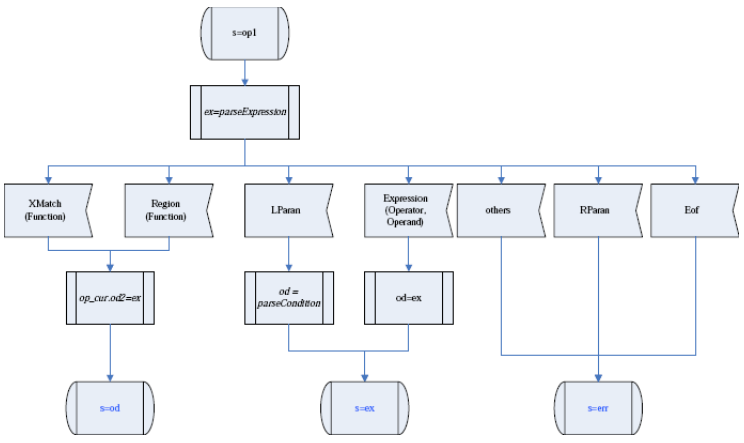

图 A.47: ADQL条件表达式状态迁移图之四

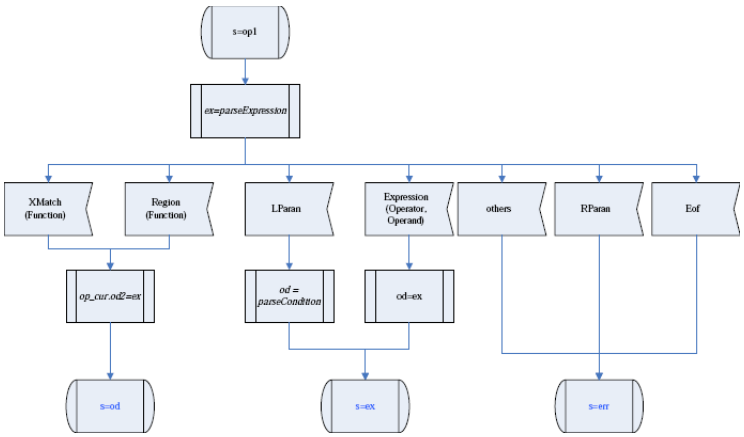

图 A.48: ADQL条件表达式状态迁移图之五

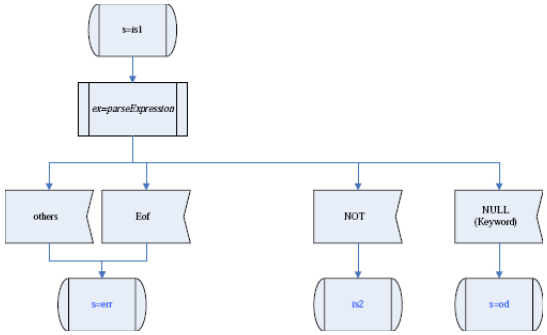

图 A.49: ADQL条件表达式状态迁移图之六

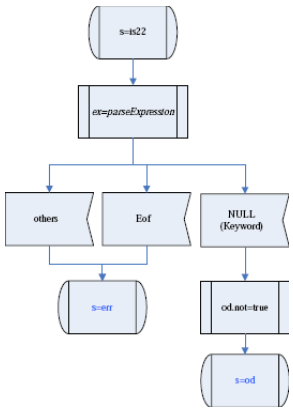

图 A.50: ADQL条件表达式状态迁移图之七

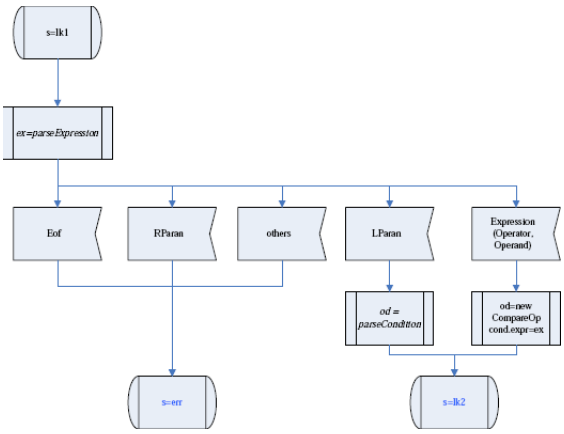

图 A.51: ADQL条件表达式状态迁移图之八

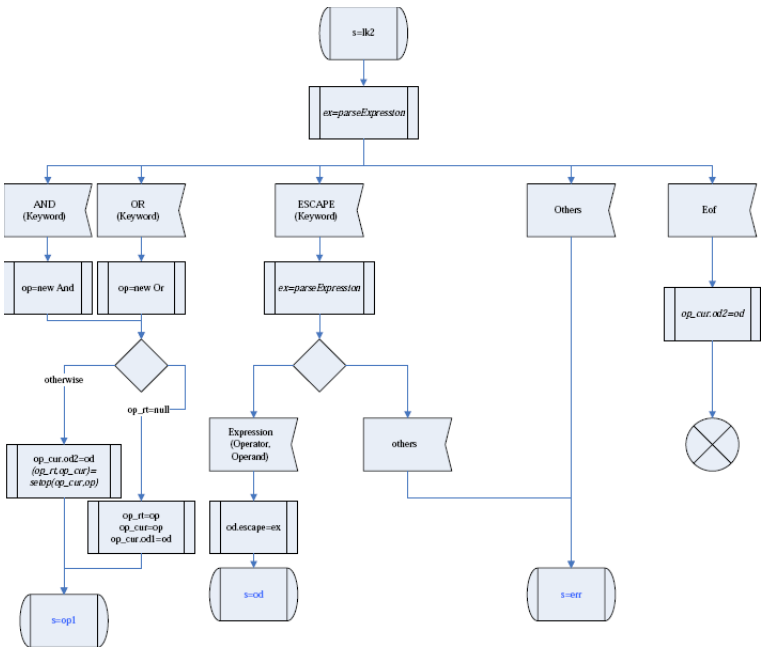

图 A.52: ADQL条件表达式状态迁移图之九

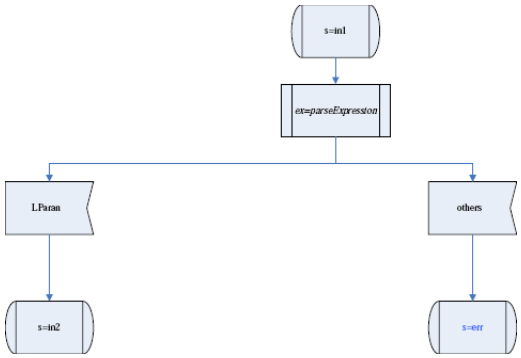

图 A.53: ADQL条件表达式状态迁移图之十

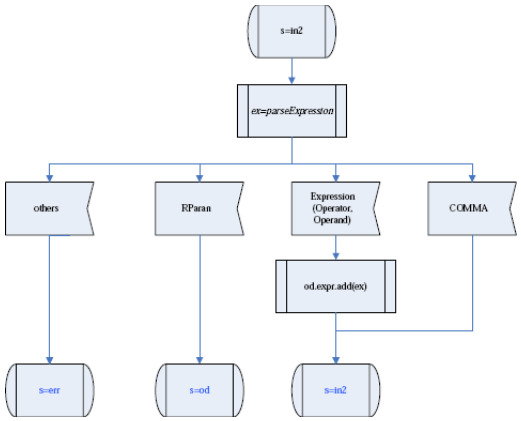

图 A.54: ADQL条件表达式状态迁移图之十一

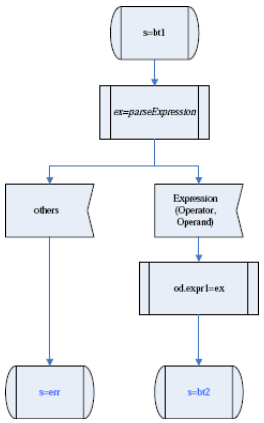

图 A.55: ADQL条件表达式状态迁移图之十二

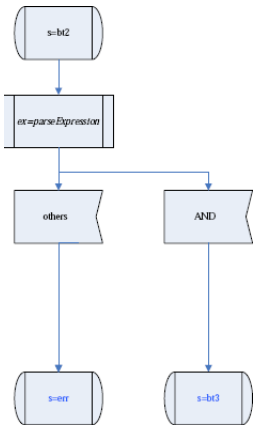

图 A.56: ADQL条件表达式状态迁移图之十三

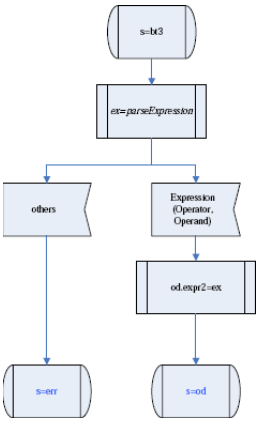

图 A.57: ADQL条件表达式状态迁移图之十四

附录 B　JDL/s的定义

/*The follwing section is the definition of JDL document structure*/

jdldocument::=project

project::=key_project space+ projectname newline (job newline)* key_end

projectname::=name

job::=description* key_job space+ jobname newline variabledeclaration* newline mainfunctiondefinition newline functiondefinition* newline key_end

jobname::=name

variabledeclaration::=description* key_var space+ variablename space* semicolon

variablename::=name

description::=digitsign string newline

/*the following section is the definition of function structure in JDL*/

mainfunctiondefinition::=key_function space+ returnvalue space* assign space* key_main leftparathesis space* rightparathesis newline (statement newline)* key_end

returnvalue::=variablename | (leftbracket variablename (space* comma space* variablename)+ rightbracket)

functiondefinition::= key_function space+ (returnvalue space* assign)? space* functionname leftparathesis space* parameterlist? space* rightparathesis newline (statement newline)* key_end

functionname::=name

parameterlist::=variablename (space* comma space* variablename)+

statement::=expression | ifstatement | switchstatement | forstatement | whilestatement | key_break | key_continue

/*The following section is the definition of expression in JDL*/

expression::=expressioncell semicolon

expressioncell::=expressioncell space* operator2 space* expressioncell | operator1 space* expressioncell | expressioncell space* operator0 | operand

operator0::=transpose | arraytranspose

operator1::=not

operator2::=bitand | assign | product | leftdivision | power | and | outerproduct | or | equals | greaterequal | greaterthan | lessequal | lessthan | subtraction | innerproduct | arrayproduct | arrayleftdivision | arraypower | arraydivision | mod | sum | division | notequal | bitor

bitand::=ampersand

product::=asterisk

leftdivision::=backslash

power::=caret

and::=doubleampersand

outerproduct::=doubleasterisk

or::=doubleverticalbar

subtraction::=minus

innerproduct::=period

arrayproduct::=periodasterisk

arrayleftdivision::=periosbackslash

arraypower::=periodcaret

arraydivision::=periodslash

mod::=persant

sum::=plus

transpose::=quotation

arraytranspose::=periodquotation

not::=tilde

operand::=nan | boolean | integer | real | complex | matrix | quotation char quotation | doublequotation string doublequotation | variable | functioncall

boolean::=true | false

integer::=(plus | minus)? digit*

real::= ((plus | minus)? Infinity) | ((plus | minus)? digit* period digit*) | ((plus | minus)? digit* (period digit*)? (#e|#E) ((plus | minus)? digit* (period digit*)?) | pi

complex ::=(real space* (plus|minus) space*)? real space* (#i | #I | #j | #J)

matrix::=(leftbracket space* row (space* semicolon space* row)* space* rightbracket) | series

series::=(integer | real) (colon (integer | real))? colon (integer | real)

row::=number (space* comma space* number)*

number::=nan | boolean | integer | real | complex | null | quotation char quotation | doublequotation string doublequotation

variable::=variablename (leftparathesis expressioncell (comma expressioncell)+ rightparathesis)?

functioncall::=functionname leftparathesis expressioncell (comma expressioncell)+ rightparathesis

/*The following section is the definitions of flow control blocks in JDL*/

ifstatement::=key_if space+ expressioncell newline (statement newline)* (key_elseif spae+ expressioncell (statement newline)*)* (key_else (statement newline)*)? key_end

switchstatement::=key_switch space+ expressioncell newline (key_case space+ expressioncell newline statement newline)*)+ (key_otherwise newline (statement newline)*)? key_end

forstatement::=key_for space+ variablename space* assign space* series newline (statement newline)* key_end

whilestatement::=key_whilespace+ expressioncell newline (statement newline)* key_end

/*The following section is the definitions of JDL keywords*/

key_break::=#b#r#e#a#k

key_case::=#c#a#s#e

key_continue::=#c#o#n#t#i#n#u#e

key_else::=#e#l#s#e

key_elseif::=#e#l#s#e#i#f

key_end::=#e#n#d

false::=#f#a#l#s#e

key_for::=#f#o#r

key_function::=#f#u#n#c#t#i#o#n

key_if::=#i#f

Infinity ::=#I#n#f#i#n#i#t#y

key_job::=#j#o#b

key_main::=#m#a#i#n

nan::=#N#a#N

null::=#n#u#l#l

key_otherwise::=#o#t#h#e#r#w#i#s#e

pi::=#p#i | #P#i | #P#I

key_project::=#p#r#o#j#e#c#t

key_switch::=#s#w#i#t#c#h

true:=#t#r#u#e

key_var::=#v#a#r

key_while::=#w#h#i#l#e

/*The following section is the definitions of all validated characters in JDL*/

ampersand::=#x26 /*&*/

assign::=#x3D /*=*/

asterisk::=#x2A /* * */

backslash::=#x5C /**/

caret::=#x5E /*$\hat{*}$/

colon::=#x3A /*:*/

comma::=#x2C /*,*/

digitsign::=#x23 /*#*/

doubleampersand::=ampersand ampersand /*&&*/

doubleasterisk::=asterisk asterisk /* ** */

doublequotation::=#x22 /*” */

doubleverticalbar::=verticalbar verticalbar /* || */

equals::=#x3D /*==*/

greaterequal::=greaterthan assign /* >= */

greaterthan::=#x3E /* > */

leftbrace::=#7B /* { */

leftbracket::=#x5B /*[*/

leftparathesis::=#x28 /*(*/

lessequal::=lessthan assign /* <= */

lessthan::=#x3C /* < */

minus::=#x2D /* - */

period::=#x2E /* . */

periodasterisk::=period asterisk /* .* */

periodbackslash::=period backslash /* .*/

periodcaret::=period caret /* .ˆ */

periodquotation:: period quotation /* .’ */

periodslash::=period slash /* ./ */

persant::=#x25 /* % */

plus::=#x2B /* + */

quotation::=#x27 /* ‘*/

rightbrace::=#x7D /* } */

rightbracket::=#x5D /*]*/

rightparathesis::=#x29 /*)*/

semicolon::=#x3B /*;*/

slash::=#x2F /* / */

tilde::=#x7E /* ~ */

tildeequal::=tilde assign /* ~= */

vertivalbar::=#x7C /* | */

char::=[#x41-#x5A] | [#x61-#x7A] | #x5F

digit::=[#x30-#x39]

name::=char (char|digit)*

string::=([#x20-#x7E] | #x0C | #x0A)+

space::=#x20+

newline::=space (#xD#xA | #xD | #xA)+

附录 C　524个候选体和对它们的证认

表 C.1: 524个候选体和对它们的证认

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
120.1	25.1	2.2402	0.55	
120.3	52.9	2.6051	0.59	
121.1	54.9	2.664	0.49	
121.1	56.3	2.0689	0.49	
121.5	40.5	2.164	0.47	
121.9	52.3	2.2392	0.54	
122.7	56.9	2.0292	0.61	
123.3	54.9	2.0147	0.56	
123.3	58.3	2.2603	0.48	
123.5	51.1	2.0118	0.5	SDSSJ0814+5105
125.1	57.5	2.0666	0.51	SDSSJ0821+5608
125.3	56.1	2.1343	0.45	
125.3	56.7	2.4567	0.52	
126.1	35.1	2.0414	0.62	
126.1	45.5	2.2614	0.5	
126.1	53.3	2.0342	0.63	
127.1	57.3	2.218	0.49	
127.9	58.5	2.0867	0.61	
128.1	49.5	2.2543	0.55	
128.7	60.5	2.0698	0.62	
129.3	49.9	2.2421	0.48	
129.5	51.9	2.2621	0.53	
129.7	46.5	2.2141	0.53	
130.3	59.7	2.1946	0.57	
130.7	25.1	2.0658	0.64	
131.3	35.1	2.0917	0.54	

表 C.2: 524个候选体和对它们的证认（续一）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
131.5	54.9	2.0577	0.51	
131.9	55.3	2.3196	0.49	
131.9	56.9	2.3267	0.58	
132.5	52.1	2.2182	0.6	
132.5	63.1	4.6871	0.48	UMa II
132.7	31.9	2.2937	0.55	
132.7	63.1	3.2752	0.48	UMa II
132.9	48.5	2.0753	0.58	
132.9	63.1	6.2763	0.5	UMa II
132.9	64.5	2.1611	0.4	UMa II
133.1	54.5	2.0385	0.46	
133.1	63.1	4.3198	0.48	UMa II
133.3	63.1	4.3376	0.47	UMa II
133.5	63.1	2.669	0.47	UMa II
133.7	41.9	2.0415	0.58	
134.5	63.1	2.1632	0.6	UMa II
134.7	61.7	2.162	0.57	
135.3	38.9	2.0422	0.53	
135.7	55.1	2.1848	0.51	
135.7	61.7	2.6249	0.54	
135.9	63.3	2.8252	0.56	
136.1	53.9	2.3953	0.58	
136.7	45.5	2.7388	0.59	
136.9	47.1	2.0059	0.48	
136.9	64.3	2.1528	0.66	
137.5	66.3	2.4537	0.54	
137.7	44.3	2.0139	0.61	

表 C.3: 524个候选体和对它们的证认（续二）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
137.7	60.1	2.4347	0.49	NGC 2768
137.9	58.9	2.2545	0.61	
137.9	65.5	2.0451	0.5	
137.9	66.7	2.6206	0.51	
138.1	49.5	2.2553	0.6	
138.1	49.9	2.3102	0.59	
138.1	66.3	2.097	0.51	
138.3	57.5	2.0091	0.57	
138.3	60.1	2.1767	0.44	NGC 2768
139.7	58.1	3.2592	0.52	
139.7	64.5	2.5099	0.49	
139.9	45.1	2.1203	0.62	
140.1	51.9	2.0633	0.64	
140.3	61.5	2.2297	0.47	
140.5	49.9	2.0779	0.58	
140.9	47.7	2.3493	0.61	
140.9	62.1	3.2006	0.62	
140.9	66.1	2.2824	0.54	
141.1	39.9	2.0278	0.63	
141.5	58.9	2.706	0.6	
141.5	66.3	2.2856	0.53	
142.5	68.1	2.1065	0.65	
142.7	56.9	2.0038	0.5	
143.1	65.5	2.2661	0.64	
143.3	68.5	2.3618	0.43	
143.5	62.9	2.098	0.57	
143.7	61.3	2.3205	0.58	
143.9	48.3	2.3476	0.58	
143.9	55.5	2.257	0.63	
143.9	58.5	2.1783	0.53	

表 C.4: 524个候选体和对它们的证认（续三）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
144.1	59.1	2.3472	0.53	
144.1	65.3	2.589	0.5	
144.3	45.1	2.0645	0.53	
144.5	51.5	2.3177	0.6	
144.5	66.9	2.0395	0.65	
144.7	46.3	2.3022	0.63	
144.7	64.5	2.111	0.56	
144.9	42.3	2.054	0.59	
145.7	36.3	2.0536	0.59	
145.7	43.1	2.0787	0.59	
145.7	56.7	2.0146	0.43	
145.9	58.1	2.2709	0.56	
146.1	55.3	2.1295	0.54	
146.1	68.1	2.8075	0.58	
146.3	25.1	2.0379	0.58	
146.3	62.1	2.0976	0.5	
146.3	69.1	2.4015	0.59	NGC 3031
146.5	65.3	2.3299	0.55	
147.1	25.1	2.0755	0.49	
147.1	58.7	2.0112	0.6	
147.1	59.3	2.0129	0.59	
147.3	69.5	2.3716	0.56	NGC 3031
147.5	59.1	2.0142	0.59	
147.7	64.9	2.0407	0.63	
147.9	25.7	2.006	0.56	
147.9	67.3	2.0556	0.59	
148.1	51.5	2.1175	0.59	
148.3	69.7	2.596	0.34	Galaxy
148.5	62.7	2.2821	0.57	
148.7	69.1	2.8112	0.38	NGC 3031

表 C.5: 524个候选体和对它们的证认（续四）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
148.7	69.3	2.6325	0.38	NGC 3031
148.9	68.9	2.8469	0.37	NGC 3031
148.9	69.1	2.5185	0.37	NGC 3031
149.1	25.1	2.0845	0.53	
149.3	69.1	2.5082	0.37	NGC 3031
149.5	51.7	2.0228	0.52	
149.7	51.7	2.0922	0.51	
149.7	64.1	2.2263	0.65	
149.7	67.5	2.2922	0.56	
149.9	63.3	2.2685	0.56	
150.1	39.1	2.1183	0.61	
150.1	57.5	2.2526	0.53	SDSSJ1000+5730
150.3	53.9	2.3457	0.56	
150.7	51.3	2.1153	0.58	
150.9	67.9	2.6959	0.5	
151.1	69.5	2.3358	0.49	
151.3	67.9	2.2549	0.5	
151.3	69.9	2.0038	0.53	
151.5	61.5	2.2899	0.61	
151.5	62.7	2.0549	0.58	
151.7	64.7	2.1252	0.57	
152.1	68.5	2.1816	0.58	
152.3	57.9	3.5006	0.62	
152.7	63.1	2.0354	0.58	
152.7	68.5	3.231	0.54	
153.1	62.9	2.1988	0.59	
153.1	68.7	2.257	0.58	
153.3	68.3	2.4748	0.58	
153.3	68.5	2.8291	0.58	
153.5	55.1	2.1005	0.57	

表 C.6: 524个候选体和对它们的证认（续五）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
153.5	68.3	2.5082	0.58	
153.5	68.5	2.8193	0.58	
153.7	45.5	2.0292	0.61	
153.7	61.9	2.4015	0.5	
153.7	68.5	3.4964	0.57	
153.9	49.5	2.0299	0.58	
153.9	62.9	2.1508	0.58	
154.1	65.1	2.0239	0.62	
154.3	40.3	2.071	0.59	
154.3	64.3	2.0787	0.58	
154.9	32.5	2.121	0.59	
154.9	46.3	2.463	0.56	
154.9	50.7	2.0954	0.57	
155.1	26.7	2.1663	0.47	
155.1	56.1	2.0552	0.59	
155.3	61.7	2.9553	0.43	
155.5	64.1	2.6193	0.57	
155.7	59.9	2.1038	0.61	
156.5	48.5	2.0779	0.57	
156.7	46.9	2.2609	0.56	
156.9	65.3	2.6655	0.61	
157.3	66.1	2.183	0.59	
157.5	66.1	2.1371	0.57	
157.7	53.3	3.1683	0.54	
157.9	66.1	3.1771	0.57	
158.1	30.7	2.3812	0.4	
158.5	51.9	2.3012	0.42	UMa I
158.7	51.9	2.4438	0.41	UMa I
158.9	51.9	2.1511	0.39	UMa I
159.3	25.1	2.4094	0.56	

表 C.7: 524个候选体和对它们的证认（续六）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
159.3	43.1	2.0202	0.47	
159.3	62.9	2.028	0.59	
159.3	66.3	2.6298	0.54	
159.9	41.5	2.0719	0.49	
159.9	54.7	2.0021	0.62	
160.1	47.9	2.1991	0.56	
160.1	51.7	2.0475	0.56	
160.1	62.5	2.0932	0.59	
160.3	64.3	2.371	0.42	
160.7	57.9	2.1413	0.53	
161.3	66.7	2.5426	0.57	
161.5	58.1	2.8449	0.5	
161.7	58.7	2.3524	0.53	
161.9	42.1	2.1525	0.42	
161.9	66.9	2.1281	0.56	
162.1	25.1	2.2403	0.55	
162.3	50.5	2.0085	0.5	
162.3	50.7	2.1843	0.53	
162.3	51.1	7.6355	0.53	Willman 1
162.5	56.7	2.3655	0.59	
162.9	36.5	2.3043	0.34	
163.5	56.3	2.6093	0.41	GCl?
163.5	65.9	2.4589	0.52	
163.7	55.1	2.1697	0.6	
163.9	67.3	2.0371	0.55	
164.1	55.1	2.0434	0.6	
164.3	53.1	2.1345	0.5	
164.5	28.7	2.0237	0.36	SDSSJ1058+2842
164.5	55.9	2.031	0.51	
165.3	41.7	2.2961	0.52	

表 C.8: 524个候选体和对它们的证认（续七）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
165.3	61.7	2.094	0.47	
165.9	42.3	2.0119	0.54	
165.9	55.5	2.0182	0.58	
165.9	67.5	2.3135	0.5	
166.1	57.3	2.0984	0.59	
166.1	58.3	2.0153	0.62	
166.3	56.1	2.1569	0.58	
166.3	59.3	2.5074	0.56	
166.5	36.9	2.0503	0.46	
166.5	51.9	2.2649	0.6	UMa I
166.9	50.1	2.0864	0.5	
167.1	45.3	2.1472	0.42	
167.3	45.7	2.1815	0.42	
167.3	53.5	2.0294	0.51	
167.3	61.3	2.0736	0.53	
167.3	66.9	2.4969	0.61	
167.5	67.3	2.2462	0.67	
167.5	67.5	2.5994	0.66	
167.7	62.5	2.5746	0.45	
167.9	50.7	2.0389	0.56	
167.9	55.7	2.0476	0.54	
168.1	66.3	2.0442	0.62	
168.3	36.1	2.1494	0.58	
168.3	43.5	2.279	0.39	NSCS J111243+433034
168.3	52.7	2.0221	0.55	
168.3	59.9	2.2697	0.55	
168.3	61.1	2.2069	0.51	
168.3	66.1	2.0168	0.64	
168.3	66.9	2.1276	0.64	
168.7	43.1	2.0503	0.38	NSCS J111243+433034

表 C.9: 524个候选体和对它们的证认（续八）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
168.7	66.3	2.267	0.53	
168.7	67.3	2.3436	0.52	
169.1	67.7	2.3008	0.51	
169.5	48.5	2.2625	0.58	
170.1	43.9	2.015	0.61	
170.1	56.1	2.1334	0.54	
170.1	65.7	2.0285	0.41	
170.1	67.9	3.4625	0.57	
170.7	41.5	2.1225	0.48	
170.9	53.1	2.1611	0.57	
171.1	53.5	2.4833	0.66	
171.1	54.3	2.3893	0.61	
171.3	60.1	2.154	0.58	
171.5	65.3	2.2474	0.55	
171.9	65.3	2.463	0.55	
172.3	28.9	3.9347	0.28	Pal 4
172.3	60.3	2.0087	0.54	
172.9	66.9	2.2254	0.57	
173.5	62.7	2.3675	0.55	
173.5	65.3	2.1182	0.52	
173.7	63.9	2.1643	0.58	
173.9	55.9	2.3718	0.6	
173.9	58.3	2.2951	0.44	
173.9	59.7	2.4149	0.6	
174.3	58.7	2.1737	0.43	
174.3	59.7	2.2154	0.6	
174.9	62.5	2.2891	0.63	
175.1	66.1	3.3833	0.59	
175.1	67.5	2.2994	0.58	
175.5	48.9	2.2243	0.54	

表 C.10: 524个候选体和对它们的证认（续九）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
175.5	61.3	2.0974	0.55	
175.5	66.1	2.3311	0.58	
176.3	47.1	2.0051	0.55	
177.1	65.1	2.2446	0.55	
177.5	62.3	2.0456	0.5	
178.3	33.3	2.0404	0.51	
178.3	66.3	2.1458	0.57	
178.5	68.1	2.0636	0.58	
178.7	66.1	2.0519	0.6	
178.9	68.1	2.235	0.51	
178.9	68.5	2.3107	0.56	
179.1	68.5	2.0315	0.55	
179.5	55.5	2.1651	0.47	NGC 3998
180.3	55.5	2.1018	0.57	NGC 3998
180.5	63.3	2.2411	0.54	
180.7	58.7	2.0011	0.62	
181.1	68.3	2.0707	0.54	
181.1	68.5	2.9122	0.54	
181.7	65.1	2.0763	0.56	
181.9	63.1	2.0341	0.58	
182.1	62.9	2.0392	0.56	
182.3	68.3	2.1873	0.54	
182.5	25.5	2.0126	0.47	
182.7	64.7	2.3154	0.57	
182.9	56.9	2.3933	0.42	
183.1	40.5	2.031	0.51	
183.9	36.3	2.5224	0.48	NGC 4214
183.9	58.7	2.7903	0.61	
184.5	47.7	2.1687	0.47	M106(NGC4258)
184.5	68.5	2.171	0.62	

表 C.11: 524个候选体和对它们的证认（续十）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
184.9	29.3	2.7083	0.21	NGC4283/NGC4278/NGC4286
184.9	68.7	2.3709	0.61	
185.1	29.3	5.0574	0.21	NGC4283/NGC4278/NGC4286
185.1	59.5	2.261	0.55	
185.7	68.5	2.3837	0.57	
185.9	25.3	2.099	0.42	
185.9	67.3	2.216	0.61	
186.1	65.1	2.0263	0.64	
186.3	25.1	2.0993	0.45	
186.3	50.3	2.0147	0.43	
186.5	25.1	2.0166	0.47	
186.5	33.5	2.7017	0.45	NGC4395
186.5	58.1	2.0349	0.6	
186.7	54.1	2.0071	0.58	
187.1	67.7	2.2534	0.43	
187.1	68.5	2.9544	0.68	
187.3	63.9	2.2928	0.58	
187.3	67.7	2.1289	0.43	
187.5	60.5	2.3133	0.59	
187.5	63.9	2.4912	0.58	
188.3	50.5	2.5351	0.55	
188.5	56.7	2.8166	0.55	
188.7	57.5	2.5524	0.55	
189.5	58.9	2.038	0.45	
190.1	68.1	2.6313	0.56	
190.3	68.5	2.3943	0.56	
190.5	41.3	2.0791	0.53	
190.9	43.1	2.0352	0.5	
190.9	68.3	2.1209	0.59	
191.1	49.1	2.1005	0.5	

表 C.12: 524个候选体和对它们的证认（续十一）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
191.3	67.9	2.215	0.59	
191.3	68.5	2.2121	0.59	
191.5	51.9	2.1645	0.53	UMa I
191.5	63.9	2.1459	0.56	
191.9	61.7	2.2995	0.53	
192.1	56.1	2.0217	0.56	
192.1	58.7	2.0814	0.63	
192.1	67.5	2.6032	0.58	
192.7	67.5	2.1438	0.63	
192.9	63.9	2.1167	0.63	
193.1	65.3	2.2684	0.66	
193.1	65.7	2.0204	0.63	
193.5	46.3	2.1785	0.44	
193.5	55.1	2.3604	0.51	
193.5	65.3	2.0862	0.65	
193.9	55.3	2.273	0.53	
194.1	63.5	2.4216	0.47	
194.3	34.3	2.4875	0.36	CV II
194.3	42.1	2.4074	0.51	
194.5	60.9	2.2476	0.52	
194.5	63.3	2.2443	0.47	
195.1	49.7	2.2567	0.52	
195.1	56.5	2.0026	0.51	
195.1	64.1	2.227	0.54	
195.3	63.5	2.4934	0.46	
195.9	25.1	2.0065	0.47	
195.9	49.1	2.3024	0.57	
195.9	67.9	2.1195	0.58	
195.9	68.1	2.4363	0.57	
196.7	65.5	2.0196	0.46	
196.9	46.5	2.371	0.41	ABELL 1682

表 C.13: 524个候选体和对它们的证认（续十二）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
197.3	68.1	2.0799	0.52	
197.7	41.9	2.0252	0.5	
197.7	60.7	2.0616	0.48	
197.9	63.3	2.1062	0.67	
198.5	51.9	2.0111	0.51	
198.5	65.9	2.4069	0.56	
198.9	42.1	2.4596	0.41	NGC5055
199.3	58.1	2.1708	0.51	
199.3	60.3	2.1106	0.57	
199.3	66.7	2.1391	0.52	
199.9	39.9	2.5418	0.51	
200.1	67.9	2.8235	0.66	
200.7	60.5	2.515	0.57	
200.9	64.3	2.0934	0.57	
201.7	32.1	2.2881	0.56	CV I
201.9	33.5	10.1027	0.26	CV I
201.9	58.5	2.0057	0.53	
201.9	62.7	2.0843	0.57	
201.9	64.7	2.2888	0.62	
202.1	33.5	11.5731	0.26	CV I
202.1	33.7	3.4221	0.25	CV I
202.3	28.7	2.0497	0.58	SDSSJ1329+2841
202.5	58.5	3.0118	0.52	
202.5	66.7	2.1133	0.58	
204.3	64.7	2.4717	0.55	
204.3	65.9	2.1016	0.49	
204.3	67.3	3.3498	0.64	
204.7	52.9	2.466	0.65	
204.9	64.9	2.6997	0.46	
205.3	28.3	8.7512	0.21	NGC 5272

表 C.14: 524个候选体和对它们的证认（续十三）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
205.3	28.5	13.0906	0.21	NGC 5272
205.3	67.3	2.1547	0.45	
205.5	28.1	2.0597	0.21	NGC 5272
205.5	28.5	6.6468	0.21	NGC 5272
205.5	59.7	2.1552	0.49	
205.7	28.3	3.1271	0.21	NGC 5272
205.7	28.5	8.777	0.21	NGC 5272
205.7	52.7	2.2302	0.58	
206.1	56.9	2.1597	0.55	
206.3	54.3	2.0167	0.49	
207.1	61.7	2.1961	0.57	
207.3	51.5	2.269	0.59	
207.5	60.1	2.3846	0.51	
207.9	62.3	2.0703	0.58	
207.9	65.9	2.0345	0.65	
208.3	37.9	2.334	0.49	
208.5	64.3	2.0264	0.58	
208.5	64.5	2.0723	0.57	
208.7	61.9	2.0339	0.57	
208.9	51.9	2.1128	0.57	
209.3	58.5	2.5246	0.65	
209.3	66.5	2.2899	0.62	
209.3	66.7	2.0188	0.64	
209.7	61.5	2.2714	0.57	
210.3	64.1	2.2275	0.58	
210.7	54.3	5.1059	0.45	NGC5457
210.7	54.5	2.3908	0.46	NGC5457
210.9	54.3	2.5766	0.45	NGC5457
211.1	28.5	4.0596	0.34	NGC 5466
211.3	28.5	34.6436	0.34	NGC 5466

表 C.15: 524个候选体和对它们的证认（续十四）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
211.3	28.7	3.1793	0.34	NGC 5466
211.3	58.3	2.4062	0.48	
211.5	25.1	2.2555	0.6	
211.5	28.5	24.6993	0.34	NGC 5466
211.5	40.9	2.0315	0.62	
211.9	45.3	2.1621	0.56	
212.5	63.1	2.1084	0.5	
212.7	58.1	2.0513	0.54	
213.1	55.5	2.3531	0.57	
213.3	43.9	2.1536	0.58	
213.3	65.3	2.4516	0.61	
214.9	51.3	2.382	0.55	
214.9	58.9	2.2514	0.52	
215.1	45.9	2.0441	0.56	
215.1	63.7	2.2702	0.5	
215.3	25.1	2.4253	0.48	
215.9	25.1	2.1706	0.64	
215.9	56.1	2.0054	0.55	
216.1	53.5	2.265	0.41	
216.1	65.3	2.1044	0.57	
217.7	54.3	2.0205	0.55	
217.9	52.9	2.1437	0.5	
217.9	64.9	2.0438	0.51	
218.1	49.7	2.719	0.55	
218.3	58.5	2.0052	0.51	
218.5	54.9	2.1155	0.62	ABELL 1936
218.7	54.3	2.1201	0.54	ABELL 1936
219.3	48.9	2.0907	0.56	
219.5	63.1	2.019	0.59	
219.7	64.3	2.1027	0.65	

表 C.16: 524个候选体和对它们的证认（续十五）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
220.7	64.3	2.1104	0.52	
220.9	49.5	2.1585	0.54	
221.9	57.9	2.1966	0.49	
223.1	51.1	2.3808	0.47	
223.5	54.1	2.0545	0.51	
224.5	46.3	2.0382	0.57	
225.1	60.5	2.1771	0.56	
225.9	56.3	2.5183	0.54	
226.5	53.5	2.5818	0.52	
227.1	41.9	2.0423	0.56	
228.1	65.1	3.7687	0.63	
228.3	65.1	2.8069	0.62	
228.3	65.3	2.4503	0.62	
228.5	42.5	2.0693	0.54	
228.5	65.1	2.2435	0.63	
228.5	65.3	2.1335	0.64	
228.9	50.3	2.2562	0.52	
229.1	43.7	2.2102	0.63	
229.3	65.1	2.1278	0.56	
229.5	61.3	2.6693	0.51	
229.9	64.7	2.009	0.5	
230.1	54.9	2.0961	0.54	
231.3	64.3	2.0964	0.53	
231.7	66.5	2.5578	0.58	
231.9	64.7	2.0964	0.56	NGC5949
232.3	51.9	2.0272	0.56	
233.7	49.7	2.0099	0.57	
234.9	58.3	2.2095	0.48	NGC 5987
237.1	61.3	2.0744	0.56	
237.1	64.3	2.1514	0.63	

表 C.17: 524个候选体和对它们的证认（续十六）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
237.3	62.5	2.1248	0.56	
237.9	62.5	2.1161	0.57	
238.1	25.1	2.0043	0.42	
238.3	25.1	2.057	0.42	
238.5	25.1	2.475	0.53	
238.5	63.7	2.0456	0.58	
238.7	60.3	2.35	0.52	
238.9	25.1	2.8199	0.53	
239.3	25.1	2.2123	0.52	
239.7	60.7	2.0497	0.55	
239.9	61.3	2.3323	0.49	
240.9	54.9	2.2037	0.53	
243.3	57.3	2.0494	0.44	
244.7	56.7	2.0029	0.51	
245.5	51.5	3.0647	0.57	
247.9	54.7	2.3501	0.55	
250.3	36.5	3.8198	0.21	NGC 6205
250.3	36.7	4.662	0.21	NGC 6205
250.5	36.3	10.3415	0.21	NGC 6205
250.5	36.5	3.1563	0.21	NGC 6205
250.5	36.7	5.5571	0.21	NGC 6205
250.7	36.3	4.1493	0.21	NGC 6205
250.7	36.5	4.3509	0.21	NGC 6205
250.9	64.1	3.6415	0.48	
251.1	64.1	3.3434	0.48	
251.9	63.3	2.3788	0.55	
252.3	52.7	2.2105	0.56	
252.5	64.1	2.1548	0.62	
256.1	25.3	2.1914	0.59	
256.1	25.7	2.2059	0.59	

表 C.18: 524个候选体和对它们的证认（续十七）

ra(J2000)	dec(J2000)	$\tilde{n}_{center}$	R	证认天体
256.1	65.3	2.013	0.58	
256.3	25.1	3.2094	0.6	
256.3	25.3	2.1448	0.58	
256.5	25.1	2.332	0.59	
256.7	65.3	2.0903	0.59	
258.9	43.1	2.0266	0.28	NGC6341
259.1	42.9	3.2472	0.28	NGC6341
259.1	43.1	8.9508	0.28	NGC6341
259.3	54.5	2.0244	0.6	
259.5	57.9	2.1619	0.33	Draco
259.7	57.9	7.0613	0.31	Draco
259.7	64.1	2.1694	0.59	
259.9	53.1	2.1076	0.48	
259.9	57.9	21.2592	0.31	Draco
260.1	57.9	26.9303	0.31	Draco
260.1	66.1	2.214	0.61	
260.3	52.9	2.0414	0.52	
260.3	57.9	16.0273	0.32	Draco
260.5	57.9	4.0301	0.31	Draco
262.9	35.3	2.034	0.56	

参考文献

[1] G. Gilmore and N. Reid. New light on faint stars. III - Galactic structure towards the South Pole and the Galactic thick disc. *MNRAS*, 202:1025–1047, March 1983.

[2] H. J. Newberg, B. Yanny, C. Rockosi, E. K. Grebel, H.-W. Rix, J. Brinkmann, I. Csabai, G. Hennessy, R. B. Hindsley, R. Ibata, Z. Ivezić, D. Lamb, E. T. Nash, M. Odenkirchen, H. A. Rave, D. P. Schneider, J. A. Smith, A. Stolte, and D. G. York. The Ghost of Sagittarius and Lumps in the Halo of the Milky Way. *ApJ*, 569:245–274, April 2002.

[3] H. J. Newberg, B. Yanny, N. Cole, T. C. Beers, P. Re Fiorentin, D. P. Schneider, and R. Wilhelm. The Overdensity in Virgo, Sagittarius Debris, and the Asymmetric Spheroid. *ApJ*, 668:221–235, October 2007.

[4] A. Naim, K. U. Ratnatunga, and R. E. Griffiths. Quantitative Morphology of Moderate-Redshift Galaxies: How Many Peculiar Galaxies Are There? *ApJ*, 476:510, February 1997.

[5] G. A. Richter. Search for optical identifications in the 5C3-radio survey. II - Statistical treatment and results. *Astronomische Nachrichten*, 296:65–81, 1975.

[6] P. J. E. Peebles. The mean mass density estimated from the Kirshner, Oemler, Schechter galaxy redshift sample. *AJ*, 84:730–734, June 1979.

[7] V. Belokurov, N. W. Evans, M. J. Irwin, P. C. Hewett, and M. I. Wilkinson. The Discovery of Tidal Tails around the Globular Cluster NGC 5466. *ApJ*, 637:L29–L32, January 2006.

[8] C. M. Rockosi, M. Odenkirchen, E. K. Grebel, W. Dehnen, K. M. Cudworth, J. E. Gunn, D. G. York, J. Brinkmann, G. S. Hennessy, and Ž. Ivezić. A

Matched-Filter Analysis of the Tidal Tails of the Globular Cluster Palomar 5. *AJ*, 124:349–363, July 2002.

[9] A. S. Szalay and R. J. Brunner. Exploring Terabyte Archives in Astronomy. In B. J. McLean, D. A. Golombek, J. J. E. Hayes, and H. E. Payne, editors, *New Horizons from Multi-Wavelength Sky Surveys*, volume 179 of *IAU Symposium*, page 455, 1998.

[10] C.-Z. Cui and Y.-H. Zhao. Architecture of Chinese Virtual Observatory. *Astronomical Research and Technology. Publications of National Astronomical Observatories of China (ISSN 1672-7673), Vol. 1, No. 2, p. 140 - 151 (2004)*, 1:140–151, June 2004.

[11] D. Egret, F. Genova, M. Allen, T. Boch, F. Bonnarel, S. Derriere, P. Fernique, M. Louys, A. Oberto, F. Ochsenbein, A. Schaaff, and M. Wenger. Aladin 1.3 at CDS: new steps towards the Virtual Observatory. In F. Combes and D. Barret, editors, *SF2A-2002: Semaine de l'Astrophysique Francaise*, page 27, June 2002.

[12] T. Boch, P. Fernique, F. Bonnarel, M. G. Allen, O. Bienaymé, and S. Derrière. Aladin, a portal for the Virtual Observatory. In D. Barret, F. Casoli, G. Lagache, A. Lecavelier, and L. Pagani, editors, *SF2A-2006: Semaine de l'Astrophysique Francaise*, page 83, June 2006.

[13] S. Kale, T. M. Vijayaraman, A. Kembhavi, P. R. Krishnan, A. Navelkar, H. Hegde, P. Kulkarni, and K. D. Balaji. VOPlot: A Toolkit for Scientific Discovery using VOTables. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 350, July 2004.

[14] M. B. Taylor. TOPCAT & STIL: Starlink Table/VOTable Processing Software. In P. Shopbell, P. M. Britton, and R. Ebert, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIV*, volume 347 of *Astronomical Society of the Pacific Conference Series*, page 29, October 2005.

[15] F. Ochsenbein, R. Williams, C. Davenhall, D. Durand, P. Fernique, R. Hanisch, D. Giaretta, T. McGlynn, A. Szalay, and A. Wicenec. VOTable: Tabular Data for the Virtual Observatory. In P. J. Quinn and K. M. Górski, editors, *Toward an International Virtual Observatory*, page 118, 2004.

[16] 崔辰州. 中国虚拟天文台系统设计. PhD thesis, 中国科学院研究生院, 2003.

[17] N. Yasuda, Y. Mizumoto, M. Ohishi, W. O'Mullane, T. Budavári, V. Haridas, N. Li, T. Malik, A. S. Szalay, M. Hill, T. Linde, B. Mann, and C. G. Page. Astronomical Data Query Language: Simple Query Protocol for the Virtual Observatory. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 293, July 2004.

[18] R. Plante, G. Greene, R. Hanisch, T. McGlynn, W. O'Mullane, and R. Williamson. Resource Registries for the Virtual Observatory. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 585, July 2004.

[19] J. C. McDowell, S. Lowe, M. Cresitello-Dittmar, J. Deponte Evans, I. Evans, A. H. Rots, and M. Harris. Spectral Data Models for the Virtual Observatory. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 269, July 2004.

[20] M. Cresitello-Dittmar, J. Deponte Evans, I. Evans, M. Harris, S. Lowe, J. C. McDowell, and A. H. Rots. Designing a Data Model for the Virtual Observatory. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 277, July 2004.

[21] A. R. Thakar, T. Budavari, T. Malik, A. S. Szalay, G. Fekete, M. Nieto-Santisteban, V. Haridas, and J. Gray. SkyQuery - A Prototype Distributed Query and Cross-Matching Web Service for the Virtual Observatory. In *Bulletin of the American Astronomical Society*, volume 34 of *Bulletin of the American Astronomical Society*, page 1274, December 2002.

[22] Y. Zhao. LAMOST project and its scientific goals. *Publications of the Yunnan Observatory*, pages 1–7, December 1999.

[23] C.-Z. Cui, M. Dolensky, P. Quinn, Y.-H. Zhao, and F. Genova. VOFilter: Bridging Virtual Observatory and Industrial Office Applications. *Chinese Journal of Astronomy and Astrophysics*, 6:379–386, June 2006.

[24] D. Wang and Y.-H. Zhao. VO-IMPAT: Image Processing and Analysis Toolkit for the Virtual Observatory of China. *Publ. Nat. Astron. Obs. China, 3, 295-303 (2006)*, 3:295–303, 2006.

[25] C. Cui, H. Sun, and H. Zhao. SkyMouse, An Integrated On-line Astronomical Information Access System. *The Virtual Observatory in Action: New Science, New Technology, and Next Generation Facilities, 26th meeting of the IAU, Special Session 3, 17-18, 21-22 August, 2006 in Prague, Czech Republic, SPS3, #55*, 3, August 2006.

[26] C. Liu, H. Tian, D. Gao, Y. Yang, Y. Lu, C. Cui, and Y. Zhao. Integrated Access of Distributed and Heterogeneous Astronomical Data Resources. *Astronomical Research and Technology. Publications of National Astronomical Observatories of China (ISSN 1672-7673), in press*, 2007.

[27] B. Liu, C.-Z. Cui, and Y.-H. Zhao. Construction of the Sky Node system for Chinese Virual Observatory. *Astronomical Research and Technology. Publications of National Astronomical Observatories of China (ISSN 1672-7673), Vol. 3, No. 4, p. 355 - 364 (2006)*, 3:355–364, December 2006.

[28] C. Liu, D. Wang, B. Liu, D. Gao, C. Cui, and Y. Zhao. An astronomical data mining application framework for virtual observatory. In *Advanced*

Software and Control for Astronomy. Edited by Lewis, Hilton; Bridger, Alan. Proceedings of the SPIE, Volume 6274, pp. 627415 (2006)., volume 6274 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, July 2006.

[29] D. Wang, Y.-X. Zhang, C. Liu, and Y.-H. Zhao. Two novel approaches for photometric redshift estimation based on SDSS and 2MASS databases. *ChJAA in press, ArXiv e-prints*, 707, July 2007.

[30] D. Wang, Y. X. Zhang, C. Liu, and Y. H. Zhao. Kernel Regression For Determining Photometric Redshifts From Sloan Broadband Photometry. *MNRAS in press, ArXiv e-prints*, 706, June 2007.

[31] D. Wang. Research on Algorithms of Estimating Photometric Redshifts Based on Large Sky Survey Databases. *PASP*, 119:1204–1204, October 2007.

[32] D. Wang, Y. Zhang, C. Cui, and Y. Zhao. Software kits for measuring photometric redshifts. In *Advanced Software and Control for Astronomy. Edited by Lewis, Hilton; Bridger, Alan. Proceedings of the SPIE, Volume 6274, pp. 627413 (2006).*, volume 6274 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, July 2006.

[33] C. Liu, J. Hu, H. Newberg, and Y. Zhao. Candidate Milky Way Satellites in the Galactic Halo. *A& A in press, ArXiv Astrophysics e-prints*, October 2007.

[34] Y. Zhao and C. Cui. Lamost Project and the Chinese Virtual Observatory. *Large Telescops and Virtual Observatory: Visions for the Future, 25th meeting of the IAU, Joint Discussion 8, 17 July 2003, Sydney, Australia*, 8:23, 2003.

[35] 张彦霞. 多波段天体物理中的自动分类方法研究. PhD thesis, 中国科学院研究生院, 2003.

[36] C. J. Grillmair and O. Dionatos. Detection of a 63° Cold Stellar Stream in the Sloan Digital Sky Survey. *ApJ*, 643:L17–L20, May 2006.

[37] D. G. York, J. Adelman, J. E. Anderson, Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. Fan, G. R. Federwitz, L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C.-h. Huang, C. Hull, D. Husby, S.-I. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Kent, R. S. J. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider, M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, S.-i. Wang, M. Watanabe, D. H. Weinberg, B. Yanny, and N. Yasuda. The Sloan Digital Sky Survey: Technical Summary. *AJ*, 120:1579–1587, September 2000.

[38] X. Fan, V. K. Narayanan, R. H. Lupton, M. A. Strauss, G. R. Knapp, R. H.

Becker, R. L. White, L. Pentericci, S. K. Leggett, Z. Haiman, J. E. Gunn, Ž. Ivezić, D. P. Schneider, S. F. Anderson, J. Brinkmann, N. A. Bahcall, A. J. Connolly, I. Csabai, M. Doi, M. Fukugita, T. Geballe, E. K. Grebel, D. Harbeck, G. Hennessy, D. Q. Lamb, G. Miknaitis, J. A. Munn, R. Nichol, S. Okamura, J. R. Pier, F. Prada, G. T. Richards, A. Szalay, and D. G. York. A Survey of z>5.8 Quasars in the Sloan Digital Sky Survey. I. Discovery of Three New Quasars and the Spatial Density of Luminous Quasars at z˜6. *AJ*, 122:2833–2849, December 2001.

[39] 田海俊. 虚拟天文台数据访问服务（vo-das）之任务调度及vo-das的应用. Master's thesis, 华中师范大学, 2007.

[40] T. Budavári, A. S. Szalay, J. Gray, W. O'Mullane, R. Williams, A. Thakar, T. Malik, N. Yasuda, and R. Mann. Open SkyQuery – VO Compliant Dynamic Federation of Astronomical Archives. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 177, July 2004.

[41] 李国杰. 信息服务网格－第三代internet. 计算机世界, 2001.

[42] M. Antonioletti, M. P. Atkinson, R. Baxter, A. Borley, N. P. Chue Hong, B. Collins, N. Hardman, A. Hume, A. Knox, M. Jackson, A. Krause, S. Laws, J. Magowan, N. W. Paton, D. Pearson, T. Sugden, P. Watson, and M. Westhead. The design and implementation of grid database services in ogsa-dai. *Concurrency and Computation: Practice and Experience*, 2005.

[43] K. Karasavvas, M. Antonioletti, M. P. Atkinson, N. P. Chue Hong, T. Sugden, A. C. Hume, M. Jackson, A. Krause, and C. Palansuriya. Introduction to ogsa-dai services. *Computer Science*, 2005.

[44] R. M. Cutri, M. F. Skrutskie, S. van Dyk, C. A. Beichman, J. M. Carpenter, T. Chester, L. Cambresy, T. Evans, J. Fowler, J. Gizis, E. Howard, J. Huchra, T. Jarrett, E. L. Kopan, J. D. Kirkpatrick, R. M. Light, K. A. Marsh, H. McCallon, S. Schneider, R. Stiening, M. Sykes, M. Weinberg, W.

A. Wheaton, S. Wheelock, and N. Zacarias. *2MASS All Sky Catalog of point sources.* The IRSA 2MASS All-Sky Point Source Catalog, NASA/IPAC Infrared Science Archive. http://irsa.ipac.caltech.edu/applications/Gator/, June 2003.

[45] 周秉峰. *UML*软件建模. 北京大学出版社, 2001.

[46] 段朝晖. Fortran语言的发展历史. 计算机世界报, 8, 1995.

[47] T. Jenness, F. Economou, R. P. J. Tilanus, C. Best, R. M. Prestage, P. Shimek, K. Glazebrook, and T. J. Farrell. Perl at the Joint Astronomy Centre. In D. M. Mehringer, R. L. Plante, and D. A. Roberts, editors, *Astronomical Data Analysis Software and Systems VIII*, volume 172 of *Astronomical Society of the Pacific Conference Series*, page 494, 1999.

[48] K. Glazebrook. Perl as an Astronomer-Friendly Language—the pgperl Experience. In G. H. Jacoby and J. Barnes, editors, *Astronomical Data Analysis Software and Systems V*, volume 101 of *Astronomical Society of the Pacific Conference Series*, page 307, 1996.

[49] N. Pirzkal and R. N. Hook. Python in Astronomy. In D. M. Mehringer, R. L. Plante, and D. A. Roberts, editors, *Astronomical Data Analysis Software and Systems VIII*, volume 172 of *Astronomical Society of the Pacific Conference Series*, page 479, 1999.

[50] 闫殿武. *IDL*可视化工具入门与提高. 机械工业出版社, 2003.

[51] L. Girardi, E. K. Grebel, M. Odenkirchen, and C. Chiosi. Theoretical isochrones in several photometric systems. II. The Sloan Digital Sky Survey ugriz system. *A& A*, 422:205–215, July 2004.

[52] M. B. Taylor. STILTS - A Package for Command-Line Processing of Tabular Data. In C. Gabriel, C. Arviset, D. Ponz, and S. Enrique, editors, *Astronomical Data Analysis Software and Systems XV*, volume 351 of *Astronomical Society of the Pacific Conference Series*, page 666, July 2006.

[53] J. C. Kapteyn. First Attempt at a Theory of the Arrangement and Motion of the Sidereal System. *ApJ*, 55:302, May 1922.

[54] H. Shapley. Studies based on the colors and magnitudes in stellar clusters. VI. On the determination of the distances of globular clusters. *ApJ*, 48:89–124, September 1918.

[55] M. R. Merrifield. The Galactic Bar. In D. Clemens, R. Shah, and T. Brainerd, editors, *Milky Way Surveys: The Structure and Evolution of our Galaxy*, volume 317 of *Astronomical Society of the Pacific Conference Series*, page 289, December 2004.

[56] W. Baade. The Resolution of Messier 32, NGC 205, and the Central Region of the Andromeda Nebula. *ApJ*, 100:137, September 1944.

[57] B. J. Bok and P. F. Bok. *The Milky Way*. Cambridge: Harvard University Press, 1981.

[58] C. C. Lin and F. H. Shu. On the Spiral Structure of Disk Galaxies. *ApJ*, 140:646, August 1964.

[59] J. N. Bahcall. Star counts and galactic structure. *ARA& A*, 24:577–611, 1986.

[60] N. Reid. Starcounts as a Probe of Galactic Structure. In S. R. Majewski, editor, *Galaxy Evolution. The Milky Way Perspective*, volume 49 of *Astronomical Society of the Pacific Conference Series*, page 37, January 1993.

[61] J. N. Bahcall and R. M. Soneira. The universe at faint magnitudes. I - Models for the galaxy and the predicted star counts. *ApJS*, 44:73–110, September 1980.

[62] O. J. Eggen, D. Lynden-Bell, and A. R. Sandage. Evidence from the motions of old stars that the Galaxy collapsed. *ApJ*, 136:748, November 1962.

[63] J. Binny and M. Merrifield. *Galactic Astronomy*. Princeton University Press, 1998.

[64] L. Searle and R. Zinn. Compositions of halo clusters and the formation of the galactic halo. *ApJ*, 225:357–379, October 1978.

[65] C. S. Frenk, S. D. M. White, M. Davis, and G. Efstathiou. The formation of dark halos in a universe dominated by cold dark matter. *ApJ*, 327:507–525, April 1988.

[66] S. R. Majewski. The Milky-Way - Clues to a Merger Past. In S. R. Majewski, editor, *Galaxy Evolution. The Milky Way Perspective*, volume 49 of *Astronomical Society of the Pacific Conference Series*, page 5, January 1993.

[67] B. Yanny, H. J. Newberg, S. Kent, S. A. Laurent-Muehleisen, J. R. Pier, G. T. Richards, C. Stoughton, J. E. Anderson, Jr., J. Annis, J. Brinkmann, B. Chen, I. Csabai, M. Doi, M. Fukugita, G. S. Hennessy, Ž. Ivezić, G. R. Knapp, R. Lupton, J. A. Munn, T. Nash, C. M. Rockosi, D. P. Schneider, J. A. Smith, and D. G. York. Identification of A-colored Stars and Structure in the Halo of the Milky Way from Sloan Digital Sky Survey Commissioning Data. *ApJ*, 540:825–841, September 2000.

[68] Ž. Ivezić, J. Goldston, K. Finlator, G. R. Knapp, B. Yanny, T. A. McKay, S. Amrose, K. Krisciunas, B. Willman, S. Anderson, C. Schaber, D. Erb, C. Logan, C. Stubbs, B. Chen, E. Neilsen, A. Uomoto, J. R. Pier, X. Fan, J. E. Gunn, R. H. Lupton, C. M. Rockosi, D. Schlegel, M. A. Strauss, J. Annis, J. Brinkmann, I. Csabai, M. Doi, M. Fukugita, G. S. Hennessy, R. B. Hindsley, B. Margon, J. A. Munn, H. J. Newberg, D. P. Schneider, J. A. Smith, G. P. Szokoly, A. R. Thakar, M. S. Vogeley, P. Waddell, N. Yasuda, and D. G. York. Candidate RR Lyrae Stars Found in Sloan Digital Sky Survey Commissioning Data. *AJ*, 120:963–977, August 2000.

[69] S. R. Majewski, M. F. Skrutskie, M. D. Weinberg, and J. C. Ostheimer. A Two Micron All Sky Survey View of the Sagittarius Dwarf Galaxy. I.

Morphology of the Sagittarius Core and Tidal Arms. *ApJ*, 599:1082–1115, December 2003.

[70] H. J. Newberg, B. Yanny, E. K. Grebel, G. Hennessy, Ž. Ivezić, D. Martinez-Delgado, M. Odenkirchen, H.-W. Rix, J. Brinkmann, D. Q. Lamb, D. P. Schneider, and D. G. York. Sagittarius Tidal Debris 90 Kiloparsecs from the Galactic Center. *ApJ*, 596:L191–L194, October 2003.

[71] V. Belokurov, D. B. Zucker, N. W. Evans, G. Gilmore, S. Vidrih, D. M. Bramich, H. J. Newberg, R. F. G. Wyse, M. J. Irwin, M. Fellhauer, P. C. Hewett, N. A. Walton, M. I. Wilkinson, N. Cole, B. Yanny, C. M. Rockosi, T. C. Beers, E. F. Bell, J. Brinkmann, Ž. Ivezić, and R. Lupton. The Field of Streams: Sagittarius and Its Siblings. *ApJ*, 642:L137–L140, May 2006.

[72] M. Fellhauer, V. Belokurov, N. W. Evans, M. I. Wilkinson, D. B. Zucker, G. Gilmore, M. J. Irwin, D. M. Bramich, S. Vidrih, R. F. G. Wyse, T. C. Beers, and J. Brinkmann. The Origin of the Bifurcation in the Sagittarius Stream. *ApJ*, 651:167–173, November 2006.

[73] D. Martínez-Delgado, J. Peñarrubia, M. Jurić, E. J. Alfaro, and Z. Ivezić. The Virgo Stellar Overdensity: Mapping the Infall of the Sagittarius Tidal Stream onto the Milky Way Disk. *ApJ*, 660:1264–1272, May 2007.

[74] L. Monaco, M. Bellazzini, P. Bonifacio, A. Buzzoni, F. R. Ferraro, G. Marconi, L. Sbordone, and S. Zaggia. High-resolution spectroscopy of RGB stars in the Sagittarius streams. I. Radial velocities and chemical abundances. *A& A*, 464:201–209, March 2007.

[75] M. Juric, Z. Ivezic, A. Brooks, R. H. Lupton, D. Schlegel, D. Finkbeiner, N. Padmanabhan, N. Bond, C. M. Rockosi, G. R. Knapp, J. E. Gunn, T. Sumi, D. Schneider, J. C. Barentine, H. J. Brewington, J. Brinkmann, M. Fukugita, M. Harvanek, S. J. Kleinman, J. Krzesinski, D. Long, E. H. Neilsen, Jr., A. Nitta, S. A. Snedden, and D. G. York. The Milky Way Tomography with SDSS. *ArXiv Astrophysics e-prints*, October 2005.

[76] H. J. Newberg and B. Yanny. The Halo of the Milky Way. In P. K. Seidelmann and A. K. B. Monet, editors, *Astrometry in the Age of the Next Generation of Large Telescopes*, volume 338 of *Astronomical Society of the Pacific Conference Series*, page 210, October 2005.

[77] Y. Xu, L. C. Deng, and J. Y. Hu. The asymmetric structure of the Galactic halo. *MNRAS*, 368:1811–1821, June 2006.

[78] S. V. Duffau, R. Zinn, G. Carraro, R. A. Méndez, A. K. Vivas, C. Gallart, and R. Winnick. Confirmation of Halo Substructure using Quest RR Lyrae Data: The New Virgo Stellar Stream (VSS). In *Revista Mexicana de Astronomia y Astrofisica Conference Series*, volume 26 of *Revista Mexicana de Astronomia y Astrofisica Conference Series*, pages 70–71, June 2006.

[79] B. Yanny, H. J. Newberg, E. K. Grebel, S. Kent, M. Odenkirchen, C. M. Rockosi, D. Schlegel, M. Subbarao, J. Brinkmann, M. Fukugita, Ž. Ivezic, D. Q. Lamb, D. P. Schneider, and D. G. York. A Low-Latitude Halo Stream around the Milky Way. *ApJ*, 588:824–841, May 2003.

[80] J. Peñarrubia, D. Martínez-Delgado, H. W. Rix, M. A. Gómez-Flechoso, J. Munn, H. Newberg, E. F. Bell, B. Yanny, D. Zucker, and E. K. Grebel. A Comprehensive Model for the Monoceros Tidal Stream. *ApJ*, 626:128–144, June 2005.

[81] D. Martínez-Delgado, J. Peñarrubia, D. I. Dinescu, D. J. Butler, and H. W. Rix. The Canis Major dwarf galaxy as the progenitor of the Monoceros tidal stream. In H. Jerjen and B. Binggeli, editors, *IAU Colloq. 198: Near-fields cosmology with dwarf elliptical galaxies*, pages 97–100, 2005.

[82] N. F. Martin, R. A. Ibata, M. Bellazzini, M. J. Irwin, G. F. Lewis, and W. Dehnen. A dwarf galaxy remnant in Canis Major: the fossil of an in-plane accretion on to the Milky Way. *MNRAS*, 348:12–23, February 2004.

[83] V. Belokurov, N. W. Evans, M. J. Irwin, D. Lynden-Bell, B. Yanny, S. Vidrih, G. Gilmore, G. Seabroke, D. B. Zucker, M. I. Wilkinson, P. C.

Hewett, D. M. Bramich, M. Fellhauer, H. J. Newberg, R. F. G. Wyse, T. C. Beers, E. F. Bell, J. C. Barentine, J. Brinkmann, N. Cole, K. Pan, and D. G. York. An Orphan in the "Field of Streams". *ApJ*, 658:337–344, March 2007.

[84] M. Fellhauer, N. W. Evans, V. Belokurov, D. B. Zucker, B. Yanny, M. I. Wilkinson, G. Gilmore, M. J. Irwin, D. M. Bramich, S. Vidrih, P. Hewett, and T. Beers. Is Ursa Major II the progenitor of the Orphan Stream? *MNRAS*, 375:1171–1179, March 2007.

[85] S. Jin and D. Lynden-Bell. Are Complex A and the Orphan stream related? *MNRAS*, 378:L64–L66, June 2007.

[86] N. F. Martin, R. A. Ibata, S. C. Chapman, M. Irwin, and G. F. Lewis. A Keck/DEIMOS spectroscopic survey of faint Galactic satellites: searching for the least massive dwarf galaxies. *MNRAS*, 380:281–300, September 2007.

[87] C. J. Grillmair. Substructure in Tidal Streams: Tributaries in the Anticenter Stream. *ApJ*, 651:L29–L32, November 2006.

[88] M. Odenkirchen, E. K. Grebel, C. M. Rockosi, W. Dehnen, R. Ibata, H.-W. Rix, A. Stolte, C. Wolf, J. E. Anderson, Jr., N. A. Bahcall, J. Brinkmann, I. Csabai, G. Hennessy, R. B. Hindsley, Ž. Ivezić, R. H. Lupton, J. A. Munn, J. R. Pier, C. Stoughton, and D. G. York. Detection of Massive Tidal Tails around the Globular Cluster Palomar 5 with Sloan Digital Sky Survey Commissioning Data. *ApJ*, 548:L165–L169, February 2001.

[89] C. J. Grillmair and R. Johnson. The Detection of a 45 deg Tidal Stream Associated with the Globular Cluster NGC 5466. *ApJ*, 639:L17–L20, March 2006.

[90] E. F. Bell, D. B. Zucker, V. Belokurov, S. Sharma, K. V. Johnston, J. S. Bullock, D. W. Hogg, K. Jahnke, J. T. A. de Jong, T. C. Beers, N. W. Evans, E. K. Grebel, Z. Ivezic, S. E. Koposov, H.-W. Rix, D. P. Schneider,

M. Steinmetz, and A. Zolotov. The accretion origin of the Milky Way's stellar halo. *ArXiv e-prints*, 706, May 2007.

[91] M. J. Irwin, P. S. Bunclark, M. T. Bridgeland, and R. G. McMahon. A new satellite galaxy of the Milky Way in the constellation of Sextans. *MNRAS*, 244:16P–19P, May 1990.

[92] R. A. Ibata, G. Gilmore, and M. J. Irwin. Sagittarius: the nearest dwarf galaxy. *MNRAS*, 277:781–800, December 1995.

[93] B. Willman, J. J. Dalcanton, D. Martinez-Delgado, A. A. West, M. R. Blanton, D. W. Hogg, J. C. Barentine, H. J. Brewington, M. Harvanek, S. J. Kleinman, J. Krzesinski, D. Long, E. H. Neilsen, Jr., A. Nitta, and S. A. Snedden. A New Milky Way Dwarf Galaxy in Ursa Major. *ApJ*, 626:L85–L88, June 2005.

[94] D. B. Zucker, V. Belokurov, N. W. Evans, M. I. Wilkinson, M. J. Irwin, T. Sivarani, S. Hodgkin, D. M. Bramich, J. M. Irwin, G. Gilmore, B. Willman, S. Vidrih, M. Fellhauer, P. C. Hewett, T. C. Beers, E. F. Bell, E. K. Grebel, D. P. Schneider, H. J. Newberg, R. F. G. Wyse, C. M. Rockosi, B. Yanny, R. Lupton, J. A. Smith, J. C. Barentine, H. Brewington, J. Brinkmann, M. Harvanek, S. J. Kleinman, J. Krzesinski, D. Long, A. Nitta, and S. A. Snedden. A New Milky Way Dwarf Satellite in Canes Venatici. *ApJ*, 643:L103–L106, June 2006.

[95] V. Belokurov, D. B. Zucker, N. W. Evans, M. I. Wilkinson, M. J. Irwin, S. Hodgkin, D. M. Bramich, J. M. Irwin, G. Gilmore, B. Willman, S. Vidrih, H. J. Newberg, R. F. G. Wyse, M. Fellhauer, P. C. Hewett, N. Cole, E. F. Bell, T. C. Beers, C. M. Rockosi, B. Yanny, E. K. Grebel, D. P. Schneider, R. Lupton, J. C. Barentine, H. Brewington, J. Brinkmann, M. Harvanek, S. J. Kleinman, J. Krzesinski, D. Long, A. Nitta, J. A. Smith, and S. A. Snedden. A Faint New Milky Way Satellite in Bootes. *ApJ*, 647:L111–L114, August 2006.

[96] D. B. Zucker, V. Belokurov, N. W. Evans, J. T. Kleyna, M. J. Irwin, M. I. Wilkinson, M. Fellhauer, D. M. Bramich, G. Gilmore, H. J. Newberg, B. Yanny, J. A. Smith, P. C. Hewett, E. F. Bell, H.-W. Rix, O. Y. Gnedin, S. Vidrih, R. F. G. Wyse, B. Willman, E. K. Grebel, D. P. Schneider, T. C. Beers, A. Y. Kniazev, J. C. Barentine, H. Brewington, J. Brinkmann, M. Harvanek, S. J. Kleinman, J. Krzesinski, D. Long, A. Nitta, and S. A. Snedden. A Curious Milky Way Satellite in Ursa Major. *ApJ*, 650:L41–L44, October 2006.

[97] C. J. Grillmair. Detection of a 60 deg-long Dwarf Galaxy Debris Stream. *ApJ*, 645:L37–L40, July 2006.

[98] V. Belokurov, D. B. Zucker, N. W. Evans, J. T. Kleyna, S. Koposov, S. T. Hodgkin, M. J. Irwin, G. Gilmore, M. I. Wilkinson, M. Fellhauer, D. M. Bramich, P. C. Hewett, S. Vidrih, J. T. A. De Jong, J. A. Smith, H.-W. Rix, E. F. Bell, R. F. G. Wyse, H. J. Newberg, P. A. Mayeur, B. Yanny, C. M. Rockosi, O. Y. Gnedin, D. P. Schneider, T. C. Beers, J. C. Barentine, H. Brewington, J. Brinkmann, M. Harvanek, S. J. Kleinman, J. Krzesinski, D. Long, A. Nitta, and S. A. Snedden. Cats and Dogs, Hair and a Hero: A Quintet of New Milky Way Companions. *ApJ*, 654:897–906, January 2007.

[99] M. J. Irwin, V. Belokurov, N. W. Evans, E. V. Ryan-Weber, J. T. A. de Jong, S. Koposov, D. B. Zucker, S. T. Hodgkin, G. Gilmore, P. Prema, L. Hebb, A. Begum, M. Fellhauer, P. C. Hewett, R. C. Kennicutt, Jr., M. I. Wilkinson, D. M. Bramich, S. Vidrih, H.-W. Rix, T. C. Beers, J. C. Barentine, H. Brewington, M. Harvanek, J. Krzesinski, D. Long, A. Nitta, and S. A. Snedden. Discovery of an Unusual Dwarf Galaxy in the Outskirts of the Milky Way. *ApJ*, 656:L13–L16, February 2007.

[100] B. Willman, M. R. Blanton, A. A. West, J. J. Dalcanton, D. W. Hogg, D. P. Schneider, N. Wherry, B. Yanny, and J. Brinkmann. A New Milky Way Companion: Unusual Globular Cluster or Extreme Dwarf Satellite? *AJ*, 129:2692–2700, June 2005.

[101] S. Koposov, J. T. A. de Jong, V. Belokurov, H.-W. Rix, D. B. Zucker, N. W. Evans, G. Gilmore, M. J. Irwin, and E. F. Bell. The Discovery of Two Extremely Low Luminosity Milky Way Globular Clusters. *ApJ*, 669:337–342, November 2007.

[102] J. D. Simon and M. Geha. The Kinematics of the Ultra-faint Milky Way Satellites: Solving the Missing Satellite Problem. *ApJ*, 670:313–331, November 2007.

[103] J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, I. K. Baldry, J. C. Barentine, T. C. Beers, V. Belokurov, A. Berlind, M. Bernardi, M. R. Blanton, J. J. Bochanski, W. N. Boroski, D. M. Bramich, H. J. Brewington, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, L. N. Carey, S. Carliles, M. A. Carr, F. J. Castander, A. J. Connolly, R. J. Cool, C. E. Cunha, I. Csabai, J. J. Dalcanton, M. Doi, D. J. Eisenstein, M. L. Evans, N. W. Evans, X. Fan, D. P. Finkbeiner, S. D. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, G. Gilmore, K. Glazebrook, J. Gray, E. K. Grebel, J. E. Gunn, E. de Haas, P. B. Hall, M. Harvanek, S. L. Hawley, J. Hayes, T. M. Heckman, J. S. Hendry, G. S. Hennessy, R. B. Hindsley, C. M. Hirata, C. J. Hogan, D. W. Hogg, J. A. Holtzman, S.-i. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Jester, D. E. Johnston, A. M. Jorgensen, M. Jurić, G. Kauffmann, S. M. Kent, S. J. Kleinman, G. R. Knapp, A. Y. Kniazev, R. G. Kron, J. Krzesinski, N. Kuropatkin, D. Q. Lamb, H. Lampeitl, B. C. Lee, R. F. Leger, M. Lima, H. Lin, D. C. Long, J. Loveday, R. H. Lupton, R. Mandelbaum, B. Margon, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, J. A. Munn, R. Nakajima, T. Nash, E. H. Neilsen, Jr., H. J. Newberg, R. C. Nichol, M. Nieto-Santisteban, A. Nitta, H. Oyaizu, S. Okamura, J. P. Ostriker, N. Padmanabhan, C. Park, J. J. Peoples, J. R. Pier, A. C. Pope, D. Pourbaix, T. R. Quinn, M. J. Raddick, P. Re Fiorentin, G. T. Richards, M. W. Richmond, H.-W. Rix, C. M. Rockosi, D. J. Schlegel, D. P. Schneider, R. Scranton, U. Seljak, E. Sheldon, K. Shimasaku, N. M. Silvestri, J. A. Smith, V. Smolčić, S. A. Snedden, A. Stebbins, C. Stoughton,

M. A. Strauss, M. SubbaRao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tegmark, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. Vanden Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, D. H. Weinberg, A. A. West, S. D. M. White, B. Wilhite, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, S. Zibetti, and D. B. Zucker. The Fifth Data Release of the Sloan Digital Sky Survey. *ApJS*, 172:634–644, October 2007.

[104] M. Fukugita, T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku, and D. P. Schneider. The Sloan Digital Sky Survey Photometric System. *AJ*, 111:1748, April 1996.

[105] D. J. Schlegel, D. P. Finkbeiner, and M. Davis. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. *ApJ*, 500:525, June 1998.

[106] W. E. Harris. A Catalog of Parameters for Globular Clusters in the Milky Way. *AJ*, 112:1487, October 1996.

[107] A. W. McConnachie and M. J. Irwin. Structural properties of the M31 dwarf spheroidal galaxies. *MNRAS*, 365:1263–1276, February 2006.

[108] I. King. The structure of star clusters. I. an empirical density law. *AJ*, 67:471, October 1962.

[109] S. M. Faber and D. N. C. Lin. Is there nonluminous matter in dwarf spheroidal galaxies. *ApJ*, 266:L17–L20, March 1983.

[110] M. Irwin and D. Hatzidimitriou. Structural parameters for the Galactic dwarf spheroidals. *MNRAS*, 277:1354–1378, December 1995.

[111] M. L. Mateo. Dwarf Galaxies of the Local Group. *ARA& A*, 36:435–506, 1998.

[112] N. F. Martin, R. A. Ibata, M. J. Irwin, S. Chapman, G. F. Lewis, A. M. N. Ferguson, N. Tanvir, and A. W. McConnachie. Discovery and analysis of

three faint dwarf galaxies and a globular cluster in the outer halo of the Andromeda galaxy. *MNRAS*, 371:1983–1991, October 2006.

[113] J. Binny and S. Tremaine. *Galactic Dynamics*. Princeton University Press, 1987.

[114] G. Mandushev, A. Staneva, and N. Spasova. Dynamical masses for Galactic globular clusters. *A& A*, 252:94–99, December 1991.

[115] J. K. Adelman-McCarthy and for the SDSS Collaboration. The Sixth Data Release of the Sloan Digital Sky Survey. *ArXiv e-prints*, 707, July 2007.

发表文章目录

1. Liu, C., Wang, D. , Liu, B. , Gao, D. , Cui, C. and Zhao, Y., An astronomical data mining application framework for virtual observatory, Advanced Software and Control for Astronomy. Edited by Lewis, Hilton; Bridger, Alan. Proceedings of the SPIE, Vol 6274, pp. 627415, 2006

2. 刘超, 田海俊, 高丹, 杨阳, 路勇, 崔辰州, 赵永恒, 异地异构天文数据资源的统一访问, 天文研究与技术in press, 2008

3. Liu, C. , Hu, J. , Newberg, H. and Zhao, Y., Candidate Milky Way Satellites in the Galactic Halo, A& A, 477, 139, 2008

4. 路勇, 刘超, 崔辰州, 赵永恒, VO-DAS Registry系统的设计与实现, 天文研究与技术, 4(4), 355-359, 2007

5. Wang, D. , Zhang, Y. , Liu, C. and Zhao, Y., Kernel Regression For Determining Photometric Redshifts From Sloan Broadband Photometry, MNRAS 382, 1601, 2007

6. Wang, D. , Zhang, Y. , Liu, C. and Zhao, Y., Two novel approaches for photometric redshift estimation based on SDSS and 2MASS databases, ChJAA in press, astro-ph/0707.2250, 2007

致谢

2004年7月的一个下午，我和赵永恒研究员的第一次会面促成了可能是我一生中最重要的决定：进入国家天文台攻读博士学位。来年的早春，在经历了近6年的社会历练之后，我重又回到了宁静的象牙塔内。每天的忙忙碌碌让三年的光阴如梭一样逝去，直到动手开始写这篇论文，过去的日子才又象电影一样，重新回放在头脑中。回顾三年的历程，我想我首先应该向给了我最大帮助三位老师：我的导师赵永恒研究员和胡景耀研究员，以及我所在的虚拟天文台课题组负责人崔辰州副研究员表达我发自肺腑的感激。

我由衷地感谢赵永恒研究员。赵老师在我心中始终是一位宽厚的长者，儒雅而博学，永远从容不迫。尽管他每天都十分繁忙，但仍然会抓住每个机会给我指导和建议。作为完全没有天文背景的“一张白纸”，我从赵老师那里得到了非常宝贵的宽容和引导，在完全没有感觉到壁垒的情况下，他让我逐渐融入到了这个令我兴奋不已的神奇领域里。每每遇到难题，和赵老师讨论以后总是能够迎刃而解。每到关键节点，他往往能够用几句话点破天机，令我视野豁然开朗。更为重要的是，我从赵老师那里得到了宽松的环境，让我可以根据自己的兴趣自由倘佯在从技术到科学的广阔空间里。

我也由衷地感激胡景耀研究员。胡老师在我心中是一位神通的智者，渊博的学识，敏锐的头脑，风趣的语言是他的标志。胡老师的主意总是层出不穷，让我佩服之致。和胡老师的讨论总是很惬意，因为那气氛总是包含着严谨的态度、洞察的眼光和灵光一现的兴奋。多么深邃的问题，让胡老师娓娓到来，似乎一下子简单了很多。胡老师不仅教会了我很多的基础知识，还让我领略了快乐的研究也体验了研究的快乐。是胡老师的言传身教，让我逐渐从技术的广厦窥见到科学的殿堂。

我还要感谢崔辰州副研究员。作为课题负责人，他总是那么兢兢业业，勤奋工作。在课题的选定和具体工作中他给了我无数的帮助和建议。他给了我很多发挥的空间，让我能够在虚拟天文台的研发上获得充分的自由，让我深为感动。

在天文台的三年学习生活中，还有很多老师和同学都给过我无私地指导和热情地帮助。

感谢Heidi Newberg博士，她在将近一年的时间里通过邮件和我做了深入的讨论，不仅让我了解了银河系结构的研究的最新进展，和我共享了很多她的宝贵经验，而且让我学会了逻辑严密、一丝不苟的治学方法。我相信正是这种治学方式，使得她在不长的时间里取得了卓越的科学成就。我也相信，依靠从她身上学到的这样的治学方法，我也可以在将来的研究中扎扎实实，一步一个脚印的走下去。

感谢邓李才研究员，他对我的工作给了很多中肯的建议和帮助。这些建议让我对自己的工作树立了信心。

感谢天津大学于策博士和熊科浪同学为我的工作提供的大力帮助，慷慨地让我使用高性能计算资源，并协助修改并行计算代码，使我得以顺利完成了并行计算实验。

感谢周旭研究员、姜碧沩教授、陈玉琴研究员、张昊彤副研究员、张彦霞副研究员、徐岩博士、高爽同学，让我从他们那里学到很多的东西。他们也给了我很多有益的建议。

感谢赵刚研究员、邹振隆研究员、王钢研究员、李宗伟教授、褚耀泉教授、张华伟博士、梁燕春副研究员、陆烨副研究员、施建荣副研究员、和他们的交流也许非常简单，但是总是让人感到愉快，并且受益匪浅。

感谢LAMOST实验室的全体老师和同学，特别是罗阿理副研究员、陈英老师、袁晖老师、王丹博士、王伟博士、王建岭博士、陈建军、何勃亮、田海俊、杨阳、杨帆、高丹、薛元、王凤飞、宋轶晗、孙士卫、罗宇、吴悦、贾磊、邹思成、王淑青。还要特别提到已经毕业离开LAMOST的吴潮博士和尹红星博士。

感谢天文台杜红荣、艾华、朱爱萍、田斌等老师的热心帮助。

最后，真诚感谢我的母亲默默地支持，真诚感谢我的妻子全力地支持，她们不仅替我承担了家庭很多的责任，最让我感动的是她们在我当初作出决定以后的三年里，始终不渝地给予我鼓励和信任。每当我凝视计算机屏幕前跃动的“虚拟”星空，心中充满无限安宁的时候，我知道那个推动我的世界转动不息的第一动力就是——她们。