第49卷第3期
2008年7月

天 文 学 报
ACTA ASTRONOMICA SINICA

Vol. 49 No. 3
Jul. ,2008

中国虚拟天文台交叉证认工具的开发和应用**

高 丹† 张彦霞 赵永恒

(中国科学院国家天文台 北京 100012)

摘要 随着空间和地面巡天项目的发展,天文数据呈指数增长,天文学已发展到全波段的天文学时代.为了应对形势的需要,虚拟天文台应运而生.为了使天文学家做起科学来更方便快捷,虚拟天文台项目开发了各种实用工具.在数据融合方面,中国虚拟天文台项目开发了交叉证认工具.该工具主要实现了两个服务:一个是服务器端两星表交叉证认;另一个是用户上传星表与服务器端星表交叉证认.前者两星表直接进行交叉证认,后者则先将本地数据自动入库再进行交叉证认.程序还实现了对交叉证认结果的分类和参数的自由选择等功能,以及与可视化工具 VOPlot 的集成.该工具以客户端/服务器端模式的 web 网页形式发布.此工具为多波段数据融合提供了便利,是对两个大星表交叉证认工作的预研究.在以后的工作中,将不断地更新和完善该工具,并将在此基础之上进一步开发统计分析和数据挖掘等工具.

关键词 方法:数据分析,方法:统计,天文数据库,星表,巡天

中图分类号:P 114;**文献标识码**:A

1 引言

随着科学技术的发展,天文学步入了数据丰富的时代,数据以 TB、甚至 PB 量级计量.天文数据覆盖了从伽玛射线、X 射线、紫外、光学、红外到射电等波段,天文学发展成为全波段天文学.各个波段的数据是高度相关的.若要获得对天体或天文现象更深入全面的认识和理解,需要在多波段的高维参数空间内进行探索和研究.融合的数据通常能促进新的天体、现象或规律的发现.例如:可见光波段与射电波段数据的融合发现了类星体.而来自不同项目、波段和时间的各种数据的共同属性只有位置,因而若想融合各个波段的数据,位置交叉证认是关键.通过证认可以对天体的物理性质、演化规律获得更全面系统的认识,加深对证认源的新的天文理解,为统计分析、数据挖掘做准备.而研究天体在多维参数空间中的分布,提取的天体的信息越多,越有利于天体的分类.另外,通过多波段交叉证认,增加了发现新天体的概率.天体多波段交叉证认是 LAMOST 项目(the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST)的科学目标的三大核心课题之一[1],也是要求数据融合的虚拟天文台项目的底层技术之一[2-3].正是基

* 2007-06-26 收到原稿,2008-01-07 收到修改稿

* 国家自然科学基金(10473013,90412016,10778724)资助项目

† gaodan2008@gmail.com

于这些因素,我们更加需要开发切实可行的、方便友好的交叉证认工具,一方面解决天文学家由于无法方便地获得大量多波段数据而难以进行多波段数据研究的课题;另一方面为中国虚拟天文台项目探讨从底层的数据到融合数据再到数据的可视化和挖掘提供技术支持.

2　国外研究现状

随着计算机、互联网技术的发展,海量的天文数据有了比较好的归档,并提供了互联网服务. 世界上已有多家天文数据中心和虚拟天文台项目在天文数据存储、查询和分析等方面做了大量的工作,并开发了一系列实用工具. 其中不少的工具可以实现交叉证认服务. 目前,国外常用的交叉证认工具有:VizieR、Simbad、Aladin、MAST、ESO、NED、OpenSkyQuery、TOPCAT 等.

VizieR 现已收集了 5000 多个星表[4],且数据还在不断更新. 在对每个星表都提供了查询服务的同时,还提供了小样本多源查询(即小样本交叉证认)服务. 但这些服务都只针对单个星表查询,其交叉证认也只是小样本,而对格式要求相对严格. 证认结果只能给出 VizieR 星表的参数,而不能提供融合两个星表参数的结果,需要用户进行二次加工.

Simbad 的小样本多源查询(即小样本交叉证认)功能类似 VizieR[5],主要提供点源的证认.

Aladin 是法国斯特拉斯堡数据中心开发的数据整合工具[6]. 它可以互动地可视化天空任何一部分图像,并可以与天文星表或用户上传文件叠加,在同一视场中还可叠加来自 SIMBAD、NED、VizieR 或其他星表的已知源. Aladin 尤其在天体的多波段交叉证认、观测准备和新数据集的质量控制等方面显示出优越性.

NVO (MAST&VizieR)可以交叉证认 VizieR 的星表和 MAST 的巡天任务[7].

CDS/ESO 目前正在开发大数据量星表交叉证认的工具[8],它可依各个参量交叉证认,而不仅限于位置参数. 使用此工具,用户可进行单源查询、多源查询,而证认结果亦可与其他星表再次交叉证认. 用户可以对所有星表同时查询,亦可选择对主相关星表或自己感兴趣的星表同时查询.

NED 提供了批处理查询功能[9],但用户需通过邮件使用此服务.

OpenSkyQuery 是美国虚拟天文台开发的交叉证认工具[10]. 用户可以通过强大的查询语言来实现星表的交叉证认;可以自己选择交叉的星表,也可以上传文件与选定的星表交叉证认. 其服务页面提供了各种数据库、交叉证认算法的模版、用户指导等. 用户只需读说明套模版选择数据源就可以交叉证认了. 由于网络和带宽等原因,该系统有查询的行数不能超过 5000 行的限制.

TOPCAT 是一个可以互动地将数据表图形化的浏览器和编辑器[11]. 可以处理天文数据的主要格式如 FITS 和 VOTable,其他的格式可以不断增加. 提供了各种方法浏览和分析数据表,包括浏览核心数据、表的原始信息和列的元数据、以及画图工具、统计计算、不同星表匹配算法. 用强大的可扩展的 Java 语言为基础的表达式建立新列,也可以选择行的子集用以独立分析. 表中的数据和元数据可以编辑,修改后的内容可以以各种形式输出.

然而，目前国外上述大数据中心提供的多源查询(即小样本交叉证认)或交叉证认服务都有其局限性，包括数据资源局限和功能局限，远远无法满足天文学家的需要.

综上所述，对目前存在的服务的优点总结如下[12]：

(1) 用户界面友好、可选择功能较为齐全；

(2) 支持的输出格式较多(如文本、网页、VOTable 等)，更新速度快.

对其缺点可总结如下：

(1) 用户只能选择一个星表与特定星表交叉证认；

(2) 对交叉证认结果没有分类；

(3) 只能是小文本和某一数据库的交叉证认；

(4) 不能直接得到融合两个星表参数的结果，需用户二次加工.

3 交叉证认

交叉证认是指将不同星表或数据库中的源按位置属性联系起来，如果存在相同位置或在一定误差半径范围内的源，就被证认为同一天体. 例如：有两个源分别在星表 a、b 中，它们的误差半径分别为 r1、r2，它们的之间的球面角距离为 d. 如果它们的关系满足 $d \leqslant 3(r1^2+r2^2)^{0.5}$，则认为它们可能是同一天体.

由于来自不同项目或仪器的数据的观测精度、信噪比与位置误差等的差异，使得交叉证认的结果可能有以下几种不同情况：一对一、一对多、多对一、一对无、无对一等(见图 1). 对于一对一的源，我们基本上认为该对应体就是同一个源；而对于一对多、多对一的源，就需要进一步运用概率的方法来确定. 目前很多工作是取角距离最近的源为对应体或取最亮的源为对应体，另外星表中的其他信息也可以利用. 例如在 USNO - A2.0 和 USNO - B1.0 中给出 B 星等和 R 星等，而 B - R 就是一个很好的参数用以区分最终的对应体. 得到对应体后，计算其概率，如果概率低，可以进行统计分析；如果概率高，则可以进行数据挖掘工作. 而对于一对无或无对一的情况，需要天文学家结合专业知识给出进一步的解释，也许能发现一些稀有的天体或天文现象.

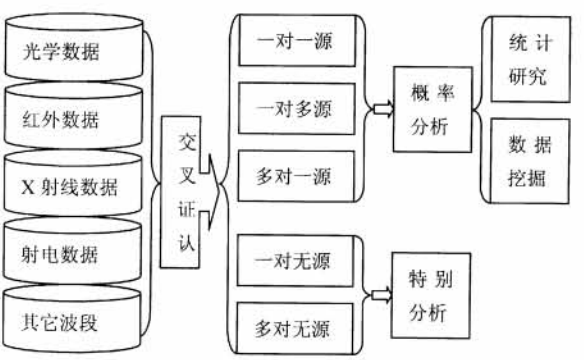

图 1 多波段数据分析流程图

Fig. 1 The scheme of multi-wavelength data analysis

4　交叉证认工具

4.1　功能和服务

在借鉴国外的交叉证认工具的优缺点的基础上,我们开发了中国虚拟天文台的交叉证认工具,其功能和服务描述如下:

(1) 实现大数据量星表交叉证认

解决了一直以来存在的大数据量星表本地、异地交叉证认困难的问题. 大数据量有助于排除由于样本小带来的泊松误差的影响,促进有趣天体或现象的发现的概率. 比如某些概率是百万分之一或一亿分之一量级的未知或稀有天体或天文现象,则有可能在大样本中发现.

(2) 用户可以任意选择两个星表交叉证认

解决了数据资源局限性的问题,实现了本地和异地的数据的交叉证认. 如果用户想交叉证认服务器端的两个星表,那么证认可以在服务器端进行,而把最终结果传给用户. 若用户要使用自己的星表与服务器的星表进行交叉证认,只需上传自己星表的元数据(ReadMe)文件和数据文件,在服务器端利用自动入库模块就可以实现星表的自动入库功能,从而可以方便地交叉证认.

(3) 用户可自由选择输出参数

为了节省存储空间,减少传输量,无须将两个星表的所有参数都传给用户,而只要把用户需要的参数输出即可.

(4) 用户可对交叉证认的结果选择

可以任意选取一对一、一对多、一对无、无对一的情况或这几种结果的任意组合.

(5) 证认结果无需二次加工

直接给出用户需要的结果,每一行包括用户选择的两个星表参数和对应源的角距离的多波段星表.

(6) 用户可以选择交叉证认的两个星表的误差半径

对于所有源误差半径都一样的星表,可选择误差半径为一常数;而对于星表中每个源有其相应的误差半径,并用星表的一列数据来表示的情况,可以选择星表相应的列名为误差半径.

(7) 输出格式的多样化

支持 VOTable、ASCII、CSV、HTML 等格式.

(8) 与 VOPlot 可视化工具的集成[13]

VOPlot 是印度虚拟天文台开发的 Java 程序,可以画不同的天文图(散点图、直方图、带误差棒的图),数据格式为 VOTable. 可以对任何星表作图,而且图可以以 eps 格式保存. 与交叉证认工具集成后,用户可以对交叉证认后的数据进行加工,例如增加列或进行一些简单的运算生成新列. 用户也可以通过 VOPlot 来可视化数据,如选择或用计算参数来做散点图、直方图等.

4.2　服务界面

下面以 X 射线波段的 ROSAT 星表和光学波段的 TYCHO 星表做交叉证认为例,说

明我们开发的交叉证认工具的功能. ROSAT 星表的误差半径选择了 PosErr 列，TYCHO 星表的误差半径选择了 5 角秒. 本工具主要实现了两个服务：一个是服务器端两星表交叉证认服务，如图 2 界面中上部的表格所示；另一个是用户上传星表与服务器端星表交叉证认服务，如图 2 界面中下部的表格所示. 用户要使用某服务只需填写相应的表格提交即可. 图 3 展示的是此工具的交叉证认结果的分类和参数的自由选择等功能，用户选择需要的参数和交叉证认的分类结果即可. 对该工具，我们提供的服务网页是：http://badc.lamost.org:8080/xmatch/ .

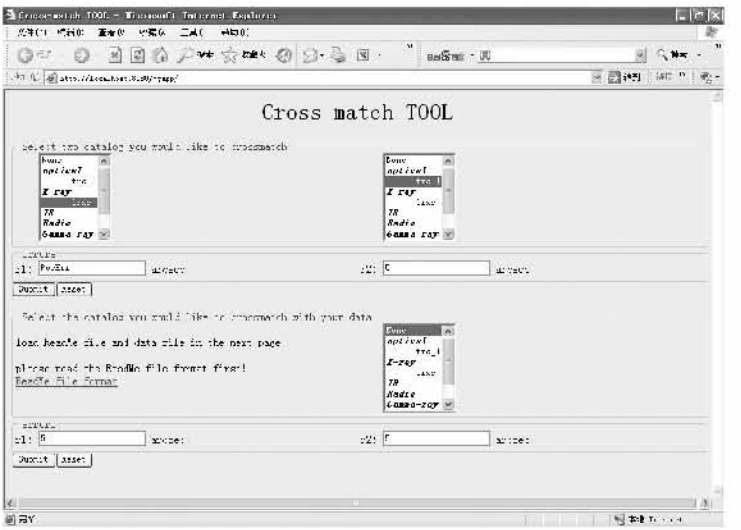

图 2 交叉证认工具主页面

Fig. 2 The main interface of cross-match tool

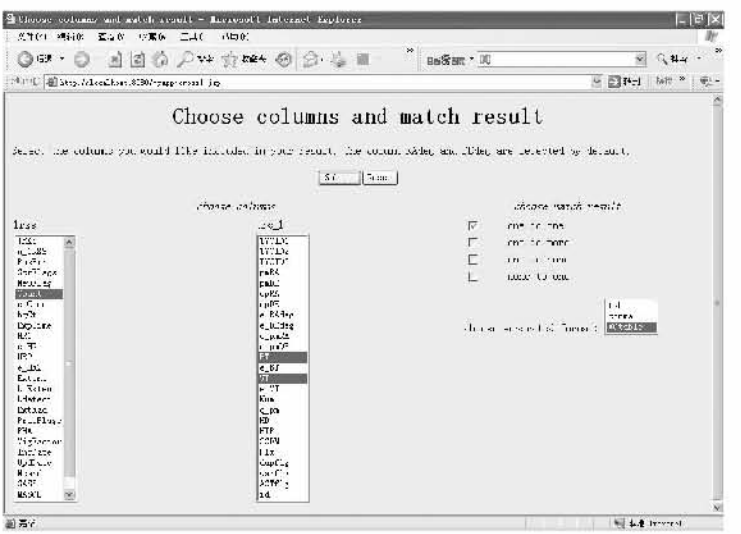

图 3 选择参数和交叉证认结果页面

Fig. 3 The interface of "choose columns and match result"

4.3　交叉证认工具程序流程图

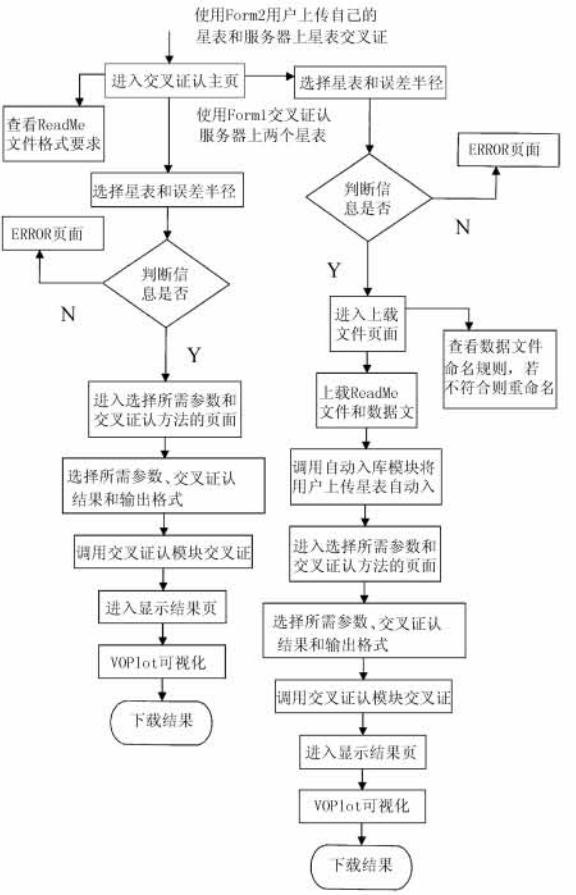

图 4　交叉证认工具程序流程图

Fig. 4　The scheme of cross-match tool

如图 4,进入交叉证认主页面后(图 2),有两个服务对应的两条流程. 一是交叉证认服务器上两个星表:用户选择星表和误差半径,提交表格;工具就会判断填写信息是否完整,若完整,则进入选择参数和交叉证认结果的页面(图 3),用户选择后,工具就调用交叉证认模块进行证认;最后,VOPlot 输出可视化结果(图 6)及下载结果(图 5). 第二个服务是用户上传自己的星表与服务器上的星表交叉证认,其与第一个服务的区别是中间多了一个上载文件页面和自动入库模块.

这个交叉证认工具的核心是交叉证认模块和自动入库模块,它们实现了主要的自动入库和交叉证认功能,这两个模块的具体细节将在本文的下一部分详细说明. 模块使用 JavaBean 组件来实现,页面则使用 JSP 来实现,而后端数据库用的是 MySQL,这样就实现了程序和页面的分离.

图 3 页面提交后下载的结果是一个. dat 文件,如图 5 所示. 图 3 中参数选择了第一个表 1rxs 的 Count 和第二个表 trc_1 的 BT、VT,而交叉证认结果选择了一对一. 此文件共有 8 列,从左至右分别是表 1rxs 的赤经、赤纬、表 1rxs 的 Count 参数、表 trc_1 的赤经、

赤纬、表 trc_1 的 BT、VT 参数、角距离 d，其中角距离 d 的单位是角秒. 图 6 是用 VOPlot 工具可视化交叉证认的结果.

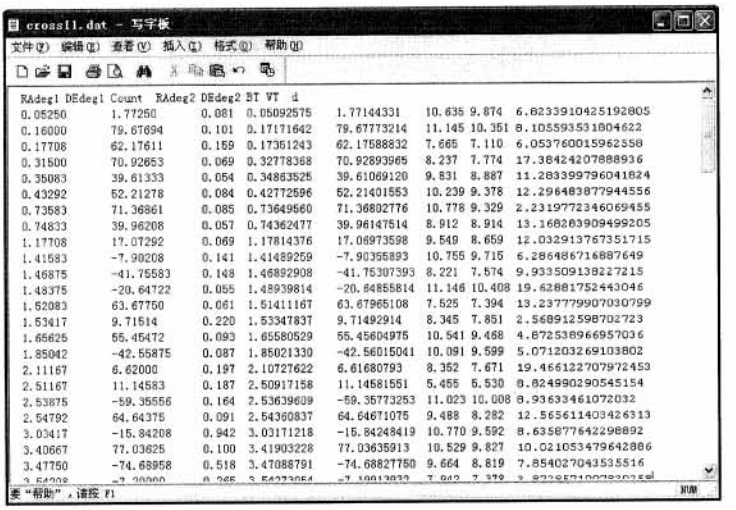

图 5 一对一的结果

Fig. 5 The result of one to one

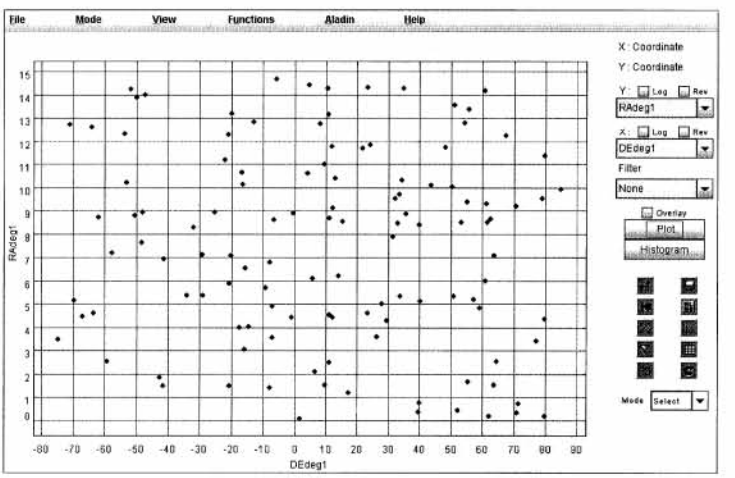

图 6 VOPlot 可视化结果

Fig. 6 Visualize the association with VOPlot

5 具体实现的步骤和原理

交叉证认工具主要包括两个重要的模块：自动入库模块和交叉证认模块. 另外，我们交叉证认的星表数据统一采用数据库的形式，并做了一些预处理工作，使交叉证认工具

可以对统一处理的数据库星表进行统一的操作. 下面分别具体介绍预处理工作和两个模块.

5.1　星表数据库预处理

随着数据量的增长,我们需要更好地管理数据和更方便的数据获取及处理服务. Jim Gray[14]对数据库格式和文本格式做了比较,很明显数据库在性能和便利上都优于文本. 我们的星表数据采用数据库的形式统一管理. 星表均由自动入库工具(交叉证认工具中的自动入库模块的用户界面版本)根据星表 ReadMe 文件自动入库,并自动进行数据库星表预处理,而交叉证认亦建立在统一管理的数据库星表之上.

天文数据主要包括星表、星图、光谱、文献资料等,其中星表是包含天体信息的数据表格,是天文学家最常用到的天文数据. 星表是经过观测、处理、整理和归档的数据. 星表数据来自于不同的观测项目、观测目标、观测仪器、观测波段、观测时间等,数据格式也各不相同,导致了星表数据的复杂性. 星表的制作者都按照一定标准制作了星表 ReadMe 文件,ReadMe 中以标准格式提供星表的有关信息,并提供数据文件中每一列的位置、名称、单位等描述. 用户从数据中心下载下来的星表数据是一个文件包,其中包括 ReadMe 文件、数据表文件、其他文件,都是文本格式.

我们入库的主要数据来源于斯特拉斯堡天文数据中心(CDS)负责收集整理的天文星表[15],这些星表包括恒星、星系以及其他银河系外的天体的数据,同时也收录了太阳系内天体数据和物理数据. CDS 的 VizieR 服务目前收录的 5000 多个星表通过 ASCII 码文本或 FITS 格式的文件发布,这些文件可以通过 FTP 下载.

以下是数据库星表预处理的步骤和目的[16−17]:

(1) 统一星表赤经、赤纬标识为度的形式:RAdeg、DEdeg,以便交叉证认工具识别和计算. 如果星表赤经、赤纬是度的形式,则统一标识为 RAdeg、DEdeg;如果星表赤经、赤纬是时分秒、度角分的形式,则统一标识为 RAJ2000、DEJ2000 或 RAB1950、DEB1950,并转换成度的形式. 在表中加入统一标识为 RAdeg、DEdeg 的两列.

(2) 按 DEdeg 排序,如果用赤经还需要考虑 0 度、360 度的特殊情况,若是赤纬就不用考虑. 排序可以提高查询计算性能.

(3) 加一列主键 id(int 逐一递增整数),以标识每一行数据(每一个源),用以先取部分参数交叉证认再调其他需要的参数,目的是节约内存.

(4) 建立多值索引(DEdeg、RAdeg),提高查询计算性能.

(5) DEdeg、RAdeg 均为非空(not null),因为数据库中索引为 not null 有利于提高性能.

5.2　自动入库模块

自动入库模块不仅实现了根据星表 ReadMe 文件自动入库,也实现了数据库星表自动预处理. 输入接口只有数据文件和星表 ReadMe 文件的路径. 如图 5,只要用户上传这两个文件,模块就会自动读取星表 ReadMe 文件并将数据表自动录入数据库,完全屏蔽掉了星表 ReadMe 文件及数据表本身的细节. 参照页面上提供的星表 ReadMe 文件最简易格式的要求,用户亦可自己编写或修改星表 ReadMe 文件. 这样已经进行了预处理的入库的星表就可直接被交叉证认模块所使用.

自动入库模块接口：

public void setFilepath(String filepath){...} 将数据文件路径输入模块

public void setReadMefilepath(String ReadMefilepath) {...} 将星表 ReadMe 文件路径输入模块

public String getMessage(){...} 实现自动入库功能，返回状态值到页面

(状态值共有五种，IO error：文件 IO 错误；Database error：数据库错误；ReadMe file Error：ReadMe 文件不符合格式要求；data file Error：上传入库的数据文件不是该 ReadMe 文件中描述的文件；putting data：自动入库成功).

5.3 交叉证认模块

天文星表的数据量大，且交叉证认的计算复杂度为 $O(N^2)$，这些交叉证认工作的技术难点带来了以下三个问题. 首先，内存不够用，为了使工作能够进行，必须设法节约内存，为此我们采用先取部分参数交叉证认，再根据 id 把其他需要参数取出的办法；其次，耗时长，必须设法减少计算量，提高速度，因此我们先确定两星表赤纬覆盖相同天区范围，再画框缩小范围，最后判断对应体，这样可以尽量早地做排除以优化交叉证认速度；最后，数据库查询计算性能低，尤其对于大数据量的查询计算，因此我们对大星表按赤纬分区交叉证认，把任务分解，前面做的数据库星表预处理工作也是为了提高数据库性能.

交叉证认模块算法解决方案：

(1) 以位置精度高的表为中心；

(2) 确定两星表赤纬覆盖相同天区范围：

DEC≤min(max(DEC_A),max(DEC_B)),DEC≥max(min(DEC_A),min(DEC_B)).

(3) 对大星表按赤纬分区交叉证认；

(4) 先取两星表赤经、赤纬、误差半径、id 参数交叉证认；

(5) 根据 id 把其他需要参数取出；

(6) 画框 :(在框的范围内才可能是对应体)

$$|RA\_A-RA\_B|<\frac{|r_1|+|r_2|}{\cos((DEC\_A+DEC\_B)/2)}$$

$$|DEC\_A-DEC\_B|<|r_1|+|r_2|$$

(7) 判断对应体 : $d^2<9(|r_1^2|+|r_2^2|)$

(8) 对交叉证认的结果分类(一对一、一对多、一对无、无对一).

交叉证认模块接口：

public void setArgs(String args) {...} 用一个字符串将交叉证认所需信息输入模块

(字符串中信息包括：表名、误差半径、所选参数、所选交叉证认结果、结果存储路径等).

public String getMessage() {...} 实现交叉证认功能，返回状态值到页面

(状态值共有三种，IO error：文件 IO 错误；Database error：数据库错误；success：交

叉证认成功).

6　总结和展望

面对海量数据,传统方式的天文研究显然已经不能应对形势的需要,天文学家必须借助于各种工具如自动入库工具、自动交叉证认工具、统计分析工具和数据挖掘工具等来进行科学研究,从而快速便捷地提高科学产出. 为此,我们实现了星表入库的自动化,并自动对星表进行了一些预处理,天文学家无须学习数据库知识即可创建自己的星表数据库. 在星表数据库基础上,我们开发了交叉证认工具,实现了两个服务:服务器端两星表交叉证认和用户自己上传星表与服务器端星表交叉证认,并提供了对交叉证认结果的分类和参数的自由选择等功能. 我们会对该工具不断地更新和完善,为提高工具的效率,改进数据索引方法如 HTM 和 kd - tree[18,19];为满足异地异构数据的访问,逐步将该工具集成在网格中;为提高工具的柔韧性,进一步革新浏览页面,优化其功能,丰富交叉证认结果的输出格式,扩展自动入库的文件格式,对交叉证认算法做进一步改进,以及对交叉证认结果的概率分析;为增加工具的可扩展性,提供与可视化工具、统计分析工具和数据挖掘工具的接口,为下一步的统计分析、数据挖掘和可视化做准备. 而且我们会针对天文学家的需求,持续地更新和完善该工具,使其真正成为天文学家的有力助手.

参 考 文 献

1　[LAMOST 的科学目标] http://www.lamost.org/xoops/modules/wfchannel/index.php? pagenum=3

2　Szalay A, Gray J. Science, 2001, 293: 203

3　赵永恒. 科学, 2002, 54(2): 13

4　[VizieR] http://vizier.u-strasbg.fr

5　[Simbad] http://simbad.u-strasbg.fr

6　[Aladin] http://aladin.u-strasbg.fr

7　[VIZIER Search] http://archive.stsci.edu/vizier.php

8　Ortiz P F, Ochsenbein F, Wicenec A et al. ESO/CDS Data-mining Tool Development Project [C]. In: ASP Conf. Ser. (Vol. 172). 1999. 379 - 382

9　[NED Batch Jobs] http://nedwww.ipac.caltech.edu/help/batch.html

10　[OpenSkyQuery] http://openskyquery.net

11　[TOPCAT] http://www.star.bris.ac.uk/~mbt/topcat/

12　高丹, 张彦霞, 赵永恒. 天文研究与技术, 2005, 3: 186 - 193

13　[VOPlot] http://vo.iucaa.ernet.in/~voi/voplot.htm

14　Jim Gray, Alex Szalay. Where the Rubber Meets the Sky: Bridging the Gap between Databases and Science. MSR - TR - 2004 - 110

15　[CDS] Centre de Données astronomiques de Strasbourg. http://cdsweb.u-strasbg.fr

16　Abel D, Devereux D, Power R, Lamb P. An O(NlogM) Algorithm for Catalogue Matching. CSIRO ICT Centre Technical Report TR - 04/1846

17　Power R, Devereux D. Benchmarking Catalogue Cross Matching. TR - 04/1848, CSIRO ICT Centre, Canberra, Australia, 2004.

18 [HTM] http://skyserver.org/HTM/
19 [kd-tree] http://kdtree.quickseek.com/

The Development and Application of the Cross-match Tool of China-VO

GAO Dan ZHANG Yan-xia ZHAO Yong-heng
(*National Astronomical Observatories, Chinese Academy of Sciences, Beijing* 100012)

ABSTRACT With the deployment and developments of various space-based and ground-based sky survey projects, astronomical data increase in the rate of the exponential growth form, and astronomy has paced into a full electromagnetic waveband era. Facing so huge amounts of data, how to save, organize, analysis and mine data is an important issue for astronomers. Under these situations, the international virtual observatory projects are being carried out. Up to now, they have developed many practical tools, such as VOPlot by India VO, OpenSkyQuery by NVO and VOSpec by ESA-VO, VOSED by Spain VO. As a member of International Virtual Observatory Alliance (IVOA), China-VO project has developed a cross-match tool, which is realized by web service. The tool can handle two catalogues from the same server, and also deal with two catalogues of which one is from the server and the other is from users. For the former one, users can execute cross-matching directly, and for the latter, users need to upload the local data into the server database by the automated software of database creation and then perform the cross-matching task. Meanwhile, the tool provides different kinds of cross-matching results and the selection of parameters. According to their own requirement, users can conveniently choose the cross-matching result and parameters. This tool is mainly used for multi-wavelength data integration, and the pre-research of cross-identification of two huge catalogues. In the near future, we will continuously enhance and update the tool, further develop statistical analysis tools and data mining tools on the basis of this work.

Key words Methods: data analysis - Methods statistical, Astronomical data bases: Miscellaneous - Catalogues - Surveys