天文资源无缝融合关键技术研究

樊东卫

指导老师：　赵永恒(研究员)　崔辰州(研究员)
中国科学院国家天文台

Co-Supervisor: Prof.Alexander Szalay, Dr.Tamás Budavári
Johns Hopkins University

目录

- 选题背景简介
- 结合天区覆盖图的星表交叉证认算法与缺失源检测及数据库实现
- 基于直线非对称几何模型的射电星表交叉证认方法
- 资源统一管理平台
- 其他工作
- 总结及未来工作设想

简介

- 天文资源多种多样
  - 天文设备、分析软件、星表、光谱、数值模拟数据等等
  - 天文开启全波段观测的同时也已经进入到了数据大爆炸时代。
- LSST每月将产生500TB的天文图像
- SDSS DR9星表包含了超过12亿天体的信息
- TMT、SKA等更大的设备正在建造中……

简介

- 如何在多个星表及其它形式数据中找到同一个天体的数据，在大数据时代将越来越困难
- 交叉证认作为一种实用的数据融合方法已经被广泛使用
- 本论文主要对Zones Algorithm进行了改进，并提出了一种新的射电星表交叉证认方法

简介

- 随着天文数据的增多，相应的天文工具、技术、服务也越来越多
- 技术门槛增高，学习成本增加

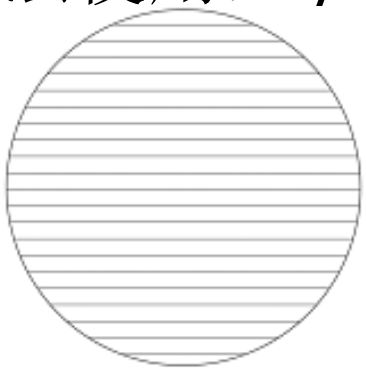
- 本论文试图建立一个资源统一管理平台
- 将众多天文数据、服务连接起来，减少使用复杂度

结合天区覆盖图的星表交叉证认算法与缺失源检测及数据库实现

- 条带算法——Zones Algorithm
- 条带算法+天区覆盖图
- 数据库实现及测试
- 缺失源检测

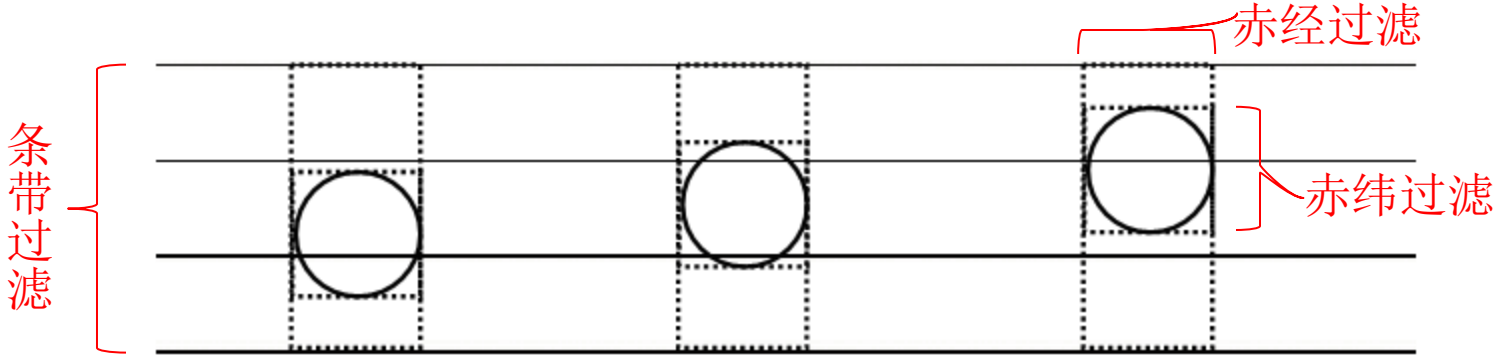
Zones Algorithm

- Zones Algorithm是最快的交叉证认算法之一
- 将整个天球划分成数个环形条带——Zones
  - 每个条带及条带内的天体均有条带编号ZoneID
  - ZoneID= $\left\lfloor \frac{\delta+90^\circ}{h} \right\rfloor$，$\delta$是赤纬，h是条带高度
  - Zones Algorithm使用(ZoneID, R.A.)对星表进行物理索引，配合算法使用，I/O效率极高



Zones Algorithm核心过滤方法

- 条带过滤:相互匹配的天体只可能出现在邻近的条带内，$ZoneID_a - \left\lfloor \frac{\theta}{h} \right\rfloor \leq ZoneID_b \leq ZoneID_a + \left\lfloor \frac{\theta}{h} \right\rfloor$

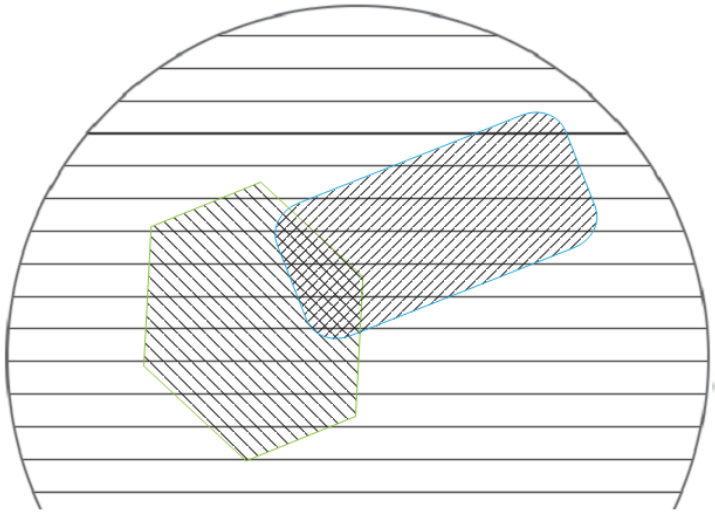


- 赤纬过滤:$\delta_a - \theta \leq \delta_b \leq \delta_a + \theta$，$\theta$为匹配边界
- 赤经过滤:$\alpha_a - Alpha_\delta \leq \alpha_b \leq \alpha_a + Alpha_\delta$

$$\mathtt{Alpha} = \left| \arctan \left(\frac{\sin\theta}{\sqrt{\cos(\delta-\theta)\cos(\delta+\theta)}} \right) \right|$$

Zones Algorithm可以更快

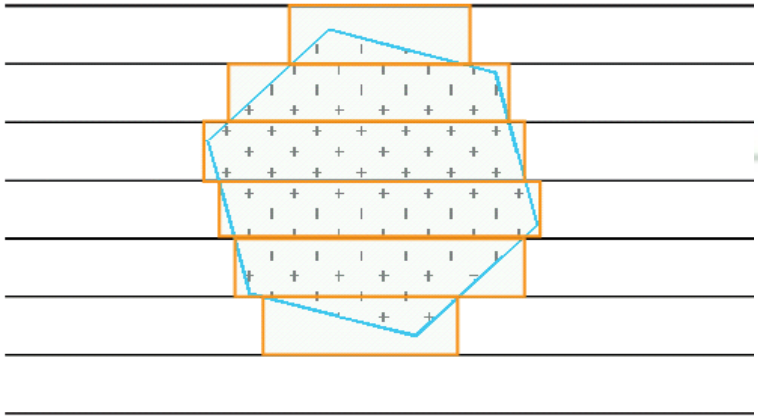
- 当两星表（A、B）重叠区域很小时，Zones Algorithm仍然要遍历几乎整个星表A



- 将星表的天区覆盖图带入交叉证认过程可减少对数据的无效遍历

在Zones Algorithm中使用星表覆盖图

- 使用条带片段来模拟星表覆盖图
  - 数据结构简单(ZoneID, RAmin, RAmax)， 易保存到数据表。且与Zones Algorithm数据索引相同
  - 条带高度h很小，如7.1”，近似程度会很好



- 直接用条带片段快速计算天区覆盖图交集

数据库实现

- Step 1: 定义条带
- Step 2: 使用条带片段模拟星表天区覆盖图
- Step 3: 计算星表天区覆盖图交集
- Step 4: 创建星表索引表
- Step 5: 创建邻近条带对比表
- Step 6: 在两个星表的交集区域做交叉证认

与Zones Algorithm大体相同

增加的步骤

与Zones Algorithm完全相同

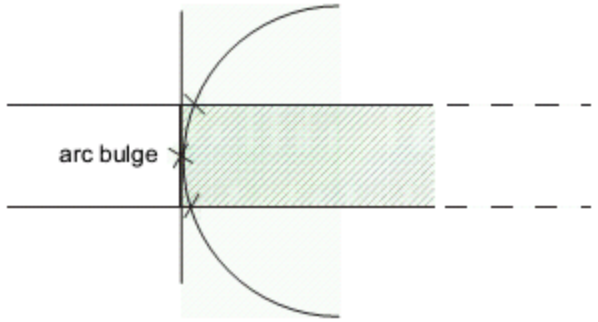
Step 1: 定义条带

- 根据给定条带高度h划分天球，将划分情况保存到数据表ZoneDef
- 计算每一个条带的Alpha值
- 使用Spherical Library定义每个条带几何信息

```
CREATE TABLE [dbo].[ZoneDef](
    [ZoneId] [int] NULL,
    [DecMin] [float] NOT NULL,
    [DecMax] [float] NOT NULL,
    [Alpha] [float] NOT NULL,
    [RegionBinary] [varbinary](600) NULL,
    [RegionBinary1] [varbinary](940) NULL,--0-180
    [RegionBinary2] [varbinary](940) NULL --180-360
)
```

Step 2:使用条带模拟星表天区覆盖图

- 取得星表天区覆盖图（Footprint）——用Spherical Library描述
- 计算每个条带与Footprint的交集
- 取得形状的各个交叉点，确定条带片段左右边界
  - 要注意边界凸出的情形
  - 合并可合并的相邻片段



Step 3: 使用条带片段计算星表天区覆盖图交集

- 假设在同一个条带上有两个星表的条带片段
- 只须一一对比各个条带片段，若有交集，取最大左边界及最小右边界
- 要注意处理Zones Algorithm中的环绕问题，即当RAmax+Alpha>360时，新增一个(RAmin-360, RAmax-360)片段到结果中

```
dbo.FootprintIntersection(
        ZoneId int,
        RAmin float,
        RAmax float,
        PRIMARY KEY(ZoneId, RAmin, RAmax) )
```

Step 4:创建星表索引表

- 与Zones Algorithm一致
- 使用聚集索引 (ZoneID, R.A.)，令索引表中的天体信息也按ZoneID、赤经的顺序在硬盘上物理存放，提高数据存取效率

Step 5:邻近条带对比表ZoneZone

- 数据库不知道哪些条带需要扫描，将扫描全表
- 邻近条带对比表ZoneZone用于提示数据库当前需要搜索哪些条带
  - Zones Algorithm中直接扫描星表本身来创建ZoneZone表，速度较慢（约1分01秒），获得的表也较大
  - 已知两个星表的天区交集，可以通过这个交集来获得一个只保存有效信息的较小的表，且速度更快（3秒）

```
ZoneZone(
        zone1 int, zone2 int, alpha float,
        primary key(zone1, zone2))
```

Step 6:在两个星表的交集区域做交叉证认

Zones Algorithm

```
SELECT c1.ObjID AS ObjID1, c2.ObjID AS ObjID2, ...
FROM Catalog1 AS c1
    INNER LOOP JOIN ZoneZone AS zz                          a
        ON zz.ZoneID1 = c1.ZoneID
    INNER LOOP JOIN Catalog2 AS c2
        ON zz.ZoneID2 = c2.ZoneID
        AND c2.RA BETWEEN c1.RA - zz.Alpha2                 b
                      AND c1.RA + zz.Alpha2
        AND c2.Dec BETWEEN c1.Dec - @theta                  c
                       AND c1.Dec + @theta
        AND ( c1.RA >= 0 OR c2.RA >= 0 )                    d
WHERE (c1.Cx-c2.Cx) * (c1.Cx-c2.Cx)                         e
    + (c1.Cy-c2.Cy) * (c1.Cy-c2.Cy)
    + (c1.Cz-c2.Cz) * (c1.Cz-c2.Cz) < @dist2
```

加入天区覆盖图

```
SELECT c1.ObjID AS ObjID1, c2.ObjID AS ObjID2, ...
FROM ZoneIntersect AS i                                     f
    INNER JOIN Catalog1 AS c1
        ON c1.ZoneID = i.ZoneID AND
           c1.RA BETWEEN i.RAmin AND i.RAmax
    INNER LOOP JOIN ZoneZone AS zz
        ON zz.ZoneID1 = c1.ZoneID
    INNER LOOP JOIN Catalog2 AS c2
        ON zz.ZoneID2 = c2.ZoneID
        AND c2.RA BETWEEN c1.RA - zz.Alpha2
                      AND c1.RA + zz.Alpha2
        AND c2.Dec BETWEEN c1.Dec - @theta
                       AND c1.Dec + @theta
        AND ( c1.RA >= 0 OR c2.RA >= 0 )
WHERE (c1.Cx-c2.Cx) * (c1.Cx-c2.Cx)
    + (c1.Cy-c2.Cy) * (c1.Cy-c2.Cy)
    + (c1.Cz-c2.Cz) * (c1.Cz-c2.Cz) < @dist2
```

a:条带粗过滤;b:赤经过滤;c:赤纬精确过滤;d:避免重复匹配;e:距离精确检查
f:新增部分，只在两星表的重叠区域内查找天体

测试

- 使用SDSS DR6 (约2.3亿天体) 及 GALEX GR3 AIS (约5.5千万天体) 主星表
- 在thumper.pha.jhu.edu上进行测试
  - Intel Xeon E5430 2.66GHz，24GB RAM
  - Windows Server 2008， Microsoft SQL Sever 2005开发版
- 两个星表的重叠区域约为3689平方度。
- 获得了约20%的速度提升。
- 通过RA或DEC裁减减小重叠区域，交叉证认的时间消耗也减小了

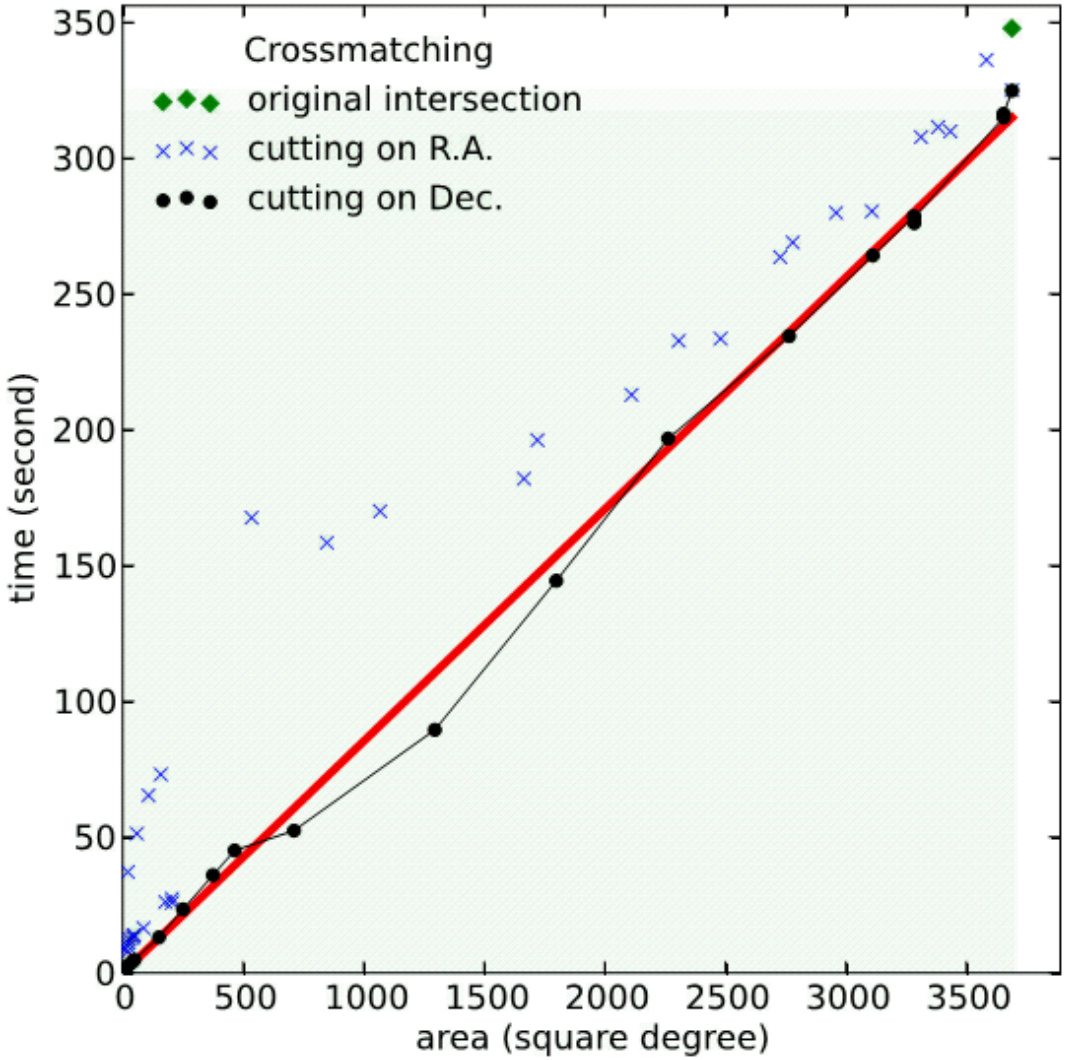


Fig. 8.— The wall-clock time of the crossmatching scales linearly with the overlapping area. Cutting on the celestial coordinates corresponds to the worst and best case, hence we expect a typical result to appear somewhere between these in real-life situations. The red line is plotted just to guide the eye.

缺失源检测

- 已知两个星表的天区交集
  - 可知O：星表A在重叠区域中的天体
- 又已知M：通过交叉证认，星表A被星表B匹配的天体。O>M
- 可得A中未被匹配天体A-M=(O-M)+(A-O)
  - O-M：在B星表的观测范围，但未能被B匹配
  - A-O：没在B星表的观测范围，所以未能被匹配
- O-M称为星表B的缺失源（dropout）

代码及测试

```
SELECT c.ObjID
FROM ZoneIntersect AS o
JOIN Catalog1 AS c ON o.ZoneID = c.ZoneID
    AND c.RA BETWEEN o.RAmin AND o.RAmax
EXCEPT
SELECT ObjID1 FROM MatchedObjects
```

位于重叠区域中的天体

已被匹配天体

- 仅需29秒可探知：约有420万个GALEX天体在SDSS的观测范围内却未被SDSS观测到

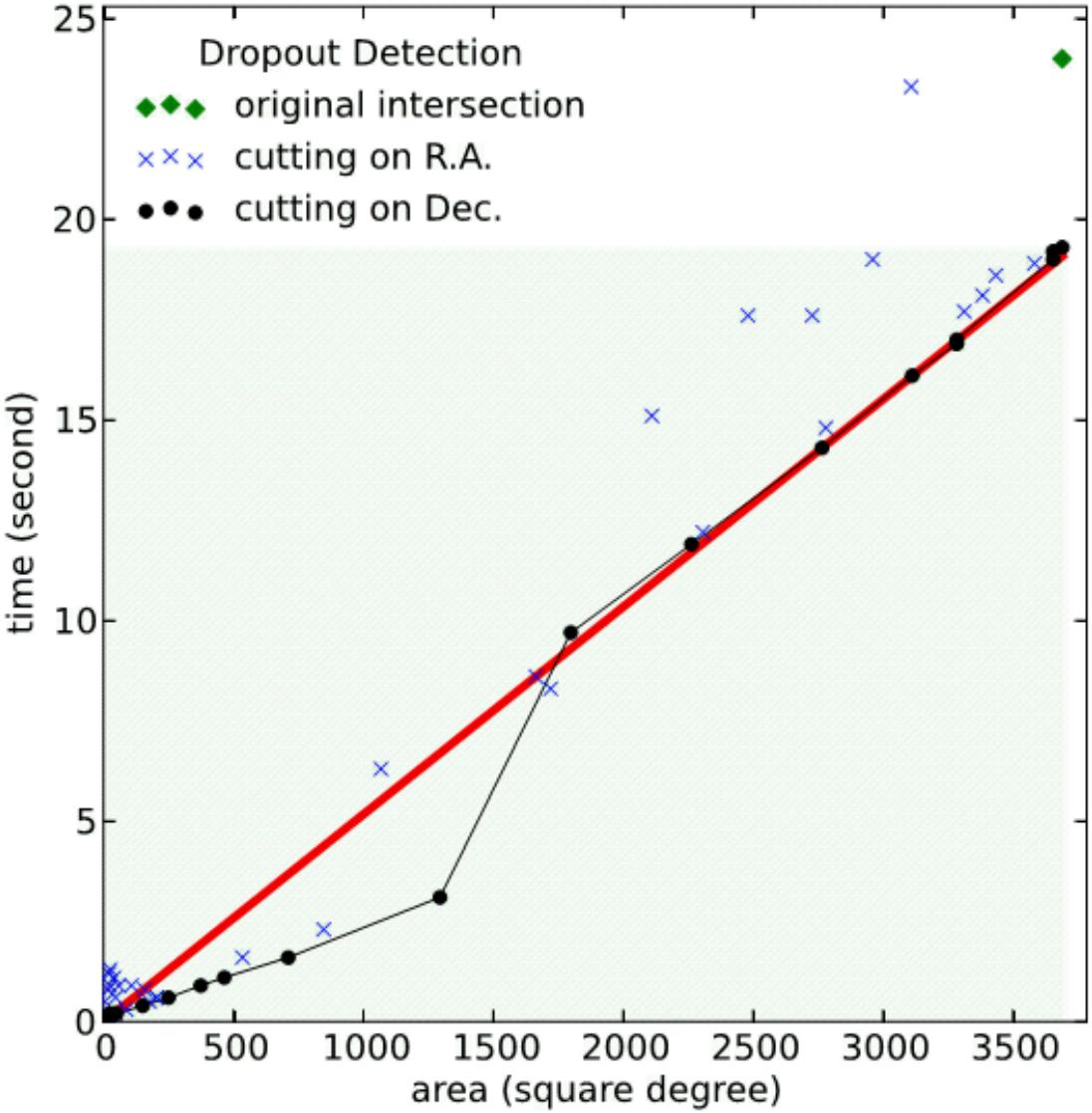


Fig. 9.— Dropout detection can be efficiently performed after crossmatching the catalogs. The time is only a fraction of the matching and it also scales linearly with the overlapping area. The worst and best cases are plotted along with a trend line that guides the eye.

小结

- Zones Algorithm非常高效，但没有利用星表天区覆盖图信息
- 通常交叉证认方法要应用天区覆盖信息，需使用与交叉证认不同的另外的一套数据索引方法
- 我们将天区覆盖信息直接加入到了交叉证认过程之中
  - 两个星表重叠区域越小速度越快
  - 自动化程度较好
  - 非常轻易地检测缺失源
- 此方法将应用到Open SkyQuery的新引擎中，也是China-VO未来类似服务的技术储备

基于直线非对称几何模型的射电星表交叉证认方法

- 贝叶斯因子
- 直线模型
- 积分计算方法及简化计算
- 实验结果分析

贝叶斯因子

- 用p(**x**|**m**,M) 表示一个准确位置为$\boldsymbol{m}$的天体在位置$\boldsymbol{x}$被观测到的概率 $\int d^3x \cdot p(\boldsymbol{x}|\boldsymbol{m}, M) = 1$
- 对于单个观测目标x_1，应用贝叶斯定理取得其验后概率密度，即其准确位置m为x_1的概率为

$$p(\boldsymbol{m}|\boldsymbol{x_1}, M) = \frac{p(\boldsymbol{x_1}|\boldsymbol{m},M)p(\boldsymbol{m}|M)}{p(\boldsymbol{x_1}|M)}\text{其中}p(\boldsymbol{m}|M) = \frac{1}{4\pi}\delta(|\boldsymbol{m}| - 1)$$

- 由全概率公式可得， x_1在天球上的验前概率

$$p(\boldsymbol{x_1}|M) = \int d^3m \cdot p(\boldsymbol{m}|M)p(\boldsymbol{x_1}|\boldsymbol{m}, M)$$

贝叶斯因子

- 给定一个假设H，所有的数据都来自同一天体m
- 相应的对立假设K，所有的观测数据都来自不同天体，且它们都不源自m
- 坐标集合$D = \{x_1, x_2, \ldots, x_n\}$对应n次不同观测，定义贝叶斯因子为$B(H, K|D) = \frac{P(H|D)/P(H)}{P(K|D)/P(K)}$
- 应用贝叶斯定理，得$B(H, K|D) = \frac{P(D|H)}{P(D|K)}$
  - 贝叶斯因子即两个假设的似然函数的比值

似然函数的参数化模型

- H假设所有观测对象都是同一目标，因而可以用一个共同位置**m**来对它进行参数化

$$p(D|H) = \int d^3m \cdot p(\boldsymbol{m}|H) \prod_{i=1}^{n} p_i(\boldsymbol{x_i}|\boldsymbol{m}, H)$$

- 对立假设K将被不同的位置$\{m_i\}$参数化，其概率为各次观测的概率密度函数的积分的乘积

$$p(D|K) = \prod_{i=1}^{n} \left[\int d^3m_i \cdot p(\boldsymbol{m_i}|K) p_i(\boldsymbol{x_i}|\boldsymbol{m_i}, K) \right]$$

- $$\begin{cases} B(H,K|D) \gg 1, H\text{成立} \\ B(H,K|D) < 1, K\text{成立} \\ otherwise, \text{需要进一步检验} \end{cases}$$

将赤道坐标转换为天球切平面坐标

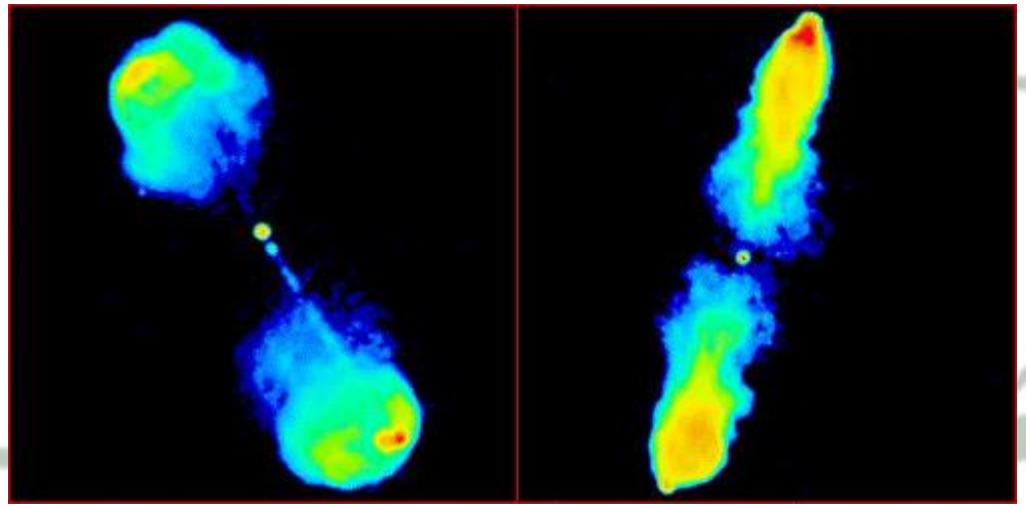
- 3维降为2维，简化几何计算及多重积分计算
  - 在光学源$x_0(\alpha,\beta)$处作天球切平面
  - 令x_0在切平面上的坐标为(0,0)
  - 其他相关坐标也可投影到这个切平面上，如点$r'(\alpha,\beta)$在x_0切平面上的坐标(p, q)的计算公式为

$$p(\alpha',\delta') = \begin{pmatrix} \cos\delta'\cos\alpha' - \cos\delta\cos\alpha \\ \cos\delta'\sin\alpha' - \cos\delta\sin\alpha \\ \sin\delta' - \sin\delta \end{pmatrix} \cdot \begin{pmatrix} -\sin\delta\cos\alpha \\ -\sin\delta\sin\alpha \\ \cos\delta \end{pmatrix}$$

$$q(\alpha',\delta') = \begin{pmatrix} \cos\delta'\cos\alpha' - \cos\delta\cos\alpha \\ \cos\delta'\sin\alpha' - \cos\delta\sin\alpha \\ \sin\delta' - \sin\delta \end{pmatrix} \cdot \begin{pmatrix} \sin\alpha \\ -\cos\alpha \\ 0 \end{pmatrix}$$

直线对称模型

- 寻找带有喷流结构的射电源
  - 其特征是可能有一个中心射电源+两侧喷流瓣
- 使用直线对称模型来模拟这一结构
  - 光学源为$\boldsymbol{m_0}$，中心点(core)矢量为$\boldsymbol{m}$，一侧瓣(lobe)矢量为$\boldsymbol{m'}$，则另一侧瓣为$\boldsymbol{m'' = 2m - m'}$



基于直线对称模型的假设计算

- 假设有四个射电源$D = \{y_0, y_1, y_2, y_3\}$，它们其中一个(CORE,LOBE,LOBE,NONE)假设的似然函数计算公式为

$$
\begin{aligned}
& p(\text{CORE}, \text{LOBE}, \text{LOBE}, \text{NONE}) \\
= & \left\{ \int d^2 m_0 \cdot p(\boldsymbol{m}_0) L_{x0}(\boldsymbol{m}_0) L_{y0}(\boldsymbol{m}_0) \right. \\
& \left. \int d^2 m_1 \cdot p(\boldsymbol{m}_1 | \boldsymbol{m}_0) L_{y1}(\boldsymbol{m}_1) L_{y2}(2\boldsymbol{m}_0 - \boldsymbol{m}_1) \right\} \\
& \left\{ \int d^2 m_2 \cdot p(\boldsymbol{m}_2) L_{y3}(\boldsymbol{m}_2) \right\}
\end{aligned}
$$

- 前面的(core, lobe, lobe)作为一个整体，而none表示y_3与它们相独立。按前述对立假设K的似然函数计算方法。两者的似然函数须相乘

基于直线对称模型的假设计算

- 四个射电源$D = \{y_0, y_1, y_2, y_3\}$还可以构成数种假设：

  (core, lobe, none, none)

  (none, lobe, lobe, none)

  (core, none, none, none)

  等等，且$D = \{y_0, y_1, y_2, y_3\}$各点的位置可以互换，以遍历各种组合情形

- 需要将这些组合一一计算出来，对比结果

概率密度函数的选择

- 对于最复杂的带有(core, lobe, lobe)的情形
- L_{x0}、L_{y0}、L_{y1}、L_{y2}、L_{y3}等概率密度函数选择二维正态分布

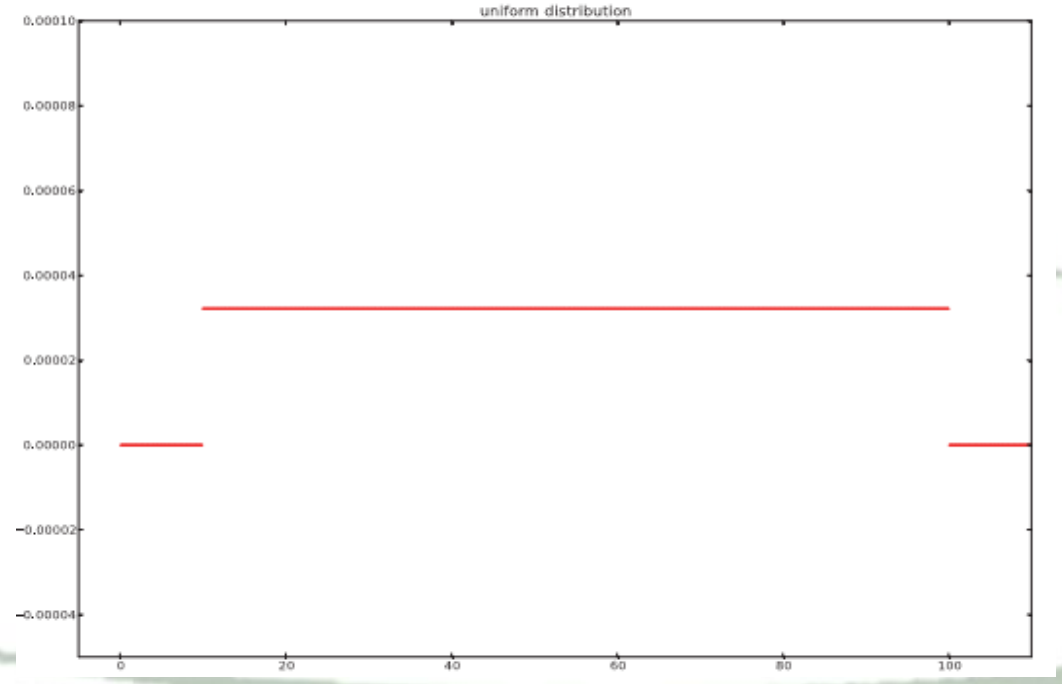
$$L_{x0}(\boldsymbol{m_0}) = g(\boldsymbol{x_0}|\boldsymbol{m_0}, \Sigma_{x0})$$

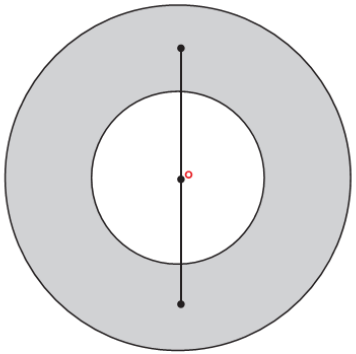
- 而验前概率密度函数$p(m_i) = \frac{1}{4\pi r^2} = 1.87 \times 10^{-12}$，即假设每平方角秒上有一个天体

概率密度函数的选择

- lobe与core的条件概率密度函数，可用均匀分布

$$p(\boldsymbol{m_1}|\boldsymbol{m_0}) = \begin{cases} [(R^2 - r^2)]^{-1}, r < |\boldsymbol{m_1} - \boldsymbol{m_0}| < R \\ 0 \qquad\qquad ,\text{其它} \end{cases}$$





Lobe与Core的条件概率密度函数的其他选择

- 瑞利分布

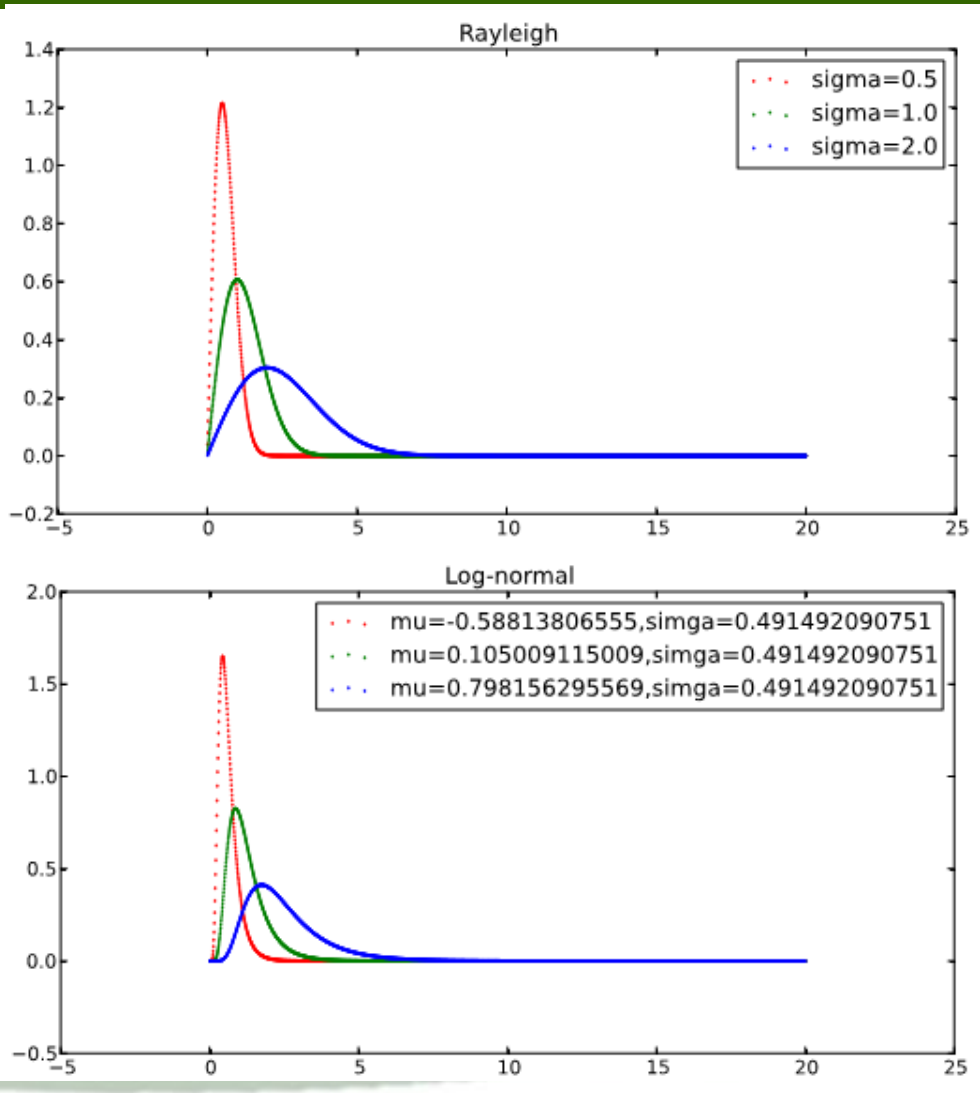
$$p(k|\sigma) = \frac{k}{\sigma^2} e^{-k^2/2\sigma^2}$$

$$mean = \sigma\frac{\pi}{2}, var = \frac{4-\pi}{2}\sigma^2$$

- 对数正态分布

$$\ln N(\mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\ln x - \mu)^2/2\sigma^2}$$

$$\begin{aligned} mean &= e^{\mu+\sigma^2} \\ var &= \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2} \\ \mu &= \ln mean - \frac{\sigma^2}{2} \\ \sigma &= \sqrt{\ln\left(\frac{var}{mean^2} + 1\right)} \end{aligned}$$



直线非对称模型

- 两侧瓣的观测位置相对于中心不一定对称
- 引入一个k因子，给予一侧瓣在直线上一定的活动范围

$$2m_0 - m_1 + k(m_0 - m_1) = (2 + k)m_0 - (1 + k)m_1$$

直线非对称模型

- 这样(core, lobe, lobe, none)的似然函数的计算公式变为

$$
\begin{aligned}
& p(\text{CORE, LOBE, LOBE, NONE}) \qquad (5.37)\\
= & \left\{ \int d^2 m_0 \cdot p(\boldsymbol{m}_0) L_{x0}(\boldsymbol{m}_0) L_{y0}(\boldsymbol{m}_0) \right.\\
& \int dk \cdot p(k) \\
& \left. \int d^2 m_1 \cdot p(\boldsymbol{m}_1|\boldsymbol{m}_0) L_{y1}(\boldsymbol{m}_1)\, L_{y2}\left[(2+k)\boldsymbol{m}_0 - (1+k)\boldsymbol{m}_1\right] \right\}\\
& \left\{ \int d^2 m_2 \cdot p(\boldsymbol{m}_2) L_{y3}(\boldsymbol{m}_2) \right\}\\
= & \left\{ \int d^2 m_0 \cdot p(\boldsymbol{m}_0) g\left(\boldsymbol{x}_0|\boldsymbol{m}_0, \Sigma_{x0}\right) g\left(\boldsymbol{y}_0|\boldsymbol{m}_0, \Sigma_{y0}\right) \right.\\
& \int dk \cdot p(k) \\
& \left. \int d^2 m_1 \cdot p(\boldsymbol{m}_1|\boldsymbol{m}_0) g\left(\boldsymbol{y}_1|\boldsymbol{m}_1, \Sigma_{y1}\right) g\left(\boldsymbol{y}_2|(2+k)\boldsymbol{m}_0 - (1+k)\boldsymbol{m}_1, \Sigma_{y2}\right) \right\}\\
& \left\{ \int d^2 m_2 \cdot p(\boldsymbol{m}_2) g\left(\boldsymbol{y}_3|\boldsymbol{m}_2, \Sigma_{y3}\right) \right\}
\end{aligned}
$$

p(k)是一维正态分布概率密度函数

重点抽样积分方法

- 大量使用了各类概率密度函数及积分
- 积分计算的时间消耗很大
- 选择使用蒙特卡罗积分方法中的一种降方差方法——重点抽样方法(Importance Sampling)
  - 通过生成特定分布的随机数来快速计算积分值
  - 主要是一、二维正态分布随机数

简化计算 1：多维正态分布概率密度函数乘积

- 多维正态概率密度函数可表示为

$$
\begin{aligned}
g(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= (2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= (2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}[-(\boldsymbol{x}-\boldsymbol{\mu})]^T\boldsymbol{\Sigma}^{-1}[-(\boldsymbol{x}-\boldsymbol{\mu})]\right\} \\
&= g(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\Sigma})
\end{aligned}
$$

x和μ可互换,变成在x而非未知的μ附近取随机点

- 两个多维正态分布概率密度函数的乘积

$$
\begin{aligned}
g_p(\boldsymbol{\mu},\boldsymbol{\Sigma}) &= g(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\cdot g(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) \\
\text{其中}\boldsymbol{\Sigma} &= \left(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}\right)^{-1} \\
\boldsymbol{\mu} &= \left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right)
\end{aligned}
$$

- 但其乘积不再是一个标准化的概率密度函数，其积分为C。需要将结果除以此常数C

$$
C=(2\pi)^{-\frac{d}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}\,|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}\,|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}\,exp\left[\frac{1}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}-\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2\boldsymbol{\mu}_2\right)\right]
$$

简化计算 2：直接计算最简单的贝叶斯因子

- 全none假设的似然函数是一个常数，可以让其它假设都除以这个常数，得到其相对于全none假设的贝叶斯因子

$$p(CORE, LOBE, LOBE, NONE) = \frac{1}{a}\{x_0y_0y_1y_2\} \cdot \frac{1}{a}\{y_3\} = \frac{1}{a^2}\{x_0y_0y_1y_2\}$$

$$p(CORE, LOBE, NONE, NONE) = \frac{1}{a}\{x_0y_0y_1\} \cdot \frac{1}{a}\{y_2\} \cdot \frac{1}{a}\{y_3\} = \frac{1}{a^3}\{x_0y_0y_1\}$$

$$p(NONE, LOBE, LOBE, NONE) = \frac{1}{a}\{x_0y_1y_2\} \cdot \frac{1}{a}\{y_0\} \cdot \frac{1}{a}\{y_3\} = \frac{1}{a^3}\{x_0y_1y_2\}$$

$$p(NONE, LOBE, NONE, NONE) = \frac{1}{a}\{x_0y_1\} \cdot \frac{1}{a}\{y_0\} \cdot \frac{1}{a}\{y_2\} \cdot \frac{1}{a}\{y_3\} = \frac{1}{a^4}\{x_0y_1\}$$

$$p(CORE, NONE, NONE, NONE) = \frac{1}{a}\{x_0y_0\} \cdot \frac{1}{a}\{y_1\} \cdot \frac{1}{a}\{y_2\} \cdot \frac{1}{a}\{y_3\} = \frac{1}{a^4}\{x_0y_0\} = \frac{1}{a^4} \cdot C_{x0y0}$$

$$p(NONE, NONE, NONE, NONE) = \frac{1}{a}\{x_0\} \cdot \frac{1}{a}\{y_0\} \cdot \frac{1}{a}\{y_1\} \cdot \frac{1}{a}\{y_2\} \cdot \frac{1}{a}\{y_3\} = \frac{1}{a^5}$$

简化后的公式

- 经过简化后的(core, lobe, lobe, none)假设与(none, none, none, none)假设的贝叶斯因子计算公式为

$$
\begin{aligned}
& B(CORE, LOBE, LOBE) \\
= \; & a^3 \int d^2m \cdot p(\boldsymbol{m}_0) \cdot g_p\left(\boldsymbol{m}_0 | \boldsymbol{x}_0, \boldsymbol{y}_0, \Sigma_{x0}, \Sigma_{y0}\right) \\
& \int dk \cdot p(k) \\
& \frac{1}{(k+1)^2} \int d^2m_1 \cdot p(\boldsymbol{m}_1 | \boldsymbol{m}_0) \cdot g_p\left(\boldsymbol{m}_1 | \boldsymbol{y}_1, \frac{k+2}{k+1}\boldsymbol{m}_0 - \frac{y_2}{k+1}, \Sigma_{y1}, \frac{\Sigma_{y2}}{(k+1)^2}\right)
\end{aligned}
\tag{5.67}
$$

在$\boldsymbol{y_1}$附近取随机点

- 由于$a = 1.87 \times 10^{12}$，数值太大，一般取结果的对数$\log_{10}$值。这样对比结果的时候不再需要用除法，而改用减法。

数据分析

表 5.2: ATLAS射电天文学家手工交叉证认结果的统计信息

类型	数量	百分比
CORE,LOBE,LOBE	10	1.647%
LOBE,LOBE	22	3.624%
CORE,LOBE	5	0.8237%
double	27	4.448%
CORE	559	92.092%
complex	11	1.812%

double是
(core, lobe) + (lobe, lobe)

- 数据来源： SWIRE 3 CDF-S vs. ATLAS
- (core, lobe, lobe)型占的比率很低，是低概率事件；(lobe, lobe)稍高；(core, lobe)或称core-jet的比率也很低；core占了绝大多数；还有部分复杂情形，ATLAS天文学家没有确定。

数据分析算法

- 人为地给较好的core设定一个标准，比如设为10，表明它是一个很好的core。这样的core不能作为任何假设的lobe。从计算结果中删除掉这些将它作为lobe的假设
- 首先取最好的(core, lobe, lobe)型，然后将涉及到它们其中任一成员的计算结果删除
- 再选最好的(lobe, lobe)及(core, lobe)，同样将涉及到它们其中任一成员的计算删除
- 剩下的结果中，寻找最好的(core)型配对

结果对比

- 经过大量积分计算及结果之后，对比程序得到的结果与澳大利亚射电天文学家手工交叉证认结果

表 5.3: 计算结果与ATLAS射电天文学家手工交叉证认结果对比

类型	计算结果	匹配数	遗漏数
triple	14	9	1
double	53	17	10
core	550	520	39

小结

- 使用直线非对称模型对光学星表与可能带有喷流射电源的射电星表的进行交叉证认的一种基于贝叶斯假设推断方法
- 此方法可应用到未来的大规模射电波段巡天（如SKA）观测结果与光学观测结果的数据融合中来
- 此方法还存在一些问题，比如直线非对称模型不能适应喷流瓣夹角较大的情形，需要进一步修正模型。

资源统一管理平台

- 自然科学基金（编号U1231108）支持项目

- 简介
- 目标
- 架构与技术选择
- 主要功能插件介绍

简介

- FITS等格式是天文界的行业规范，应用领域窄，通用软件不支持
- 专业软件不注重管理
- 软件、服务很多，使用门槛高
- 数据过多，不易查找

目标

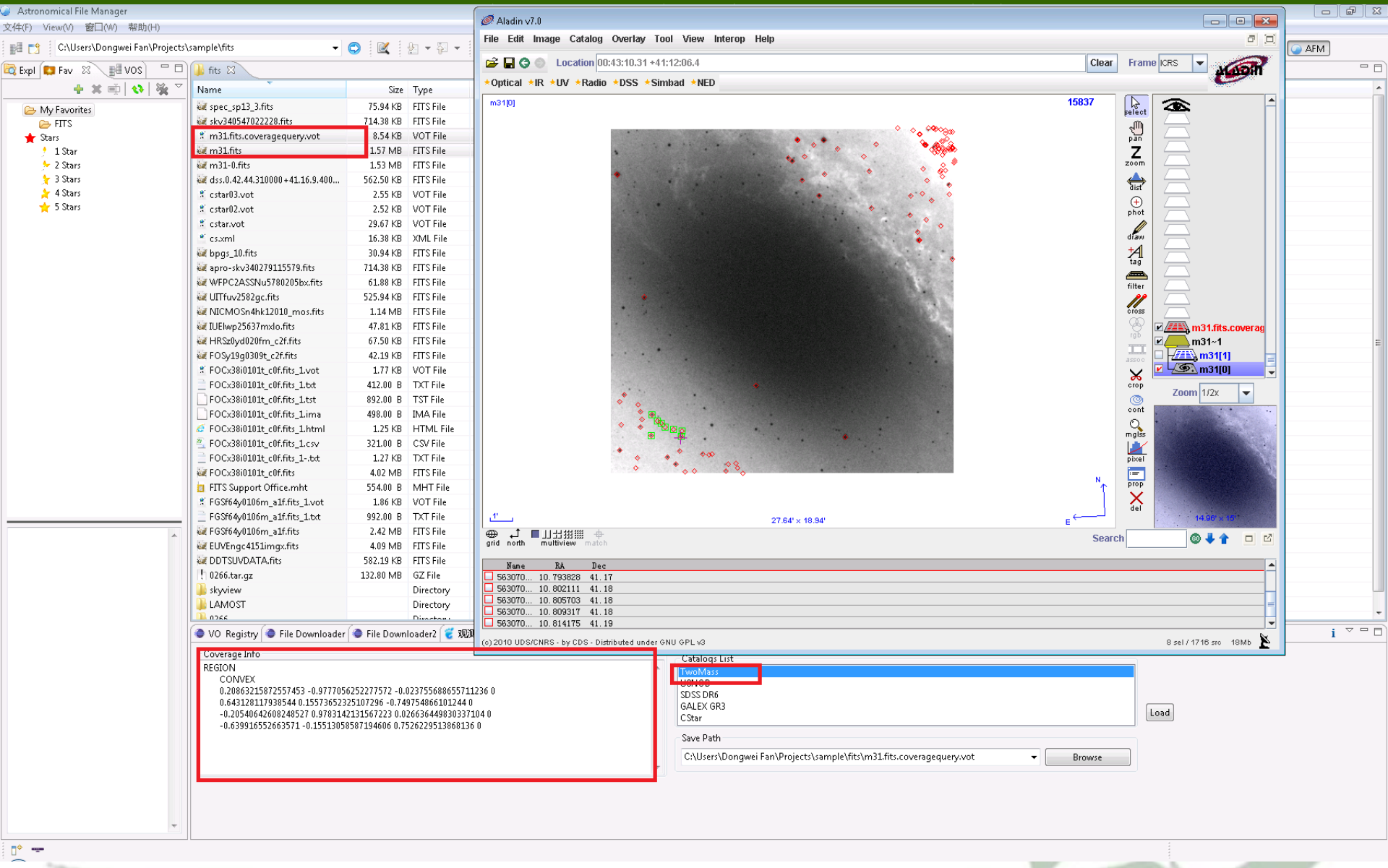
- 天文资源管理器的关注点在于如何方便天文学家获取数据、访问服务
- 不直接处理数据，而是将数据转交给专门的数据处理工具
- 重点在于管理与整合本地、远程数据及服务
- 简化一些工具的使用流程，使天文学家无须花费精力了解过于细节的技术问题

架构与技术选择

- Eclipse RCP富客户端平台
  - Java跨平台，天文界有多种Java函数库可用
  - 插件化，所有功能通过插件完成，插件间通过接口沟通数据
  - 可添加、去除插件，建立个性化工作平台
  - 面向接口编程，低耦合，高内聚。适合小团队开发
  - SWT/JFace比AWT/Swing更优异的使用体验
  - 完善的帮助系统、插件安装及更新系统、软件多语言本地化支持
- 使用纯Java跨平台嵌入式关系数据库Derby

主要功能插件

- 文件列表插件
  - 核心插件，定义了多个接口及数据格式，是整个软件的中心
- 天区覆盖图插件——与Tamas Budavari合作
  - 将FITS图像文件天区覆盖信息上传到Footprint Service
- 覆盖区域查找插件
  - 在星表中查找FITS图像文件所覆盖区域内的天体


Astronomical File Manager
C:\Users\Dongwei Fan\Projects\sample\fits
Name
Size
Type
m31.fits.coveragequery.vot
8.54 KB
VOT File
m31.fits
1.57 MB
FITS File
Aladin v7.0
File Edit Image Catalog Overlay Tool View Interop Help
Location 00:43:10.31 +41:12:06.4
Clear
Frame ICRS
Optical IR UV Radio DSS Simbad NED
m31[0]
15837
27.64' x 18.94'
Search
Coverage Info
REGION
CONVEX
0.20863215872557453 -0.9777056252277572 -0.023755688655711236 0
0.6431281179385 44 0.15573652325107296 -0.749754866101244 0
-0.2054064260824 8527 0.9783142131567223 0.026636449830337104 0
-0.639916552663571 -0.15513058587194606 0.7526229513868136 0
Catalogs List
TwoMass
SDSS DR6
GALEX GR3
CStar
Load
Save Path
C:\Users\Dongwei Fan\Projects\sample\fits\m31.fits.coveragequery.vot
Browse

主要功能插件

- VOSpace插件——与Dmitry合作
  - 支持基于REST技术的虚拟天文台“云存储”
- 服务检索插件
  - 在软件内直接检索Simbad、ADS、arXiv、SkyMouse、DataScope、天文名词等等服务
- 资源管理器及收藏夹插件
  - 与大多数文件管理工具类似的树状文件展示及文件收藏功能

主要功能插件

- 应用程序通信插件——SAMP
- 内嵌专业工具
  - 内嵌ds9、fv、Aladin、Topcat、VOSpec等专业天文软件。其他软件也可通过类似方式加入
- FITS头查看插件
  - 树状列表展示FITS头信息
- 文件检索插件
  - 通过文件名或FITS头信息查找文件

主要功能插件

- 天文每日一图插件
  - NASA“Astronomy Picture of the Day”展示及桌面壁纸设置等
- 天文文件批量下载插件
  - 通过给定的坐标列表等信息到指定服务网址批量下载文件

小结

- 专注于对天文数据的管理而不是处理
- 通过多种方式整合各个本地应用程序以及网络服务
- 跨平台插件化体系，有利于调动社区力量
- 可以集合专门领域的插件，组织成一个性化的个人工作平台
- 有望作为天文领域云的客户端

其他工作

- Spherical Toolkit C++版本
- 国家天文台在线直播系统
  - 天文讲座视频亦是非常重要的天文资源，有效地存档有助于天文工作者更好的学习、工作。
  - 提出一整套的低成本、易携带的软硬件方案
  - 硬件：使用家用DV + USB采集卡 + 笔记本电脑 + 网络服务器
  - 软件：Windows Media Encoder + Media Service + SQL Server + IIS + 基于ASP.net的在线管理系统

其他工作

- 基于微软World Wide Telescope、Google Earth、Google Map (Sky)的LAMOST Footprint试验
- 参与微软WWT在中国的多项事务，创建、维护WWT在全球第一个社区：WWT北京社区
- 中国天文数据中心在线可视化工具的调研、应用试验
- 升级维护SkyMouse、FitHAS等China-VO成果

总结

- 研究了天文资源无缝融合的几个关键技术，包括星表交叉证认技术、天文资源的统一管理平台以及非结构化的天文讲座等视频资源的管理。主要成果与创新点包括以下几项：

- 改进条带算法，将星表天区覆盖信息直接带入到交叉证认过程，并使用同样的索引方式对数据进行组织。
- 快速的星表缺失源检测

总结

- 提出光学星表与射电星表进行交叉证认的一种可行方法
- 跨平台、插件化的天文数据、服务集成平台

文章发表

- **Fan, Dongwei**; Budavári, Tamás; Szalay, Alexander S.; Cui, Chenzhou; Zhao, Yongheng. Efficient Catalog Matching with Dropout Detection. 2013. PASP, Volume 125, issue 924, pp.218-223
- Cui, Chenzhou; **Fan, Dongwei**; Zhao, Yongheng; Kembhavi, Ajit; He, Boliang; Cao, Zihuang; Li, Jian; Nandrekar, Deoyani. Enhanced Management of Personal Astronomical Data with FITSManager. 2012. New Astronomy,Volume 17, Issue 2, p. 167-174
- **樊东卫**; 崔辰州; 赵永恒. FITS文件管理器设计与实现, 2011, 天文研究与技术, Volume 8, No.3
- 李建; 崔辰州; 何勃亮; 赵永恒; 曹子皇; **樊东卫**; 李长华; 谌悦. 天文数据库回顾与展望. 2013. 天文学进展. Volume 31, No.1
- 崔辰州; 李建; 蔡栩; 范玉峰; 王锋; 曹子皇; 苏丽颖; **樊东卫**; 乔翠兰; 何勃亮; 李长华; 赵永恒; 谌悦; 王传军; 辛玉新; 白金明; 季凯帆. 程控自主天文台网络的发展. 2013. 天文学进展. Volume 31, No.2

未来工作设想

- 继续专研交叉证认算法，应用到星表数据及其他数据的融合中，丰富China-VO服务，也作为本人主要研究方向之一
- 继续发展资源统一管理平台，完成自然科学基金项目，与天文领域云等信息化平台整合起来，逐渐形成一个可靠性好的实用平台
- 发掘自己与天文学研究更多的结合点，如可视化、高性能计算、数据挖掘、数值模拟等。除了技术之外，还要在理论方面充实自己

谢谢