



大数据时代的天文学研究

崔辰州^{①*}, 于策^②, 肖健^③, 何勃亮^①, 李长华^①, 樊东卫^①, 王传军^{④⑤}, 曹子皇^⑥, 范玉峰^④, 洪智^③, 李珊珊^①, 米琳莹^①, 袁实^③, 万望辉^{①⑦}, 王建国^④, 王甲卫^①, 尹树成^③, 郝晋新^⑧, 薛艳杰^⑧, 刘梁^⑨, 陈肖^⑩, 张海龙^⑪, 谌俊毅^⑫, 乔翠兰^⑦, 苏丽颖^⑬

① 中国科学院国家天文台中国虚拟天文台, 北京 100012;

② 天津大学计算机科学与技术学院, 天津 300072;

③ 天津大学软件学院, 天津 300072;

④ 中国科学院云南天文台丽江天文观测站, 昆明 650011;

⑤ 中国科学院大学, 北京 100049;

⑥ 中国科学院国家天文台郭守敬望远镜运行与发展中心, 北京 100012;

⑦ 华中师范大学物理科学与技术学院, 武汉 430079;

⑧ 中国科学院国家天文台基础科学研究部, 北京 100012;

⑨ 中国科学院紫金山天文台信息化建设中心, 南京 210008;

⑩ 中国科学院上海天文台计算中心, 上海 200030;

⑪ 中国科学院新疆天文台计算机技术室, 乌鲁木齐 830011;

⑫ 中国科学院云南天文台信息中心, 昆明 650011;

⑬ 北京工业大学机电学院, 北京 100024

* 联系人, E-mail: ccz@bao.ac.cn

2014-09-11 收稿, 2014-11-17 接受, 2015-01-20 网络版发表

中国科学院信息化专项(XXH12503-05-05)、国家自然科学基金委员会与中国科学院天文联合基金(U1231108)、科技部科技基础性工作专项(2012FY120500)和北京市教委青年拔尖人才培养计划资助

摘要 天文学已经进入数据密集型时代或者说大数据时代。面对海量天文数据在存储、计算、网络、软件、算法乃至工作模式等方面的需求和挑战, 天文学家连同计算机和信息技术领域的专家正努力使基于科学数据的知识发现过程变得更加容易。虚拟天文台旨在实现科学数据的互操作, 打造一个全球性的数据网格。天文信息学则从分支学科的高度来考虑天文学的长远发展。数据挖掘和知识发现在数据密集型时代大有可为, 自身也必将获得长足发展。本文简要论述天文学研究在数据密集型时代所面临的挑战, 介绍虚拟天文台理念和最新进展, 探讨天文信息学发展的必要性和所包含的研究内容, 阐明数据挖掘和知识发现的必要性和发展方向。

关键词

虚拟天文台
天文信息学
大数据
数据挖掘
知识发现

2014年7月21日, 中国科学院云南天文台2.4 m光学天文望远镜2014年下半年观测时间正式面向国内外天文学家开放申请。与以往通过发送电子邮件来提交申请材料不同, 此次观测时间的申请只有一个入口, 即“中国虚拟天文台”。这个看似微小的变化却意味着我国天文学研究领域一个新时代的到来, 即从传统的工作模式走向泛在融合的信息化时代。

1 数据密集型时代

天文学, 与许多学科类似, 正在信息与计算技术(ICT)等新兴科技的驱动下发生着根本性的变革。地基和空基的望远镜等观测设施的工作波段已经覆盖了整个电磁波谱(从射电一直到 γ 射线)以及多个非电磁窗口(比如宇宙线、中微子、引力波等), 收集的数据经过规范处理后进入数据库, 为下一步的科学分

引用格式: 崔辰州, 于策, 肖健, 等. 大数据时代的天文学研究. 科学通报, 2015, 60: 445–449

Cui C Z, Yu C, Xiao J, et al. Astronomy research in big-data era (in Chinese). Chin Sci Bull, 2015, 60: 445–449, doi: 10.1360/N972014-00839

析做好准备. 天文学数据正在以前所未有的速度从数据量、增长量和复杂性等各方面快速发展. 这些丰富的数据资源大大加深了人类对宇宙的认识.

2009年正式投入运行的LAMOST光谱巡天望远镜每晚拍摄数万条天体的光谱, 产生20 GB的光谱数据. 地面广角相机阵GWAC每天的观测数据量可达7.4 TB(1 TB=1024 GB). “天籁计划”大型射电干涉仪阵列一期96面天线的数据流量为4.8 GB/s, 二期1000面天线的数据流量为3.2 TB/s. 国际上, 大口径全天巡视望远镜(LSST)2018年投入运行后预计每晚可产生15 TB的原始观测数据, 低频射电阵(LOFAR)每秒产生上百千兆字节的科学数据, 一平方千米天线阵(SKA)整体运行后每秒钟收集的原始观测数据更是高达千万亿字节. 海量而复杂的图像、光谱、星表、时序等数据就像一个深邃的数字宇宙, 为天文学家提供了广阔的挖掘空间. 例如, 通过对收录着数十上百亿天体的多波段巡天数据库的融合来深入研究宇宙的大尺度结构和银河系的精细结构, 在海量数据中借助高级的数据挖掘手段发现稀有和新的天体类型, 在宇宙时标尺度上揭示类星体、星系和星系团的演化, 将大规模数值模拟结果与精细的观测数据进行比对分析等.

天文数据是宝贵的科研资料, 是人类了解和研究宇宙奥秘的第一手素材. 在人类探索宇宙奥秘和自身起源强烈的好奇心驱使下, 随着科学和技术的不断进步, 特别是望远镜设计与制造技术、探测器设计与制造技术、信息与计算技术等的发展, 天文学已经进入海量数据时代, 或者说“大数据”时代. 天文学研究成为一项以数据为中心, 为数据所驱动的科研活动.

驱动这场变革的背后推手除了有不断丰富数据, 还有互联网带来的便捷的数据访问和资源共享. 现代天文数据库中蕴含的信息内容十分丰富, 档案研究和数据挖掘不仅是必要的而且是必需的. 大多数情况下, 数据的拥有者发掘的信息只是这些数据库科学价值的很小一部分.

与此同时, 数值模拟不再仅仅是理论分析的一种辅助手段, 而变成研究许多复杂现象(比如宇宙大尺度结构的形成与演化、星系间的相互作用、恒星系统的形成)主要乃至唯一的途径, 这些数值模拟程序也产生大量的数据. 换句话说, 理论家不再仅是用公式来表达自己的思想, 还有大量的数据. 实验(观测)和理论都在用大量的数据来推动人类知识的发展. 现代

科学与传统科学无论是从量上还是从质上都发生了根本的变化, 成为一门数据驱动的科学, 这就是“第四范式”^[1]. 基于数据的科学发现已成为继实验(观测)、理论分析、科学计算之后, 科学发展的另一个重要支柱.

天文学是数据密集型科学的领头羊, 这主要归因于下面3个因素^[2]. 首先, 天文学最早采用(某种程度上是发展了)现代数字探测器, 例如CCD和数字相干器, 并把科学计算作为数据处理的手段, 把数值模拟作为一种科研工具. 国际天文学领域中e-Science的文化理念早在20世纪80年代, 在英特网(WWW)和商业数据库诞生之前, 就被培育起来. 天文数据集的体量从最初的千字节到兆字节, 20世纪80年代末发展到千兆字节, 90年代中期到万亿字节, 如今则是千万亿字节. 天文学家早在20世纪80年代初就设计了领域内通用的数据交换标准, 即FITS^[3].

第二个重要的因素是美国宇航局(NASA)等空间机构为其空间科学计划建立起一批数据中心, 在一定的保护期后把科学数据向全社会开放共享. 天文学家的这种做法不但推动了数据库和数据管理工具的发展, 也逐渐培育出科学数据开放共享和重复利用的科学文化. 这些数据中心成为今天虚拟天文台的发祥地和重要基础.

第三个重要因素是大型数字巡天计划的出现并成为天文数据的主要来源. 利用照相底片做巡天观测, 通过扫描实现数字化, 这样的传统巡天工作在20世纪90年代便寿终正寝. 传统巡天计划造就了第一个万亿字节量级的天文数据集, 即数字化帕洛玛巡天(DPOSS)^[4]. 这个记录很快便被斯隆数字巡天(SDSS)等纯数字的巡天计划打破^[5]. 除了取得瞩目的科学成果, 现代数字巡天计划还改变了天文学的研究模式和天文学家的思维模式. 基于现代巡天数据库, 科研人员不依赖于望远镜也能做出漂亮的研究成果. 数字巡天时代的天文学发展不但需要天文学家的个人智慧, 更需要大型科研团队的协同创新.

数据密集型时代的天文学研究给天文学家和技术专家带来诸多机遇的同时也带来非常多的挑战. 数据密集型环境下, 天文学研究所需的资源不但包括数据和文献, 更是需要数据库、分布式存储、高性能计算、数据挖掘和知识发现工具、创新的可视化环境等. 不同波段、时刻、空间尺度的数据融合把这些挑战又提升到一个新的高度. 针对数据的采集、归档、管理、访问、处理、挖掘、展现, 这些科研活动

环节,在数据量不断增大、数据结构越来越复杂的大数据时代,传统的方式和手段已不再适用和满足天文学研究的需求。

2 虚拟天文台

虚拟天文台就是要把天文学从老的数据贫乏的研究体制变革到新的数据极大丰富的研究体制,充分利用信息技术来解决新的、数据密集型天文学时代的挑战。

虚拟天文台(virtual observatory, VO)^[6,7]是通过先进的信息技术将全球范围内的天文研究资源无缝透明连结在一起形成的数据密集型网络化天文学研究和科普教育环境。其将全球的天文数据库连接起来形成一个多波段的数字星空,一个全球性的天文数据网格,让科学家和普通用户能够基于数据发现、高效数据访问和互操作,以各种创新的方式进行检索、展现和分析,打造创新型的科学研究和资源使用环境。

由各国虚拟天文台项目组成的国际虚拟天文台联盟(简称IVOA)致力于为实现数据的互操作制订有关的标准和规范,使数据产品生成、管理和使用的各个环节都在标准的框架下进行。虚拟天文台的诞生消除了各个数据库系统访问标准不统一的问题,使得星表数据交叉认证、异构图像和光谱数据的分析处理、多波段数据的可视化等以往棘手的工作可由程序和软件来完成。

传统研究模式为人类留下了千兆字节量级的数据遗产,巡天模式为人类留下了万亿字节量级的数据遗产,虚拟天文台模式则把人类带入千万亿字节量级的数据密集型时代。这3种模式并不是一种取代另一种,而是同时存在于现代天文学研究活动中。

中国虚拟天文台旨在完成国际虚拟天文台宏伟构想中的中国部分,正在以国内核心天文观测设备的时间申请、审批,数据汇交、共享、使用,课题设计、开展为线索,融合天文观测和科研活动所需的科学数据、科技文献、高性能计算、软件和实用工具等资源,打造一个物理上分散、逻辑上统一的网络化科学研究平台;基于虚拟天文台技术和云计算技术实现一个全生命周期数据管理与开放共享平台^[8,9]。中国虚拟天文台是一个数据驱动的科研信息化环境,基于标准、完整、有质量保障的元数据和科学数据系统,通过具备互操作能力的软件、工具和服务,为天文学家等科学用户打造一个泛在融合的信息化科研

新模式。同时,这是一个开放的平台,通过标准的接口和协议与国际上的资源和服务实现无缝融合。

3 天文信息学

通过虚拟天文台连通起来的全球天文数据网格让天文数据的发现和访问变得空前便捷,从理论上讲这至少能够适应下一代巡天和空间计划的需求。但这只是万里长征迈出的第一步,其中缺少的一个关键环节便是使用方便、适应性强、可以从海量数据集中提取信息和知识的工具库,特别是那些可以在万亿字节量级数据集上开展数据挖掘和分析的工具。目前这方面的工作主要靠个别的研究小组以及一些天文台和巡天计划来开展。

要全面应对海量科学数据时代天文学研究的挑战,天文学家们需要更宽广的视野和长久的策略,需要一个重要的新学科,即天文信息学(Astroinformatics)(<http://www8.nationalacademies.org/astro2010/DetailFileDisplay.aspx?id=455>)。天文信息学旨在为天文学和信息技术以及计算机应用科学搭建桥梁,以基于VO框架建立起来的数据网格为基础,为数据密集型天文学的研究者们提供一个更广阔的社区。

虚拟天文台的核心目标是制订一套完整的标准来实现对全球天文信息资源的发现、访问和互操作,这是天文信息学的基础和出发点。为了应对数据密集型的天文学研究,还需要更多系统性的研究和开发,把计算和分析的科学工具应用到天文学领域,从海量的数据中甄别出新的模式和新的发现。天文信息学代表了数据密集型天文科学研究的一种新模式,它涵盖一系列内在相互关联的领域,包括数据组织、数据描述、天文分类学、天文概念语义、数据挖掘、机器学习、可视化、天文统计学等。研究的内容包括:数据模型、数据转换和归一化方法、索引技术、信息提取和整合方法、知识发现方法、基于内容和基于语境的信息呈现、一致化的语义描述、分类学、天文本体论等。

4 数据挖掘和知识发现

随着数据密集型时代的到来,天文学研究的模式正在从以往的小样本向着大数据模式转变,从个人研究到团队协作转变。海量的天文数据给天文学家带来巨大的机遇和挑战,天文学研究越来越离不开大数据集的统计分析,即数据挖掘和知识发现。

大多数传统的数据分析方法都无法适应万亿字节量级或者高维度的数据,必须开发和采用现代的数据挖掘和数据库知识发现技术^[10]。人工审查海量数据已无可能,必须采用机器学习技术来自动化的完成这项工作。当前大多数的数据挖掘工具虽然可以应对较大的数据量但无法应对很高的数据维度,需要研究能适应现代海量复杂数据集的数据挖掘工具和方法。

以新一代概要巡天观测为代表的新兴观测手段让时域天文学获得长足发展,同时带来许多新的需求和挑战^[11]。新一代概要式巡天能够实现对大面积天区的快速、多次扫描,从而产生很多倍于传统数字巡天的数据量。新一代概要式巡天计划的实施把数据处理和分析的对象从海量数据集变为海量数据流,研究工作的复杂度进一步提升。很多物理过程和事件持续的时间很短,要求近乎实时地完成目标认证、特征提取、天体分类、随动观测优先级确定等工作。这与传统的数据处理和分析要求很不相同。受限于可调动的观测资源,概要式巡天计划探测到的暂现事件中只能有很小一部分能够得到随动观测。现有的分类方法会被用来为这些暂现事件自动判定随动观测优先级。事件分类和可用资源都在动态地发生着变化。前者归因于新的测量结果,后者则受到资源

分配、天气、昼夜交替等多种因素的限制。为能应对这些棘手的挑战,必须通过创新的手段实现对海量数据流的鲁棒、灵活、实时处理和挖掘。

有效的可视化是数据探究、分析和理解的重要环节,也是数据挖掘不可分割的一部分。可以说,可视化是科学数据所蕴含的定量内容和对其直观理解的桥梁。对不同波段、不同天区的精彩图片进行展现是天文学独特的优势,但对高维参数空间的海量现代天文数据进行可视化仍然需要很多创新的手段和方法。这是当代“大数据”科学所面临的一个共同挑战。人类的视觉感知和模式识别能力是很有限的。把几个维度的数据通过一些技巧来呈现在一个伪3D的图像上还是可以接受的,但现代科学数据集已经具有数百上千的维度。可视化是制约数据密集型科学研究的关键瓶颈之一^[12]。

5 结论

这是一个令人兴奋的时代。新技术、新方法层出不穷,天文学家随时都有可能从海量数据中取得新的发现。探测器、传感器、高性能计算集群等快速积累着各种数据,对网络、存储、计算等信息化基础设施提出几乎无穷尽的需求。数据挖掘、知识发现、数据可视化专家需要不停地思考如何让研究工具更加强壮。

致谢 本文得到中国天文数据中心提供的数据资源支持。中国天文数据中心得到国家科技基础条件平台建设项目“地球系统科学数据共享平台”和“基础科学数据共享网——理化天文空间生物”的资助。

参考文献

- 1 Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research, 2009
- 2 Longo G, Brescia M, Djorgovski S G, et al. Data driven discovery in astrophysics. In: the Proceedings of ESA-ESRIN Conference: Big Data from Space 2014, Frascati, Italy, November 12–14, 2014. arXiv:1410.5631
- 3 Wells D, Greisen E, Harten R. FITS—A flexible image transport system. *Astron Astrophys Suppl Ser*, 1981, 44: 363
- 4 Djorgovski S G, Gal R R, Odewahn S C, et al. The Palomar Digital Sky Survey (DPOSS). In: Colombo S, Mellier Y, Raban B, et al., eds. *Wide Field Surveys in Cosmology*. Gif sur Yvette: Editions Frontieres, 1998. 89
- 5 Djorgovski S G, Mahabal A, Drake A, et al. Sky surveys. In: Bond H, ed. *Astronomical Techniques, Software, and Data*. Berlin: Springer Verlag, 2012
- 6 Cui C Z, Zhao Y H. Worldwide R&D of virtual observatory. In: Jin W J, Platais I, Perryman M A C, eds. *A Giant Step: From milli- to micro-arcsecond Astrometry*. Proceedings of the International Astronomical Union (2007) Symposium S248, 2008, 3: 563–564
- 7 Brunner R, Djorgovski S G, Szalay A. Virtual Observatories of the Future. Provo, UT: Astronomical Society of the Pacific, 2001. 225
- 8 Cui C Z, Xue Y J, Li J, et al. Virtual observatory, an e-science environment for astronomy (in Chinese). *Bull Chin Acad Sci*, 2013, 28: 511–518 [崔辰州, 薛艳杰, 李建, 等. 虚拟天文台——天文学研究的科研信息化环境. 中国科学院院刊, 2013, 28: 511–518]
- 9 Cui C Z, Xue Y J, Li J, et al. e-Science Infrastructure for Astronomy Research (in Chinese). China's e-Science Blue Book 2013. Beijing: Science Press, 2013 [崔辰州, 薛艳杰, 李建, 等. 天文学研究的科研信息化环境. 中国科研信息化蓝皮书 2013. 北京: 科学出版社, 2013]

- 10 Mahabal A A, Djorgovski S S, Donalek C, et al. Classification of optical transients: Experiences from PQ and CRTS surveys. In: Turon C, Meynadier F, Arenou F, eds. *Gaia: At the Frontiers of Astrometry*. EAS Publ Ser 45. Paris: EDP Sciences, 2010. 173
- 11 Bloom J, Richards J. Data mining and machine-learning in time-domain discovery & classification. *Adv Mach Learn Data Min Astron*, 2011, arXiv/1104.3142
- 12 Goodman A. Principles of high-dimensional data visualization in astronomy. *Astron Nachrichten*, 2012, 333: 505

Astronomy research in big-data era

CUI ChenZhou¹, YU Ce², XIAO Jian³, HE BoLiang¹, LI ChangHua¹, FAN DongWei¹,
WANG ChuanJun^{4,5}, CAO ZiHuang⁶, FAN YuFeng⁴, HONG Zhi³, LI ShanShan¹, MI LinYing¹,
QIU Shi³, WAN WangHui^{1,7}, WANG JianGuo⁴, WANG JiaWei¹, YIN ShuCheng³, HAO JinXin⁸,
XUE YanJie⁸, LIU Liang⁹, CHEN Xiao¹⁰, ZHANG HaiLong¹¹, CHEN JunYi¹², QIAO CuiLan⁷ &
SU LiYing¹³

¹ Chinese Virtual Observatory, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China;

² School of Computer Science and Technology, Tianjin University, Tianjin 300072, China;

³ School of Computer Software, Tianjin University, Tianjin 300072, China;

⁴ Lijiang Observatory, Yunnan Astronomical Observatory, Chinese Academy of Sciences, Kunming 650011, China;

⁵ University of Chinese Academy of Sciences, Beijing 100049, China;

⁶ LAMOST, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China;

⁷ College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China;

⁸ Basic Science Research Division, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China;

⁹ Information Center, Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210008, China;

¹⁰ Computing Center, Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China;

¹¹ Information Center, Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi 830011, China;

¹² Information Center, Yunnan Astronomical Observatory, Chinese Academy of Sciences, Kunming 650011, China;

¹³ College of Mechanical Engineering and Applied Electronics Technology, Beijing University of Technology, Beijing 100124, China

Astronomy research has entered an increasingly data-intensive, or “big data” era. Exponential growth in the size of astronomical archives brings with it completely new requirements and challenges for data storage, computing power, networks, software, algorithms, and even research methods. Astronomers and experts in information and computation technology (ICT) are working together to simplify knowledge discovery within these massive data sets. Worldwide astronomical archives are connected through the virtual observatory (VO) framework, which provides interoperability standards and services that have allowed the creation of a global astronomical data grid. Astrodynamics, a bridge between astronomy and ICT and applied computer science, aims to engage a broader community of researchers both as contributors to and as consumers of the new methodology for data-intensive astronomy, thus building upon the data-grid foundations established by the VO framework. Data mining (DM), knowledge discovery in databases (KDD), machine learning, and visualization techniques have become a necessity. In this paper, the challenges facing modern astronomical research are briefly described, the concepts underlying the VO and astrodynamics and the latest progress in these fields are introduced, and the need for DM and KDD techniques in astronomy is discussed.

virtual observatory, astrodynamics, big data, data mining, knowledge discovery in databases

doi: 10.1360/N972014-00839