Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

CIShell基本原理及其应用初探

□ 张金柱 / 中国科学院国家科学图书馆 北京 100190

摘要：科研环境正在向协同、共享、多学科、多数据流、动态配置等转变，CIShell提供的从数据到知识流的协同工作平台成为可能的解决方案。文章对CIShell的理论基础和技术实现基础进行介绍，并对其系统结构、算法设计、数据格式转换进行说明，借鉴在图书情报、政策分析、医学领域的应用，提出一种基于CIShell的情报分析集成框架的雏形，实现算法、工具、数据集的动态更新、配置，对其数据处理流程和Bundles结构设计进行论述。

关键词：CIShell，科研环境，情报分析，集成框架
DOI：10.3772/j.issn.1673—2286.2012.06.008

1 引言

当前，科研环境正在发生着巨大的变化：

1) 从依靠著名科学家到依靠研究团队：以往，科技进步往往是由一些著名的科学家推动；如今，科技进步往往由来自不同学科领域的专家组成的高效合作团队来推动[1]，需要依靠团队成员的协同合作来完成少数科学家不可能完成的任务，甚至已经发展到需要多个合作团队的协同合作完成。

2) 从单纯的用户向贡献者转变[2]：Web2.0等技术使任何人都可以向Wikipedia贡献词条或者在Flickr和Youtube上分享图片和视频。用户与贡献者之间的界限已经逐渐模糊，海量用户的专业知识已经开始得到分享，有重要的科研价值和应用价值，并已经有相关的研究。如Wikispecies、Wikiprofessional等把维基内容和语义技术进行组合，支持对科学数据进行实时的基于社区的注释，并进行语义检索[3]。

3) 从单一学科到多学科交叉转变：最好的科研、工具往往频繁地借鉴、融合、改进了其他不同学科的方法和技术，这样的例子比比皆是。以比较熟悉的情报分析方法为例，它已经从统计学方法逐渐加入复杂网络的相关分析方法，对数据进行深层次的挖掘，描述科学知识的结构、演化和动力学过程，找到新兴研究热点和趋势。同时，不同学科或国际研究团队的研究者、实践者和相关领域的专家来共同解释分析结果，每一个研究人员均有其擅长的方向，因此，从不同人员的角度、不同学科的角度对结果进行全方面分析是必要的。

4) 从单一样本到数据流：单一的样本数据已经不能满足当前科研的需求，与样本数据相关的研究数据能提供更多的语义等上下文信息，因此，更多的研究人员需要理解多种类型的数据，如出现在科研的各个过程中的不同格式的、动态的和原始的数据。

5) 从静态的工具到可进化的动态配置化转变：现在的软件需要更多的灵活性和可扩展性，可以针对科研的不同需求来进行改变、扩展。通过插件和微服务的方式向越来越多的人提供科研支持将是未来发展方向[4]。

与上述情况对应的是，情报分析也正面临着类似的转变：依靠研究团队的集体智慧；每一名用户可以方便地贡献自己的数据、方法、工具；已经并继续从其他相关学科借鉴、融合、改进研究方法和技术；不仅仅关

注文献、专利等科技信息，还需要利用市场、经济、战略、社交等信息；情报分析方法众多，但其分享、重用性并不好，很多重复工作一直在进行，需要对这些方法提供统一平台、标准，实现动态更新、配置。

面对这些问题，CIShell（Cyberinfrastructure Shell，http://cishell.org/）为上述需求提供了相关的解决方案。CIShell是开源的、社区驱动的平台，用来集成和利用数据集、算法、工具和计算资源的框架。其理论基础是CI（Cyberinfrastructure），CI是把通信、信息、社会、计算、协同、文化六大组成要素整合为一体，把数据引向知识的基础设施[5]。实现的思想则是依照OSGi的面向服务和插件的理念。

在情报分析领域，虽然已有较多的基于CIShell开发的协同平台来提供不同领域的算法和工具集，如NWB（Network Workbench，http://nwb.cns.iu.edu/）主要提供社会科学、物理学方面的算法和工具集，EpiC（Epidemics Cyberinfrastructure，http://epic.slis.indiana.edu/）主要提供医学方面的算法和工具集，IVC（Information Visualization Cyberinfrastructure，http://iv.slis.indiana.edu/）则更多地是关注可视化方面的算法和工具集。然而，摆在我们面前的问题是，如何将已有的各种算法、工具集快速地集成到CIShell框架中，为机构中其他用户所用，并让机构其他用户能够根据相应的流程快速将算法、工具集进行改造并集成进来。

本文在对CIShell基本原理、系统结构、实现方案进行说明的前提下，综述其已有的多个应用，试图提出一种融合现有数据、工具、算法的一个情报分析集成框架的雏形。

2 CIShell基本原理

CIShell是Cyberinfrastructure for Network Science Center（http://cns.iu.edu/）的一个子项目，是一个集成数据集、算法、工具和计算资源的框架，它有众多的特性[2]，其理论基础是CI，而实现方式则是基于OSGi。

1) 核心框架和插件填充：计算机科学家需要设计一个标准化、模块化、易维护和可扩展核心框架，在此即为CIShell。而数据和算法则是以插件的方式进行填充，这些插件的提供者可能不是计算机科学家，而更多的是领域专家。

2) 易于使用：由于插件的提供和使用可能大部分来自非计算机科学的领域专家，因此，在使用插件时需要保证不再编写任何代码。用户能够根据向导说明和其他说明文档，从插件库中组织适合自己需求的数据处理流程。

3) 模块化：将功能分解，减少耦合性，从而在替换某个模块达到质量或效率的提升时，不会改变整个结构，只需更改相应的模块。合理的设计，保证封闭、开放、效率的平衡，既在一定程度上能保证易于修改，也就是易于优化和扩展；也在一定程度上保持良好的性能，增强软件的灵活性。

4) 开放源代码和数据：为了科研目的，使任何人都可以对代码和数据进行下载、改进、重写和复制。

2.1 CI (Cyberinfrastructure)

CI是一个把通信、信息、社会、计算、协同、文化六大组成要素整合为一体，把数据引向知识的基础设施。它支持数据到知识转变过程的数据流，并把这些纳入到统一框架中，如图1所示。CI包括几个核心要素：分布计算、每秒万亿次浮点的运算速度、海量的存储、高速网络、灵活的数据集成、灵活的分析软件集成等。根据已有的研究以及CI实践来看，还包括以下几项特征：社会导向、分布式的合作、虚拟组织、跨学科、跨地域、互操作性、支持数据/运算密集的应用程序、支持桌面操作的终端、异构性、复杂性、可重复使用[6]。美国、英国等国家相当重视CI的研究与实际应用，投入大量资金并在生态研究、环境科学、大气科学、地震模拟、生物信息等学科领域进行了试验[7]。

图2则表示CI的整体结构，主要包括三个方面：

1) 硬件：计算、存储和通信技术等硬件支持。

2) 应用层：软件程序、服务、仪器、数据、信息、知识，适用于特定课题的社会实践、学科和工作人员。

3) 共享层：网格服务、算法、中间件、机构和个人。这个层次将为具有授权的机构的研究者们提供高效的共享平台，改进他们的工作内容、工作方式和工作伙伴。

2.2 OSGi

OSGi（Open Service Gateway Initiative）是Java中目前唯一的模块化、动态化的规范。OSGi规范对于

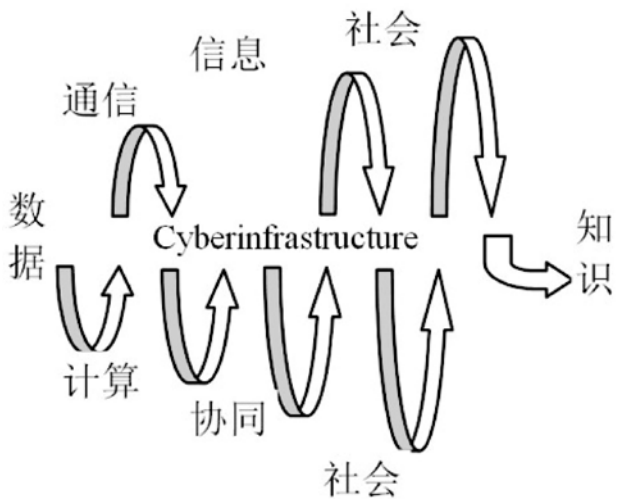

图1 CI数据流

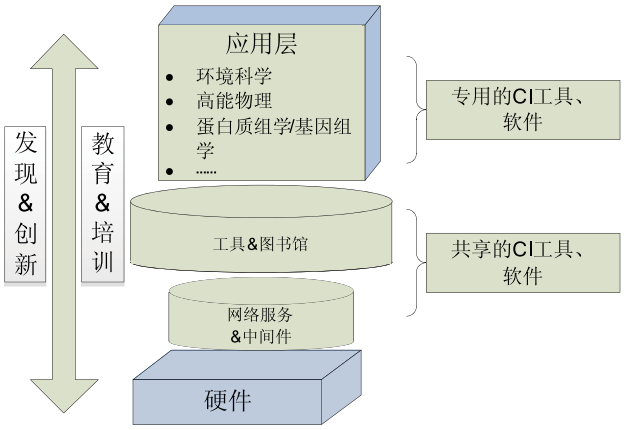

图2 CI整体结构

模块的物理隔离、模块的交互、多版本控制方面都有非常完善的机制，并且得到了几乎所有的App Server厂商或开源社区的认可。OSGi联盟是于1999年成立的非盈利国际组织，旨在建立一个开放的服务规范，为通过网络向设备提供服务建立开放的标准，是开放业务网关的发起者，旨在成为一个轻巧的、松耦合的、面向服务的应用程序开发框架。Eclipse3.0之后也放弃了其自身的插件机制，而开始使用更为规范的OSGi插件体系。事实证明Eclipse在采用OSGi架构后，其性能、可扩展性等方面都取得了巨大的改善。

基于OSGi的应用程序，其基本组成部分为Bundle，每一个Bundle对应到一个插件，一个插件则表示一个算法、工具或数据集。其基本特点是可动态更改运行状态和行为，每个Bundle都是可热插拔；稳定高效，不会因为局部的错误导致全局系统的崩溃；可复用性强，OSGi框架本身可复用性极强，很容易构建真正面向接口的程序架构，每一个Bundle都是一个独立可复用的单元。

OSGi的开源框架主要有Equinox（http://www.eclipse.org/equinox/）、Knopflerfish（http://www.knopflerfish.org/）和Felix（http://felix.apache.org/site/index.html）。

3 CIShell系统结构和实现方案

3.1 CIShell的系统结构

CIShell是一个算法驱动的、以算法为中心的扩展结构，通过输入数据对象（Data）、用户参数（Dictionary）和上下文信息（CIShellContext），处理后输出数据对象（Data）[8]。如图3所示。

1) 红色方框1表示算法。

2) 红色方框2表示输入数据，以Data表示，BasicData对Data进行了实现，可以对CIShell中支持的数据类型进行实例化。

3) 红色方框3为CIShell提供的基本服务，它提供数据集格式转换、GUI创建、初始值设定和日志记录等服务，日志记录控制显示在界面控制台的内容。

4) 红色方框4表示用户输入的参数，参数基本设置（如名称、描述、数据类型）均以界面提供。

3.2 CIShell的算法设计

算法设计分为创建基于Java的算法和非Java的算法，在算法完成后导出为jar包的形式，以插件的形式提供给其他算法或程序调用，如CIShell中及相关软件中调用。以创建基于Java算法为例，其生成全流程如图4所示。

3.3 数据格式和转换

不同的算法、工具提供的数据描述格式可能不同，CIShell提供了一些常用的数据格式转换功能，这些保证其他算法、工具能够方便、快速地集成进来。如GraphML、XGMML、NWB（*.nwb）、Pajek .NET（*.net）、Pajek .MAT（*.mat），这5种格式均可实现

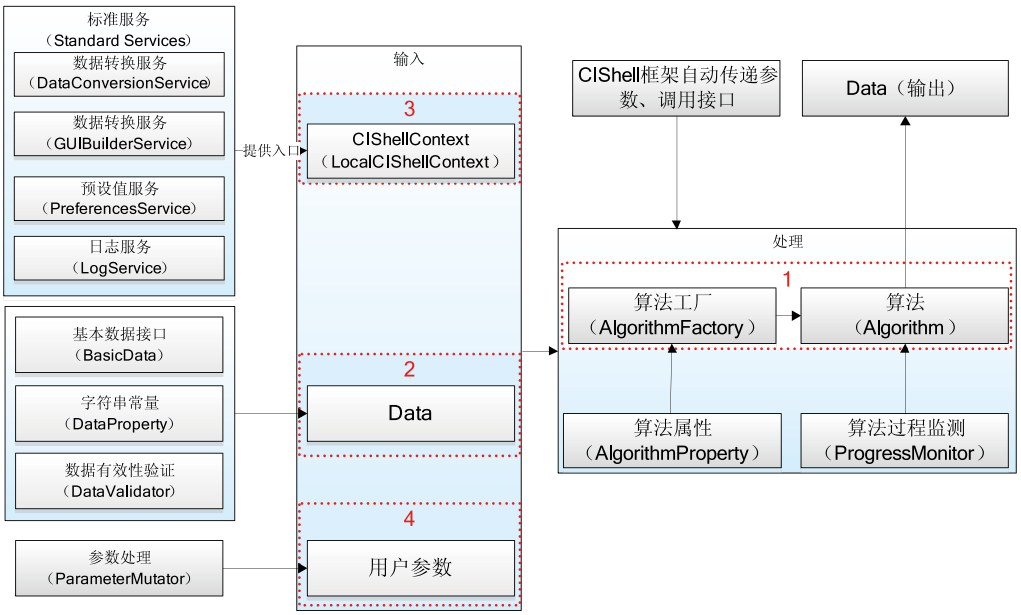

图3 CIShell整体结构

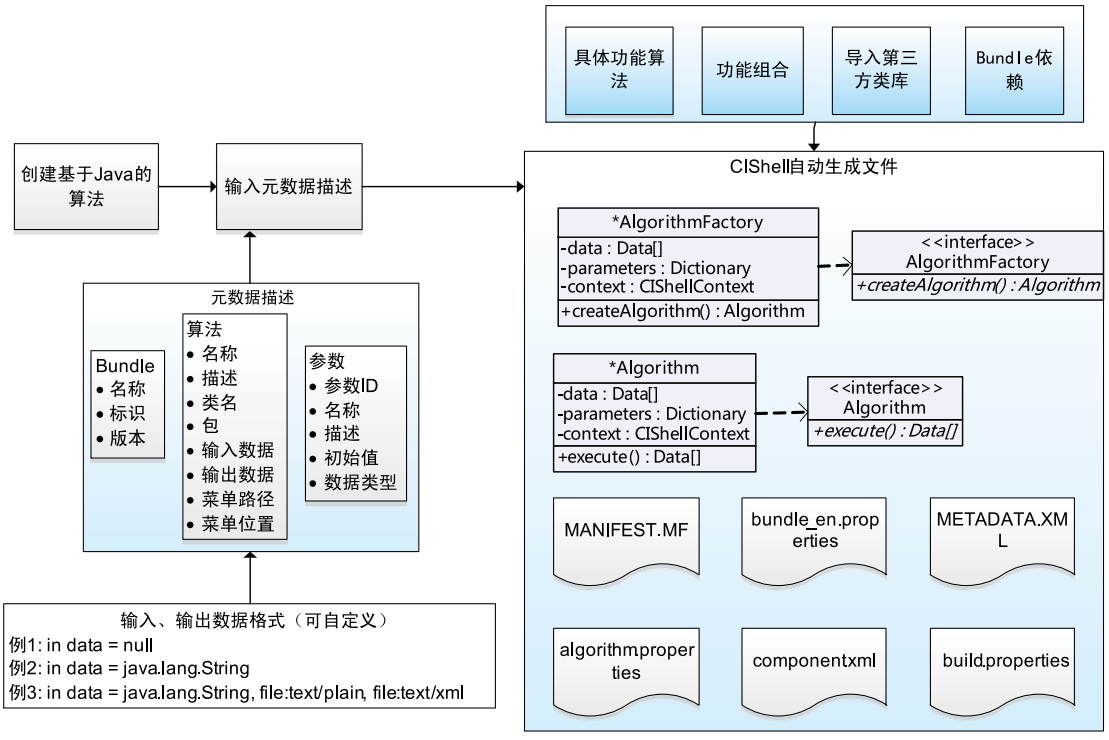

图4 算法生成流程

互相转换，其中，GraphML、XGMML（*.xml）、NWB（*.nwb）三种格式间的转换是无损耗的，而和其他两种的转换则是有损耗的。以NWB文件[9]为例，存储的信息包括节点、弧（有向图）、边（无向图）的信息，分别以*Nodes、*DirectedEdges、*UndirectedEdges表示，每个节点、弧、边均可以包含任意多个属性信息。

图5中节点代表数据格式，而弧代表转换的方向和路径，它们共同形成多个有向图，代表着不同的格式之

图5 CIShell及相关软件中数据格式转换图

间的转换，NWB中的数据转换服务便是依据此有向图来提供。在实际计算中，还会考虑转换过程的数据损耗、时间和空间复杂度等问题，在图上以弧的属性来表示[10]。

4 CIShell在多个领域的应用

基于CIShell的工具软件在图书情报、政策、医学方面的分析、可视化方面发挥着重要的作用，并有一些专门针对某一领域开发的工具，如传染病分析、政策分析方面。

4.1 图书情报领域

基于CIShell的NWB工具包含了众多的科学计量学方面的算法和数据集。

Börner对化学领域[11]的的SCIE和SSCI中的7,227种期刊进行了知识图谱分析，得到671个期刊聚类，并进一步进行聚类，得到14个研究领域，最后对这14个研究领域的变化情况进行了定量说明，对其未来发展方向给出了建议。其后，Börner对学科交叉情况进行了分析[12]，从概念网络、合作网络、期刊网络和知识流网络几个层次对学科交叉情况进行了知识图谱分析。

4.2 政策分析领域

基于CIShell开发的Science Policy Tool（SciPol）提供政策分析的工具。

Börner在其项目申请书中提到[13]，当今社会学术数据、信息、知识、经验无处不在，因此，科技政策及其他决策制定者都迫切需要新的工具来确定出版物、专利、技术要求、基金建议书和其他一些重要或有潜在效益的发展进度报告。同时，他们还需工具来对大量相关数据进行分析和挖掘，并以可视化的形式把这些综合分析结果展示出来，来辅助他们进行定性的决策支持。从这个意义上讲，政策制定者需要的不再仅仅是一个工具，还需要一个概念框架，能够把他们的信息需求同其他数据、分析结果和指标联系起来。项目要解决三个方面的问题：1、对有代表性的政策制定者和机构进行调研分析，这些机构包括NSF、NIH、DOE、OSTP以及其主要成员；2、对政策制定者的工作要求进行概念层的分析，并把这些要求以工作流程的形式固定下来；3、设计一个原型工具，用来可视化那些直接观察难以理解的大量复杂数据的结构、模式、趋势和异常情况。

Hidalgo通过对一个国家不同类型的产品进行分析[14]，发现越复杂的产品会集中于中心，而一些简单产品则会分布在外围，从而对国家的经济表现进行评价，解释贫穷国家为什么尽管出口多却不能反映到收入水平

的提高上。

4.3 医学领域

Yldrm使用基于CIShell的NWB工具，对药物针对的蛋白质和人类蛋白质相互作用产生的致病基因之间的关系进行了分析和可视化[15]。

Epidemics（http://epic.slis.indiana.edu/）是由NIH（National Institute of Health）资助的项目，它支持不同层次的复杂数据集的集成和分析，这些层次包括人口密度、病历记录和社会行为；同时能够对人口众多的区域出现的非线性和复杂现象进行分析、建模和可视化。

Colizza(2007)认为[16]，日益复杂的社会关系和交通设施是流行病传播的关键因素，而现行的计算机技术和信息分析、可视化技术可以对这些网络复杂性进行定量分析，并对这些现象进行模拟，最后总结了在流行病传播中影响其传播的多个指标，并希望以此对流行病进行动态监测。

5 基于CIShell的情报分析集成框架雏形

CIShell的框架具有如下特点：定义标准的算法接口和数据规范；提供基本日志、数据转换等服务；能根据用户参数配置生成参数输入界面；组织参数、输入数据自动调用Bundle中的接口，并把处理中间结果在相应区域显示。不仅如此，已经形成了大量基于CIShell开发的算法、工具，可以快速集成到情报分析集成框架中进行利用。这些使开发者根据CIShell提供的规范，专注于自己的算法，其他的工作均由CIShell完成。下面主要对机构内已完成的系统、算法、工具如何集成到情报分析框架中进行说明，找出其中的共性特点，希望找到一些通用流程。当然，在具体实施时，还需要根据实际情况进行扩展。

5.1 数据处理流程

在情报分析中，一类重要的分析过程是利用WoS（Web of Science）提供的引文等数据进行相关分析，其分析过程的数据处理流程如图6所示。

在CIShell及其相关软件中，科学计量方面提供的算法最开始的输入数据主要为isi文件，预处理后输出主要为csv文件；在建模、分析阶段，输出均为标准格式描述的网络文件，这些文件可以相互转化，如Graphml、nwb文件格式等等；可视化过程中使用网络文件为输入，输出相应的可视化图形。这个过程利用CIShell框架和基本算法可以快速构建。

5.2 Bundles结构设计

基于上述数据处理流程，需要对Bundle进行相应的设计。Bundle是CIShell框架的最基本组成部分，起到最核心的作用，每个Bundle可以是一个算法，也可以是一个工具、软件，这些算法、工具也可以通过组合形成新的Bundle加入到Bundles池，并为其他Bundles所用。

不同的Bundle归入到不同的类中，如预处理、分

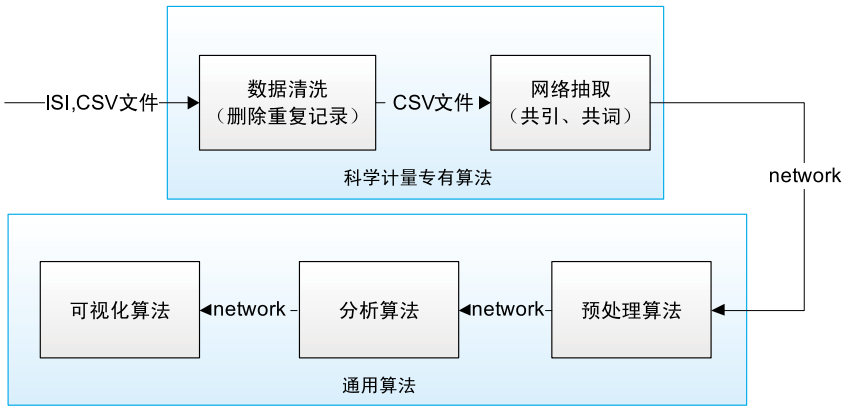

图6 基于CIShell的情报分析数据处理流程

析、可视化等，按流程配置时从每类中选择一个或多个算法，如在预处理中可能会选择删除部分数据、去重等算法组配。前提是不同Bundle间数据交换格式相同，前一个Bundle的输出可以作为下一个Bundle的输入。Bundle的粒度一般以基本功能为单位，便于其复用。这些基本功能的Bundle可以通过组合的方式形成更大粒度的Bundle。Bundles的设计结构如图7所示。

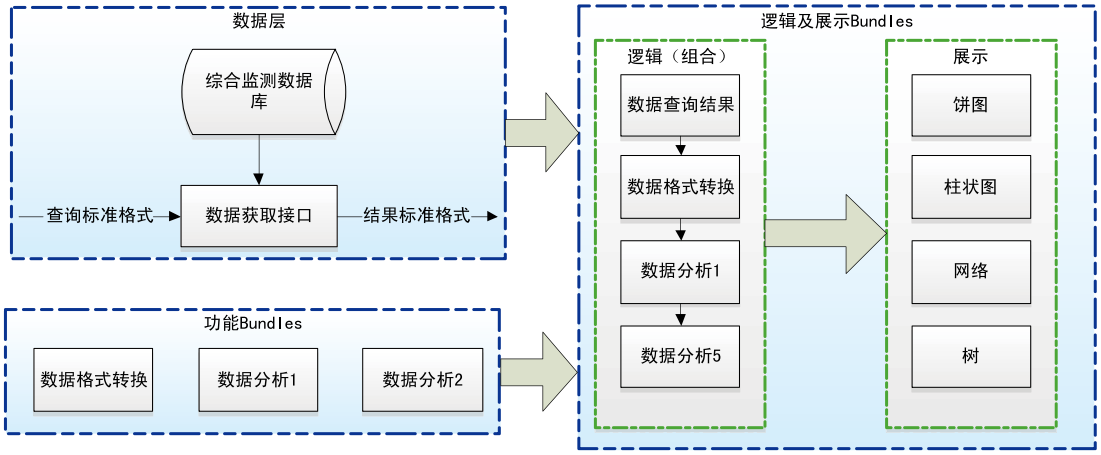

图7 Bundles结构设计

6 结语

从以上的叙述可以看到，CIShell为不同结构、层次的科研人员的协同合作提供了一个良好的框架，为简化科研流程、减少重复开发工作做出了一定的贡献。不仅开发人员可以进行相关算法的研发，情报分析人员经过培训后，也可以根据自己的需求对算法池中已有的算法进行组合，形成新的算法，用户能够根据向导说明和其他说明文档，从插件库中组织适合自己需求的数据处理流程；分析人员也补充情报分析的核心算法，使不同的情报分析方法得到重用和组合。需要注意的是，CIShell涉及的算法均以英文数据为例，在应用到中文信息处理时，还需进一步的测试，并作相应的修改。

尽管如此，采用CIShell或OSGi框架进行开发也是有难度的，因为科研机构或企业大多已经积累一些可复用的工具箱程序，而采用OSGi架构需要重新对这些遗留系统进行封装，更有可能的是需要把整个体系架构打散，进行重新的架构和排列。这个开发成本不能说不高，但也是值得的，因为从此以后企业可以利用OSGi独特的特性，将重复的知识轻易地过滤掉。对于新的开发，可以从企业的Bundles库中精简出可复用的模块，量身定做新的Bundles，最大限度地利用以前的积累，更好地支持科研工作。

参考文献

[1] BÖRNER K, et al. Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams [J]. Complexity: Special Issue on Understanding Complex Systems, 2005, 10(4): 57-67.
[2] BÖRNER K. Plug-and-Play Macroscopes [J]. Communications of the ACM, 2010, ACM Press.
[3] MONS B, et al. Calling on a Million Minds for Community Annotation in WikiProteins [J]. Genome Biology, 2008, 9(5): R89.
[4] HERR II B W, et al. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence [M]. Boston, MA: Annals of the New York Academy of Sciences, 2007: 161-179.
[5] CRAWFORD D. Charting our cyberinfrastructure future [C]// NSF cyberinfrastracture working group, 2003: 06-06.
[6] 董锐.高校虚拟科研组织中Cyberinfrastructure建设的研究[D].浙江大学,2009.

[7] 中国科学技术信息研究所.网格技术可实现计算资源共享[J/OL]. http://www.51cto.com/art/200602/21524.htm, 2006.
[8] Indiana University. Cyberinfrastructure Shell Core Specification 1.0.
[9] NWB File Format [EB/OL]. [2012-04-01]. http://nwb.cns.iu.edu/Docs/CR_NWB_File_Format_May_18_2007.rtf.
[10] BÖRNER K, et al. Network Workbench Tool User Manual 1.0.0 [EB/OL]. [2012-04-01]. http://nwb.cns.iu.edu/Docs/NWB-manual-1.0.0beta.pdf.
[11] BOYACK K W, BÖRNER K, KLAVANS R. Mapping the Structure and Evolution of Chemistry Research [J]. Scientometrics, 2009, 79(1): 45-60.
[12] BÖRNER K, BOYACK K W. Mapping Interdisciplinary Research (Sidebar, Systems Science Section) [M]. New York: Oxford University Press, 2010: 457-460.
[13] BÖRNER K, SCHARNHORST A. Special Issue on the Science of Science: Conceptualizations and Models of Science [M]. 2009: 161-172.
[14] HIDALGO C, et al. The product space conditions the development of nations [J]. Science, 2007, 317(5837): 482.
[15] Y LD R M M, et al. Drug-target network [J]. Nature Biotechnology, 2007, 25(10): 1119.
[16] COLIZZA V, et al. Epidemic modeling in complex realities [J]. Comptes Rendus Biologies, 2007, 330(4): 364-374.

作者简介

张金柱（1983-），男，中国科学院文献情报中心图书馆学在读博士，研究方向：文献计量和自然语言处理。E-mail: zhangjinzhu@mail.las.ac.cn

An Exploration on Principles and Applications of CIShell

Zhang Jinzhu / National Science Library, Chinese Academy of Sciences, Beijing, 100190

Abstract: The research environment is changing to collaboration, sharing, mutidisciplinary, muti-datastream and dynamic configuration. CIShell provides a solution which offers a cooperation workbench from data to knowledge. It introduces rationale, technology implementation, system structure, algorithm design and data form transformation of CIShell, then takes examples from the applications of Library and Information Science, policy analysis and medicine analysis. At last, it proposes a prototype of information analysis integration framework which discusses data processing flow and bundles structure design.

Keywords: CIShell, Research environment, Information analysis, Integration framework

（收稿日期：2012-04-07）

业界动态

美法官准许Google图书扫描案立为集体诉讼案

【搜狐IT消息】北京时间6月1日消息，据国外媒体报道，美国联邦巡回法院法官陈卓光(Denny Chin)当地时间周四裁定，数以千计的作者可以以集体诉讼的方式起诉Google扫描图书、建立数字图书馆的计划。

陈卓光裁定，集体诉讼"比要求数以千计的作者分别起诉更有效，而且效率更高"。他说，要求作者分别起诉Google，面临几乎完全相同的法律导致不同审判结果的危险，而且诉讼成本将呈几何级数增长。

美国作家协会过去曾要求使该案成为集体诉讼。Google已经扫描了逾2000万册图书。陈卓光驳回了Google提出的将美国作家协会剔出该案的请求。

Google周四发表声明称，该公司"坚信Google Books完全符合版权法"。(阳光)

（来源：http://it.sohu.com/20120601/n344606531.shtml，查询日期：2012-06-07）