大数据环境下科技情报研究的新模式

陈伟[1],杨锐[1],何涛[1],王朔[1],陈江萍[2]

1. 中国科学院武汉文献情报中心,武汉 430071
2. 美国北德克萨斯大学信息学院,美国丹顿 76203

摘要 大数据时代为科技情报研究与服务带来了重大的机遇和挑战,迫切需要发展新的数据驱动型情报研究模式来变革数据治理和工作流程,提高情报研究和咨询服务的质量。本文概述了传统的人力驱动型科技情报工作模式,分析了存在的问题和局限性;综述了海量异构数据集成、数据管理与分析方法和工具的开发进展;提出了建设数据驱动型科技情报研究模式的整体架构,展望了未来研究的重点。

关键词 科技情报研究;大数据;数据驱动;数据集成;数据分析

科技情报研究是现代图书情报机构的核心知识服务之一,需要通过对海量信息的检索、采集、处理与解释,分析特定技术领域的发展现状和未来发展方向,为科技政策决策者提供咨询参考。一般科技情报研究包括技术发展趋势分析、新兴技术主题监测、科技竞争力与合作分析、循证型科技战略与政策分析等。

传统的科技情报研究框架包括6个连贯且迭代的阶段:情报分析方案规划、多源异构信息采集、信息分类手工处理、信息定量定性分析、情报产品编制与传播,以及支撑决策的成效评估与反馈。每个阶段的任务主要由科技情报研究人员人工实施,最大的问题是每个阶段需要耗费大量的时间和人力工作,特别是在信息检索采集、信息集成和信息分析阶段。从而导致科技情报研究的效率和时效性受到较大的负面影响。

大数据时代的来临不仅为加速科学进步提供了前所未有的机遇,还使得创建数据驱动型知识发现新模式成为可能[1]。科学研究正在经历数据密集型范式转变[2]。作为支撑科技决策的耳目、尖兵和参谋,大数据时代的科技情报研究需要通过知识分析和知识发现服务提供及时、精准和全面的情报分析[3]。为应对这一挑战,迫切需要发展新的科技情报研究模式加快大数据治理与工作流程,提供高质量的决策咨询服务。

一个集成了一系列合适的分析工具、架构完善的数据治理体系有助于更高效地开展科技情报研究工作。本研究目的即是通过改造传统的科技情报研究框架,增加数据集成管理和分析能力,重新设计数据驱动型科技情报研究新模式。新的模式有望推动实时信息采集与分析,使情报研究人员能够快速获取所需的情报,并通过一系列内嵌的分析方法开展深度情报分析。本文首先剖析传统的人力驱动型科技情报工作流程,分析其存在的问题和局限性,综述海量异构数据集成、数据管理与分析方法和工具的研究进展。基于此,提出新的数据驱动型科技情报研究模式的整体架构。

收稿日期:2018-06-30;修回日期:2018-08-13
基金项目:中国科学院文献情报能力建设专项课题(Y7KZ131001);中国科学院青年创新促进会项目(2017221)
作者简介:陈伟,副研究员,研究方向为能源科技战略情报、知识管理与信息服务,电子信箱:chenw@whlib.ac.cn
引用格式:陈伟, 杨锐, 何涛, 等. 大数据环境下科技情报研究的新模式[J]. 科技导报, 2018, 36(16): 78-85; doi: 10.3981/j.issn.1000-7857.2018.16.009

1　传统科技情报研究模式分析

1.1　人力驱动型模式分析

科技决策本质上是一个信息汇聚的过程。为推动这一过程,国内文献情报机构已开展了一系列探索性工作,根据决策者的需求建立了情报服务模型。传统人力驱动型科技情报研究模式可分为6个连贯且迭代的阶段(图1)。

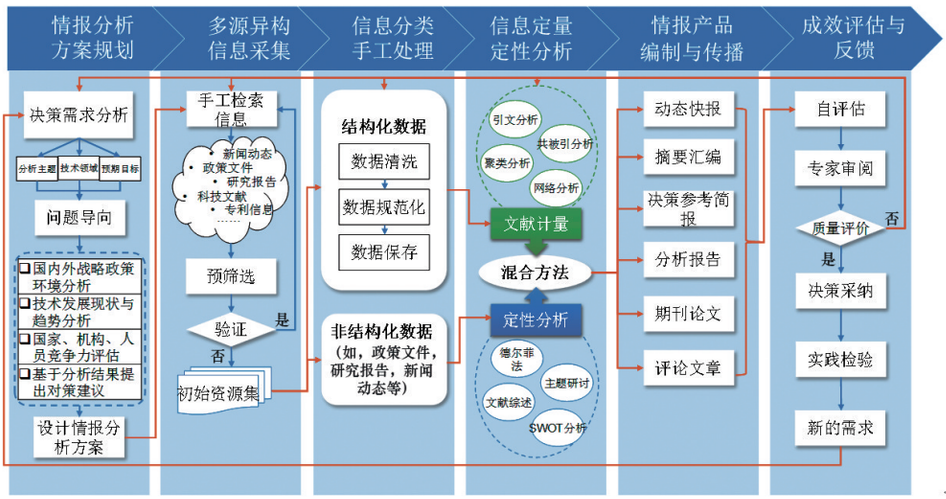

图1　传统人力驱动型科技情报研究模式

Fig. 1　Traditional human-driven mode of scientific information analysis

1)情报分析方案规划阶段。研究人员基于决策者的需求以确定分析主题、涉及技术领域和预期目标,以问题为导向设计情报分析方案。一般需求是调研一个特定的技术领域,包括国内外战略规划图景,技术发展现状与趋势分析,国家、机构、科学家各层面科研竞争力评估,并基于上述分析结果提出对策建议。

2)多源异构信息采集阶段。根据分析方案,情报研究人员从不同信息源手工检索多种类型信息,包括论文、专利、报告、统计信息等。通过预筛选和信息验证,将相关信息归类为原始资源集,保存在分散的个人文件系统中。

3)信息分类处理阶段。包括数据分类、元数据抽取、数据清洗、数据规范化和数据保存。利用德温特数据分析器(Derwent Data Analyzer™,DDA)等商业软件和CiteSpace等开源软件处理从论文和专利数据库下载的原始结构化数据[4-5]。但由于缺少合适的方法和工具,需手工处理如战略政策和报告文件等非结构化数据。

4)信息定量定性分析阶段。这一阶段应用定量分析和定性分析方法来整合数据,发现新的知识。目前定量方法主要限于文献计量方法,通过分析科技文献和专利发现及评估技术发展与演变态势、科研竞争力以及合作网络等。定性方法如专家德尔菲法、文献综述、主题研讨、SWOT分析等多用于分析文本数据。

5)情报产品编制与传播阶段。研究人员将分析结果编辑成文,根据决策者的需求和传播的要求,生成各种类型的情报产品,包括快报、汇编、决策参考简报、分析报告、展示幻灯片、期刊论文、评论等。

6)支撑决策的成效评估与反馈阶段。情报分析产品完成后,情报研究人员首先进行自评估,并征求领域专家或用户方的反馈。高质量的研究成果被决策者采用并付诸实践或作为进一步决策的支撑,而质量不高的成果基于反馈结果重复上述阶段修正。有时决策者会根据实践中的变化或新出现的形势在已有情报成果基础上提出新的情报需求,使得研究人员完成各阶段的迭代更新。

1.2　存在问题与局限性

在大数据时代,决策层对多源异构数据实时分析和深度挖掘的需求日益强烈。数据的体量和类型已经远远超出手工分析的能力[6]。由于情报任务通常有固定的完成期限,需要有良好组织的知识管理能力和合适的分析方法能够在有限的时间产出高质量的情报研究成果,从而支撑高效科学的决策[7]。显然,目前的人

力驱动型科技情报研究模式存在诸多问题和局限性,无法适应不断变化的科研和决策环境要求,主要存在以下4个方面的问题。

1)过程耗时。多个阶段需要大量的时间和人力工作,特别是在信息检索采集、信息集成处理和信息分析阶段。这些任务还严重依赖于手工收集、处理、集成和解读大量的信息。

2)知识发现能力有限。由于在情报任务中采集和储存的多数数据是多属性和非结构化格式的文本信息,情报研究人员能够有效分析的数据只占较小比例。

3)数据管理与共享问题。战略政策和报告数据集通常储存在分散的个人文件系统中,没有合适的基础设施来共享和集成相关数据,因而不能有效地管理和利用。

4)方法学问题。大部分的情报成果是描述性、小规模的分析,缺乏理论框架和量化内容分析的方法学和研究模型。

目前的情报研究模式还属于描述型信息分析,注重通过挖掘历史数据来理解以往的经验和实践成效,研究其背后的影响因素。尽管这一分析模式对于决策而言仍有一定的价值,但由于其受限于手工数据采集和分析能力而缺乏前瞻性,另一方面越来越多的决策需求需要通过集成和分析海量的多源异构数据以获得预见性判断来满足,因此发展基于大数据的预测型分析模式乃至解决方案型分析模式,从而能够利用有限的资源做出更好的决策和行动建议。将是未来科技情报研究的大势所趋。

2 大数据分析机遇

目前的科技情报研究工作模式可以通过集成大数据方法和技术加以改进。大数据的"4V"特征,即海量的数据规模(volume)、快速的数据流转和动态的数据体系(velocity)、模态繁多的数据类型(variety)和巨大的数据价值(value),对数据管理和分析提出了新的挑战[8-10]。大数据已经引起了产业界、学术界、政府机构等各创新单元的高度重视,对于其能够产出丰硕的成果给予很高的期望[11],普遍认为在数据获取、分享、集成、分析及建立数据预测模型等方面的能力提升能够推动各个学科新的知识发现不断涌现[1]。大数据范式有潜力将不完美的、复杂的以及通常是非结构化的数据转换为切实可行的情报,并且为提升科学研究、商业活动、健康医疗、公共管理以及国家安全等关键领域的战略决策能力创造了经济可行的机遇[6,12]。

情报研究人员越来越需要将不同来源、不同类型的数据集成到数据分析过程中,而主要限制因素不仅是需要分析的数据规模,更主要的是异构数据的多样性[13]。为解决这些挑战,学术界和产业界提出了多种大数据集成和分析方法与工具。

2.1 数据集成

集成多样化的数据和方法使我们能够发展预测性分析的能力以发现新的知识。由于传统的数据集成方法在大数据环境下效率低下,探索如何开发新的数据关联和集成方法来最大程度地提高大数据的价值成为一个热门的研究课题,特别是数据的深度集成仍是一个难题。除了已有多个昂贵的数据集成商业化平台外[14],近年来产业界和学术界还开发了一系列用户友好、功能丰富的数据集成开源工具[15-16]。其中有许多工具,如Kettle[17]和Talend Open Studio[18],具有直观的图形化用户界面和易于使用的拖放功能,能够兼容多个运行平台/操作系统,并且能够进行自定义的部署配置。这些高效低成本的解决方案能够探索用于开发多样化的大数据应用。

斯坦福大学InfoLab实验室开发了一个开源的知识抽取系统DeepDive[19],能够从非结构化信息(如文本)创建结构化数据,并将这类数据集成到现有的结构化数据库。DeepDive充分利用统计推断和机器学习的效率和有效性用于复杂的抽取任务,已在药物基因组学、古生物学、反人口贩卖执法等一系列领域获得了应用[20]。

还有相当多的研究人员在这一领域开展了大量工作。美国亚利桑那大学[21]开发了用于情报与安全信息学的数据基础设施,主要关注于数据采集、数据管理和数据获取。这一基础设施由在线存档和分析工具组成,集成了大批的开源数据,使研究者能够更方便地与同行开展合作。Ma等[22]基于统一概念模型(UCM)提出了一个数据集成框架,解决现实世界中汽油和天然气安全性监管的问题。通过UCM的结构对齐,将不同来源的数据自动转换成实例数据,存储在图数据库中,并通过语义相似度计算指标建立相互关联。Daraio等[23]提出了基于本体的数据管理(OBDM)方法集成异构数据,包括学术大数据(如论文和引文等)支持科研评估和开发科学学政策模型。Meng等[24]建立了以作者为中

心的计算机科学学科中文文献集成系统ScholarSpace(C-DBLP),支持按研究者、研究领域和研究主题等类别的学术信息分面检索。Williams等[25]开展了数字图书馆与学术文献搜索引擎CiteSeerχ的案例研究,集成了网络上的海量文献数据,并进行了自动抽取、聚类、实体链接和人名消歧等数据处理。

2.2　数据分析

数据分析是大数据价值链上最后和最重要的一个环节,目的是提取有用的价值,提出建设性结论和/或支撑决策。一般而言,按照数据类别可将数据分析分为6种类型(表1,根据文献[26]~[29]修改),结构化数据分析、文本分析、网站数据分析、多媒体数据分析、网络数据分析以及移动数据分析[26]。大部分数据分析属于描述性分析或预测性分析,近年来在决策过程中后者受到了越来越多的重视。根据Wlodarczyk等的定量分析结果[27],综合运用大数据和预测性分析技术的趋势显示,大数据是预测性分析背后的主要驱动因素。

目前,有许多大数据挖掘和分析工具在线可供使用,包括昂贵的商业软件/平台和Weka、KNIME等开源工具[29],其中大部分基于Java,且各平台通用。但数据分析是一个很宽泛的领域,包含不同的情景变化并且极其复杂。根据不同的数据特征和应用场景需求,数据分析算法的时空复杂性大相径庭[27]。虽然研究人员已建立了各种框架解决从数据中抽取有用知识的问题,但通常只限于有限的数据类型或特定的应用场景。因此,需要对目前的数据集成与分析方法进行详细评估和测试,经过定制化改造后,才能解决分析多学科、动态和复杂数据的挑战,从而在大数据环境下的科技情报研究新模式中灵活应用。

表1　大数据分析技术(按数据类型分类)

Table 1　List of big data analysis techniques

分析技术	数据源	研究方法	开源工具
结构化数据分析	统计数据 科学数据	聚类分析 因子分析 关联分析 回归分析 统计分析	Excel SPSS Rapidminer KNIME Weka
文本数据分析	文档 报告 论文、专利、网页中的文本信息 日志	文档表示 自然语言处理 信息抽取 主题建模 总结摘要 分类 聚类 问答系统 观点挖掘	NLTK OpenNLP GATE LingPipe Weka Rapidminer
网站数据分析	网页	网站内容挖掘 网站结构挖掘 网站用量挖掘	KXEN LIONsolver Dataiku
多媒体数据分析	图片 音频 视频	总结摘要 标注 索引和检索 推荐 事件监测	OpenIMAJ LIRE ImageTerrier
网络分析	社交网络 文献	关联预测 社团发现 社会网络演化 社会影响分析 关键词检索 分类 聚类 迁移学习	Cytoscape Gephi Cuttlefish UciNet NetDraw Pajek
移动数据分析	移动端APP 传感器 射频识别(RFID)	监控 地域挖掘	Flurry Analytics Countly Google Analytics

3　数据驱动型科技情报研究新模式

提出一种数据驱动型科技情报研究新模式的概念框架,解决传统情报研究模式的问题和局限性。这一概念框架主要利用大数据管理和分析方法改革现有的耗时耗力、依赖手工收集和分析信息的方式,能够智能获取、存储、检索、组织、处理、分析与可视化呈现海量异构数据,利用新技术建立不同数据集间的数据关联,集成和综合分析结构化和非结构化数据,从而发现有价值的知识。

3.1　整体架构

数据驱动型科技情报研究新模式的设计应考虑在大数据环境下协同工作的功能性、灵活性和可用性,设计能够收集大批量的政策、科技和产业等类型数据,包括战略规划、政策、路线图、经费预算、项目、机构、人员、研究设施、科技文献、专利、分析报告、新闻动态、统计数据等;提供特定研究领域发展趋势的精准分析和可视化呈现;具有高度灵活性可定制的资源描述、数据模型和算法,开展信息发现、遴选、组织和分析。该模

式通过采用大数据架构和工具,设计和建造从数据获取到数据存储、处理、检索和分析的全套解决方案,使情报研究人员能够快速获取所需信息,并灵活调用各种分析方法开展深度情报分析,整体架构如图2所示。

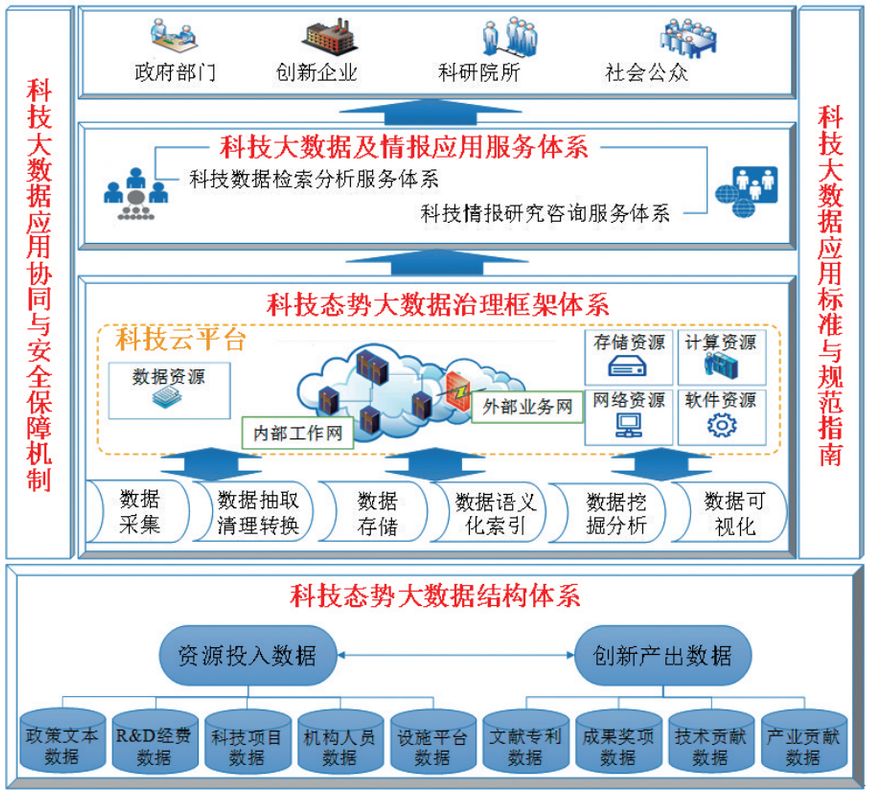

图2 数据驱动型科技情报研究新模式整体架构

Fig. 2 Overall architecture of new data-driven mode of scientific information analysis

3.2 科技态势大数据结构体系

围绕科技创新决策需求研究构建科技态势大数据结构体系,支撑科研态势分析感知环境建设,重点研究内容包括2个方面。

1)确定科技态势基础源数据构成。通过对各种异构的权威网站和数据库资源梳理分析,以及开展文献资料调研和专家咨询等方式,在政策、经费、项目、机构、人才、装备、论文、专利、成果、奖项、评价指标、产业经济、资源生态、社会环境等方面发现、遴选和评价不同来源、不同类型的高质量科技态势基础源数据。

2)从资源投入数据和创新产出数据2个维度来构建科技态势大数据体系。其中,资源投入数据维度包括:科技创新战略环境、R&D经费投入、R&D机构人员投入、R&D设施平台投入;创新产出维度包括:科技论文、发明专利、技术贡献、产业贡献等。

3.3 科技态势大数据治理框架体系

基于大数据生态架构和机器学习关键技术开发科技态势数据采集、抽取与融合、存储、索引和数据分析等先进方法,形成科技大数据治理模型和框架体系,实现对科技战略政策文本数据、科技投入产出数据、技术经济数据、环境社会影响数据等的智能采集、语义化知识组织和定量可视化分析,重点研究内容包括3方面。

1)科技态势数据采集和加工处理。对3个层面的数据资源进行收集整理,一是已经建成的科技领域专业平台系统数据库和非结构化、半结构化和结构化数据的采集和集成;二是动态科技监测实时流数据的采集;三是科技领域开放数据的采集等。通过数据清洗、格式转换、实体和关系抽取、数据汇聚和关联、有效性效验等数据处理工作,加工数据进入科技大数据云存储中心。

2)科技态势数据分布式存储。建设基于云计算的科技大数据分布式云存储系统,以支持海量科技数据资源的存储扩展。深入研究大数据文件系统的海量数据管理规范,通过海量异构数据的抽取、映射、收割、导

入等集成方法,形成清晰的大数据存储结构。

3)科技态势数据挖掘应用。研究在大数据环境下构建多种微服务集群,提供多种大数据处理架构下机器学习、数据挖掘算法和计算模型支持,对多源异构科技数据进行政策文本计算、比较分析、聚类分析、因果分析、关联分析、趋势预测等分析,实现数据深度挖掘,为科技前沿识别、态势刻画、趋势预测以及技术评估提供数据分析支撑。

3.4　科技大数据及情报应用服务体系

完善科技大数据及情报应用服务体系设计,通过构建覆盖多部门、多层次的协同服务体系,开展数据驱动型情报咨询服务,发布系列化数据分析与情报研究报告等决策支持产品,重点研究内容有3个方面。

1)建设用户情景导向的科技大数据及情报应用服务产品体系。研究构建国家政府部门、一流科研院所、创新科技企业、社会公众不同层级的典型需求模型,设计相应的精细化加工数据产品、情报报告和服务体系。

2)建设科技态势大数据及情报应用服务云平台,对内建立完善的科技数据与情报成果管理和共享机制;对外提供科技态势大数据多维检索与分析,以及情报研究咨询定制服务,提高科技大数据及情报应用服务的便利性、规范性和权威性。

3)探索研究主动对外服务和社会化传播模式。丰富科技云平台的数据在线服务和情报产品个性化定制服务,提供数据规范应用程序编程接口(API),并依托学术期刊、报纸、微信新媒体等平台,传播科技态势大数据与情报成果。

3.5　科技态势大数据运维保障支撑体系

科技态势大数据运维保障支撑体系重点研究内容包括3个方面。

1)制定完善的科技态势大数据体系标准规范,保证数据集群应用过程中各个环节正规有序,对科技态势大数据集群涉及的元数据标准、数据存储、数据共享和重用以及合理使用数据问题能够快速响应,并合理解决。

2)形成科技态势数据资源可持续收集汇聚标准流程,保障海量异构数据资源通过定期下载、采集、收割等资源获取方法形成科技态势重要方向结构化、半结构化和非结构化数据资源的完整性和有效性,构建完善的科技态势数据深加工机制。

3)建立可靠的技术支撑和支持保障机制,形成稳定的科技态势数据资源获取、数据资源组织、数据资源存储到数据资源分析应用等一系列工作的长期可持续服务机制,依托平台形成完备科技态势数据资源基础服务环境,保障服务平台的稳定运维。

4　结论

提出了一种数据驱动型科技情报研究新模式的概念框架,以建立科技大数据及情报应用服务体系为目标,以形成完备的科技大数据结构体系和有效的科技大数据治理框架体系为基础,以大数据生态圈信息技术和服务平台为支撑,以打造科技数据与情报服务产品为抓手,以数据标准和工作规范为机制保障,丰富化科技大数据资源和情报应用服务产品,全面提升科技大数据与情报应用服务的能力和水平。未来研究将进一步优化系统设计,并在解决实际决策问题的现实环境中进行评估。

1)通过开展差异化的精准用户画像和开发相应的海量异构数据治理模型,优化系统设计。长期以来,传统科技情报研究模式习惯于利用单一的数据治理模式应对所有类型用户的需求。随着大数据时代科研范式的转变,科技决策者的需求因时而异。科技大数据及情报应用服务体系需要探索建立用户画像模型库,明确界定不同决策情景下的数据需求,以及相应的数据收集和分析模型,这将有助于情报研究人员便捷调用适用于不同领域和不同用户需求的大数据分析方法和开发环境。

2)综合评估现有的数据集成、数据分析方法和开源工具,避免误用和滥用。科技大数据及情报应用服务体系需要集成多种数据集成和数据分析方法及工具,不同的方法、工具在不同数据规格和应用情景中能够发挥的功能大相径庭,如应用不当产生的分析结果反而会误导决策。为解决数据敏感性和应用场景适用性问题,需要详细调研各种方法工具适用的用户情景和数据规范标准,从用户需求和方法工具供给两方面实现适配管理。

3)探索自动构建垂直科技领域知识图谱的方法。通过提供有价值的背景领域知识,垂直科技领域知识图谱能够极大地提高传统信息处理任务(如信息抽取、检索、推荐、问答系统等)的有效性,因此对于科技情报研究而言有着重要意义。为应对决策者的不同需求,

科技大数据及情报应用服务体系需要在领域专家辅助下利用大量文本语料丰富垂直科技领域知识图谱,从而提高知识服务的效率和质量。

参考文献(References)

[1] Honavar V. The promise and potential of big data: A case for discovery informatics[J]. Review of Policy Research, 2014, 31(4): 326-330.

[2] Hey T, Tansley S, Tolle K. The fourth paradigm: Data-intensive scientific discovery[M]. Redmond, Washington: Microsoft Research, 2009.

[3] 张志强. 论科技情报研究新范式[J]. 情报学报, 2012, 31(8): 788-797.
Zhang Zhiqiang. New paradigm for S & T intelligence studies [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(8): 788-797.

[4] Clarivate Analytics. Derwent data analyzer[EB/OL]. [2018-05-06]. https://clarivate.com/products/derwent-data-analyzer.

[5] Chen Chaomei. CiteSpace[EB/OL]. (2016-10-03)[2018-05-06]. http://cluster.cis.drexel.edu/~cchen/citespace.

[6] Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making[J]. Big Data, 2013, 1(1): 51-59.

[7] Lee S, Mortara L, Kerr C, et al. Analysis of document-mining techniques and tools for technology intelligence: Discovering knowledge from technical documents[J]. International Journal of Technology Management, 2012, 60(1/2): 130-156.

[8] IDC. Big data analytics: Future architectures, skills and roadmaps for the CIO[EB/OL]. [2018-04-05]. http://triangleinformationmanagement.com/wp-content/uploads/2013/12/bigdata-idc-wp.pdf.

[9] IBM. Extracting business value from the 4 V's of big data[EB/OL]. [2017-10-26]. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data.

[10] National Institute of Standards and Technology (NIST). NIST big data interoperability framework: Volume 1, definitions[EB/OL]. (2015-10-22)[2017-09-26]. https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1.pdf.

[11] Power D. Using "Big Data" for analytics and decision support [J]. Journal of Decision Systems, 2014, 23(2): 222-228.

[12] Hilbert M. Big data for development: A review of promises and challenges[J]. Development Policy Review, 2016, 34(1): 135-174.

[13] Hendler J. Data integration for heterogeneous datasets[J]. Big Data, 2014, 2(4): 205-215.

[14] Gartner. Magic quadrant for data integration tools[EB/OL]. [2017-10-20]. https://www.gartner.com/doc/3393017/magic-quadrant-data-integration-tools.

[15] Walsh C, Rodrigue B, Mummadi Y. Data and analytics: Open source data integration tool comparison[EB/OL]. [2017-10-20]. https://www.excella.com/wp-content/uploads/2016/03/Open-Source-DI-Tool-Comparison_March2016.pdf.

[16] Hassani P. Best open source data integration tools[EB/OL]. (2017-03-25)[2017-10-08]. https://blogs.systweak.com/2017/03/best-open-source-data-integration-tools.

[17] Pentaho Corporation. Data integration-kettle[EB/OL]. [2017-10-12]. http://community.pentaho.com/projects/data-integration.

[18] Talend. Talend open studio for data integration[EB/OL]. [2017-10-12]. https://www.talend.com/download/talend-open-studio/#t4.

[19] Stanford University. DeepDive[EB/OL]. [2017-10-16]. http://deepdive.stanford.edu.

[20] Zhang C, Ré C, Cafarella M, et al. DeepDive: Declarative knowledge base construction[J]. Communications of the ACM, 2017, 60(5): 93-102.

[21] The University of Arizona. Data infrastructure buildings blocks (DIBBs) for intelligence + security informatics (ISI) research and community[EB/OL]. [2017-10-20]. https://ai.arizona.edu/research/dibbs#introduction.

[22] Ma B, Jiang T, Zhou X, et al. A novel data integration framework based on unified concept model[J]. IEEE Access, 2017, 5: 5713-5722.

[23] Daraio C, Lenzerini M, Leporelli C, et al. Data integration for research and innovation policy: An ontology-based data management approach[J]. Scientometrics, 2016, 106(2): 857-871.

[24] 孟小峰, 杜治娟. 大数据融合研究:问题与挑战[J]. 计算机研究与发展, 2016, 53(2): 231-246.
Meng Xiaofeng, Du Zhijuan. Research on the big data fusion: Issues and challenges[J]. Journal of Computer Research and Development, 2016, 53(2): 231-246.

[25] Williams K, Wu J, Choudhury S, et al. Scholarly big data information extraction and integration in the CiteSeerχ digital library[C]//Proceeding of IEEE 30th International Conference on Data Engineering (ICDE). Piscataway NJ: IEEE 2014: 68-73.

[26] Hu H, Wen Y, Chua T, et al. Toward scalable systems for big data analytics: A technology tutorial[J]. IEEE Access, 2014, 2: 652-687.

[27] Chen M, Mao S, Liu Y. Big data: A survey[J]. Mobile Networks and Applications, 2014, 19(2): 171-209.

[28] Kambatla K, Kollias G, Kumar V, et al. Trends in big data analytics[J]. Journal of Parallel and Distributed Computing, 2014, 74(7): 2561-2573.

[29] Yaqoob I, Hashem I A T, Gani A, et al. Big data: From beginning to future[J]. International Journal of Information Management, 2016, 36(6): 1231-1247.

[30] Wlodarczyk T, Hacker T. Current trends in predictive analytics of big data[J]. International Journal of Big Data Intelligence, 2014, 1(3): 172-180.

New mode for scientific information analysis in the big data era

CHEN Wei[1], YANG Rui[1], HE Tao[1], WANG Shuo[1], CHEN Jiangping[2]

1. Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan 430071, China
2. Department of Information Science, University of North Texas, Denton, Texas 76203, USA

Abstract The era of big data brings both opportunities and challenges to the scientific information analysis (SIA) and the intelligent information services. It is an urgent task developing a new SIA framework to reform the data governance and the workflow, and to improve the quality of the information services. This paper describes the traditional SIA mode currently applied. It also reviews the progresses in the fields of the massive heterogeneous data integration, the data management and the analytics methods, and their applications. An integration-based conceptual framework for the SIA is proposed through an examination of the limitations of the traditional mode. The new framework is characterized by the development of an Intelligent Decision-making Support System based on the Big Data that can store, organize, process, and visualize heterogeneous data. Next, the functions and the characteristics of the proposed framework are explained. The paper concludes with a discussion of the future research.

Keywords scientific information analysis; big data; data-driven; data integration; data analytics

（责任编辑　傅雪）