● 赵建保[1]，黄晓斌[2]
（1. 广东农工商职业技术学院 计算机系，广州 510507；2. 中山大学 资讯管理学院，广州 510006）

基于 Citespace 的大数据研究可视化分析

［关键词］大数据；CiteSpace；可视分析；知识图谱

［摘 要］以 ISI Web of Knowledge 数据库中 2008～2014 年间大数据为主题的 1547 条引文为研究对象，并以 CiteSpace 作为信息可视化工具，绘制了国家、机构和研究热点知识图谱，揭示了大数据的学科属性、研究力量、研究演进和研究热点。

［中图分类号］G250.252；G255.76
［文献标志码］B
［文章编号］1005－8214(2015)10－0054－04
DOI:10.14064/j.cnki.issn1005-8214.2015.10.018

随着移动互联网、物联网、社交网络等技术和应用的兴起，信息化与工业化的深度融合，数据产生已经从被动转向了自动阶段，数据源越来越多，数据精度越来越高，数据呈现了规模性（Volume）、多样性（Variety）、高速性（Velocity）、真实性（Veracity）、价值性（value）、汇聚性（Aggregate）的特征，大数据必将广泛应用于金融、商业、科学研究、消费行业等领域。已有的数据集成、数据存储、数据分析模式已难以满足大数据的需求，理清学界业界近几年大数据研究力量、研究路径和研究热点，对科研管理、决策和开发尤其必要。

1 文献检索与计量分析

2014 年 8 月 27 日使用检索式为“TOPIC:（big+data） Timespan：2008-2014. Indexes: SCI-EXPANDED, CPCI-S, CPCI-SSH.”对 Web of Science 进行主题检索，2008~2014 年共发表 1547 篇文献；其中 2008~2011 年 72 篇，2012~2014 年 1475 篇；2012 年 233 篇，2013 年 859 篇，2014 年 383 篇，从 2012 年以来大数据研究力量骤增，研究成果较 2011 年增长了 9 倍多。

从 WoS 提供的研究领域划分看，计算机科学 881 篇，工程 536 篇，电信 125 篇，说明大数据学科性质是计算机科学技术。从文献类型方面会议论文（PROCEEDINGS PAPER） 807 篇，期刊论文（ARTICLE） 472 篇，其他类型文献 279 篇。

2 大数据研究力量分析

设置 CiteSpace 参数生成 2008～2014 年间国家合作图谱，显示了大数据研究主要有美国（572 篇）、中国（248 篇）、德国（72 篇）、英国、韩国、澳大利亚、日本等，美国和中国大数据研究起步较早，发文量较大。从国家合作看，国家间合作普遍开始于 2013 年之后，国家间合作呈现非网络结构，说明国家间合作以单边合作为主，多边合作较少。

设置 Citespace 参数生成机构合作图谱，显示国内外主要大数据研究机构有中国科学院、麻省理工学院、南加利福尼亚大学和加州大学洛杉矶分校等，研究机构发文量统计如表 1 所示。

表 1 研究机构发文量排名

发文量	机构名称（中英文）
31	Chinese Acad Sci（中国科学院）
21	MIT（麻省理工学院）
16	Univ SO CALIF(南加利福尼亚大学)
15	Univ Calif Los Angeles（加州大学洛杉矶分校）
14	Stanford Univ(斯坦福大学)
13	Harvard Univ（哈佛大学）
12	Univ Technol Sydney（悉尼科技大学）

可划分为以中国科学院、麻省理工学院和南加利福尼亚大学三大学术合作群体。其中，中国科学院与北京大学、北京理工大学等研究机构开展了合作，麻省理工学院与加州理工学院（CALTECH）、卡内基梅隆大学（Carnegie Mellon University）等研究机构开展了合作，南加利福尼亚大学跟加州大学洛杉矶分校（Univ Calif Los Angeles）等研究机构开展了合作。

［基金项目］本文系 2010 年国家社会科学基金项目“网页内容分析与挖掘的企业竞争情报方法研究”（项目编号：10BTQ034），广东省教育科学“十二五”规划教育信息技术研究专项课题“构建适应项目化教学的网络课程系统研究”（项目编号：12JXN020）的成果之一。

从大数据研究的代表人物看，排前3位的分别是Jeffrey Dean、Tom White和Angela Hung Byers。Jeffrey Dean是Google公司Knowledge Group研究员，2009年当选美国工程院院士，研究方向为大规模分布式系统、信息检索、机器学习等。1999年加入Google后参与了Google广告服务系统、Google爬虫、索引和查询服务系统、MapReduce、BigTable等众多Google的核心产品设计和实现。主要学术研究成果有*Mapreduce: Simplified data processing on large clusters*，*MapReduce: a flexible data processing tool*和*Bigtable: A Distributed Storage System for Structured Data*等。其中，*Mapreduce: Simplified data processing on large clusters*的谷歌学术显示的被引数高达11505次，影响力极高。Tom White是畅销书*Hadoop: The Definitive Guide*的作者，从2007年2月担任Apache Hadoop项目负责人，是Apache软件基金会的成员之一。Angela Hung Byers是2011年麦肯锡全球研究院调研报告《大数据：创新、竞争和生产力的下一个新领域》的项目负责人。

3 大数据研究演进分析

演进路径是研究领域的知识基础和前沿随时间演进的动态过程。知识基础以经典文献和关键文献为骨架构成，为研究领域演进提供动力和基础。2008～2013年经典文献如表2所示。

表2 大数据研究领域经典文献

年份	作者	标题	频次
2008	Dean J	Mapreduce: Simplified data processing on large clusters	112
2008	Howe D	Big data: The future of biocuration	20
2008	Lynch C	Big data: How do your data grow?	18
2009	White T	Hadoop: The Definitive Guide	24
2009	Hey Tony	The Fourth Paradigm: Data-Intensive Scientific Discovery	18
2009	Schatz MC	Cloudburst: highly sensitive read mapping with mapreduce	14
2010	Schadt EE	Computational solutions to large-scale data management and analysis	22
2010	Ekanayake J	Twister: A runtime for iterative MapReduce	17
2010	Dean J	MapReduce: a flexible data processing tool	16
2011	Byers A H	Big data: the next frontier for innovation, competition, and productivity	40
2011	Lavalle S	Big data, analytics and the path from insights to value	11
2011	Trelles O	Big data, but are we ready?	10
2012	White T	Hadoop—The Definitive Guide: Storage and Analysis at Internet Scale	17
2012	Chaudhuri S	What Next? A Half-Dozen Data Management Research Goals for Big Data and the Cloud	11
2012	Chen HC	Business Intelligence And Analytics: From Big Data To Big Impact	11
2013	Murdoch TB	The Inevitable Application of Big Data to Health Care	12
2013	Marx V	The Big Challenges Of Big Data	9
2013	Cukier K	Big Data: A Revolution That Will Transform How We Live, Work and Think	8

结合WoS大数据文献分布规律，参照新兴技术研究的特点和发展范式，可把2014年之前的大数据研究划分为萌生期（1980～2008）和快速发展期（2009～2013）二个阶段。

萌生期（1980～2008年）。1980年3月，Alvin Toffler在《第三次浪潮》（*The Third Wave*）一书中预言大数据将是“第三次浪潮的华彩乐章”。2008年1月，Google公司Jeffrey Dean和Sanjay Ghemawat在*Communications of the ACM*发表了*Mapreduce：Simplified data processing on large clusters*，以谷歌大数据处理为例介绍了MapReduce编程模型在处理各种大数据任务的可用性及数据处理模式，即程序员通过指定Map函数和Reduce函数，底层系统会自动实现大规模集群的并行计算，并自动处理机器故障和调度机间的通信，有效地利用网络和磁盘资源。[1] 9月Nature推出了大数据专刊*Big Data: Science in the Petabyte Era*正式提出了大数据概念，[2] Doug Howe等在专刊中发表*Big data: The future of biocuration*文章，提出应对生物学大数据的3项行动倡议，即出版物和数据库之间的数据交换、建立权威的数据标准和设置数据策划岗位。Clifford Lynch专刊中发表*Big data: How do your data grow?*评论，阐述了实现数据重用的前提是保存数据，讨论了数据管理的体制与机制。[3] 12月，Bryant、Katz和Lazowska三位信息领域资深科学家联合“计算社区联盟（Computing Community Consortium）”发表了《大数据计算：商务、科学和社会领域的革命性突破》（*Big-Data Computing：Creating revolutionary breakthroughs in commerce，science and society*）白皮书，阐述了在数据驱动的研究背景下，解决大数据问题所需的技术以及面临的一些挑战。由此可见，在大数据萌生期主要研究重点是大数据的应用前景及面临的技术问题。

快速发展期（2009～2013年）。2009年6月，Schatz在*Cloudburst：highly sensitive read mapping with mapreduce*中介绍了基于MapReduce的CloudBurst并行算法用于分析人体基因组数据的良好性能；10月，Hadoop开源项目负责人Tom White著《Hadoop权威指南》（*Hadoop: The Definitive Guide*），全面介绍了MapReduce编程技术及部署要求，为MapReduce的后续研究和应用提供了权威指导；同月，微软研究院副总裁Tony Hey博士在*The Fourth Paradigm: Data-Intensive Scientific Discovery*一书中通过分析众多数据密集型科学研究实例提出了科学研究的第四范式，即科学研究将从以计算为中心转变到以数据处

理为中心；2010 年 1 月，Jeffrey Dean 在 *MapReduce: a flexible data processing tool* 中阐述了 MapReduce 在大数据处理中具有良好的容错性、异构存储系统加载和处理数据的便捷性以及为执行复杂函数提供了良好的架构；6 月，Ekanayake 在 *Twister: A runtime for iterative MapReduce* 中提出了支持跌代计算的 MapReduce 编程模型 Twister 及体系结构，并比较了 Twister、Hadoop 与 DryadLING 在海量数据并行处理的性能。9 月 Schadt 等发表 *Computational solutions to large- scale data management and analysis* 文章，以生命科学中基因组大数据为例提出了云计算和异构计算来处理海量和高维数据集的方案。2011 年 2 月 Science 杂志出版专刊 *Dealing with Data*，主要讨论了科学研究中大数据的问题及其重要性。[4] 3 月 Trelles 等发表文章 *Big data, but are we ready?* 指出计算节点间的数据通信将成为生物信息学研究中瓶颈，提出了通过云计算和异构框架克服硬件瓶颈（如开发高速并行 I/O 来缩短存储与计算间的路径，整合光电通信技术提高高维数据传输速度），而通过多处理器来克服软件瓶颈。[5] 5 月麦肯锡全球研究院 Byers 等发布调研报告《大数据：创新、竞争和生产力的下一个新领域》（*Big data：the next frontier for innovation，competition，and productivity*），分析了大数据的影响、关键技术和应用领域，明确提出了政府和企业决策者应对大数据发展的策略。同年 5 月 EMC 公司董事长兼首席执行官乔图斯在 EMC World 2011 拉斯维加斯大会主题为“云计算适逢大数据”，阐述了云计算与大数据的理念和技术趋势。6 月由 EMC 赞助的 IDC 数字宇宙研究《从混沌中提取价值》（*Extracting Value from Chaos*）提到三点重要论断：全球数据量大约每两年翻一番；2010 年全球数据量跨入 ZB 时代，预计 2011 年全球数据量将达到 1.8ZB；未来全球数据增速将会维持，预计到 2020 年全球数据量将达到令人恐怖的 35ZB。[6] 10 月 Gartner 将大数据列入 2012 年十大战略新兴技术。2012 年 1 月，瑞士达沃斯世界经济论坛发布报告《大数据，大影响》（*Big Data, Big Impact*）指出数据已经成为一种新的经济资产类别。2012 年 3 月美国奥巴马政府推出了大数据研究和发展计划（*Big Data Research and Development Initiative*），投资两亿多美元推动大数据相关的采集、组织、分析、决策工具及技术研究，计划将大数据技术用于高科技领域。5 月，Tom White 在 *Hadoop—The Definitive Guide: Storage and Analysis at Internet Scale* 书中介绍了构建可靠、可扩展的 Apache Hadoop 分布式系统，为程序员分析数据和管理员配置和运行 Hadoop 集群提供了权威指导。在第三版中也增加了 MapReduce API、MapReduce2 和 YARN 的部分。5 月微软研究院的 Surajit Chaudhuri 在 *What Next? A Half-Dozen Data Management Research Goals for Big Data and the Cloud* 中描述了基于大数据和云计算的数据管理研究面临隐私保护（Data Privacy）、近似查询结果（Approximate Results）、数据探索与分析（Data Exploration To Enable Deep Analytics）、企业数据集成（Enterprise Data Enrichment）、面向租户进行性能隔离（Performance Isolation For Multi-Tenancy）的 6 个挑战。12 月，Chen 等在 *MIS QUARTERLY* 发表文章 *Business Intelligence and Analytics: From Big Data to Big Impact*，采用文献计量学研究了商务智能分析领域的演进、应用、前沿及研究框架。2013 年 3 月，Cukier 在 *Big Data: A Revolution That Will Transform How We Live, Work and Think* 一书中，前瞻性地指出大数据带来的信息风暴正在变革我们的生活、工作和思维，分三个部分讲述了大数据时代的思维变革、商业变革和管理变革。明确指出放弃对因果关系的渴求而关注相关关系，大数据的核心就是预测。书中展示了谷歌、微软、亚马逊、IBM 等大数据先锋们最具价值的应用案例。4 月，Murdoch 在 *The Inevitable Application of Big Data to Health Care* 中讨论大数据在卫生保健中的应用，借助经济模型强调了应用中将面临的机遇和挑战，建议通过加强病人和医生数据的收集来提高卫生保健的服务质量和效率。6 月，Marx 在 *the Big Challenges of Big Data* 中介绍了生命科学大数据的增长态势，指出了存储和分析异构复杂数据面临的挑战以及云计算在生命科学大数据的应用。由此可见，在大数据快速发展期主要研究重点是大数据处理的生态系统构建及业界学界的行业产业应用实践。

历经 Toffler 的大数据预言，Dean、White、Byers、Murdoch 等一大批研究者的研究探索，大数据研究主题以大数据的应用前景、大数据概念、大数据生态系统构建和业界学界应用落地为主线，呈现了大数据研究与大数据应用交织演进的态势。可以预见，2014 年后，大数据研究开始转向行业领域应用系统集成、大数据分析、管理及生态系统优化方向。

4 大数据研究热点分析

研究热点可通过引文的主题词出现频率来探测。

设置 CiteSpace 参数生成 2012～2014 大数据研究热点图谱（见下图）。

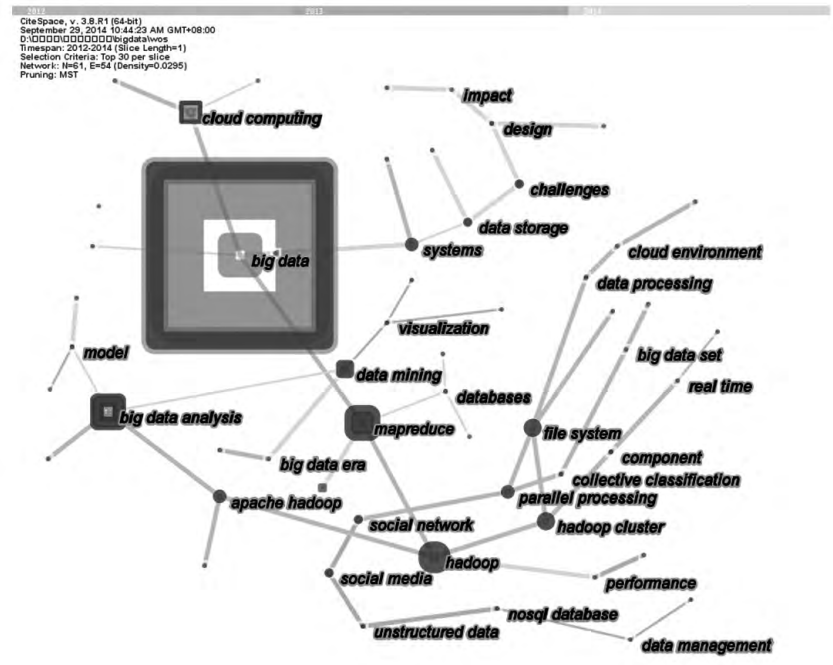

图　2012～2014 大数据研究热点图谱

图谱中的方形结点表示主题词，文字是主题词标签，节点的大小代表出现的频次。从研究热点的年度分布看，2012 年大数据研究的热点是 hadoop 生态系统，2013 年度热点是异构数据的管理和可视化技术，2014 年研究热点是大数据分析及生态系统的完善和体系化。将热点主题词进行同义词合并，得出大数据研究主要主题词排序，依次是大数据（big data）、大数据分析（big data analytics）、云计算（cloud computing）、mapreduce、数据挖掘（data mining）、hadoop、大数据应用（big data application）、模型（model）、机器学习（machine learning）、大数据时代（big data era）、系统（systems）和社交媒体（social media），big data（大数据）的节点最大，这跟本身是检索主题词有关。（见表 3）。

表 3　2012～2014 大数据研究热点

频率	热点词	频率	热点词
1003	big data	40	big data application
150	big data analytics	38	model
123	cloud computing	38	machine learning
102	mapreduce	37	big data era
74	data mining	36	systems
60	hadoop	35	social media

热点词 big data analytics（大数据分析）指根据分析主题需求，基于云计算技术，采用数据挖掘、机器学习、统计分析等数据分析方法，发现大数据价值的过程。从大数据分析支撑技术来看，大数据中绝大部分都是半结构化和非结构化的数据，传统的关系型数据库缺乏可扩展性已经无法进行分析处理，而以 mapreduce 实现分析处理和以 GFS、HDFS 为代表的分布式文件系统具有良好的横向扩展能力，现已成为大数据分析的主流技术。大数据分析是整个大数据处理流程的核心，通过分析过程发掘大数据价值并将其应用到推荐系统、商业智能、决策支持等诸多领域。热点词 cloud computing（云计算）为大数据存储、管理以及数据分析等提供支撑和基础平台。云计算是一种大规模的分布式模型，通过网络将抽象的、可伸缩的、便于管理的数据能源、服务、存储方式等传递给终端用户，[7] 最典型的就是以分布式文件系统 GFS、批处理技术 mapreduce、分布式数据库 BigTable 为代表的大数据处理技术以及在此基础上产生的开源数据处理平台 Hadoop。云计算从技术层面强调单个节点的计算能力最大化，大数据从效用层面强调数据价值最大化。热点词 mapreduce 是 Google 公司和 Hadoop 开源软件框架共有的核心计算模型。大数据处理模式主要有流处理和批处理两种，流处理是直接处理，而批处理则是先存储后处理。流处理应用场景主要有网页点击数的实时统计、传感器网络、金融中的高频交易等，比较代表性的开源系统如 Twitter 的 Storm、Yahoo 的 S4 以及 Linkedin 的 Kafka 等。批处理模式应用场景主要有离线和近线处理，mapreduce 是最具代表性的批处理模式，其核心思想在于“分而治之”，把计算推到数据而不是把数据推到计算，有效地避免数据传输过程中产生的大量通信开销。mapreduce 将运行大规模集群上的复杂的并行计算过程高度地抽象为 Map 和 Reduce 两个函数，mapreduce 模型首先将用户的原始数据源进行分块，然后分别交给不同的 Map 任务区处理。Map 任务从输入中解析出链/值（Key/Value）对集合，然后对这些集合执行用户自行定义的 Map 函数得到中间结果，并将该结果写入本地硬盘。Reduce 任务从硬盘上读取数据之后会根据 key 值进 （下转第 95 页）

并针对该课题提供信息检索、信息推送、参考咨询等定题服务。③在学生就业阶段各对口学科服务小组要全程跟进，上网搜索并时时关注学院等各类招聘信息，及时利用“学科微博”或“学科微信”平台发出相关招聘单位的信息。要特别关注弱势学生的应聘情况，辅导他们撰写简历，进行面试培训。对自信心不足的学生要一对一地进行面试辅导和培训，发现适合他们的招聘单位要在第一时间内告之，全力帮助他们找到工作。

[参考文献]

[1] 黎清．高校图书馆个性导向服务趋势变革与策略提升[J]．图书馆理论与实践，2015(3)：74－79.

[2] 刘建国．高校图书馆弱势群体信息资源配置研究[J]．图书馆理论与实践，2013(8)：76－78.

[作者简介] 刘恒波(1970－)，女，宁夏财经职业技术学院图书馆馆员。

[收稿日期] 2015－05－26 [责任编辑] 菊秋芳

(上接第57页) 行排序，将具有相同Key值的组织在一起，最后用户自定义的Reduce函数会作用于这些排好序的结果并输出最终结果。[8] data mining(数据挖掘)是数据分析师针对业务分析需求，利用各种分析工具从海量数据中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。数据挖掘的任务有分类与回归、聚类、关联规则、时序模式、偏差检测五个方面。数据挖掘过程包括定义挖掘目标、数据取样、数据探索、预处理、模式发现、模型构建、模型评价七个步骤，常用的数据挖掘工具有SAS Enterprise Miner、SPSS Clementine、IBM Miner、MATLAB、WEKA。[9] 热点词Hadoop是目前最为流行的大数据处理平台，已经发展成为包括文件系统(HDFS)、数据库(HBase)、数据处理(MapReduce)等功能模块在内的完整生态系统(Ecosystem)，Hadoop已经成为大数据处理工具事实上的标准。

从大数据处理流程来看，大数据处理流程可划分为数据采集、数据处理与集成、数据分析和数据解释四个阶段，[10] 研究热点中大数据分析、云计算、mapreduce和数据挖掘都属于大数据分析环节。从大数据生态系统来看，大数据分析和数据挖掘都属于大数据分析的范畴，是实现大数据价值的前提，云计算和mapreduce都属于云计算的范畴，为大数据提供了存储和分布式计算，由此说明，支撑大数据系统的基础平台和大数据分析是大数据研究的最热门主题。

[参考文献]

[1] Dean J，Ghemawat S．Mapreduce：Simplified data processing on large clusters[J]．Communications of The ACM，2008，51(1)：107－113.

[2] Nature．Big data：Science in the petabyte Era[EB/OL]．[2014－10－13]．http://www.nature.com/nature/journal/v455/n7209/edsumm/e080904-01.html.

[3] Lynch C．Big data：How do your data grow?[J]．nature，2008(455)：28－29.

[4] Science．Special online collection：dealing with big data[EB/OL]．[2014－10－13]．http://www.sciencemag.org/site/special/data/.

[5] Trelles O，et al．Big data，but are we ready?[J]．Nature Reviews Genetics，2011(12)：224.

[6] IDC．Extracting Value from Chaos[EB/OL]．[2014－09－18]．http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[7] Foster I，et al．Cloud computing and grid computing 360-degree compared[C]//Proceedings of the Grid Computing Environments Workshop 2008(GCE '08)．Austin：IEEE，2008：1－10.

[8] 孟小峰，慈祥．大数据管理：概念、技术与挑战[J]．计算机研究与发展，2013，50(1)：146－169.

[9] 张良均，等．数据挖掘：实用案例分析[M]．北京：机械工业出版社，2013，6.

[10] 刘智慧，张泉灵．大数据技术研究综述[J]．浙江大学学报(工学版)，2014，40(6)：957－972.

[作者简介] 赵建保(1978－)，男，广东农工商职业技术学院计算机系讲师，研究方向：可视化、可视分析和Web工程；黄晓斌(1961－)，男，中山大学资讯管理学院教授，博士生导师，研究方向：竞争情报、网络信息开发利用。

[收稿日期] 2014－11－17 [责任编辑] 刘 丹