

大数据研究的现状、热点演进与前沿分析

刘飞¹, 牛丽慧², 金昌博³

(1. 新疆医科大学图书馆, 乌鲁木齐 820054; 2. 山西大学经济与管理学院, 太原 030006;

3. 广州医科大学, 广州 511436)

摘要: 搜集2009-2014年Web of Science核心合集的引文数据库中大数据领域的文献, 进行计量及可视化分析。利用CiteSpaceII对该领域文献做作者合作分析、共词分析和共被引分析, 结果显示不同国家之间的研究差距较大, 美国是该领域的中心国家及主要力量, 中国在该领域有一定的积累, 但仍待进一步发展。大数据相关技术、大数据带来的变革性影响及大数据引起的负面效应是该领域的主要研究前沿。

关键词: 大数据; 可视化; Map-Reduce; 个人隐私

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2015.06.010

1 引言

大数据是指一类大规模的数据组, 其获取、存储、管理和分析的能力通常超越那些传统的数据分析工具^[1]。2009年在甲型H1N1流感爆发前几周, 谷歌基于大数据技术准确预测出流感的来源, 这一年“大数据”也逐渐成为互联网信息技术行业的流行词汇。2012年3月22日, 美国政府宣布投资2亿美元拉动大数据相关产业, 将“大数据战略”上升为国家战略^[2]。为了解大数据领域的研究现状, 本文检索Web of Science引文数据库中大数据领域的文献并进行可视化分析, 了解该领域的研究现状与趋势, 为该领域的研究提供理论参考。

2 数据来源和方法

选择Web of Science核心合集的引文数据库, 以“big data”为检索词, 检索2009-2014年有关大数据的文献, 共2626篇。应用信息可视化工具CiteSpaceII^[3]对所采集的数据进行分析。

利用该软件中的分治策略, 将数据分为6个时间段, 即每年为一个时间切片, 以提高软件的运行效率以及识别出学科前沿的时态模式。进而, 分别采用共词分

析、共被引分析、合作分析, 对数据进行挖掘, 考察该领域的知识基础和研究前沿。可视化计算过程中, 选择夹角余弦算法计算节点连接强度, 筛选出引用频次超过50的节点, 并用最小生成树算法对网络进行关键信息筛选。

3 结果与分析

3.1 大数据研究领域的文献分析

3.1.1 趋势分析

本文共检索到2009年1月1日至2014年12月31日发表的关于大数据的文献2626篇, 总被引频次2678次, 每年平均引用次数382.57次, 各年发文量见表1。从总体发文趋势看, 大数据研究领域的国际核心期刊论文量呈快速增长趋势。

表1 2009-2014年大数据研究领域发表文献的年代分布

年份	2009	2010	2011	2012	2013	2014
发文量(篇)	16	13	28	262	1038	1269

3.1.2 国家和地区分布

在检索结果中, 2009-2014年间共有61个国家和地区发表了大数据研究领域的相关文献, 排序前10位的国家见表2。其中美国1029篇, 居于第1位, 占全部文献量的39.185%; 中国446篇, 排名第2位, 占全部文献量的16.984%; 其后, 依次英国162篇, 德国141篇, 澳大利亚103篇。在大数据研究领域不同国家之间的研究差距较大, 其中美国的发文成果超过全部发文量的1/4。根据日本学者汤浅光朝的观点, 即当某个国家的科学论文作者人数和科学成果数超过全世界的1/4, 则科学活动的中心转入该国^[5]。据此, 美国是该领域的中心国家及主要力量, 中国在大数据研究领域有一定积累, 但仍待进一步发展。

表2 2009-2014年大数据研究领域
发表文献量排名前10位的国家

排序	国家	发文量	所占比例
1	美国	1029	39.185%
2	中国	446	16.984%
3	英国	162	6.169%
4	德国	141	5.369%
5	澳大利亚	103	3.922%
6	韩国	94	3.580%
7	加拿大	89	3.389%
8	日本	89	3.389%
9	意大利	66	2.513%
10	荷兰	62	2.361%

3.1.3 研究机构分布

大数据领域的研究共涉及1556家机构, 发文量在10篇以上的研究机构有54家, 占有所有研究机构的3.47%, 这54家研究机构共发文852篇, 占总发文量的32.4%。

表3为发文排名前10位的研究机构, 主要分布在美国和中国, 美国占8所, 中国占2所。中科院、麻省理工学院、哈佛大学位列前3。

3.1.4 作者分布

在搜集到的2626篇文章中共有7958个作者, 其中

表3 2009-2014年大数据研究领域发表
文献量排名前10位的研究机构

排序	机构	国家	发文量	所占比例
1	中科院 (CHINESE ACAD SCI)	中国	1029	2.133%
2	麻省理工学院 (MIT)	美国	446	1.409%
3	哈佛大学 (HARVARD UNIV)	美国	162	1.257%
4	南加利福尼亚大学 (UNIV SO CALIF)	美国	141	1.142%
5	斯坦福大学 (STANFORD UNIV)	美国	103	1.028%
6	加州大学洛杉矶分校 (UNIV CALIF LOS ANGELES)	美国	94	1.028%
7	加州大学圣地亚哥分校 (UNIV CALIF SAN DIEGO)	美国	89	0.914%
8	清华大学 (TSINGHUA UNIV)	中国	89	0.876%
9	华盛顿大学 (UNIV WASHINGTON)	美国	66	0.762%
10	明尼苏达大学 (UNIV MINNESOTA)	美国	62	0.724%

表4 2009-2014年大数据研究领域发表
文献量排名前10位的高被引作者

排序	作者	被引频次
1	DEAN J	311
2	MANYIKA J	106
3	WHITE T	92
4	ZAHARIA M	71
5	STONEBRAKER M	60
6	GHEMAWAT S	58
7	MAYER-SCHONBERGE V	56
8	THUSOO A	56
9	CHANG F	55
10	SHVACHKO K	49

发文量最多的作者是悉尼科技大学工程学院的Jinjun Chen, 发文量11篇。进一步, 对作者被引用频次进行统计, 表4为被引用频次最多的前10位作者。被引次数最高的作者DEAN J系美国计算机科学家和软件工程师, 谷歌系统和基础设施集团的高级研究员, Map-Reduce、BigTable 等系统的创造者, 2008年与其合作者Ghemawat S发表的“Map-Reduce: simplified data processing on large clusters”一文, 为大数据技术的发展做出了不可磨灭的贡献。

3.2 大数据研究领域的知识基础、研究热点及前沿分析

3.2.1 大数据研究领域的知识基础

一个学科的知识基础是对应于研究前沿的所有前期文献集合,研究前沿和知识基础之间的关系本质上是对应的。不提及对方,它们任何一方都无法充分定义自我^[4]。有关大数据领域在2009-2014年所收集到2626篇文献共有193861篇有效参考文献,其被引频次 ≥ 5 的有68篇,被引频次 ≥ 20 的有20篇。为了进一步了解大数据研究前沿的知识基础,使用Citespace对有效参考文献进行文献共被引分析,选择“cited reference”作为节点类型,生成共被引网络图谱,见图1。被引次数最高的4篇文献见表5。

图1节点最大的是被引次数最高的文献,达182次,是大数据领域最经典的文献,即:Dean J,及Ghemawat S 2008年在Communications of the ACM期刊上发表的题为“Map-Reduce: simplified data processing on large clusters”的论文^[6],该文献详细解释了Map-Reduce的工作方式,对大数据研究产生了重大影响,是

大数据技术发展的奠基性文献。

被引次数排名第二的文献是Manyika J, Chui M, Brown B等2011年6月发表在Mckinsey Global Institute上题为“Big data: The next frontier for innovation, competition, and productivity”^[1]。该文献被引次数为80次,论述了大数据在卫生保健、公共管理、商品零售、制造业、个人数据方面带来的巨大影响,并提出大数据在五个方面创造价值;被引次数排名第三位的是Mayer-Schönberger V与Cukier K 2013年在Houghton Mifflin Harcourt上发表的论文《Big data: A revolution that will transform how we live, work, and think》^[8],该文献论述了大数据带给人们生活,工作与思维的影响与变革。第四位高被引文献是Bell G Hey T和Szalay A 2009年在Science期刊上发表的题为“Beyond the data deluge”的文献^[9],论述了大数据对各学科发展的影响。

3.2.2 大数据研究领域的研究热点

利用CiteSpace软件,将网络节点选为keyword,得到关键词共词网络图谱,见图2。图中共212个节

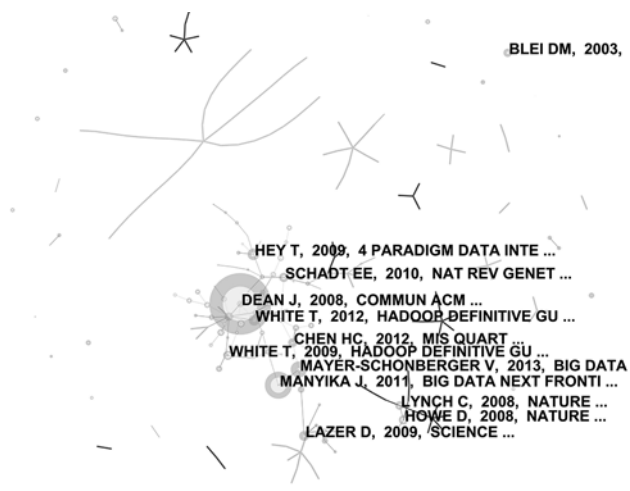


图1 2009-2014年大数据研究领域共被引网络图谱

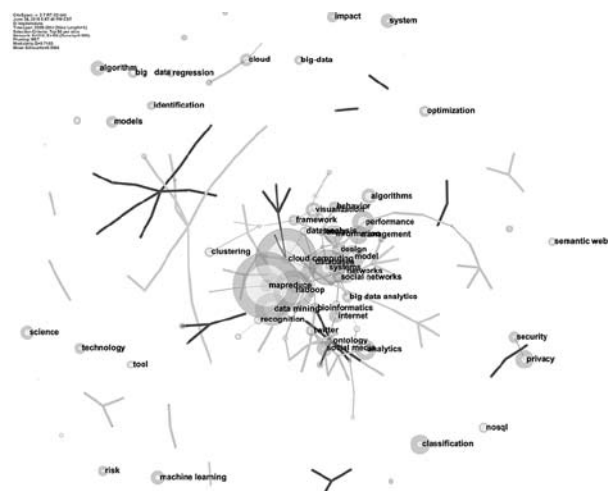


图2 2009-2014年大数据研究领域关键词共词图谱

表5 2009-2014年大数据研究领域高被引文献

排序	被引次数	年份	经典文献	所在期刊
1	182	2008	Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters	Communications of the ACM
2	80	2011	Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity	Mckinsey Global Institute
3	48	2013	Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think	Houghton Mifflin Harcourt
4	36	2009	Bell G, Hey T, Szalay A. Beyond the data deluge	Science

点, 413条连线。较大的关键词节点主要有: 云计算 (cloud computing)、分布式计算 (Map-Reduce)、hadoop、数据挖掘 (data mining)、模型 (model)、隐私 (privacy)、社交媒体 (social media)、算法 (algorithms)、机器学习 (machine learning) 等。

为了更清晰地分析出2009-2014年大数据领域的热点的演进情况, 分别截取各年度Top1高频关键词及与每个关键词共现的词进行统计分析。每年Top1高频关键词见表6。大数据的研究热点演进情况可总结如下: 2009年主要关注的研究热点有云计算、算法、生物信息学、本体及网络等。2010年大数据关注点聚焦在疾病、教育、合作应用以及云计算、分布式计算系统等方面。2011年研究热点是与大数据技术相关的系统、网络、设计和通过大数据进行人类行为研究等。2012年主要关注大数据技术及算法, 高频关键词有分布式计算、Hadoop、数据挖掘、大数据模型及分析等。2013和2014年开始关注大数据带来的负面效应, 即个人隐私, 在技术方面有机器学习、系统优化及建模等高频词。

表6 2009-2014年大数据研究领域各年份Top1高频词

关键词	频次	年份
Cloud computing	134	2009
disease	17	2010
systems	77	2011
Map-Reduce	150	2012
privacy	42	2013
security	26	2014

3.2.3 大数据研究领域的研究前沿

1994年, 加菲尔德指出“研究前沿是被引频次最高的核心文献与引证这些核心文献的来源文献的集合, 前沿的名称可以从来源文献标题中出现频次最高的单词或词组中提取出来”^[11]。在CiteSpaceII中, 对研究前沿的侦测分析是基于从题目、摘要、系索词 (标引文献的单元词或词组) 和文献记录的标示符中提取出的突变专业术语 (burst terms) 而确定的^[12]。下面分别通过探测高突变专业术语及高突现性论文来分析大数据领域的研究前沿。图3为截取涌现值排序前5位突变专业术语, 分别有: 大数据集 (big data set)、高性能计算 (high performance computing)、Hadoop 集群 (Hadoop cluster)、数据分析 (data analytics)、半导体物理学 (semiconductor physics) 等, 从图中可以看出各突变词首次出现的年份以及其突变的时间段。

Top 5 term with Strongest Citation Bursts

term	Year	Strength	Begin	End	2009-2014
big data set	2009	2.8016	2009	2010	■■■■■■■■■■
high performance computing	2009	2.4166	2012	2012	■■■■■■■■■■
hadoop cluster	2009	2.0875	2012	2012	■■■■■■■■■■
data analytics	2009	2.0875	2012	2012	■■■■■■■■■■
semiconductor physics	2009	1.9603	2010	2010	■■■■■■■■■■

图3 2009-2014年大数据研究领域前5位高突现词

在CitespaceII软件中对于突现性文献含义有如下描述: “突现性高的节点意味着这些文献在相应的时间区里受到了格外的关注, 一定程度上代表了该学科在

表7 2009-2014年大数据研究领域高突现性文献

排序	涌现值	突现时间段	突现文献	所在期刊
1	2.5218	2013-2014	Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity	Mckinsey Global Institute
2	2.5072	2013-2014	Dean J, Ghemawat S. MapReduce: a flexible data processing tool	Communications of the ACM
3	2.4667	2013-2014	Olston C, Reed B, Srivastava U, et al. Pig latin: a not-so-foreign language for data processing	Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM
4	2.1149	2012	Golder S A, Macy M W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures	Science
5	2.0448	2012	White T. Hadoop: the definitive guide: the definitive guide	O'Reilly Media, Inc

相应时间区间的研究前沿和热点问题”^[4]。通过对高突现性文献及相关文献分析,本文将大数据前沿归纳为以下三方面。

(1) 大数据相关技术研究

大数据技术主要围绕云计算、Hadoop展开。学者 Golder S A与 Macy M W详细介绍了Hadoop技术的原理及应用^[14]。Hadoop是google的云计算系统的开源实现,用Java开源实现了Mapreduce模型。Map-Reduce作为云计算主流并行编程模式之一,主要用于将大规模数据集并行运算。Map-Reduce的研究热点主要集中在这几方面:①基于Map-Reduce模型的应用。如Afrati F N等讨论了基于Map-Reduce的多路连接算法的最优实现^[15]。Dean J等介绍了Map-Reduce的功能与计算过程^[6]。②Map-Reduce模型的改进。如威斯康星大学的Clustera与加利福尼亚大学欧文分校的Hyracks,将Map-ReduceMap函数的输出结果传递到Reduce函数的两步处理扩展成任何函数集组合,可避免过多的通信开销^[16,17]。学者Olston C等,在SQL语言与Map-Reduce的基础上提出了一种基于数据流的大规模分析处理语言Pig Latin,可减少用户开发和执行数据分析的时间^[18]。

(2) 大数据带来的变革性影响研究

Mayer-Schönberger V等在其论文中从公共卫生、商业、思维等方面论述了大数据带来的影响,认为大数据的核心是预测,可以应用在疾病诊断、推荐治疗,甚至识别潜在犯罪分子等方面^[8]。Einav等提出大数据集改变了经济学家进行实证研究使用的统计方法和数据收集来源^[20]。Manyika J等论述了大数据在卫生保健、公共部门管理、零售商、制造业、个人数据方面带来的巨大影响,并提出大数据在五个方面创造价值:第一,大数据可以高效率提高信息的透明度和可用性;第二,组织利用大数据提高其预测结果,及时调整商业手段;第三,允许更窄的细分客户,提高个性化产品和服务;第四,利用大数据精密的分析能够实质性改善领导决策;最后,大数据可用于改善下一代产品的开发和服务^[21]。

(3) 大数据引起的负面效应研究

个人隐私和过度依赖数据是大数据引起的两方面负面效应。Mayer-Schönberger V等认为大数据的优势在于潜在价值的挖掘,但是深度的数据分析很容易追溯到个人的行为及喜好,使得目前用以保护隐私的法律手段和技术失效,个人隐私面临被泄露的风险;通过大数据预测未来行为而据此采取惩罚措施,否认了人们的自由选择的能力;管理者单纯依赖大数据进行决

策可能会做出错误的决定^[19]。此外,Sung-Hwan Kim等提出利用属性选择方法确保大数据的安全^[22]。

由上述分析可以看出,大数据领域相关技术研究及其应用不仅是目前的学者高度关注的热点,更是未来研究趋势所在,说明大数据技术的研究不仅是该领域的核心,更是向其它研究领域拓展及应用的基础。

4 结语

(1) Web of Science核心合集在2009-2014年共收录大数据领域文献2626篇,共有61个国家和地区发表相关文献,其中美国1029篇,居第1位,中国发文446篇,排名第2位,其后依次是英国162篇,德国141篇,澳大利亚103篇。大数据领域的研究共涉及1556家研究机构,前10名美国占8家,中国占2家。前3位分别为中科院,麻省理工学院,哈佛大学。在搜集到的2626篇文献中共有7958个作者,其中发文量最多的作者是悉尼科技大学工程学院的Jinjun Chen。

(2) 大数据研究领域的知识基础由Jeff Dean发表的奠基性文章“MapReduce: simplified data processing on large clusters”提出。2009年到2014年大数据的研究热点演进情况为:云计算(cloud computing)、云(cloud)、系统(systems)、分布式计算(Map-Reduce)、隐私(privacy)、安全(security),其中相关技术的研究一直是每年持续的研究热点。大数据领域的研究前沿主要有三个方面:大数据相关技术研究,大数据带来的变革性影响研究及大数据引起的负面效应研究。

5 结语

通过分析大数据领域相关文献发现美国是该领域的中心国家及主要力量,中国在大数据研究领域具有一定积累,但仍待进一步发展。大数据技术研究、大数据带来的变革性影响及大数据引起的负面效应三个方面的内容,成为今后大数据领域的研究方向。

参考文献

- [1] Big data: The next frontier for innovation, competition, and productivity[EB/OL].[2015-05-01]. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

- [2] 大数据堪比隐形金矿奥巴马政府称其为“未来的新石油”[EB/OL].[2015-05-16]. http://news.xinhuanet.com/2013-04/23/c_124617746.htm.
- [3] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science & Technology, 2006, 57(3): 359-377.
- [4] 陈悦, 陈超美, 胡志刚, 等. 引文空间分析原理与应用: CiteSpace 实用指南[M]. 北京: 科学出版社, 2014: 63-69, 27-44, 117-140.
- [5] 邱均平. 信息计量学[M]. 武汉: 武汉大学出版社, 2007: 158-202.
- [6] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. OSDI' 04 Proceedings of conference on Symposium on Operating Systems Design & Implementation, 2004, 51(1): 107-113.
- [7] Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work and think[M]. Michigan: Hodder & Stoughton, 2013.
- [8] Bell G, Hey T, Szalay A. Computer science. Beyond the data deluge[J]. Science, 2009, 323(5919): 1297-1298.
- [9] Bailon-Moreno R, Jurado-Alameda E, Ruiz-Banos R, et al. Analysis of the field of physical chemistry of surfactants with the Unified Scientometric Model. Fit of relational and activity indicators[J]. Scientometrics, 2005, 63(2): 259-276.
- [10] Garfield E. Research Fronts[J]. Current Contents Editions, 1994(41): 3-7.
- [11] Chen C, Paul R J. Visualizing a knowledge domain's intellectual structure[J]. Computer, 2001, 34(3): 65-71.
- [12] 姚宏宇, 田溯宁. 云计算: 大数据时代的系统工程[M]. 北京: 电子工业出版社, 2013(1): 1-20.
- [13] Golder S A, Macy M W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures[J]. Science, 2011, 333(6051): 1878-1881.
- [14] Afrati F N, Ullman J D. Optimizing joins in a map-reduce environment[J]. EDBT, 2010, 23(9): 99-110.
- [15] DeWitt D J, Paulson E, Robinson E, et al. Clustera: an integrated computation and data management system[C]. Proceedings of VLDB conference, 2008, 1(1): 28-41.
- [16] Borkar V, Carey M, Grover R, et al. Hyracks: A flexible and extensible foundation for data-intensive computing[C]. Data Engineering (ICDE), 2011 IEEE 27th International Conference on. New York: IEEE, 2011: 1151-1162.
- [17] Olston C, Reed B, Srivastava U, et al. Pig latin: a not-so-foreign language for data processing[C]. Proceedings of the 2008 ACM SIGMOD international conference on Management of data. New York: ACM, 2008: 1099-1110.
- [18] Einav L, Levin J. Economics in the age of big data[J]. Science, 2014, 346(6210): 715.
- [19] Big data: The next frontier for innovation, competition, and productivity[EB/OL].[2015-05-16]. http://biojobsblog.com/wp-content/uploads/2013/07/MGI_big_data_full_report.pdf.
- [20] Sung-Hwan Kim, Nam-Uk Kim, Tai-Myoung Chung. Attribute Relationship Evaluation Methodology for Big Data Security [C]. IT Convergence and Security (ICITCS). 2013 International Conference on, Macao: IEEE, 2013: 1-4.

作者简介

刘飞, 男, 1981年生, 新疆医科大学图书馆副研究馆员, 研究方向: 信息服务。

牛丽慧, 女, 1991年生, 山西大学经济与管理学院情报学专业硕士研究生, 研究方向: 自然语言处理, 通讯作者, E-mail: 871298207@qq.com。

金昌博, 男, 1986年生, 广州医科大学妇产科学专业博士研究生, 研究方向: 医学情报。

Status, Focuses Evolution, and Fronts of Big Data Research

LIU Fei¹, NIU LiHui², JIN ChangBo³

(1. Xinjiang Medical University, Wulumuqi 820054, China; 2. Department of Economic and Management, Shanxi University, Taiyuan 030006, China;

3. Guangzhou Medical University, Guangzhou 511436, China)

Abstract: This paper did bilio-metric and visual analysis for documents of big data field in Web of Science during 2009 to 2014 year. Using CiteSpace II to analyse distribution of authors and their cooperation, institutions and countries as well as map up keywords co-occurrence analysis and references co-cited analysis, the results pointed out that big data research field is in a boom period, large gap exists between different countries, America is the center and main force of the field, China is relatively substantial in this field, but still to need further development. The hot points vary from cloud computing in 2009 to personal privacy in 2014, and there are three aspects of frontiers, i.e. big data related technical research, the transformative effects of big data research and the negative effects study big data caused.

Keywords: Big Data; Visual Analysis; Map-Reduce; Personal Privacy

(收稿日期: 2015-05-16; 编辑: 王立学)