

从文献计量分析看 2008—2014 年大数据研究

朱向林, 程如烟

(中国科学技术信息研究所, 北京 100038)

摘 要: 大数据作为一个近几年新兴的研究领域, 2008 年以后在全球范围内得到快速地发展和广泛地应用。以美国科学引文索引数据库 SCI 收录的 2008—2014 年间的 1171 篇大数据论文为研究对象, 对这些论文的产出规模、影响力、国家及机构分布、作者合著情况等方面进行了计量分析和可视化展示, 并从关键词入手对大数据领域发展趋势作出预测。

关键词: 大数据; SCI 论文; 文献计量; 关键词分析

中图分类号: G327.1 **文献标识码:** A **DOI:** 10.3772/j.issn.1009-8623.2015.06.008

1 引言

1974 年, “大数据” 就已经被预言家提出来了, 但是直到 2008 年, 当互联网时代真正进入海量存储, 海量信息处理的时候, 对于它的研究才开始从技术、方法、理论等各方面缓慢进行^[1]。

2012 年 1 月世界经济论坛在瑞士达沃斯举行, “大数据” 是其中重点讨论的主题之一, 论坛的报告中指出 “数据已经成为一种新的经济资产类别, 就像货币和黄金一样”^[2]。同年, 美国政府开始针对 “大数据” 有所动作, 白宫网站上发布了《大数据研究和发展倡议》的倡议书, 美国联邦政府承诺投资超过 2 亿美元, 共涉及 6 个联邦政府部门, 将在今后几年大力推动和改善大数据的提取、存储、分析、共享和可视化。2013 年初, 美国著名科学家维克托·迈尔-舍恩伯格的著作《大数据时代(Big Data)》^[3] 在全球范围内引起轩然大波, 开创了国外大数据系统研究的先河。可见, 各行各业都对大数据有足够的重视, 大数据已经成为了近几年学术研究的热点问题。鉴于此, 本研究通过基于事实的、可循证的文献计量研究, 对 2008—2014 年大数据领域发表的 SCI 论文进行深入剖析, 以不同纬度

揭示最新的该领域的发展脉络, 并以可视化展示, 以供国内学者参考借鉴。

2 数据来源与研究方法

本文选用 Thomson Reuters 公司的 Web of Science 中的引文数据库 SCI^[4] 为数据来源。为了构建 “big data” 研究领域的数据集, 本文将研究领域限定为当前的研究热点 “big data” 本身, 所以检索词仅限于 “big data” (加引号), 检索类型为 “主题” 检索。由于 2008 年 9 月《Nature》杂志开设了 “大数据” 专刊, “big data” 在学术领域才有实质性的发展, 所以, 此次检索的时间跨度为 2008—2014 年, 共得到 1 171 篇 (检索时间为 2015 年 4 月 13 日) 大数据领域的相关文献。

本文利用 Web of Science 中自带的文献计量工具 Endnote 对检索结果的各字段进行计量统计, 然后用网络可视化工具 Echarts 将结果可视化, 用 Gephi 绘制合著作者和高频关键词的网络关系图。

3 大数据研究基本概况

3.1 产出规模及影响力分析

2008—2014 年间, 大数据领域的相关论文共

第一作者简介: 朱向林 (1990—), 男, 中国科学技术信息研究所 2014 级研究生, 主要研究方向为科技政策。

收稿日期: 2015-05-26

1 171 篇, 产出规模不算很大。从论文产出时间序列分析, 2008—2010 年, 大数据研究产出较少, 每年都不会超过 10 篇, 但是 2008 年的这 10 篇论文从大数据的技术、发展方向以及对人类的挑战等多个角度介绍了大数据的相关问题, 这些文献被后来的文章多次引用, 揭开了国际上研究大

数据热潮的序幕^[2]。从 2011 年开始, SCI 中有关大数据的论文开始逐年增加, 并且增长速度逐步加快, 2014 年的论文总量是之前六年数量总和的近 3 倍。可以看出, 大数据领域发展迅速, 但并不成熟。

论文影响力方面, SCI 中收录的 1 171 篇大数

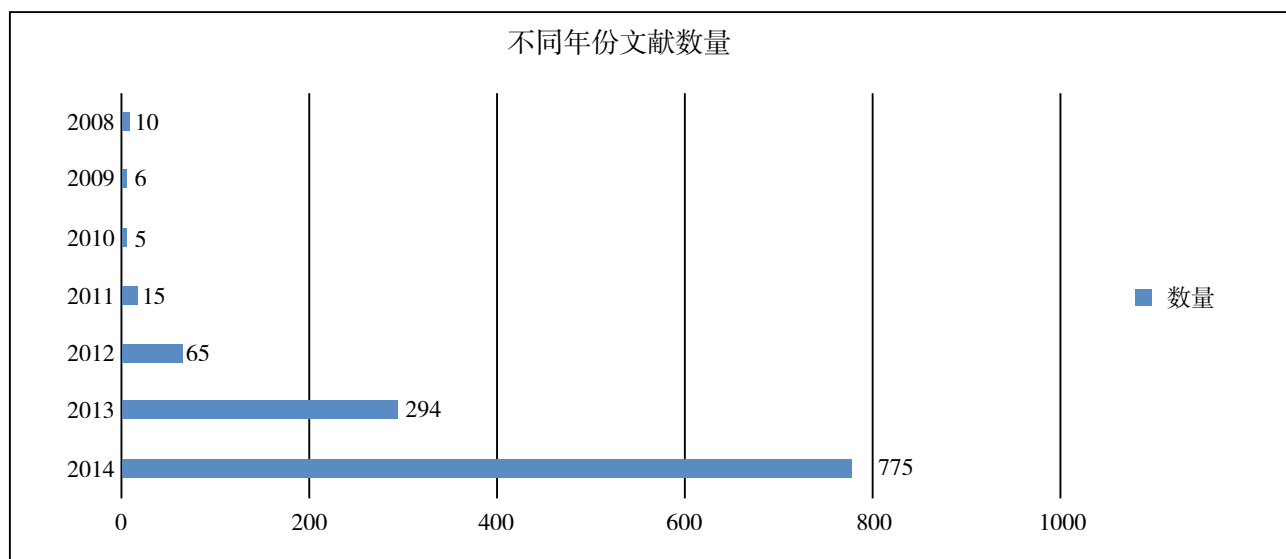


图1 2008—2014年间SCI大数据论文发表情况

数据来源: 对从 Web of Science 核心合集中检索出的相关数据进行统计整理而成。

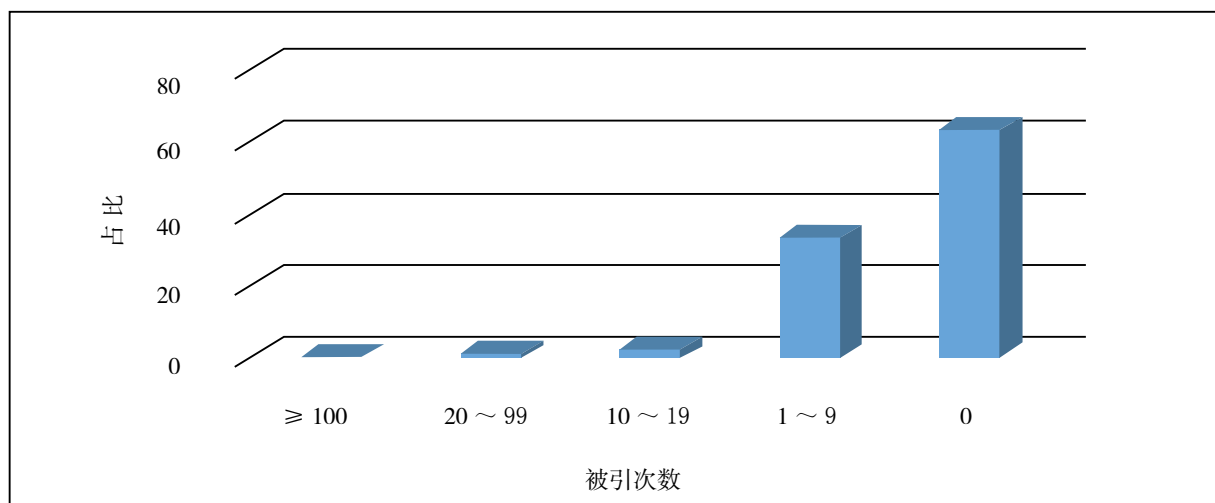


图2 2008—2014年间SCI大数据论文被引情况

数据来源: 对从 Web of Science 核心合集中检索出的相关数据进行统计整理而成。

据论文的平均被引用次数为 2.25, 其数值偏低是与 2014 年的论文量偏多且 2014 年的论文的 0 被引率高达 77% 有关。H 指数为 20, 即 20 篇论文的被引用次数在 20 次及以上。被引用次数不少

于 100 次的论文数量很少, 仅有 3 篇, 由美国科学家 Duncan, Laramie E 和 Keller, Malthew C 共同发表在《American Journal of Psychiatry》2011 年第 10 期上的“A Critical Review of the First 10 Years of

Candidate Gene-by-Environment Interaction Research in Psychiatry”一文，被引次数最高，达到了 184 次。被引用次数在 20 ~ 99 和 10 ~ 19 的占比分别为 1.4% 和 2.3%，被引用次数在 1 ~ 9 的论文数量占比相对高一些，为 33%，另外还有 695 篇文章没有得到引用，其中有 539 篇是 2014 年新发表的，因此尚未得到引用也属于正常现象。

3.2 国家（地区）及机构分析

在区域分布方面，共有 65 个国家或地区在大数据研究方面形成了 SCI 论文产出。与其他领域比较来说，参与国家并不算多，其中，论文产出量在 30 篇以上的国家有 125 个，用网络制图工具 Echarts 可以很直观地将其地理位置进行显示（如图 3），美国是这一领域发文量最多的国家，其次是中国。

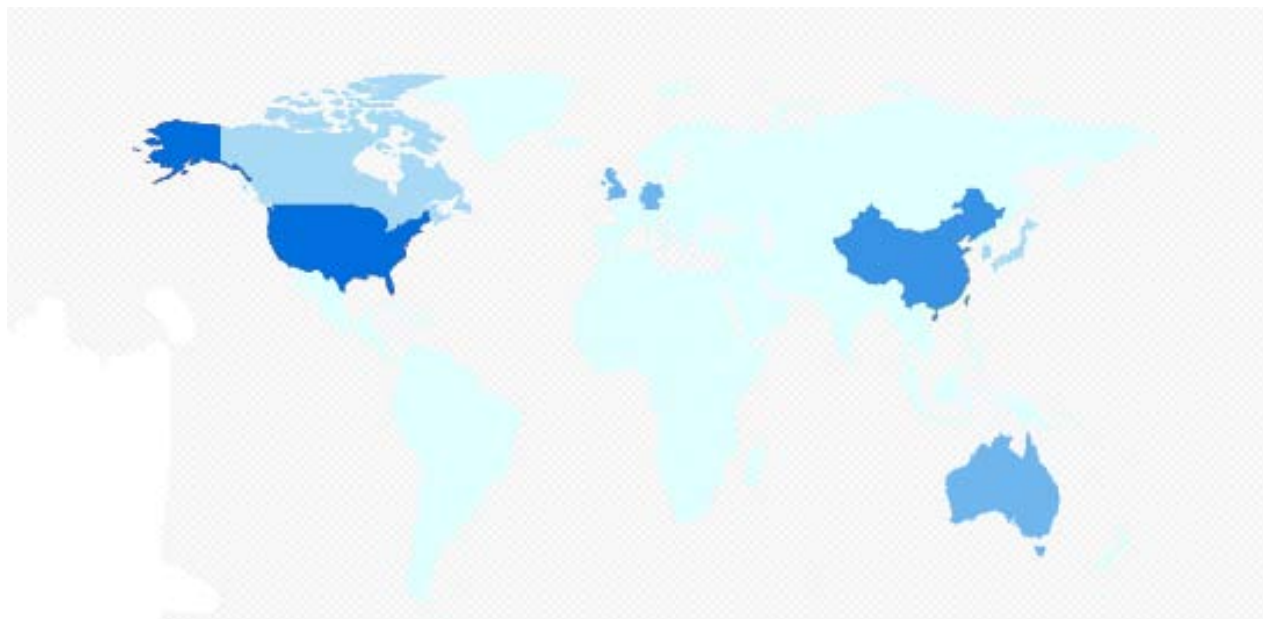


图 3 2008—2014 年 SCI 大数据论文数量大于 30 篇的国家地理分布

数据来源：对从 Web of Science 核心合集中检索出的相关数据进行统计整理而成。

具体对比来说（见表 1），这些国家发表的论文去重之后共计 971 篇，约占发表总量的 83%。美国、中国、英国、德国、澳大利亚等是大数据研究比较活跃的国家，其中，美国的表现非常突出，其论文出产数量远远高于其他国家，约占全部论文产出的 49.53%；中国的表现也比较抢眼，其论文产出量达到 197 篇；其他国家（如英国、德国等）的论文产出量均未超过 100 篇。就论文平均被引次数来看，美国、英国、德国、澳大利亚、加拿大、日本、意大利、荷兰、瑞士以及法国的论文平均被引次数在 1 171 篇论文的总平均被引次数之上，德国和日本表现最为出色。从论文的平均被引次数以及 H 指数两项指标所综合反映的学术影响力来看^[5]，美国相关论文的 H 指数达到 18，远高于其他所列国家，但其相关论文平均被引次数并没有占据主导地位。

就机构来看，大数据研究领域一共涉及到 414

个机构。根据普赖斯定律，发表论文数为 N 篇以上的机构为核心发文机构。核心发文机构的下限为： $N=0.749$ ，其中 $\eta_{\max}$ 为最高产机构所发的论文数^[6]。根据统计结果得知， $\eta_{\max}=37$ ，因此核心发文机构的发文量应为 5 篇及以上，符合条件的机构一共有 95 个，即全球大数据研究的核心机构有 95 个，其发文量去重后为 546 篇。本文列出了发表量前 10 的机构（见表 2）。从表中可以明显的看出，发文量排名前 10 的机构几乎都是高校，可见高校是研究大数据的最重要的阵地。在这 9 个机构中，美国在数量上占据绝对地位，除中国科学院外，其他 8 个机构都是来自美国。

3.3 著者合著情况分析

论文的合著情况在一定程度上能够反映某个领域的科研交流与合作情况。本文在衡量有关大数据的研究论文合著时，引入了“合作度”和“合作率”两个指标，“合作度”是指在一定时间段内某一学

表 1 2008—2014 年 SCI 大数据论文数量大于 30 的国家或地区

国家 / 地区	数量	占比 (%)	论文平均被引次数	H 指数
美国	580	49.53	3.13	18
中国	197	16.82	1.24	9
英国	91	7.77	2.6	8
德国	70	5.98	5.13	7
澳大利亚	58	4.95	2.53	7
加拿大	51	4.36	3.18	7
韩国	50	4.27	0.72	3
日本	39	3.33	5.33	4
意大利	38	3.25	3.84	6
荷兰	36	3.07	2.97	5
瑞士	32	2.73	2.69	5
法国	30	2.56	2.5	5

数据来源：对从 Web of Science 核心合集中检索出的相关数据进行统计整理而成。

表 2 2008—2014 年 SCI 大数据论文发表核心机构统计表

机构名称	篇数	国家
中国科学院	38	中国
哈佛大学	29	美国
斯坦福大学	26	美国
麻省理工学院	23	美国
加州大学洛杉矶分校	20	美国
加州大学圣地亚哥分校	18	美国
南加利福尼亚大学	17	美国
华盛顿州立大学	15	美国
美国密歇根大学	15	美国
明尼苏达大学	15	美国

数据来源：对从 Web of Science 核心合集中检索出的相关数据进行统计整理而成。

科学学术期刊每篇论文的平均作者数，“合作率”是指在一定时间段内某一学科学术期刊合著论文占全部论文总数的比例^[2]。合作度和合作率共同反映了一种期刊论文作者合作能力发挥程度，数值越高，合作能力发挥越充分。2008—2014 年大数据研究合著情况如表 3、表 4 所示。

去除没有署作者名的论文之后，一共有 1 156 篇论文，中国有 197 篇文章。根据表 4 的统计数据可知，大数据研究的论文中合著论文共计 902 篇，占全部论文数的 78.03%，可见大多数的论文都是由学者合作完成的。另外，合作度 3.85 表明平均每篇论文几乎是由 4 位学者共同完成的。在合著的

表 3 2008—2014 年大数据研究合著情况表

独立著者 (篇)	2 人	3 人	4 人	5 人	6 人及以上	著者记录数	全部篇数	合作度	合作率 (%)
254	212	224	172	97	197	4 447	1 156	3.85	78.03

数据来源：对从 Web of Science 核心合集中下载的有关“big data”的 SCI 论文进行统计整理而成。

表 4 2008—2014 年中国大数据研究合著情况表

独立著者(篇)	2人	3人	4人	5人	6人及以上	著者记录数	全部篇数	合作度	合作率(%)
6	27	34	49	33	48	952	197	4.83	97

数据来源：对从 Web of Science 核心合集中下载的有关“big data”的 SCI 论文进行统计整理而成。

论文中,以3人合著的论文居多,但是值得注意的是,6人及以上合著的论文量达到了197篇,合著作者数最多的文章是由44位专家共同完成的,这么大规模的合著在国内是罕见的,中国最多有22人合著,反映了国际学者合作意识之强,合作规模之大^[6]。而让笔者意想不到的是中国的论文合著率高达97%,合作度也在国际平均水平之上。

3.4 关键词热点分析

本文对检索到的 1 171 篇相关论文进行了关键词的统计,得到 2 213 个关键词。由于关键词是未被规范过的自然词语,所以有可能出现重复出现的

现象, 比如“big data”和“big-data”, 其实际意义是一样的。笔者对这些词进行了人工预处理, 以最大限度减少误差。经过处理, 最终得到 2 011 个关键词, 对这些关键词进行词频统计, 发现频次为 1 的关键词共有 1 745 个, 占总数量的 87%; 频次为 2 的关键词共有 155 个; 频次为 3 的关键词共有 47 个; 频次为 4 的关键词共有 22 个; 频次为 5 及以上的共有 42 个, 占总数量的 2.1%。为了能够直观地表示这些关键词之间的关系, 本文用 Gephi 软件绘制了关键词的网络图(见图 4)。

图中点的大小代表着关键词出现次数的多少,

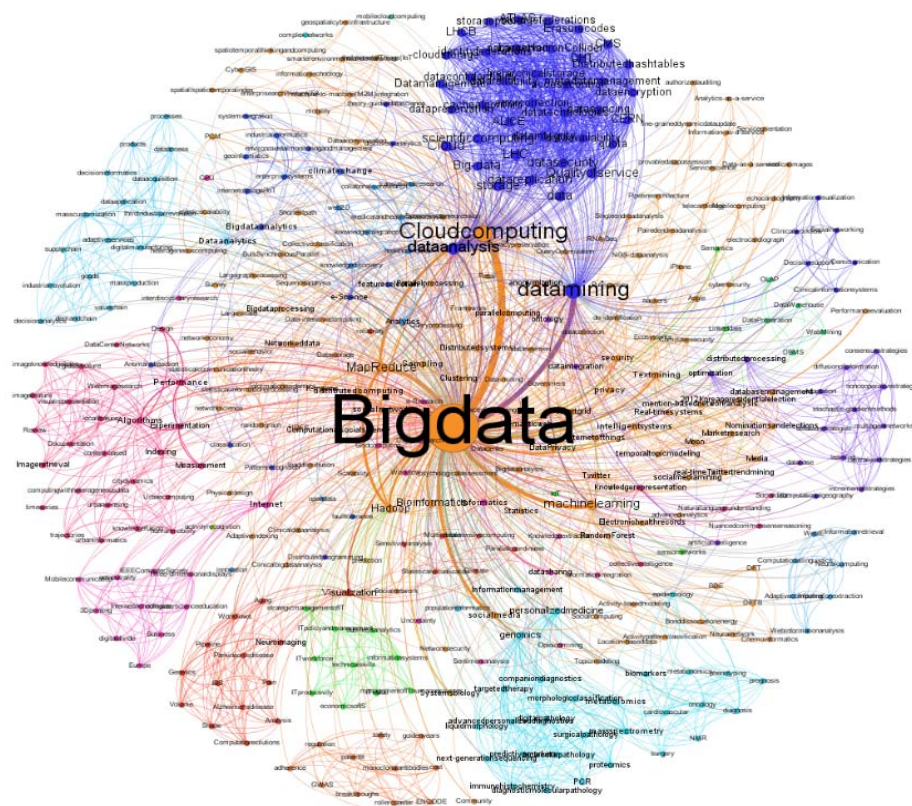


图 4 2008—2014 年大数据相关论文关键词网络图

数据来源：对从 Web of Science 核心合集中下载的有关“big data”的 SCI 论文进行统计整理而成。

两个关键词之间连线的粗细,代表两个关键词的关系紧密程度,由于论文涉及的关键词太多,笔者又认为应该有一个整体的图形表示,所以图形相对较

大。图中所显示的一共有 154 个子群，最大的子群以“Bigdata”为中心，与它有密切关联的“云计算”、“数据挖掘”、“大数据分析”、“Hadoop”等，

几乎都是高频词。对关键词网络进行分析,结果显示凝聚子群密度值为 0.048,比较接近于 0,因此,关键词共词网络趋向于随机分布,也就是说,相关大数据论文的关键词并不是太紧凑,研究状态还比较松散,若以库恩范式理论来衡量,研究尚未形成范式^[7]。从图中也可以看出,关键词网络虽然大,包括范围广,但是很大的子群几乎没有,词与词之间相对比较分散。唯一看着比较聚集的子群是有关

物理粒子研究的,关注的重点是核、信息种子、粒子加速器等。

除了关键词之间的网络图外,本文又对关键词随时间的变化做了统计(见表 5)。将其分为 4 个时间段,2008—2010 年、2011—2012 年、2013 年和 2014 年,我们可以看到关键词随着时间的变化所呈现的词频的变化,从此可以分析出大数据领域研究热点的变化。

表 5 关键词频次统计表

2008—2010 年		2011—2012 年	
关键词	次数	关键词	次数
大科学	1	大数据	12
Web 2.0	1	可视化	3
数据压缩	1	Web2.0	3
数据加密	1	量度	2
循环冗余	1	本体论	2
2013 年		2014 年	
关键词	次数	关键词	次数
大数据	51	大数据	212
云计算	10	MapReduce	32
MapReduce	10	云计算	28
绩效	7	数据挖掘	24
数据挖掘	6	机器学习	18
算法	5	社交媒体	15
基因组学	5	Hadoop	12
Hadoop	5	生物信息学	9
绘制	4	网络安全技术	8
社会网络	4	大数据分析	7

数据来源:对从 Web of Science 核心合集中下载的有关“big data”的 SCI 论文进行统计整理而成。

2008—2010 年:这个时间段正是大数据被学者重视,被大家逐渐认识的过程,关于它的研究还是集中在对大量数据研究的基础上,而多数研究是关于技术方面的,如,数据加密、数据压缩等,在这段时间,“大数据”一度被“大科学”替代,研究方向相对来说很有限。

2011—2012 年:“可视化”在这个时间段在大数据研究中开始被重视,citespace、Ucinet、Gephi 等一大批可视化工具开始用于学术,学者开始认识到可视化在大数据时代给研究带来的便利。有关大

数据研究的方法论在这一段实践也相继提出,包括数据的构建、数据的度量方法、Web 2.0 时代大数据如何发展等。

2013 年,这一年被称为“大数据元年”,大数据在这一年开始应用于各个学科,生物、医学、计算机等学科都出现了大数据的身影。而学者研究最多的还是大数据在技术方面的突破,“云计算”时代还没有到来就被“大数据”所替代,大数据的发展速度超过了大多数学者的预期。

2014 年:“大数据”开始逐渐应用到社会生活

中，电子商务、影像分析、网络安全、智能机器人等，人们也真正感受到了大数据带来的便利。

大数据让许多不定的因素变成确定，让概率事件变成确定事件，天气预报几乎实时并且精准、网上推送系统更加了解用户的心理、计算机网络更加安全。

3.5 知识基础

论文中引用的参考文献反映了研究工作的知识来源，是支撑研究进行的知识基础^[8-9]。本文利用 Excel 软件以及陈超美博士开发的 CitesPace 软件，来研究大数据领域的知识基础。按照以上规则进行分区，对数据进行筛选得出合格参考文献共计 32 056 篇。再从各分区里筛选出被引次数排在

前 30% 的被引文献，对各分区筛选出的文献进行聚类，选择关键路径算法得到合并后的网络节点数为 2 330 个，链接数为 7 391 个。

根据统计结果，我们得到以下信息：在所统计的高频被引参考文献中，2000 年之前发表的文献为 257 篇，2000—2010 年的文献有 990 篇，剩下 1 083 篇文献是 2010 年以后发表的，其中最早的被引文献是 Pearson K 发表于 1901 年的一篇文章，参考文献数最多的一篇有 440 篇。高被引参考文献中被引次数 ≥ 5 次的有 185 篇，它们都是大数据领域研究的重要知识来源^[10]。其中，被引频次 ≥ 15 次的参考文献一共有 14 篇，笔者对其一一列出，如表 6 所示。

表 6 被引频次 ≥ 15 次的参考文献

序号	被引频次	作者	出版年	发表论文 / 著作	备注
1	60	Dean J	2008	<i>Communications of the ACM</i>	期刊
2	34	Manyika J.	2011	<i>Big data: The next frontier for innovation, competition, and productivity</i>	报告
3	24	Schadt EE	2010	<i>Nature Reviews Genetics</i>	期刊
4	22	Mayer-Schonberger V.	2013	<i>Big Data</i>	书
5	21	Hey T.	2009	<i>The Fourth Paradigm: Data-Intensive Scientific Discovery</i>	书
6	21	Howe D	2008	<i>Nature</i>	期刊
7	21	Marx V	2013	<i>Nature</i>	期刊
8	20	Lynch C	2008	<i>Nature</i>	期刊
9	18	Murdoch TB	2013	<i>JAMA</i>	期刊
10	17	Chen HC	2012	<i>MIS Quarterly</i>	期刊
11	17	Schatz MC	2009	<i>Bioinformatics</i>	期刊
12	16	Tibshirani R	1996	<i>J. Royal Statist. Soc. B</i>	期刊
13	15	Ginsberg J	2009	<i>Nature</i>	期刊
14	15	Lazer D	2009	<i>Science</i>	期刊

数据来源：对从 Web of Science 核心合集中下载的有关“big data”的 SCI 论文进行统计整理而成。

4 结论与展望

对 2008—2014 年大数据研究 SCI 论文的产出规模、部分国家的研究情况、关键词热点以及知识基础等相关信息进行计量分析后，我们可以看到，近 7 年以来在全球范围内，学者对大数据的研究呈现出比较明显的快速增长趋势，其成果产出不断丰富，研究的主题无论从规模还是范围来看都在不断扩大^[10]。总体来说我们可以作出以下结论并对其

进行展望：

(1) 从论文产出上看，过去 7 年，大数据研究开始逐渐活跃^[10]。具体来说，SCI 论文数量从 2011 年开始逐年大幅度增加，每年的论文量都是过去一年的几倍之多，这说明，未来大数据领域研究具有较好的发展潜力。当然，大数据从 1974 年被科学家提出，真正被人们熟知只是从 2008 年开始，当年，《Nature》开设了“大数据”专栏，让大数据真正进入人们的视野，从发展年限上来说，

大数据研究还是一个新领域。与其他学科相比,大数据研究的SCI论文发表规模不算很大,其高被引论文所占的比例不高,说明该领域研究的高质量产出不算很丰富,这可能在很大程度上与该领域的跨学科特性有关。大数据是一个很大的领域,可以覆盖每一个学科,这就对研究人员的研究能力有了一定的要求,需要研究人员能够突破单一的研究学科,从计算机建模、方法论、物理学、生物医药、信息系统、信息传播以及其他相关研究主题等不同的学科视角来阐释大数据问题^[11]。而具备这样能力的研究队伍很少,即使能够具备这样的能力,其研究队伍也很难形成庞大的规模,因此,该领域研究的高质量产出相对较少也是易见的。未来大数据研究,很可能会打破这种研究模式,从更专一的角度来研究^[12]。

(2) 从国家和机构的分析来看,7年以来共有65个国家参与了大数据的研究,美国、中国、德国、英国、澳大利亚等国是该领域研究的领头羊。相比较其他国家,美国对该领域的研究占据绝对优势,无论是论文发表数量还是论文的影响力都有很明显的优势,这一点和大多数领域的研究是相同的。中国在论文发表的数量上占据第二位,但是其论文的影响力比较弱,相比较而言,日本和德国的论文平均被引次数占据了绝对的优势,其中,日本的论文平均被引次数更是高达5.33次。所以,综合来说,在大数据的研究上美国应该是第一位的,而中国、德国应该归入第二梯队。和其他领域相似,中国在大数据领域论文发表数量方面有一定优势,但论文的影响力不足,因此,未来中国学者在提高研究质量方面还需进一步努力。

(3) 作者的合著情况能够反映学者之间的交流情况,合著的论文在学术交流的贡献上是强于独著论文的。大数据领域的SCI论文有78.03%是合著的,说明这个领域学者之间的交流很频繁。最多的合著有44位学者共同完成,在大数据领域中国没有那么大的合著规模,而且在中国产出的197篇文章中,最多的合著有22位学者共同完成。从合著角度看,中国无论从合作度还是合作率上都在世界平均水平之上,合作度甚至高达97%,较高的合作度和合作率反映出中国学者在研究大数据领域的时候注重合作,充分发挥集体智慧,交流频繁,

对这个新的领域足够重视,这对大数据在中国的发展无疑是一件好事。

(4) 对关键词的分析结果看出,对于大数据的研究目前还是不太成熟,其网络子群过于分散,尚没有形成几个核心子群,说明这个领域的研究还有巨大的潜力。大数据目前应用最多的领域是计算机技术,包括云计算、数据挖掘、机器学习、网络安全等。从关键词逐年的变化中我们可以看出,大数据的应用领域越来越广泛,并且研究成果逐渐应用到了商业管理、农业、科学研究等领域中,很大程度上实现了生活的智能化^{[11][13]}。值得我们注意的是,机器学习、生物技术、社交媒体、网络安全技术是今后几年大数据可能的发展方向,各子群之间如何建立起联系也是一个值得思考的问题。

(5) 对论文参考文献进行分析显示,大数据论文的篇均参考文献达到27篇左右,这表明大数据研究拥有比较丰富的知识基础^[10]。进一步对高被引文献进行分析可知,大数据研究的经典之作是:Jeffrey Dean教授发表在《Communications of the ACM》上的《MapReduce: 简化数据处理的大型集群》(MapReduce: simplified data processing on large clusters)、全球管理咨询公司麦肯锡发布的研究报告《大数据: 创新、竞争和生产力的下一个前沿领域》(Big data, The next frontier for innovation, competition, and productivity)和大数据研究的权威专家维克托·迈尔-舍恩伯格的《大数据时代》(Big Data)等著作,这些也是大数据研究的重要知识来源。这些高被引文献的作者较为分散,其中多位研究人员来自美国。一方面,显示了美国在大数据研究领域的翘楚地位;另一方面,也说明目前大数据研究领域最重要的创新思想发源地和知识来源地相对比较有限^[10]。而与此同时,笔者发现大多数参考文献有被引时期短的现象,说明大部分文献不具有较长的生命周期,在这个领域中能够形成持久的、有影响力的文献只是很少数的,大部分是在知识消费的过程中被逐渐消磨甚至摒弃。■

参考文献:

- [1] 城田真琴. 大数据的冲击 [M]. 北京: 人民邮电出版社, 2013: 107-110, 132.
- [2] 薛辰. 国际大数据研究论文的计量分析 [J]. 现代情报,

- 2013, 09: 129–134, 139.
- [3] 维克托 – 迈尔 – 舍恩伯格, 肯尼思 – 库克耶. 大数据时代 [M]. 浙江: 浙江人民出版社, 2012: 134–145.
- [4] Web of Science 中的 SCI 数据库 [DB/OL]. [2015-05-26]. <http://apps.webofknowledge.com/>.
- [5] 刘泽照, 张谦. 国外大数据研究特征及趋势预测 [J]. 图书馆论坛, 2014, 08: 103–108.
- [6] 杨绎. 基于文献计量的“大数据”研究 [J]. 图书馆杂志, 2012, 09: 29–32, 37.
- [7] Uprichard, Emma. Being stuck in live time: the sticky sociological imagination [J]. Sociological Review, 2012, 60(1): 124–138.
- [8] 陈超美. CiteSpace II: 科学文献中新趋势与新动态的识别与可视化 [J]. 陈悦, 侯剑华, 梁永霞译. 情报学报, 2009, 28(3): 401–421.
- [9] Chen Chaomei, Ibekwe-SanJuan F, Hou Jianhua. The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis [R/OL]. (2011-06-15) [2013-11-06]. <http://arxiv.org/ftp/arxiv/papers/1002/1002.1985.pdf>.
- [10] 刘娅, 佟贺丰. 从文献计量分析看 1981–2011 年全球系统动力学研究 [J]. 全球科技经济瞭望, 2014 (5): 69–76.
- [11] 涂子沛. 大数据: 正在到来的数据革命, 以及它如何改变政府、商业和我们的生活 [M]. 桂林: 广西师范大学出版社, 2012: 57.
- [12] E. Welsh. Dealing with data [J]. Science, 2011, 331(6018): 639–806.
- [13] McAfee, Andrew; Brynjolfsson, Erik. Big Data: The Management Revolution [J]. Harvard Business Review, 2012, 90(10): 60.

A Bibliometric Study of Big Data Research During 2008–2014

ZHU Xiang-lin, CHENG Ru-yan

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Big data, as an emerging research field in recent years, has been rapidly developed and widely used around the world since 2008. This paper takes the 1,171 literatures on big data between 2008 and 2014 derived from The Science Citation Index (SCI) database as research objects. According to the number of published papers, their impact, distribution of countries and institutions and coauthor condition, this paper has a bibliometric analysis and a visual display on these literatures. And then, this paper gives a prediction about the development trends of big data based on key words of these literatures.

Key words: big data; SCI papers; bibliometrics; keyword-analysis