

科技发展研究

第 11 期

(总第 365 期)

上海科技发展研究中心

2014 年 04 月 21 日

编者按：继上期，本期简报基于上海市软科学研究基地—前沿技术发展研究中心对论文和专利数据库的知识图谱研究成果，对大数据研究的进展、分布和前沿进行介绍和分析。供参考。

大数据技术发展态势跟踪（中）

——全球大数据研究的进展、分布和前沿

二十年来，全球大数据研究经历了一个从起步到活跃的过程。基于大数据相关的 4573 篇文献和 8571 项专利，对其研究进展、分布和前沿进行分析，可以得出如下结论：**1、美国是大数据研究的中心地带，技术创新活跃，国际间合作频繁。2、我国对大数据研究的资助力度较大，学术论文较多，但与国外创新合作较少。3、系统、性能和算法是大数据研究的重点方向 and 核心基础。4、大数据产业创新不仅聚焦软件技术研发，还在硬件技术上重点布局。**

一、大数据研究文献的国别和机构分布

1、美国是大数据研究的中心地带，我国紧随其后。美国是研究者最多的国家，约占总数的 34%，中国紧随其后，占 23%，美中两国

合计占到总数的一半以上。其余的前 10 位国家包括：德国、英国、日本、印度、加拿大、法国、澳大利亚和韩国（图 1）。

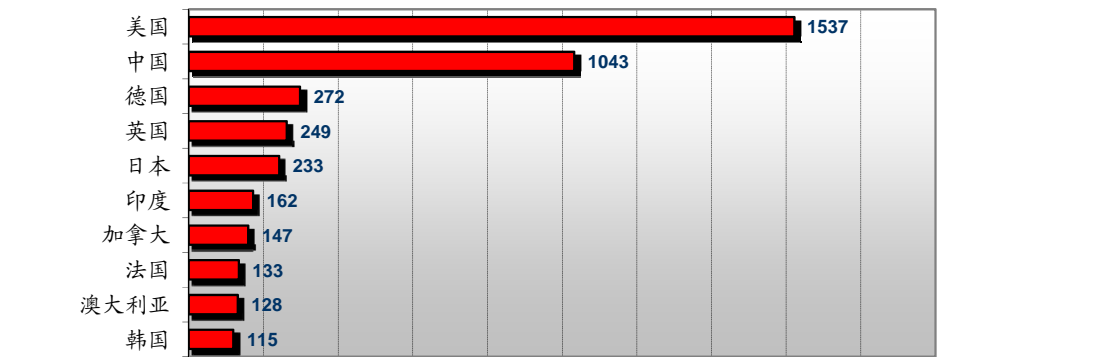


图 1. 大数据论文作者前 10 位的国家/地区分布情况（数据来源：Web of Science）

2、大数据研究文献发展的“三阶段”特征显著。第一阶段是 1994-2001 年，年均文献不超过 50 篇，研究文献主要为“美国籍”。第二阶段为 2002-2010 年，年均文献超过 100 篇，且以年均 20% 的幅度稳步增长，突出特点是我国文献开始出现，并在 2008-2009 年超过了美国。第三阶段为 2011-2013 年，研究文献出现爆发式增长，3 年文献共达到 2053 篇，占 20 年累计总数的 44.9%；主要原因是美国文献的再次崛起，3 年内贡献了 735 篇，而我国为 489 篇（图 2）。

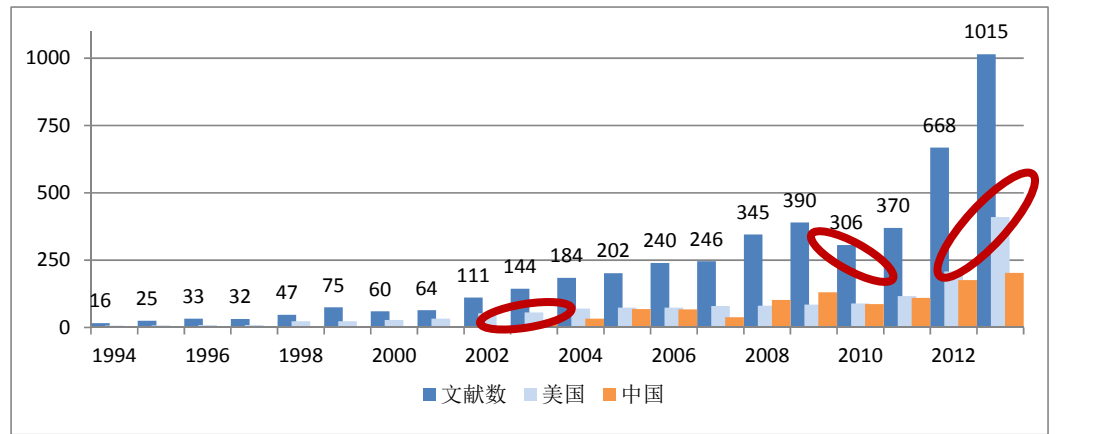


图 2. 大数据文献 20 年增长情况和中美两国文献情况（数据来源：Web of Science）

3、我国政府对大数据研究的资助力度较大。文献数前 25 位的研

究机构中（图 3），美国 16 所大学上榜，共发表论文 559 篇，占美国文献总数的 36.4%。中国则有 6 所大学机构上榜，其中中科院发表论文数居世界第一，达到 109 篇，之后依次是清华大学、上海交通大学、哈尔滨工业大学、浙江大学和华中科技大学。6 所大学机构共计发表论文 244 篇，约占我国大数据文献总数的四分之一。

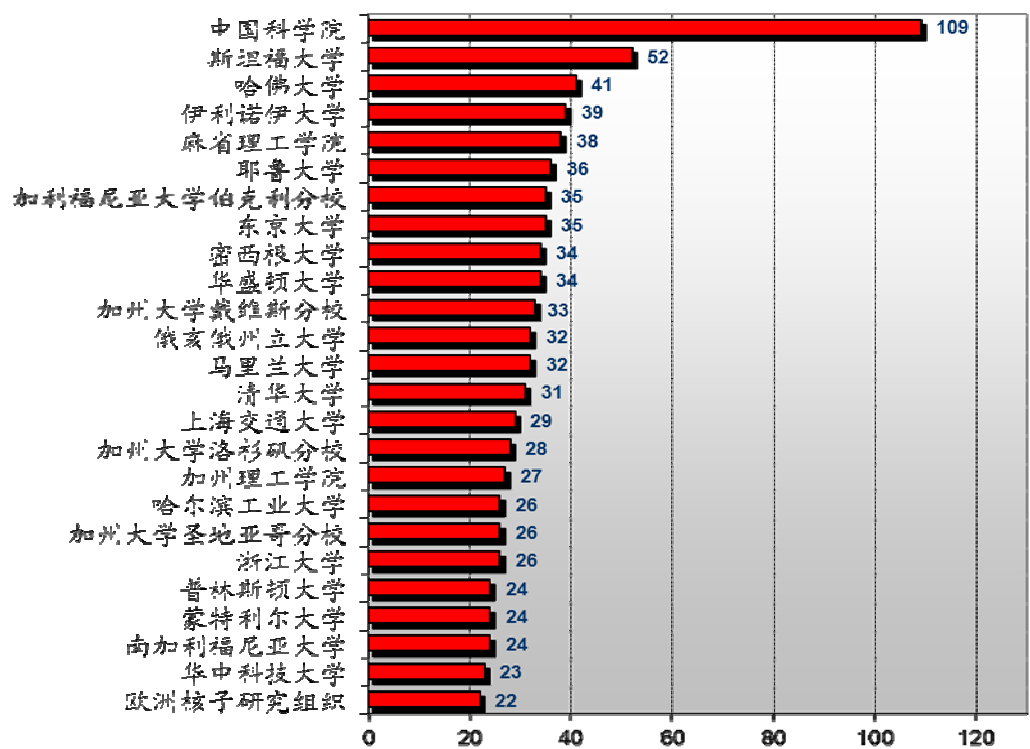


图 3. 大数据文献前 25 位的研究机构情况（数据来源：Web of Science）

从资助机构上看，资助 5 篇以上的机构一共有 36 家，大多为国家基金组织和政府部门。其中，中国国家自然科学基金、美国国家科学基金和国立卫生研究院是三家资助发表文献最多的机构，分别达到 122 篇、109 篇和 57 篇。而企业更多聚焦专利领域进行布局，仅有谷歌和微软两家公司资助的研究文献超过 5 篇。

4、我国大数据研究的质量有待进一步提升。一方面，从合作关系上看（图 4），我国的大数据研究与世界联系不多，仅与台湾地区、德国有少量合作，而美国与韩国、澳大利亚、法国、瑞典、瑞士、土

耳其等国家保持着密切的合作关系,德国、加拿大、英国之间也有不少合作。另一方面,从文献被引频次来看,研究文献的质量低于美国,美国 1537 篇论文平均引用次数达到 14.20 次,高引用指数为 66,而我国文献的高引用指数仅有 20 (表 1)。

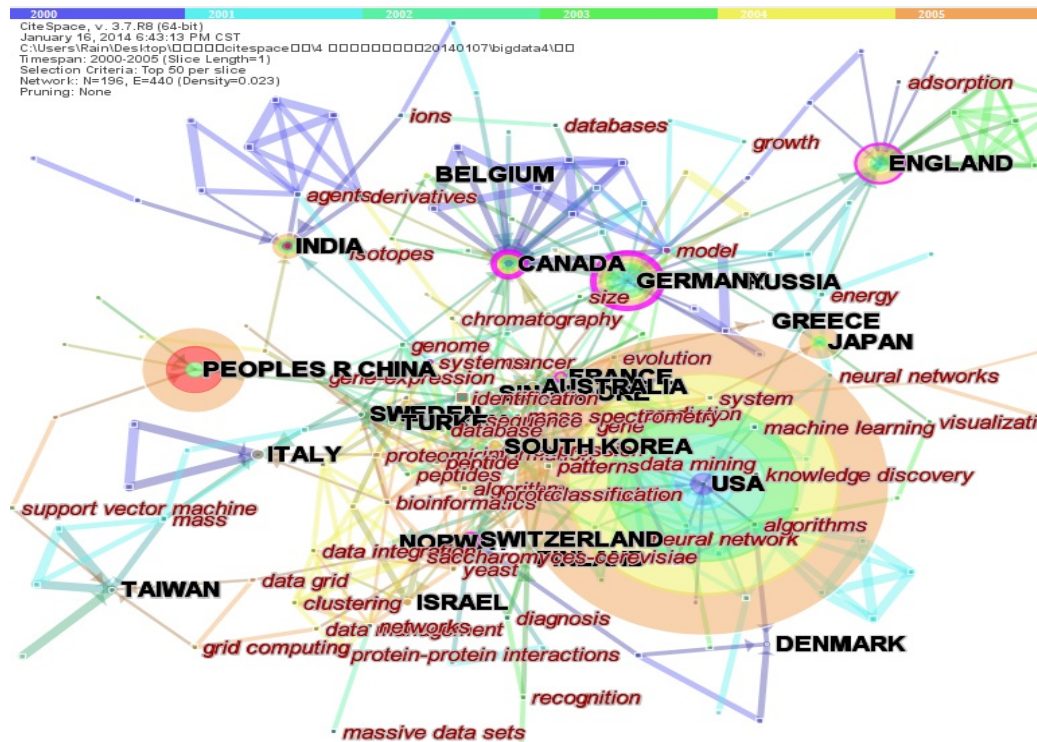


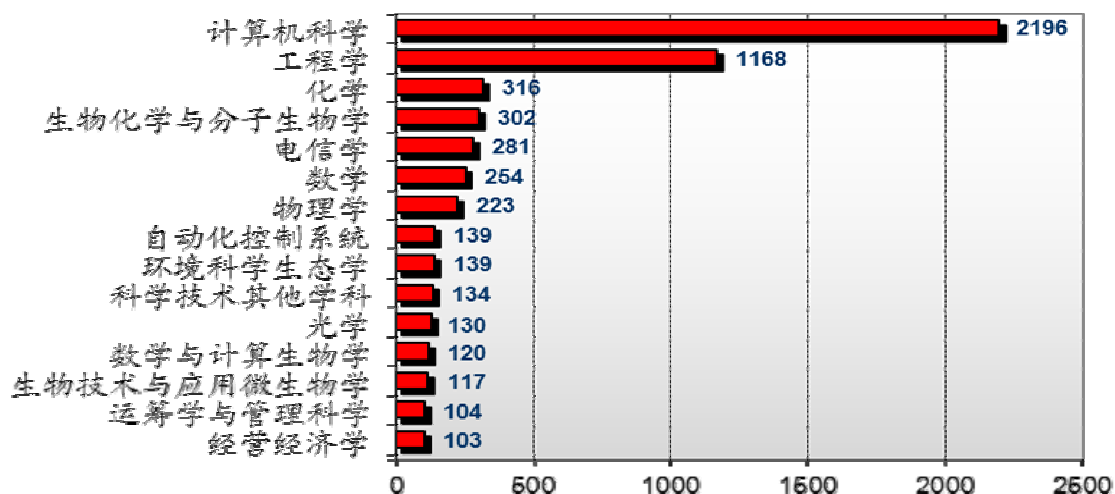
图 4. 大数据论文的国家/地区合作情况 (数据来源: Web of Science)

表 1. 中美两国引文数据对比 (数据来源: Web of Science)

	论文数	被引频次总计	去除自引的被引频次总计	施引文献	去除自引的施引文献	每项平均引用次数	高引用指数 h-index
美国	1537	21821	21600	21029	20857	14.20	66
中国	1043	1729	1687	1690	1651	1.66	20

二、大数据研究的学科领域分布

1、大数据研究开始渗透进入应用领域。文献涉及的学科领域超过 100 个。在排名前 15 位的学科领域中（图 5），除了数学、物理学等基础学科外，更是出现了微生物学、环境生态学、运筹学与管理科学等应用学科，说明大数据技术已经渗透进入各个基础和应用学科领域。



2、系统、性能和算法是大数据研究的重点方向。从大数据涉及的主要学科领域分布中可以看到，大量文献集中在数据处理的系统、性能和算法上，如数据挖掘、机器学习、主成分分析与分类等方向位于核心层，其次为神经网络、降维运算、数据存储、关联规则、数据集等（图6）。

3、核心基础技术文献被大量引用。以谷歌公司的 Mapreduce 为

例¹，作为大数据的基础技术，该文献后续被近 700 篇论文所引用，切实推动了分布式计算、Hadoop 等热点研究的开展（图 7）。

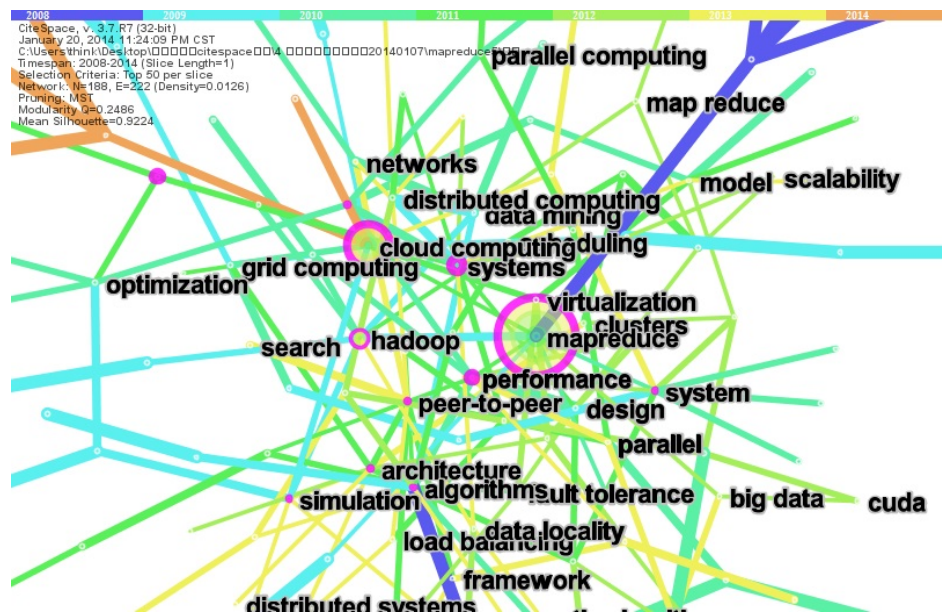


图 7. Mapreduce 技术引用文献研究的热点分布（数据来源：Web of Science）

三、大数据产业技术创新的重点方向

与学术文献研究不同的是，从企业专利布局角度出发，更有利于分析大数据产业技术创新的重点方向。为此，基于 Orbit 专利数据库和 VOSviewer 软件，对 14 家国际性的 IT 企业、互联网企业以及新兴大数据企业自 2006 年以来的 8571 项专利进行分析²：

1、大数据软件技术已趋于体系化。在数据的基本框架、采集传输、存储、处理分析等各个环节上，都有不同于以往抽样、封闭、小规模条件的新技术予以支撑，大数据软件技术体系逐步趋于完整（表 2）。在此基础上，传统的 IT 企业、互联网企业纷纷开发了基于大数据的行业解决方案和商业应用平台，一批新兴大数据企业加快发展并进行商业模式创新，产业创新生态系统所需的各项条件日趋完备。

1 Dean, Jeffrey; Ghemawat, Sanjay. Mapreduce: Simplified data processing on large clusters. COMMUNICATIONS OF THE ACM. 2008

2 由于专利没有规范的关键词，所以关键词搜索的结果并不准确。故本研究采用重点企业专利排查的方法，选取了 IBM、微软、谷歌、EMC、Oracle、HP、Amazon、Facebook、SAP、SPLUNK、Evernote、ATTIVIO、nirvanix、Cloudera 等 14 家企业。

表 2. 大数据软件技术体系 (资料来源: 上海科技情报所整理)

环 节	重 点 方 向
基础框架	分布式系统基础框架、新型文件处理系统、大规模并行数据处理框架、流处理框架
采集传输	数据融合和集成、非结构化数据、空间信息、数据识别、点对点技术
存 储	分布式存储系统、非关系型数据库、新型关系数据库、列存储数据库、数据仓库、内存数据库、可扩展技术、数据集
处 理	Pig、实时监控、数据流处理技术、大规模并行计算、分布式计算、任务执行和调度、虚拟技术、网格技术、合作/工作流
分 析	机器学习、自然语言和语义数据处理、商业智能、可扩展图计算、数据可视化、社会网络分析、关联规则挖掘、分类、数据聚类等无监督式学习、数据挖掘、遗传算法、神经网络、优化、模式识别、预测模型、信号处理、空间分析、降维运算、统计工具
服务/安全	解决方案和行业应用方法、分析平台、数据安全

2、企业在大数据硬件技术上的布局不亚于软件。与学术研究侧重于软件技术不同，企业十分重视相关硬件设备的技术布局。一方面，按照国际专利进行分类，**数据处理、存储和相关设备依然是专利较为集中的领域**，如大数据专利数量较多的 G06F-017、G06F-015、G06F-007、G06F-003、G06F-009、G06F-021 等分类号（图 8），均侧重于计算机数据的处理、存储、控制、接口、安全等硬件电路、设备和零部件。另一方面，8571 项专利集聚形成几个较大的专利群，包括采集显示和传输、处理分析、记录存储、数据管理、共享与协作等（图 9），其中**规模最大的专利群，依然是与采集、显示和传输相关的外围信息硬件设备。**

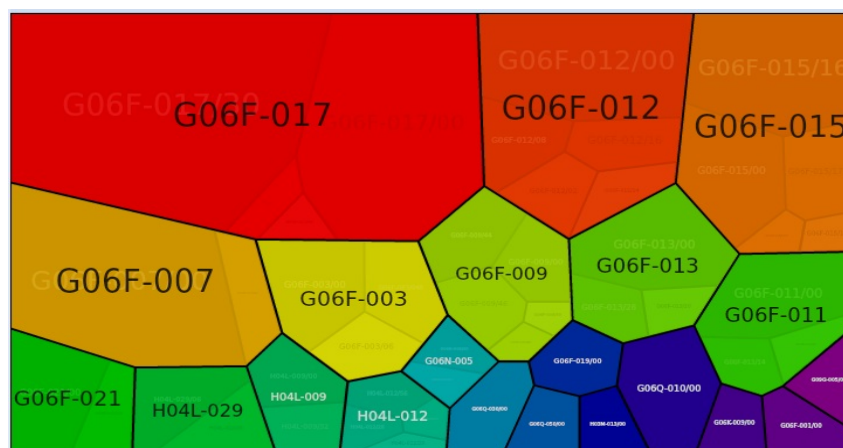
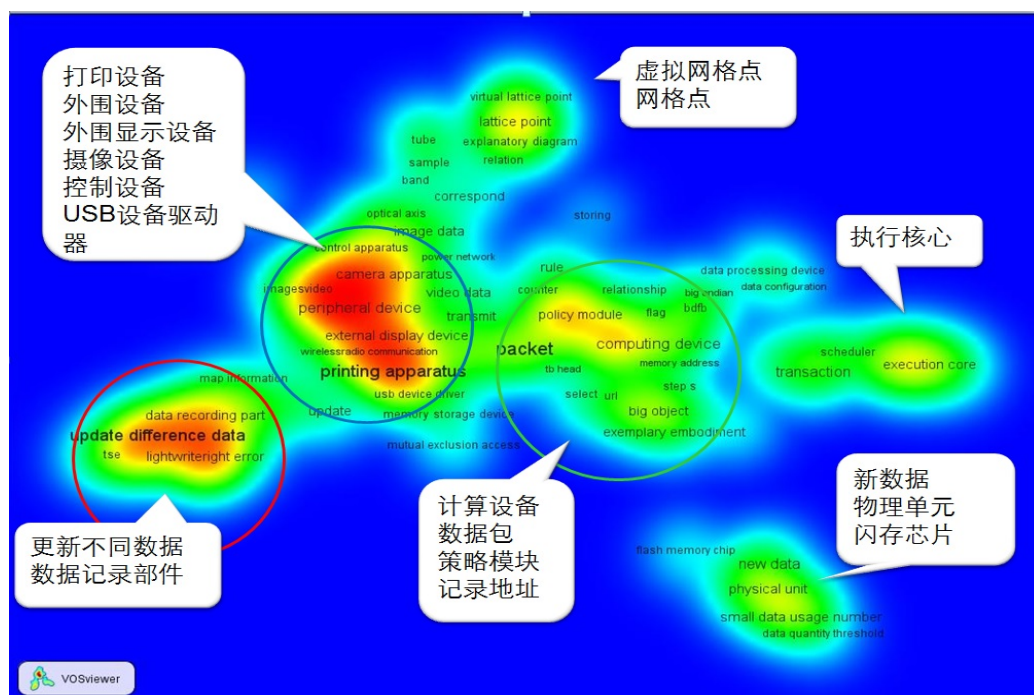


图 8. 大数据专利国际分类 (数据来源: Orbit 专利数据库)



值得一提的是，14 家国际性企业大数据技术专利的一部分贡献来自于其在华分公司。8571 项专利中，公开国为中国的有 1757 项，优先权国在中国的有 170 项，说明了我国研究人员在大数据领域研究水平的不断提升。