

大数据研究前沿、热点与合著模式的图谱分析

胡泽文¹ 李玉平² 张静¹ 吴先华¹

1. 南京信息工程大学 管理工程学院 南京 210044;

2. 南京苏高专利商标事务所 南京 210000

摘要 以可视化图谱的方式全面直观展示当前新兴的热点主题——大数据的国内外研究全貌，为领域研究者和从业者提供一定的指引。因此，本文融合 Web of Science 和中国知网中“大数据”研究文献数据，借助 CiteSpace 可视化分析软件绘制出国内外大数据领域的科学知识图谱——作者、国家和机构的科研合作网络，关键词共现聚类网络，文献共被引聚类网络，揭示出国内外大数据研究概貌。研究发现，国内大数据方面科研社区较多，但规模较小，而国际大数据科研社区较少，但规模较大，规模最大研究社区，达到 30 多人。国家和机构之间的科研合作极少，国际 TOP30 科研机构中，大部分为美国高校，中国机构较少。国内外大数据研究都涉及的热点有：基于大数据的云计算；基于 MapReduce 和 Hadoop 的海量数据分布式处理研究；大数据在数据挖掘、社会网络、互联网金融、识别与预测等领域的应用研究。而国内涉及较少的研究热点有：大数据模型和算法研究、大数据分类研究和大数据相关系统研究。

关键词：大数据；科学知识图谱；可视化；CiteSpace；知识基础；研究前沿与热点

中图分类号：G353.1

开放科学（资源服务）标识码（OSID）



Knowledge Mapping Analysis on Frontier, Hotspots and Co-authorship Patterns of Research from Big Data Domain

HU Zewen¹ LI Yuping² ZHANG Jing¹ WU Xianhua¹

基金项目：国家社科基金重大项目“基于大数据融合的气象灾害应急管理研究（16ZDA047）”；国家自然科学基金重大研究计划培育项目（91546117）；国家自然科学基金资助（71603128）；江苏省自然科学基金资助课题（BK20160974）；江苏高校品牌专业建设工程资助项目。

作者简介：胡泽文（1985-），南京信息工程大学管理工程学院副教授，主要研究方向：科学知识图谱分析、科学计量与评价，E-mail: huzewen915@163.com。吴先华（1977-），南京信息工程大学，教授，主要研究方向：气象灾害应急管理、产业经济系统分析、管理定量分析。

1. School of Management Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China;

2. Nanjing Sugao Patent & Trademark Firm, Nanjing 210000, China

Abstract The purpose of this paper is to provide help for researchers and workers of “big data” field through presenting the whole research landscape comprehensively and intuitively by using a series of knowledge mappings. This paper utilized CiteSpace II to make a series of knowledge mappings, visa co-authorship network, cooperative network of countries and institutions, co-occurrence mapping of keywords and co-cited cluster network of literature, which revealed the whole research landscape in big data field based on literature data related to “big data” from Web of Science and CNKI. Then, the results indicated that there are more but smaller research communities in China comparing to fewer but larger international research communities, with more than 30 members in largest international research communities. Fewer research cooperation happened among countries and institutions, most of the top 30 international research institutions are universities in America, while fewer domestic institutions were detected. Furthermore, both of domestic and international studies of big data gave more attention on the following topics, e.g., cloud computing based on big data, distributed processing of big data using MapReduce and Hadoop, application of big data in data mining, social network, online finance, identification and prediction. However, some international research hotspots including big data model and algorithm, classification of big data, processing system of big data have not attracted more attention from domestic scholars yet.

Keywords: Big data; scientific knowledge map; visualization; CiteSpace; knowledge basis; frontier and hotspots

1 前言

近几年,随着互联网时代的飞速发展,各行业互联网+的触网,使得各行业产生的数据量急剧增长,互联网时代也逐渐步入大数据时代。目前大数据领域已经涌现大量研究文献,Web of Science 核心库显示 2005-2015 期间有 4288 篇与大数据相关的文献,中国知网期刊数据库显示 2000-2015 期间,国内学者发表了 16,798 篇大数据方面的文献。然而目前利用科学知识图谱全面解读国内外大数据领域知识基础、科研合作、研究热点与前沿的文献还相对较少。大部分领域科学知识图谱分析的文献集中于其他领域。如陈超美博士(2010)利用自己

开发出来的 CiteSpace II 绘制出国际情报学研究的知识图谱,识别出了国际情报学的研究热点与知识基础^[1]。胡泽文和武夷山(2013)利用科学知识图谱分析工具 Citespace II 分析了数字资源保存的研究进展、热点与前沿^[2]。同年,胡泽文、孙建军和武夷山详细对比分析了 Citespace II、Ucinet、VOSviewer、Bibexcel、SPSS 和 Histcite 等目前常用知识图谱分析工具的功能及其优缺点,以及绘制图谱的方法流程及其应用领域^[3]。此后,胡泽文和武夷山(2016)融合知识图谱和其他可视化分析软件分析了零被引研究领域的 TOP 作者、机构及他们之间的合作情况,以及零被引研究的历史脉络、前沿和热点^[4]。

在大数据科学知识图谱研究方面, Manyika 和 Chui (2011) 以图表的形式展示了大数据对不同领域, 如健康护理、制造业、零售、计算机和电子产品等的价值和作用^[5]。王新才和丁家友 (2013) 采用可视化工具 Gephi, 在前期文献调研的基础上, 利用可视化图谱的方式初步探索了大数据研究的发展脉络和应用范围, 但分析的不够系统全面^[6]。党倩娜和罗天雨 (2015) 通过专利地图的可视化方式, 并结合产业动态, 全面分析大数据领域的技术创新演化路径、前沿方向和主要驱动因素^[7]。然而他们对国内外大数据领域 TOP 作者、国家、机构及其之间的合情情况, 以及大数据领域研究热点、前沿与知识基础的研究和分析较少。

因此, 本文利用 CiteSpaceII 文献计量与可视化分析软件^[8-9], 利用共被引分析、共现分析和词频分析等可视化分析方法, 选择目前知识图谱分析文献涉及较少的热点领域, 全面展示了国内外大数据研究的主要科研团队、发文国家和机构, 以及他们之间的合作情况, 同时展示出国内外大数据研究的热点主题、前沿热点与领域核心知识基础。一方面, 可以为读者和研究人员展示当前大数据领域研究的全貌, 为他们今后的研究方向和选题提供一定的参考和借鉴, 另一方面, 通过国内外大数据研究概况

进行一个对比分析, 可以促进我国大数据研究体系的完善。

2 研究方法和数据来源

2.1 研究方法

本文主要通过 Citespace II 可视化分析软件, 基于共现分析方法和共被引聚类方法, 首先绘制出大数据领域的作者科研合作网络, 揭示出国内外大数据研究领域科研团队的数量和规模情况, 以及这些科研团队形成的合作模式; 并绘制出国家和机构的科研合作网络, 揭示出国际大数据研究领域主要发文国家、机构及其之间的合作情况; 然后可视化出领域关键词的共现网络, 识别出国内外大数据研究的高频主题词及其之间的相互关系; 最后, 绘制出引文的共被引聚类网络, 揭示出国际大数据研究领域的奠基性文献。Citespace II 由美国德雷塞尔大学的陈超美博士^[10]开发, 该软件可以在科学文献中识别并显示领域最新动态和前沿, 可以显示一定时间内某一学科或知识领域的趋势和动向, 以及研究前沿的演变过程。

大数据研究知识图谱分析过程中, CiteSpace II 软件参数的设置方法如表 1 所示。

表 1 软件运行的参数名称及参数值

参数设置	参数值
时间间隔 (Time Slicing)	国内的时间范围为 (2000-2015); 国外的时间范围为(2005-2015) 时间间隔都为: 1年
网络节点类型 (Node Types)	Author(作者), Institution(机构), Country(国家), Keyword(关键词), 被引文献 (Cited Reference)
连接强度 (Links Strength)	Cosine
Top N per slice	30
网络约简 (Pruning)	网络关键路径发现 (Pathfinder) 及约简 (Pruning SlicedNetworks) 算法
可视化 (Visualization)	静态聚类视图 (Cluster View-Static) 和融合网络 (MergedNetwork)
术语和文献标签	按中心度 (Centrality) 标注

2.2 数据来源

国内大数据研究的文献来源于中国知网，检索时间：2016年1月10日，专业检索式为（TI=‘大数据’+‘海量数据’）OR（KY=‘大数据’+‘海量数据’），文献的时间范围为2000–2015年，共检索出16,798篇大数据相关文献。国际大数据研究文献来源于Web of Science核心库。以TI='large data' OR 'big data' OR 'massive data'为检索式进行检索，检索文献的时间范围是2005–2015。共检索出4288篇与大数据相关的文献。大数据研究领域文献的时空分布趋势，如图1和图2所示。

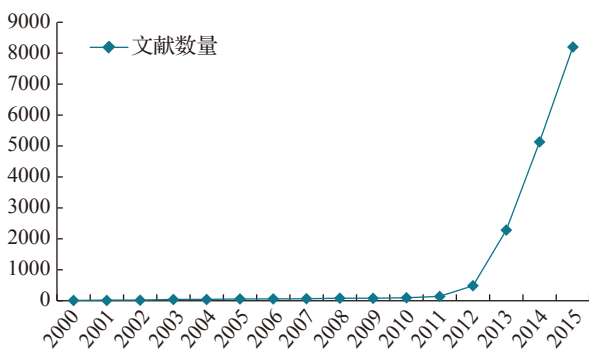


图1 国内大数据研究文献的时间分布趋势图

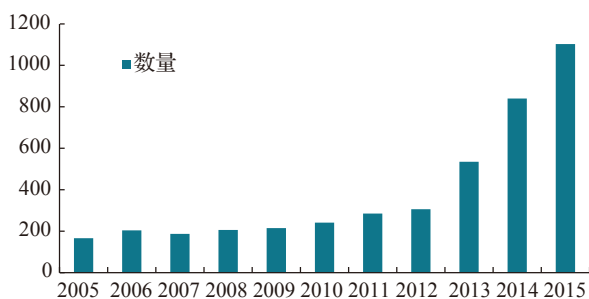


图2 国际大数据研究文献的时间分布趋势图

从图1可以看出，国内大数据研究爆发于2011年，2011年之前的11年，国内学者发表的大数据方面文献年均51篇。到2011年，国内学者发表了139篇，然后开始以指数级的速度增长，到2015年，发文高达8000多篇。

然而，国外大数据方面的研究文献相对较少，2005–2015的11年间，出版了4288篇文献，不及国内同时期出版论文数量（16671篇）的三分之一。国际大数据研究从2012年开始呈现出爆发性增长趋势，从2012年的306篇，增长到2015年的1000多篇。

3 国内大数据研究的科学知识图谱分析

3.1 国内大数据方面的主要科研人员及其合作情况

大数据方面的作者群体是大数据研究发展的一个很重要的基础要素，科研实力较强的作者在大数据研究方面做到开辟性和引领性作用，因此大数据方面高频作者及其之间的合作情况研究是非常必要的^[11]。因此本文通过对近15年（2000–2015年）国内大数据方面文献作者及其发文情况进行计量分析，展示大数据方面高频作者及其合作情况，以便读者能够直观了解到大数据方面的主要科研力量。表2展示了发文数量大于7的高频作者，图3展示了高频作者之间的科研合作图谱。

作者的发文数量能够体现出作者在该领域的学术生产力^[13]，从上表中我们可以看出：谢然和王伟是大数据研究领域实力最强的学者，发表了18篇关于大数据方面的研究文献。谢然的研究主要关于大数据应用案例、大数据的核心价值、大数据的可视化之美及可视化工具介绍等，主要侧重理论探讨^[12-14]。王伟的大数据研究主要侧重大数据技术在不同领域（如智能电网、铁路信号的具体应用方案及模型^[15-16]。

其次是发表了 17 篇的俞国强，发文数量仅次于同为第一名的谢然和王伟。

表 2 2000-2015 年作者发表文献数量

排名	文献数量	作者	排名	文献数量	作者	排名	文献数量	作者
1	18	谢然	12	11	李伟	23	8	程平
2	18	王伟	13	10	陈刚	24	8	秦建秀
3	17	喻国明	14	10	郭嘉凯	25	8	王磊
4	15	马晓亭	15	10	张鹏	26	8	王明宇
5	14	姜奇平	16	10	张勇	27	8	杨树强
6	13	乌贺栓	17	9	黄欣荣	28	8	张贝贝
7	13	李建中	18	9	王军	29	8	张楠
8	12	王鹏	19	9	张莉	30	8	孟小峰
9	12	李璐	20	9	周涛	31	8	何清
10	12	刘洋	21	9	刘新海			
11	11	甄峰	22	8	龙赢			

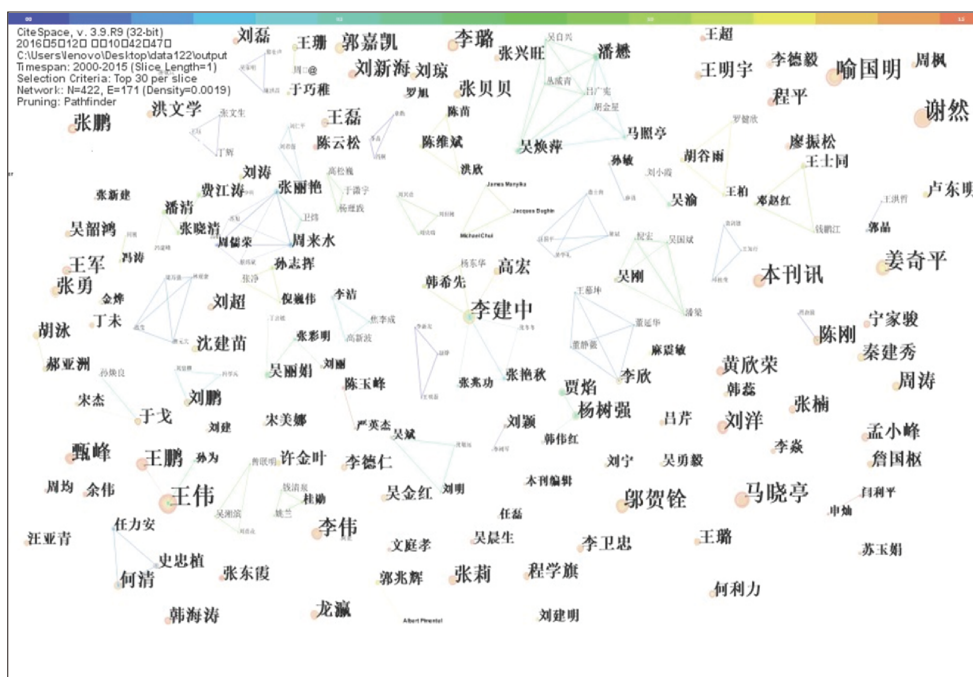


图 3 大数据领域高频作者的科研合作图谱

通过在 CitespaceII 软件中将节点类型设置为作者，我们获得大数据领域作者的科研合作图谱，如图 3 所示。图 3 共展示 422 个作者节点和 171 条连线，节点越大表示该作者在大数据方面的发文数量越多，此图中的节点相对较

多，但它们之间的连线数量较少，这表示，尽管大数据方面的作者较多，但是他们之间的科研合作相对较少，一些实力较强作者之间未有科研合作。从图中可以看出，最大的五人科研合作团队有五个，分别由潘懋、吴焕萍、周来水、

李建中和张丽艳等几位作者为核心形成，他们的研究主题主要为海量数据处理方法、模型和工具；其次，规模为4人的大数据科研团队也有五个，分别由威清、吕广宪、吴自兴、周儒荣和李欣等几位作者为核心形成，研究主题主要为海量数据属性数据的分析与处理。而规模为3人的科研合作团队数量高达23个，分别胡金星、马照亭、吴刚和董华延等作者为核心形成。

3.2 国内大数据方面主要科研机构及其之间的合作情况

大数据方面主要科研机构及其之间合作情况研究可以直观了解大数据方面最强的科研机构，以及有效把握各机构研究重心，机构之间联系的紧密程度，各个机构在大数据研究方面的重要地位和地位。表3展示了大数据方面发文数量TOP20研究机构及其发文数量。图4展示了大数据方面TOP研究机构之间的科研合作图谱。

从2000-2015年各机构发表的文献数量，可以看出，发文数量100篇以上的研究机构都为双一流大学和双一流学科大学，中国人民大学和武汉大学分别发表180篇和166篇文献，分别排名第一和第二。此外，TOP20研究机构中，基本上都为高校，仅有一家企业：中国电信股份有限公司，发文数量达到52篇，与兰州商学院和华中科技大学并列第十五位。

从图4中可以看出，图中共有368个机构节点，然而368个机构节点之间，仅有36条连线，可见各个研究机构之间的合作非常少。图中节点的大小表示研究机构的发文数量，节点越大表示研究机构的发文数量越多，节点越小

表示发文数量越少。图中，中国科学院大学所属几个机构之间有合作，并且他们的科研团队数量目前是最大的。其次为哈尔滨工业大学计算机科学与技术学院、中国人民大学新闻学院和中国工程院等为核心机构的合作。

表3 2000-2015年期间大数据方面TOP20研究机构的文献数量统计

排名	出现次数	机构名称
1	180	中国人民大学
2	166	武汉大学
3	148	清华大学
4	134	北京大学
5	111	中国传媒大学
6	107	中国科学院大学
7	88	南京大学
8	69	同济大学
9	67	北京邮电大学
10	65	上海交通大学
11	63	华东师范大学
12	60	复旦大学
12	60	上海大学
13	54	南京政治学院
14	53	河南大学
15	52	兰州商学院
15	52	华中科技大学
15	52	中国电信股份有限公司
16	51	河北大学
17	50	四川大学
18	49	北京师范大学
18	49	吉林大学
18	49	郑州大学
19	48	南京邮电大学
20	45	浙江大学

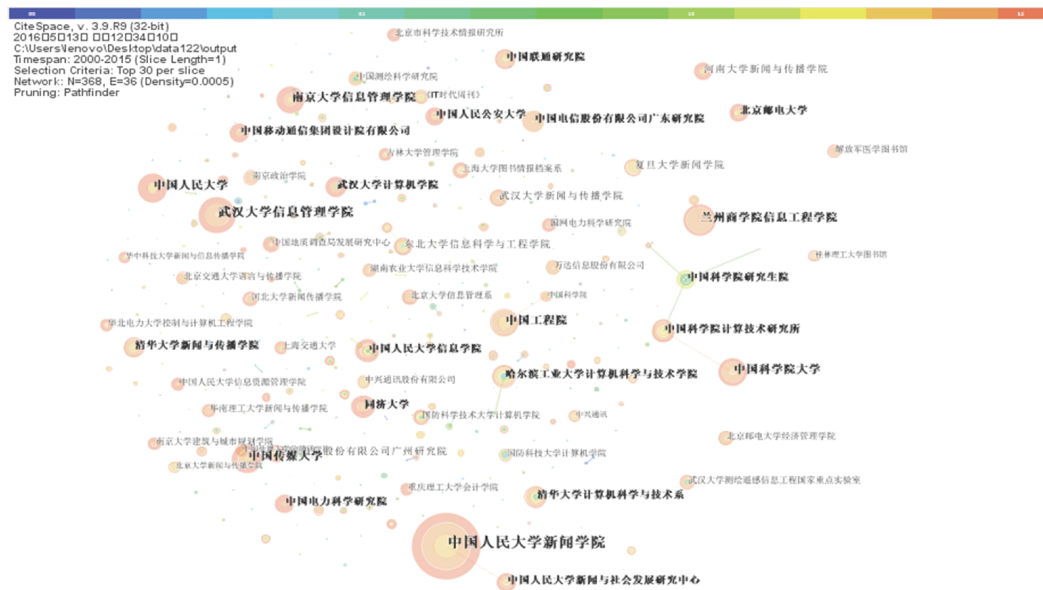


图4 大数据方面 TOP 研究机构之间的科研合作图谱（图中节点类型为“Institution”，其他参数和表1一样）

3.3 基于高频关键词的研究前沿与热点探测

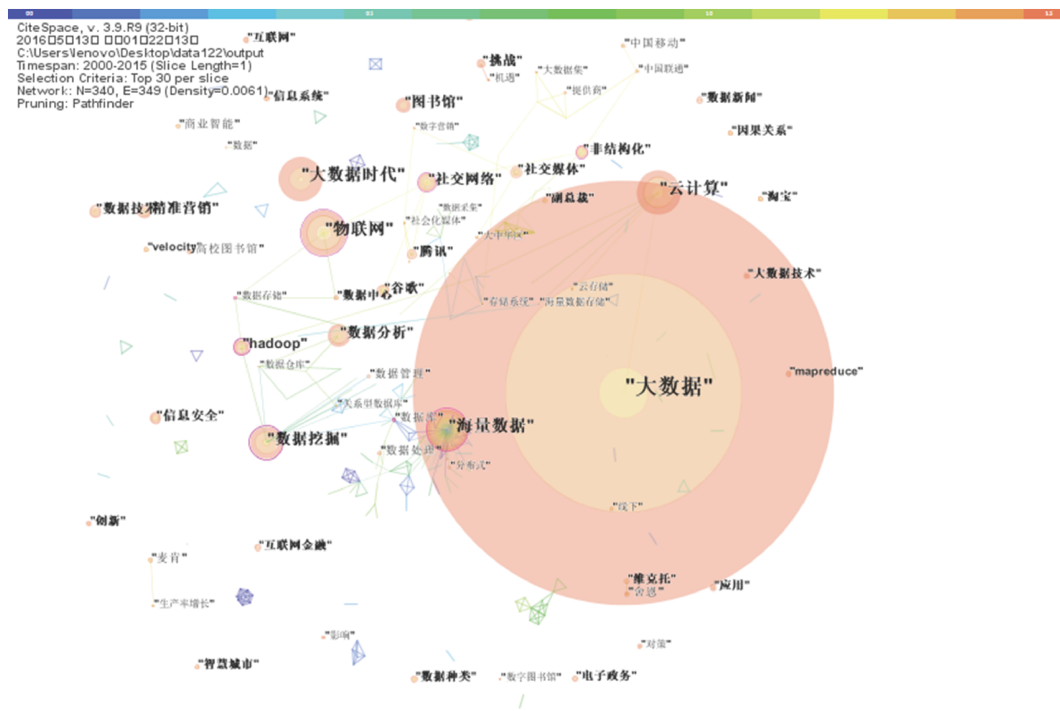
高频关键词及其之间的关系能够代表大数据方面的研究前沿和热点。表4列出了大数据

方面出现频次不少于76的高频关键词。图5展示了大数据方面 TOP 高频关键词之间的相互关系。

表4 2000-2015 期间大数据方面出现频次不少于76的高频关键词

排名	出现次数	关键词	排名	出现次数	关键词
1	4975	大数据	19	120	隐私保护
2	558	物联网	20	114	挑战
3	540	海量数据	21	99	应用
4	536	云计算	22	98	信息服务
5	527	大数据时代	23	91	数据种类
6	417	数据挖掘	24	89	MapReduce
7	282	数据分析	25	86	互联网金融
8	226	社交网络	26	82	信息系统
9	205	信息技术	27	81	大数据技术
10	200	Hadoop	28	80	电子政务
11	195	图书馆	29	80	数据新闻
12	167	精准营销	30	78	信息化
13	161	非结构化	31	77	互联网
14	157	数据技术	32	77	数据中心
15	154	信息安全	33	76	大数据分析
16	150	谷歌	34	76	淘宝
17	149	社交媒体	35	76	智慧城市
18	121	腾讯			

比如：大数据时代出现 527 次；物联网方面的研究出现 558 次。从关键词的频次可以看出：物联网，云计算，数据挖掘，数据分析，社交网络，信息技术，Hadoop, 图书馆是当今大数据方面研究的热点所在。



信息安全、智慧城市和互联网金融等新前沿
热点研究。

4.1 国际大数据方面的主要科研人员及其合作情况

表5展示了2005–2015年期间发文数量不少于9篇的高频作者,图6展示了大数据方面高频作者之间的科研合作图谱。

表 5 2005–2015 年作者发文数量统计

排名	发文量	作者	排名	发文量	作者	排名	发文量	作者
1	13	Lizhe wang	11	9	A.Bocci	21	9	J.Chudoba
2	11	Jinjun chen	12	9	Jianzhong Li	22	9	A.Lister
3	11	Rajiv Ranjan	13	9	Chang Liu	23	9	A.Clark
4	11	E.Rossi	14	9	C.Bromberg	24	9	A.Annovi
5	10	Francisoco Herrera	15	9	A.Filipic	25	9	A.Boveia
6	10	Xuyun Zhang	16	9	J.Wang	26	9	A.Cerri
7	9	E.lipeles	17	9	A.Limosani	27	9	A.Robson
8	9	A.Mehta	18	9	B.Stelzer	28	9	A.ManousakisKatsikakis
9	9	E.Thomson	19	9	A.Loginov	29	9	C.Liu
10	9	D.Whiteson	20	9	A.Warburton			



图 6 大数据领域高频作者的科研合作图谱

图 6 展示了大数据领域 311 个作者节点和 541 条连线,说明大数据领域作者之间的合作非常频繁和密切。结合表 5 和图 6,可以看出,大数据领域高频作者主要有:Lizhe wang、Jinjun chen、Rajiv Ranjan、E.Rossi、Francisoco Herrera 和 Xuyun Zhang, 分别

发表文献 10 篇以上,Lizhe wang 发表文献最多,达到 13 篇。其中 Lizhe wang、Jinjun chen、Rajiv Ranjan 和 Xuyun Zhang 等高产作者形成一个规模达到 14 人的较大科研社区,他们的研究主题主要为云计算方法、系统和应用。而作者 E.Rossi 是一位独立性较强

的高产作者，仅有一位低产合作者。高产作者 Francisoco Herrera 形成了 6 人研究社区。此外，表 5 所示发表文献为 9 篇的所有作者形成了一个最大的大数据研究社区，规模达到 30 多人，该社区中的大部分研究者科研实力水平极其相似，是一个合作非常密切，实力均等的研究社区。该社区的研究主题主要为大数据处理。

4.2 大数据研究的主要国家（地区）、科研机构及其之间的合作情况

4.2.1 国际大数据研究的机构及其之间的合作情况

表 6 展示了大数据方面 2005-2015 年期间发文数量不少于 17 篇的 TOP 研究机构及其发文数量。图 7 展示了大数据方面 TOP 研究机构之间的科研合作图谱。

表 6 大数据方面 TOP 研究机构及其发文数量

排名	发文量	机构	译名	排名	发文量	机构	译名
1	72	Chinese Acad Sci	中国科学院	16	22	Columbia Univ	哥伦比亚大学
2	52	Harvard Univ	哈佛大学	17	22	Univ california, los angeles	加州大学洛杉矶分校
3	41	Tsinghua Univ	清华大学	18	21	Wuhan Univ	武汉大学
4	37	Stanford Univ	斯坦福大学	19	21	Univ Melbourne	墨尔本大学
5	32	Russian Acad Sci	俄国科学院	20	21	Univ Edinburgh	爱丁堡大学
6	31	Univ Calif Berkeley	加利福尼亚大学伯克利分校	21	21	Univ Tokyo	东京大学
7	29	Univ Washington	华盛顿大学	22	20	Univ Michigan	密歇根大学
8	28	Univ Illinois	伊利诺伊大学	23	20	Duke Univ	杜克大学
9	27	Univ Calf San Diego	圣地亚哥大学	24	20	Zhejiang Univ	浙江大学
10	26	Univ Minnesota	明尼苏达大学	25	20	Ohio State Univ	俄亥俄州立大学
11	25	New York Univ	纽约大学	26	19	Univ Wisconsin	威斯康星大学
12	24	Univ Oxford	牛津大学	27	19	Univ Toronto	多伦多大学
13	24	Univ Maryland	马里兰大学	28	18	Univ Helsinki	赫尔辛基大学
14	24	Univ southern california	南加利福尼亚大学	29	18	Nanyang Technol Univ	南洋理工大学
15	23	Massachusetts Institute of Technology	麻省理工大学	30	18	Univ Utah	犹他大学

表 6 显示，2005–2015 年期间，大数据方面 30 个 TOP 研究机构中，大部分为美国高校，共 16 所，占比高达 53%，其中哈佛大学和斯坦福大学分别发文 52 和 37 篇，位居第二和第四。中国占四席，中国科学院和清华大学分别发文 72 篇和 41 篇，分别位居第一位和第三位。俄罗斯科学院发表了 32 篇，位居第五。图 7 展示了

199 个机构节点和他们之间科研合作的 42 条连线，相对节点数量，节点之间的连线非常少，表明机构之间的科研合作较少，高产机构之间的合作更少，几乎没有合作。

4.2.2 国际大数据研究的国家（地区）及其之间的合作情况

国家之间的科研合作图谱可以展示大数据

方面的国际交流与合作情况。表7展示了大数据方面 2005-2015 年期间发文数量不少于 17 篇

的 TOP 发文国家及其发文数量。图 8 展示了大数据方面 TOP 发文国家之间的科研合作图谱。

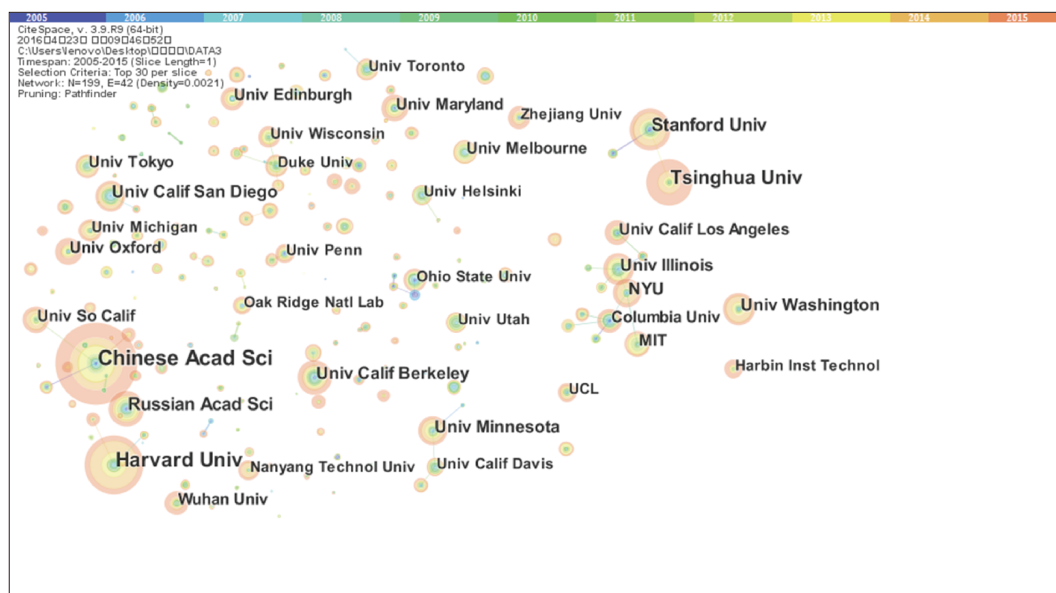


图 7 大数据方面 TOP 研究机构之间的科研合作图谱

表 7 2005-2015 年期间大数据方面 TOP 发文国家 (地区) 及发文量

排名	发文量	国家（地区）	译名	排名	发文量	国家（地区）	译名
1	1547	UNITED STATES of AMERICA	美国	17	50	SWEDEN	瑞典
2	557	PEOPLES R CHINA	中国	18	42	FINLAND	芬兰
3	322	UNITED KINGDOM	英国	19	40	BELGIUM	比利时
4	294	GERMANY	德国	20	39	NORWAY	挪威
5	177	FRANCE	法国	21	37	BRAZIL	巴西
6	168	ITALY	意大利	22	33	AUSTRIA	奥地利
7	167	JAPAN	日本	23	32	SINGAPORE	新加坡
8	162	AUSTRALA	澳大利亚	24	31	GREECE	希腊
9	145	CANADA	加拿大	25	29	ISRAEL	以色列
10	116	SPAIN	西班牙	26	28	POLAND	波兰
11	102	SOUTH KOREA	韩国	27	27	DENMARK	丹麦
12	100	NETHERLAND	荷兰	28	21	CZECH REPUBLIC	捷克
13	91	SWITZERLAND	瑞士	29	19	PORTUGAL	葡萄牙
14	63	INDIA	印度	30	19	IRAN	伊朗
15	61	TAIWAN	台湾	31	18	TURKEY	土耳其
16	59	RUSSIA	俄国	32	17	IRELAND	爱尔兰

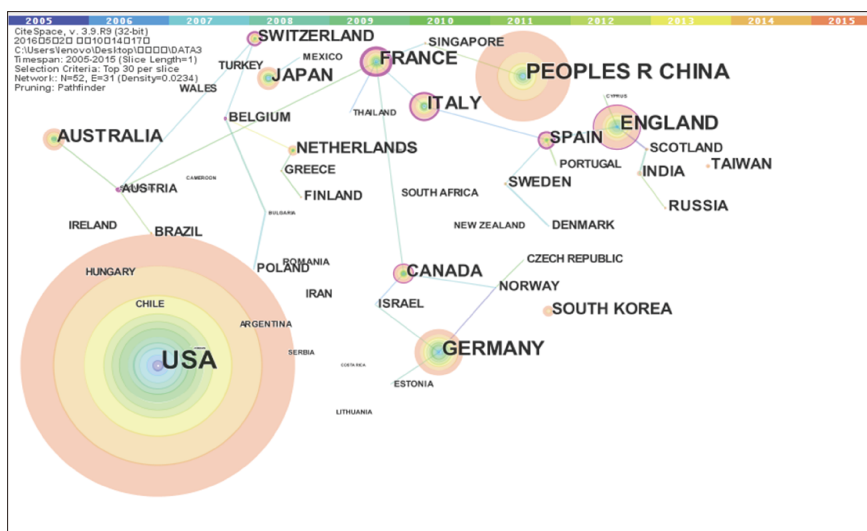


图 8 2000–2015 年期间大数据方面 TOP 发文国家之间的科研合作图谱（图中节点类型为“Country”，其他设置和表 1 一样）

结合表 7 和图 8，可以看出，大数据方面发文最多的四个国家分别为美国、中国、英国和德国，分别发文 1547 篇、557 篇、322 篇和 294 篇。其中美国在大数据方面的科研实力最强，远远超过其他国家或地区。然而国家之间的科研合作较少，图 8 共显示 52 个国家节点，但仅有 31 条科研合作连线。此外，高产国家美国和中国极少与其他国家合作。相对来说，德国、英国、法国、加拿大与其他国家科研合作相对较多。

5 国际大数据研究热点和前沿

5.1 基于关键词共现网络的研究热点探测

图 9 展示了大数据方面 348 个关键词节点及 308 条关键词共现连线的可视化图谱，图中节点大小表示关键词出现频次，节点之间的连线表示它们之间的共现关系。表 8 列出了大数据方面高频关键词的出现频次及中心度。表 9 列出了大数据方面爆发词的爆发强度、爆发开

始和结束时间。爆发词是在某时间段突然涌现的关键词，代表某时间段的前沿热点。

表 8 2000–2015 年期间大数据方面高频关键词的出现频次和中心度

主题名词	出现年代	出现频次	中心度
Big data（大数据）	2013	287	0.01
Model（模型）	2005	128	0.24
Classification（分类）	2005	101	0.13
Algorithm（算法）	2005	94	0.09
Models（模型）	2006	92	0.10
System（系统）	2006	92	0.03
Systems（系统）	2007	83	0.03
MapReduce	2011	79	0.02
Identification（识别）	2005	78	0.12
Prediction（预测）	2005	73	0.17
Networks（社交网络）	2005	72	0.05
Data mining（数据挖掘）	2005	67	0.08
Risk（风险）	2008	67	0.04
Algorithms（算法）	2005	64	0.05
Performance（性能）	2009	63	0.06
Management（管理）	2008	63	0.04
Evolution(进化)	2005	61	0.08

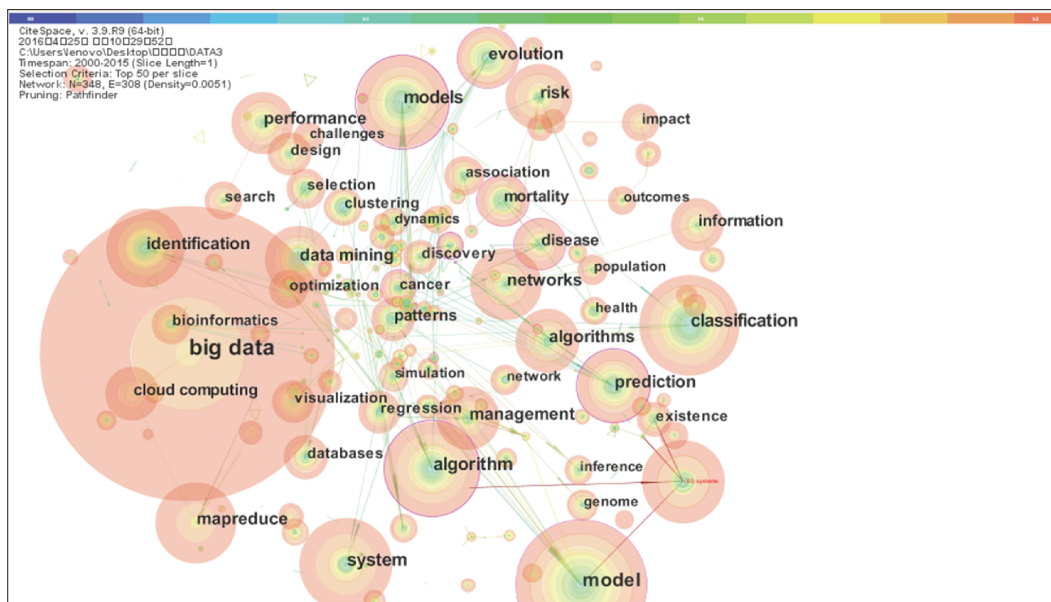


图 9 2000-2015 年期间大数据方面高频关键词的共现图谱

结合图9和表8,可以看出,大数据方面的热点研究主题主要有:大数据模型研究(共出现220次,中心度之和达到0.34);大数据相关算法研究(共出现158次,中心度之和达到0.14);大数据分类研究(共出现101次,中心度之和达到0.13);大数据相关系统研究(共出现175次,中心度之和达到0.06);大数据分布式处理工具,如MapReduce研究(出现79次,中心度达到0.02);基于大数据的识别与预测研究,其中关键词—识别和预测分别出现了78次和73次,中心度分别为0.12和0.17。此外,基于大数据的云计算研究为近期兴起的前沿热点方向。

利用陈超美提出的爆发词探测算法从文献关键词、题目和摘要分词中获得能够反映研究前沿的爆发词或突变词，如表 9 所示。从表 9 可以看出，大数据方面的研究前沿主要有：医疗领域中基于大数据分析的 Mortality(死亡率)

研究；基于大数据的可视化、优化和聚类研究，爆发强度分别为 4.37、4.3 和 4.28；大数据密度研究，爆发强度达到 3.52；基于大数据的监控研究，爆发强度达到 3.4。

表 9 大数据方面爆发词的爆发强度、爆发开始和结束时间

名称	Strength	begin	end
Mortality（死亡率）	4.54	2008	2009
Visualization（可视化）	4.37	2010	2012
Optimization（优化）	4.3	2010	2011
Clustering（聚类）	4.28	2000	2005
Density（密度）	3.52	2010	2012
Monitoring（监控）	3.4	2012	2013
Health（健康）	3.24	2008	2009
Microarray(微阵列)	3.20	2000	2009
Expectation Maximization Algorithm（演算法）	3.19	2006	2008
Evolution（进化）	3.07	2010	2011
Maximum-likelihood(最大似然)	2.98	2006	2009
Inference（推理）	2.91	2010	2013

5.2 基于文献共被引的经典文献探测

图 10 展示了 2005–2015 年期间大数据方面文

献共被引聚类的时间演化图谱。表 10 列了大数据方面爆发引文的爆发强度、爆发开始和结束时间。

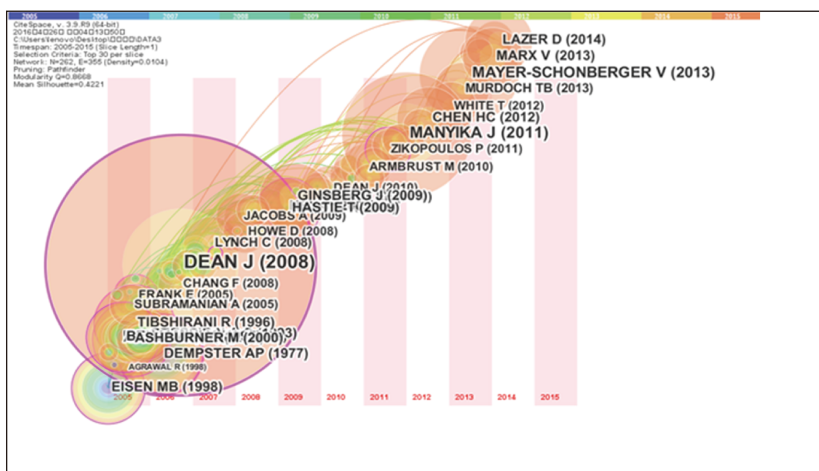


图 10 文献共被引聚类的时间 (Time Zone) 演化图谱 (图中节点类型为“Cite Reference”其他参数和表 1 一样)

通过图 10, 可以发现大数据方面被引较多的经典文献和奠基性文献, 这些文献主题也反映出大数据方面曾经的研究热点和近期的热点前沿。被引频次最多的是 Dean, Jelffrey; Ghemawat, Sanjay 在 2008 年发表的名为 *MapReduce: Simplified data processing on large clusters*, 被引频次高达 108 次, 从文献标题和内容可以看出, 作者研究的是大数据分布式并行处理工具: MapReduce。被引频次排名第

二的是 Manyika J, Chui M, Brown B, et al 在 2011 年发表的名为 *Big data: The next frontier for innovation, competition, and productivity*, 被引频次为 52 次, 该论文主要介绍了下一个创新、竞争和生产力前沿 -- 大数据。被引频次排名第三的是 MAYER SCHONBERGER 在 2013 年发布的 *Biology: The big challenges of big data*, 被引频次为 40 次, 作者介绍了生物学领域面临的大数据挑战与机遇。

表 10 大数据方面爆发引文的爆发强度、爆发开始和结束时间

References	Year	Strength	Begin	End
DEMPSTER AP, 1977, J ROY STAT SOC B MET, V39, P1	1977	5.3853	2005	2008
EISEN MB, 1998, P NATL ACAD SCI USA, V95, P14863	1998	5.1512	2005	2009
GUHA S, 1998, P ACM SIGMOD INT C M, V, P73	1998	4.4623	2005	2011
ZHANG T, 1996, P 1996 ACM SIGMOD IN, V,, P	1996	4.1013	2005	2011
SCHOLKOPF BERNHARD, 2002, LEARNING KERNELS, V, P	2002	3.4815	2005	2009
IHMELS J, 2002, NAT GENET, V31, P370	2002	3.2429	2005	2009
KOHONEN T, 1995, SELF ORG MAPS, V, P	1995	3.0291	2006	2007
CRESSIE N A C, 1993, STAT SPATIAL DATA, V, P	1993	2.9461	2006	2008
PLATT JC, 1999, ADVANCES IN KERNEL METHODS, V, P185	1999	2.6812	2005	2010
BENJAMINI Y, 1995, J ROY STAT SOC B MET, V57, P289	1995	2.5145	2007	2008

从表 10 可以看出,大数据方面爆发的引文中,大部分发表在 2002 之前,而爆发开始于 2005 年及其之后,这些爆发性引文是大数据研究参考的热点文献和重要知识基础。其中爆发强度最高的是 DEMPSTER AP 在 1977 年发表的名为 *Maximum Likelihood from Incomplete Data via the EM Algorithm* 的文献,这篇文献主要通过最大期望算法分析了基于不完整数据的最大似然估计,这也是当前大数据算法研究的核心基础文献。此外,爆发强度排名第二的是 Eisen MB 和 Spellman PT 等学者在 1998 年发表的名为 *Cluster analysis and display of genome-wide expression patterns*,此文主要探讨了全基因组表达模式的聚类分析与显示,既然作为大数据方面研究的核心知识基础,可见大数据方面的应用研究已经拓展到生物学:全基因组表达模式聚类分析领域。爆发强度排名第二的是 GUHA S 在 1998 年发表的名为 *CURE: an efficient clustering algorithm for large databases* 文献,文中阐述了一种高效的,面向大型数据库的聚类算法,此聚类算法也是面向大数据的新型聚类算法研究的核心知识基础。

6 结论

通过知识图谱,对国内(2000–2015)和国际(2005–2015)大数据方面的科研合作、研究热点与前沿进行可视化分析。分析结果显示:

(1) 国内大数据方面科研社区较多,但规模较小。最大的科研社区有五个,分别由潘懋、吴焕萍、周来水、李建中和张丽艳等几位作者为核心形成,他们的研究主题主要为海量数据

处理方法、模型和工具。国际大数据科研社区数量较少,但规模较大。其中最大的大数据研究社区,规模达到 30 多人,研究主题主要为大数据处理。

(2) 在大数据方面 TOP 研究机构、国家及其科研合作方面,国家之间,尤其高产国家之间的科研合作极少,国内外研究机构之间的合作都非常少,在 TOP30 国际科研机构中,大部分为美国高校,共 16 所,占比高达 53%。中国占四席,中国科学院和清华大学分别发文 72 篇和 41 篇,分别位居第一位和第三位。

(3) 国内大数据方面的研究热点主要集中于:基于大数据的云计算和云存储研究;海量数据分布式处理研究;基于 Hadoop 分布式处理平台的海量数据挖掘和数据分析研究;围绕物联网的数据存储、数据中心和社交网络研究;大数据应用研究,涵盖:基于谷歌和腾讯互联网平台的社会化媒体和数字营销研究,以及基于大数据技术的精准营销、信息安全、智慧城市和互联网金融等新前沿热点研究。而国际大数据方面的研究热点主要集中于:大数据模型研究;大数据相关算法研究;大数据分类研究;大数据相关系统研究;大数据分布式处理工具,如 MapReduce 研究;大数据应用研究,如:基于大数据的识别与预测应用研究。此外,基于大数据的云计算研究为近期兴起的前沿热点方向。

通过分析对比,我国大数据研究热点与国际大数据热点比较一致的有:基于大数据的云计算;海量数据分布式处理研究(MapReduce 和 Hadoop);大数据应用研究。而国内涉及较少的研究热点有:大数据模型和算法研究、大

数据分类研究、大数据相关系统研究等。

参考文献

- [1] Chen C, Ibekwe-Sanjuan F, Hou J. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis[J]. Journal of the American Society for Information Science & Technology, 2010, 61(7):1386-1409.
- [2] 胡泽文, 孙建军, 武夷山. 国内知识图谱应用研究综述[J]. 图书情报工作, 2013, 57(3):131-137.
- [3] 胡泽文, 武夷山, 孙建军. 数字资源保存的研究进展、热点与前沿[J]. 数字图书馆论坛, 2013(2):24-38.
- [4] 胡泽文, 武夷山, 袁军鹏. 零被引研究文献的知识图谱分析——历史发展脉络、主体和主题[J]. 情报科学, 2016, 36(3):85-91.
- [5] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity[J]. Analytics, 2011.
- [6] 王新才, 丁家友. 大数据知识图谱:概念、特征、应用与影响[J]. 情报科学, 2013(9):10-14.
- [7] 党倩娜, 罗天雨, 曹磊. 多维视角下大数据领域技术创新演进、前沿与特性[J]. 科学学与科学技术管理, 2015, 36(8):49-60.
- [8] 陈超美, 陈悦, 侯剑华,等. CiteSpace II :科学文献中新趋势与新动态的识别与可视化[J]. 情报学报, 2009, 28(3):401-421.
- [9] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the China Society for Scientific & Technical Information, 2009, 57(3):359-377.
- [10] Chen C. Searching for intellectual turning points: Progressive knowledge domain visualization [J]. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 2004, 101(1):5303-5310.
- [11] 薛云, 余丰民, 汤江明. 国内知识图谱学术研究脉络:2004-2010[J]. 数字图书馆论坛, 2012(6):42-50.
- [12] 谢然. 大数据应用案例TOP 100[J]. 互联网周刊, 2015(6):52-65.
- [13] 谢然. 大数据的七大核心价值[J]. 互联网周刊, 2015(8):54-60.
- [14] 谢然. 大数据可视化之美[J]. 互联网周刊, 2014(11):32-34.
- [15] 王伟, 蒋菱, 王峥,等. 基于向量评估遗传算法的智能电网大数据交易模型研究[J]. 电网与清洁能源, 2016, 32(10):1-8.
- [16] 王伟, 廖正宇, 张辉,等. 基于大数据的铁路信号系统数据存储与分析系统设计与实现[J]. 信息网络安全, 2017(1):29-37.