

数据共享论文的文献计量分析

朱艳华¹ 胡良霖¹ 高瑜蔚¹ 于铁强²

(1. 中国科学院计算机网络信息中心, 北京 100190; 2. 北京软件和信息服务业交易所有限公司, 北京 100086)

摘要: 及时了解和数据分析数据共享方面的研究文献, 对指导数据管理和共享活动具有重要意义。文章对利用中国知网学术文献总库检索到的数据共享研究领域的论文共2067篇(截至2017年5月)进行文献计量分析。结果显示, 论文呈现逐年增长的趋势, 不同文献类型的论文关注数据共享的不同方面, 数据共享研究与数据政策的实施呈现正相关性, 自然科学研究领域更重视数据共享研究等。

关键词: 数据共享; 科学数据; 数据政策; 文献计量

中图分类号: G35

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2018.05.013

Analysis of Papers on Data Sharing Based on Bibliometric

ZHU Yanhua¹, HU Lianglin¹, GAO Yuwei¹, YU Tieqiang²

(1. Computer Network Information Center of CAS, Beijing 100190; 2. Beijing Software and Information Services Exchange co.,Ltd, Beijing 100086)

Abstract: Based on research results from National Knowledge Infrastructure (CNKI), we count those papers that contain data sharing in the title of each article by May 2017. The result shows that the number of articles is 2,076 and presents a growing trend. Different literature types focus on different aspects of data sharing, data sharing research is positively related to data policy. Research on data sharing is more important in natural science research field.

Keywords: data sharing, scientific data, data policy, literature bibliometrics

数据共享已经得到各国政府和国际组织的共识, 全面公开地获取数据不仅能使其在应用过程中增值, 也是推动科学认识突破的重要条件。因此, 及时关注、了解和分析科研人员在数据共享方面的研究, 对于指导数据管理和共享活动具有重要的现实意义。本文将对数据共享研究的论文

进行文献计量分析, 以揭示数据共享研究的现状和发展趋势。数据主要来自中国知网学术文献总库, 以“数据共享”为检索词, 在文章篇名中进行精确检索, 检索时间截至2017年5月8日。选取的文献类型包括期刊、报纸、硕士论文、博士论文、国内会议和国际会议^[1]。一共检索得

作者简介: 朱艳华(1982—), 女, 中国科学院计算机网络信息中心高级工程师, 主要研究方向: 数据库技术与标准规范、数据应用服务; 胡良霖(1973—), 男, 中国科学院计算机网络信息中心正研级高级工程师, 主要研究方向: 数据库技术与标准规范、数据质量与数据服务; 高瑜蔚(1989—), 女, 中国科学院计算机网络信息中心工程师, 主要研究方向: 数据库技术与标准规范; 于铁强(1976—), 男, 北京软件和信息服务业交易所有限公司标准研究中心总经理, 工程师, 主要研究方向: 软件和信息服务业交易服务、数据流通、人工智能。

基金项目: 国家重点研发计划“分布式科技资源体系及服务评价技术研究”(2017YFB1400200); 中国科学院“十三五”信息化专项项目“科学大数据工程”(XXH13505)。

收稿时间: 2017年9月22日。

到 2067 篇论文，每条记录包括题名、作者、来源、发表时间、来源数据库、被引次数、下载次数等基本信息。本文将分别从论文的学科、发表年度、论文类型、文献来源、作者机构、资助基金、研究层次等方面进行统计分析。

1 论文的发文量、类型和来源

最早的一篇论文是《I/O 重定向与管道技术及其在实现数据共享方面的应用》，于 1986 年发表在《交通与计算机》上，介绍了输入/输出重定向和管道技术，是 UNIX 操作系统和由其演变而来的 XENIX 操作系统中很有特色的计算机管理技术；发表论文最多的年份是 2016 年，共 213 篇。历年的发文量，如图 1 所示。

从图 1 可以看到，2002—2004 年以及 2014—2016 年，这两个时间段发文量有明显的增长趋势。首先，科学数据共享工程的启动带来了 2002—2004 年数据共享论文量快速增长的第一个阶段。2001 年年底，科学数据共享工程启动第一个试点——气象科学数据共享试点，然后在资源环境、农业、人口与健康、基础与前沿等领域共 24 个部门开展了科学数据共享工作^[2]。此后 3 年，与科学数据共享工程相关的数据共享论

文数量持续增加。其次，大数据的快速发展与相关政策密集出台带来了 2014—2016 年快速增长的第二个阶段。2015 年 8 月，国务院印发《促进大数据发展行动纲要》，提出大数据成为推动经济转型发展的新动力、重塑国家竞争优势的新机遇以及提升政府治理能力的新途径^[3]。2016 年 3 月，根据两会授权，新华社在线发布了《中华人民共和国国民经济和社会发展第十三个五年规划纲要》，明确提出建设包括大数据应用在内的 8 项信息化重大工程^[4]。随后，国家发展改革委、工业和信息化部、科技部等相继启动大数据发展计划和研究项目。国家发展改革委发布《国家发展改革委办公厅关于组织实施促进大数据发展重大工程的通知》^[5]、科技部发布云计算和大数据重点专项 2016 年度项目申报指南^[6]、国家自然科学基金委发布大数据驱动的管理与决策研究重大研究计划 2016 年度项目指南等^[7]。多个省市也按照行动纲要要求公开公布相关科学数据，以上海市、贵州省为代表的部分省市先后规划、部署、实施大数据工作；阿里巴巴、百度、华为等国内产业巨头纷纷布局大数据研究和应用生态链建设。受到国家政策和相关资助项目的鼓励和支持的情况下，研究数据共享的学者们给予数据共

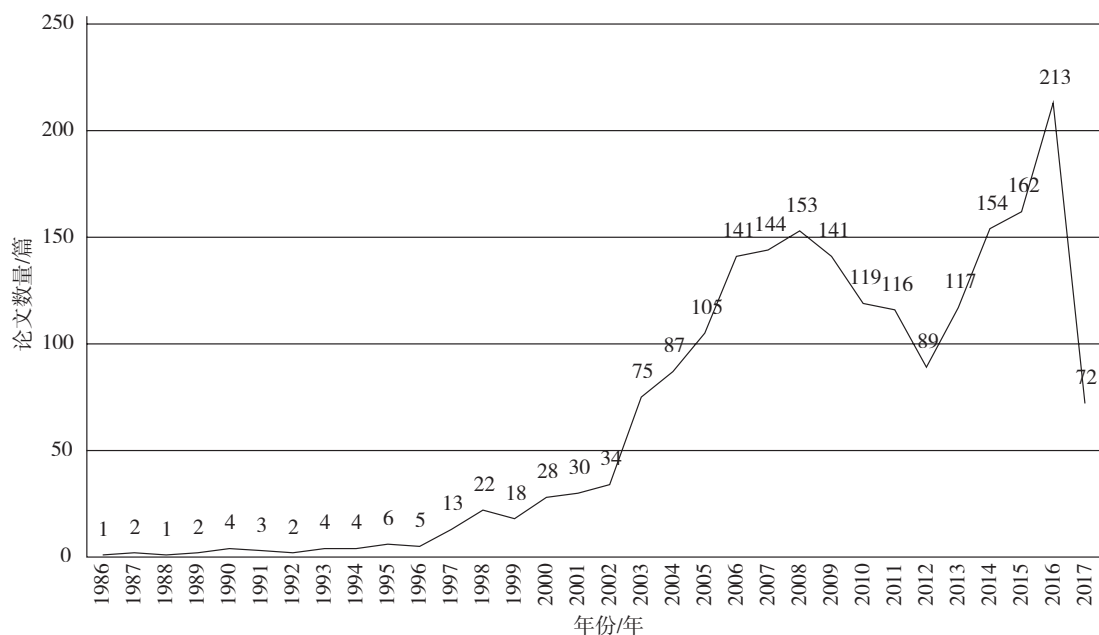


图 1 论文的年度发文量

享的大数据研究充分的关注和解读。此外,数据共享研究资助基金情况也在一定程度上体现了国家的数据政策导向。

在2067篇论文中,期刊论文1401篇,占论文总量的67.78%;报纸文章343篇,占论文总量的16.59%;硕士学位论文207篇,占论文总量的10.02%;博士学位论文11篇,占论文总量的0.53%;国内会议89篇,占论文总量的4.31%;国际会议16篇,占论文总量的0.77%。详见图2。

按文献来源的发文量进行排序,排名前10的期刊和报纸分别为:科技日报(24篇,14.29%)、中国气象报(23篇,13.69%)、中

国基础科学(23篇,13.69%)、测绘与空间地理信息(19篇,11.31%)、贵阳日报(17篇,10.12%)、电脑知识与技术(14篇,8.33%)、地球信息科学学报(14篇,8.33%)、计算机工程(12篇,7.14%)、微计算机信息(11篇,6.55%)、计算机应用与软件(11篇,6.55%)。详见图3。

2 论文的学科、领域与研究热点

2067篇数据共享论文中,标引学科与领域的共计1944篇论文。其中,有1465篇属于自然科学研究层次(其中,工程技术、基础与应用

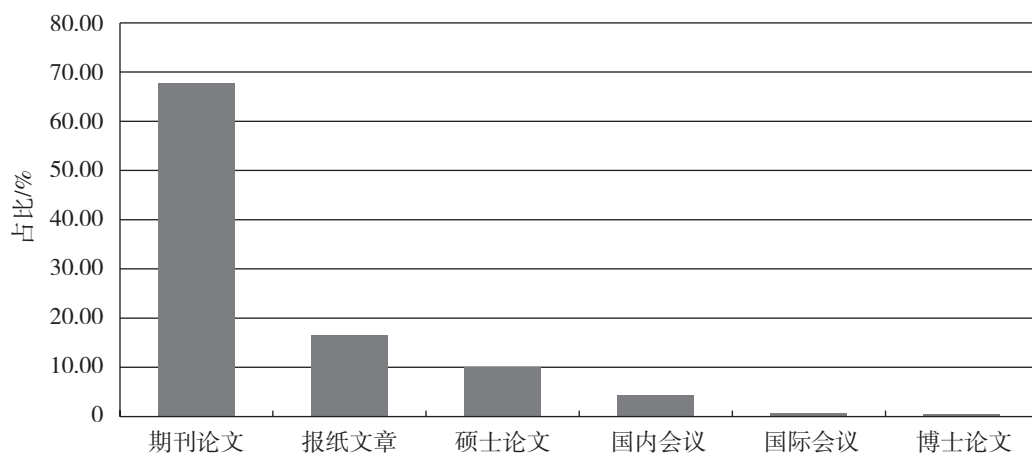


图2 论文的类型分布

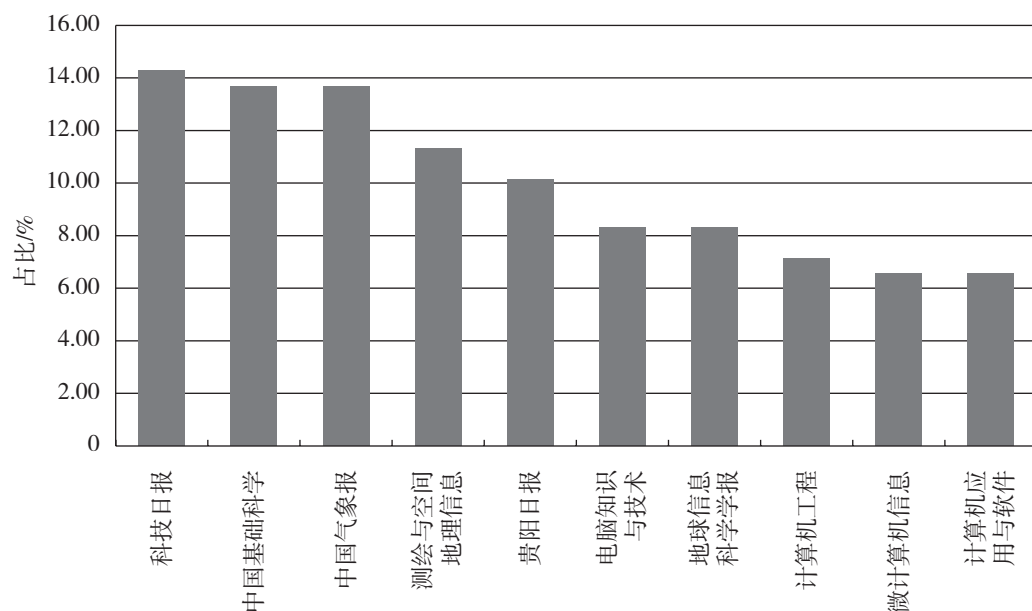


图3 论文文献来源分布

基础研究、行业技术指导、专业实用技术、政策研究、标准与质量控制分别是 976 篇、368 篇、68 篇、33 篇、18 篇、2 篇), 占标引总量的 75.36%; 454 篇属于社会科学研究层次(其中, 行业指导、基础研究、政策研究、职业指导分别是 207 篇、144 篇、75 篇、28 篇), 占标引总量的 23.35%; 25 篇属于其他研究层次(其中, 大众科普、高等教育、经济信息、大众文化、基础教育与中等职业教育分别有 11 篇、9 篇、3 篇、1 篇、1 篇), 占标引总量的 1.29%。详见图 4。

其中, 发表论文最多的前 10 门学科的分布及文章数量分别是: 计算机软件及计算机应用(851 篇, 50.41%)、自然地理学和测绘学(227 篇, 13.45%)、互联网技术(152 篇, 9.00%)、自动化技术(92 篇, 5.45%)、科学研究管理(69

篇, 4.08%)、图书情报与数字图书馆(67 篇, 3.97%)、地球物理学(59 篇, 3.50%)、地质学(58 篇, 3.44%)、信息经济与邮政经济(57 篇, 3.38%)、气象学(56 篇, 3.32%)。详见图 5。

数据共享论文排名前 10 的关键词分别为: 数据共享(624 篇, 52.35%)、共享(100 篇, 8.39%)、元数据(97 篇, 8.14%)、科学数据(86 篇, 7.21%)、XML(68 篇, 5.71%)、科学数据共享(53 篇, 4.45%)、数据库(52 篇, 4.36%)、WebGIS(38 篇, 3.19%)、共享平台(37 篇, 3.10%)、空间数据(37 篇, 3.10%)。这些关键词涉及数据共享、元数据标准、科学数据共享、数据共享与建库技术、领域数据共享等内容主题。详见图 6。

此外, 中国知网还对关键词共同出现的情

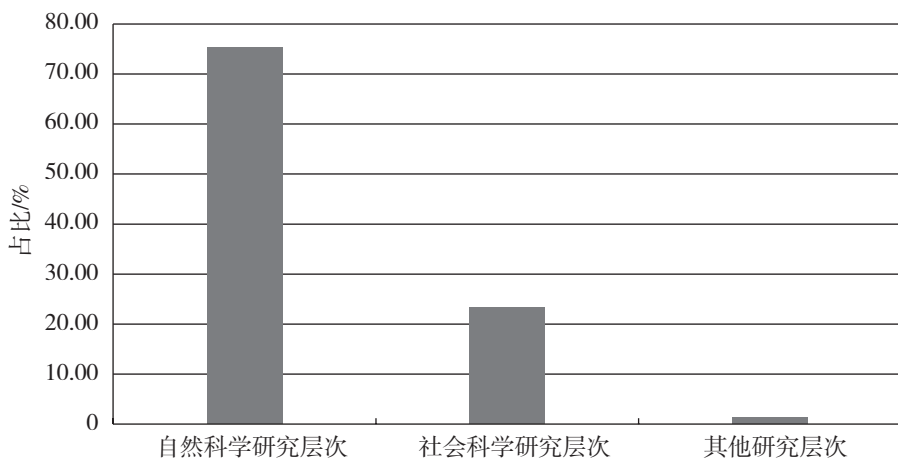


图4 论文的学科分布

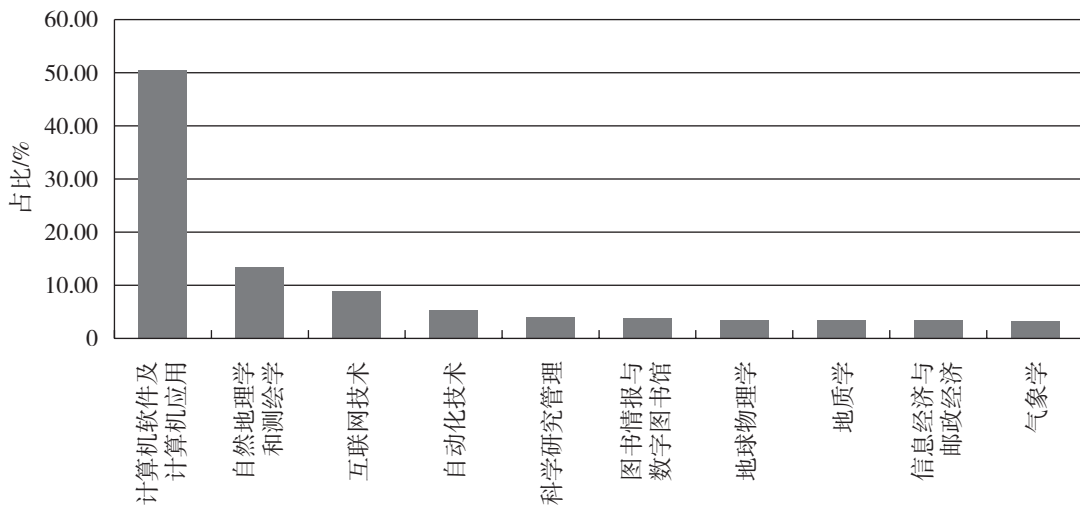


图5 论文所属学科前十名排名

况进行了共现矩阵分析。其中,与“数据共享”同时出现的关键词前10个分别为:元数据(46次)、科学数据(31次)、XML(30次)、数据库(21次)、Web GIS(12次)、数据交换(13次)、GIS(13次)、空间数据(12次)、大数据(11次)、Web Service(10次)。由此可见,研究数据共享的论文主要集中在元数据标准、科学数据共享、数据共享技术、领域数据共享等方面。

3 论文的作者机构

根据论文第一作者所在单位的行业性质划

分,发表数据共享论文的作者单位包括高等学校、科研机构、政府管理部门、公司企业等。其中,发表论文最多的10个单位分别为:中国科学院(56篇,25.81%)、武汉大学(43篇,19.81%)、中国农业科学院农业信息研究所(18篇,8.29%)、浙江大学(18篇,8.29%)、解放军信息工程大学(16篇,7.37%)、南京大学(15篇,6.91%)、华中科技大学(14篇,6.45%)、电子科技大学(13篇,5.99%)、中国水利水电科学研究院(12篇,5.54%)、河海大学(12篇,5.54%)。详见图7。

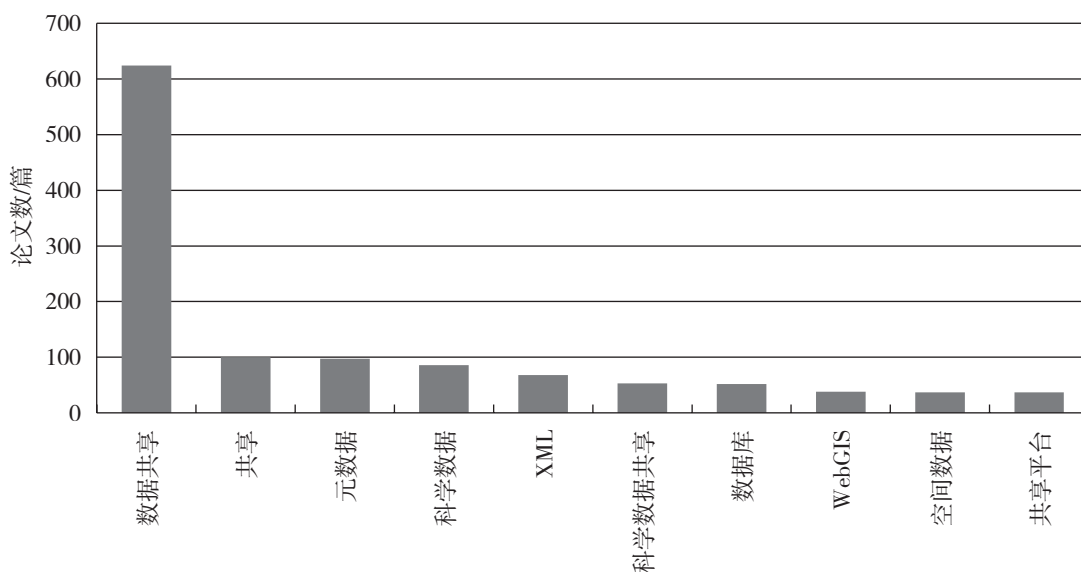


图6 数据共享论文关键词前十名分布

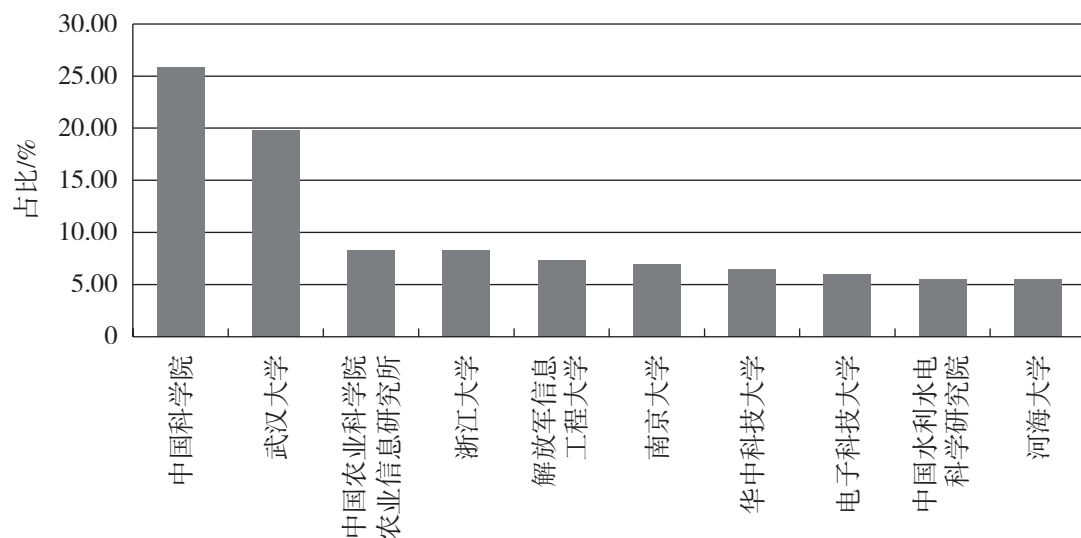


图7 论文作者机构分布

4 论文资助基金

2067 篇论文中共得到 341 项基金支持。其中, 国家部委基金 278 项, 占基金总量的 81.52%; 地方政府基金 43 项, 占基金总量的 12.61%; 高校基金 10 项, 占基金总量的 2.93%; 科研院所基金 8 项, 占基金总量的 2.35%; 企业基金 2 项, 占基金总量的 0.59%。详见图 8。

5 结论

本文通过中国知网学术文献总库, 检索到 1986—2017 年数据共享研究领域论文 2067 篇, 对这些论文进行了文献计量分析, 得到如下分析结果。

(1) 不同文献载体关注数据共享的不同方面。数据共享论文主要发表在学术期刊上, 占了论文总量的一半以上; 报纸文章也在宣传数据共享政策和共享现状方面发挥了重要作用, 发文量位居第二; 学位论文中的硕士论文和博士论文占论文总量的 10.55%。针对数据共享的研究主题, 3 种文献载体各有侧重。期刊论文关注数据共享学术研究, 如数据共享政策/模式、数据共享和管理技术、数据共享标准规范体系、数据共享服务效果评估、数据共享平台建设和服务等。报纸文章则侧重报道和宣传国内外数据共享政策和相关学术活动、科学数据共享平台建设进展和成效等, 如(地方)政府数据共享开放(特别是

以上海市、贵州省为代表的政府数据共享开放路径可行性探索)、政府数据与智慧城市、科研数据共享平台、商业数据共享等。学位论文的研究重点聚焦在以下几个专题: 数据共享平台研究与实现、数据查询与交换技术、异构数据融合与集成、元数据标准与管理、数据共享机制与策略研究等方面。

(2) 载文的期刊既有计算机工程、测绘与空间地理信息、中国医药导刊这类专业性期刊, 也有中国基础科学、电子科技大学等综合性期刊, 内容涉及基础科学、计算机、医学、气象、交通、农业、医学、海洋等学科领域; 报纸既有科技日报、人民日报、中国信息报等全国性报纸, 也有贵州日报、济南日报、银川日报等地方性报纸, 前者主要关注全国的数据共享政策和方针, 后者则重点关注当地政府的数据共享进展和现状。

(3) 数据共享研究与数据政策和资助力度呈现正相关性。2002—2004 年以及 2014—2016 年两个时间段的论文数量增长趋势明显, 这是因为国家启动了数据共享项目和颁布了大数据政策, 以及多个省市也按照行动纲要要求公开公布了相关科学数据。此外, 国家部委基金项目 and 各级地方政府基金的支持, 也是导致论文增长的原因。从论文的资助基金分布来看, 在研究数据共享论文的基金项目中, 80% 是通过国家部委基金项目的支持, 其中, 国家自然科学基金、国家科技基

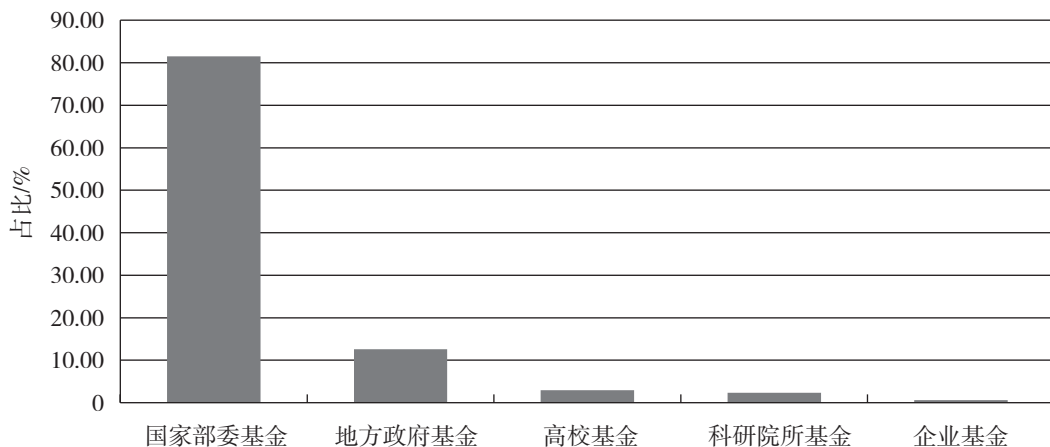


图 8 论文支持基金类型分布

基础条件平台建设计划、国家高技术研究发展计划排在前列。此外,各级地方政府基金也积极支持数据共享的研究,在资助基金中所占比例位居第二。

(4) 论文主要集中在自然科学研究领域,将近占论文总量的80%,而社会科学研究也占了一定比例。其中,自然科学领域中占份额最多的是“工程技术”共976篇,占该研究层次的66.62%;社会科学研究中所占份额最多的是“行业指导”共207篇,占该研究层次的45.60%。论文的专业主要集中在计算机软件及计算机应用领域、互联网技术、自动化技术等学科,这些领域重点研究数据共享的技术及其应用。论文涉及的专业还聚焦在地理学和测绘学、地球物理学、地质学、气象学等具体学科领域,这些领域重点关注学科数据共享平台建设、开放服务等研究主题。论文的研究热点包括数据共享模式、元数据标准规范、科学数据共享平台、数据共享技术和领域数据共享等方面。

(5) 论文的作者机构既有研究单位、政府部门,也有公司企业,体现了数据共享研究的核心团队。发表论文的机构首位是中国科学院,以下依次是武汉大学、中国农业科学院农业信息研究所、浙江大学、解放军信息工程大学、南京大学、华中科技大学、电子科技大学等,分别发文56篇、43篇、18篇、18篇、16篇、15篇、14篇、13篇等。这些机构的研究人员是国内数据共享研究的核心力量。其中,中国科学院一直高度重视科学数据在科研发现、信息化建设中的创新及应用。20世纪70年代,中国科学院开始建设专业数据库。经过数十年的持续建设,截至“十二五”项目结束,“科技数据资源整合与共享工程”系统地整合了58家单位的科学数据库,可共享数据量达655TB^[8]。

(6) 当前,缺乏有关数据管理的研究,而这

方面的研究将是为有关部门制定数据共享政策和法规提供理论基础的,还有待进一步加强。2018年1月23日,中央全面深化改革领导小组第二次会议23日审议通过了《科学数据管理办法》,并强调加强和规范科学数据管理,要适应大数据发展形势,积极推进科学数据资源开发利用和开放共享。该办法的出台和实施,将进一步加强和规范科学数据管理,保障科学数据安全,提高开放共享水平,也将极大地推动我国各领域科学数据之间、科学数据与其他领域数据之间的整合和共享,为科技创新和经济社会发展提供有力支撑,让科研人员和普通公众从获取数据中获益。

参考文献

- [1] 中国知网学术文献总库平台[EB/OL].[2017-05-08]. <http://www.cnki.net/>.
- [2] 科学数据共享工程启动[EB/OL]. [2017-05-15]. http://www.most.gov.cn/ztl/kjzg60/kjzg60hhcj/kjzg60j-cyj/200909/t20090911_72832.htm.
- [3] 国务院关于印发促进大数据发展行动纲要的通知[EB/OL]. [2017-05-25]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.
- [4] 中华人民共和国国民经济和社会发展第十三个五年规划纲要[EB/OL]. [2017-05-25]. http://news.xinhuanet.com/politics/2016lh/2016-03/17/c_111836632.
- [5] 国家发改委办公厅关于组织实施促进大数据发展重大工程的通知[EB/OL]. [2017-05-25]. <http://bigdata.sic.gov.cn/News/510/6456.htm>.
- [6] 科技部关于发布国家重点研发计划精准医学研究等重点专项2016年度项目申报指南的通知[EB/OL]. [2017-05-25]. http://www.most.gov.cn/tztg/201603/t20160308_124542.htm.
- [7] 大数据驱动的管理与决策研究重大研究计划2016年度项目指南[EB/OL]. [2017-05-25]. <http://www.nsf.gov.cn/publish/porta10/zdyjjh/2016/info52608.htm>.
- [8] 中国科学院数据云[EB/OL]. [2017-05-08]. <http://www.csdb.cn/>.