

·专题: 大数据研究·

基于知识图谱的国内大数据研究 可视化分析



□马续补 刘 玮 陈 颖

[西安电子科技大学 西安 710071]

[摘 要] 大数据被应用到更多领域, 已经成为社会发展的重要推动力。通过对CNKI和CSSCI数据库中大数据库领域文献的计量研究, 对文献分布特征、研究领域、机构与著者等外部特征进行统计描述; 利用CiteSpace和VOSviewer对高被引文献之间的共现关系进行聚类辨识经管领域大数据研究的知识基础, 对高频关键词的共现关系进行历时分析识别经管领域大数据研究热点及其演化过程。研究发现, 国内大数据领域文献数量多, 但国际影响力不强; 大数据研究的脉络沿着2012~2014年爆发期形成的知识基础稳步发展; 国内经管领域的大数据库研究知识基础由管理学、计算机科学、图书情报学、传播学和教育学领域构成; 研究热点受到大数据技术发展的影响, 也明显受到国家政策影响。

[关键词] 大数据; 共现分析; 知识图谱; 综述

[中图分类号] G353.11

[文献标识码] A

[DOI] 10.14071/j.1008-8105(2018)-1006

Knowledge Mapping Analysis of Domestic Research on Big Data

MA Xu-bu LIU Wei CHEN Ying

(Xidian University Xi'an 710071 China)

Abstract Big data has been applied in various fields and it has become the most important impetus for the social development. In this paper, a review on some significant characteristics, emerging trends and intellectual bases has been studied based on Chinese document databases and software tools. The research draws the following conclusions. Firstly, even though there are many literatures in the area of big data, most of them are lack of important impacts internationally. Secondly, big data research develops steadily following research trends created in 2012–2014. Thirdly, for big data research of economy and management field, management, computer science, information science, communications and education are major literatures bases according to citation and co-citation footprint in scientific literature. Finally, both technical development and public policies affect research trends in China.

Key words big data; co-citation; knowledge mapping; review

我国在2015年8月通过《关于促进大数据发展的行动纲要》, 2015年11月《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》中首次提出推行国家大数据战略。《关于全面推进政务公开工作的意见》《政务信息资源共享管理暂行办法》等一系列政策陆续发布。2018年初《公共信息

资源开放试点工作方案》确定在北京、上海、浙江、福建、贵州开展公共信息资源开放试点。政府信息公开网站如bjdata.gov.cn, datashanghai.gov.cn, gzdata.gov.cn等已经开始运行。大数据以人工智能、公共信息资源开放、智慧城市等多种形式改变着社会。

[收稿日期] 2018-03-08

[基金项目] 国家自然科学基金青年项目“基于公众网络参与的民生公共政策第三方动态评估机理与方法研究”(71503195); 陕西省社会科学基金青年项目(2015R020)。

[作者简介] 马续补(1983-)男, 博士, 西安电子科技大学经济与管理学院讲师; 刘玮(1975-)女, 博士, 西安电子科技大学经济与管理学院副教授; 陈颖(1997-)女, 西安电子科技大学经济与管理学院本科生。

大数据应用于政务活动中,使得政府部门在节约经费、提高生产率和创新方面更加有效^[1];大数据分析技术给健康、自然灾害等重大问题决策提供了低成本的机遇^[2];大数据推进个性化医疗、降低临床风险、预测分析和患者数据自动报告等^[3];大数据技术可以将信息和服务推送到最精准的时间、地点和目标客户,如电商平台的用户画像。物联网与大数据协同工作,通过物联网跟踪各种基础设施状态^[4],如智慧城市建设中各种管网检测。

大数据已经成为研究和应用领域关注的重要主题,相关文献处于快速增长期^[5],因此有必要从文献角度梳理我国大数据研究和应用现状,探讨我国大数据研究的知识基础和发展脉络、研究热点和演进过程,以供研究者们更深入的探索。

一、数据来源和研究方法

(一) 数据来源

文献[6]中,采用“大数据”“海量数据”“大规模数据”和“海量信息”等术语表示“大数据”概念^[6];但更多的文献^[7-10]采用“大数据”一个术语表达“大数据”概念。考虑到2011年以后“大数据”作为一个概念已经被广泛认可和使用,且相关文献在2011年以前数量少,因此本文选择“大数据”作为检索词,数据库检索时间是2018年1月22日。文献数据来自CNKI和CSSCI,构造了五组检索策略以获得不同相关程度的数据集。检索策略1是在CNKI一框式检索中主题包括“大数据”的全部结果,检索策略2是在CNKI一框式检索中标题或关键词中包括“大数据”的全部结果,检索策略3是在策略2的基础上要求文献来源是较高水平的期刊(SCI来源期刊或EI来源期刊或核心期刊或CSSCI期刊),检索策略4是在策略3的基础上要求文献分类目录为“经济与管理科学”。检索策略1~4逐步精炼检索结果,对应数据集1~4。检索策略5是在CSSCI数据库中查找标题或关键词中包括“大数

据”的引文数据,主要用于参考文献共被引分析,对应数据集C。

(二) 研究方法

文献计量和科学计量方法已经被广泛应用于科学学和科研管理等研究领域。本文主要采用以下方法分析我国大数据研究领域的研究主题、知识基础、研究热点和演化路径等。

引文分析:引文分析是文献计量和科学计量的经典方法,通常认为引用与被引用可以解释知识的演化规律和过程,被引用可以表达知识影响力。本文通过引用路径和高被引分析以发现知识的演化过程和领域高影响力文献。

共现分析:共现关系能够表现信息的内容特征和外部特征,不仅具有语义内涵而且相互关联,这些特征共同构成了文本知识关联揭示和知识挖掘的基础^[11]。本文主要采用关键词共现分析以发现研究热点和演变过程,参考文献共被引分析以发现大数据研究的知识基础,作者/机构合作分析以发现研究团队。

本文使用了CiteSpace^[12]和VOSviewer^[13]两个文献计量可视化分析软件,CiteSpace知识可视化软件是目前国内最流行的知识图谱绘制工具之一,VOSviewer也是近年来一款流行的信息可视化分析工具。考虑到二者在分析结果上的差异^[14],本文同时采用了两个工具。

二、文献分布特征

(一) 文献数量分布及增长规律

在文献计量研究中,文献数量变化可以揭示一个研究领域的发展阶段。张姣姣等分析了1990~2015年WOS数据库中大数据应用领域的文献数量,认为该领域的文献在2012~2015年符合普赖斯指数增长规律,处于快速发展期^[5]。本文设计的五组检索策略得到相关文献数量分布见表1。不同检索策略得到的文献集合数量变化趋势相同,由此可以将国内大数据研究分为三个阶段:

表1 不同检索策略的命中文献数量分布

	-2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	合计
数据集1	3858	598	762	1824	6929	15439	25110	32535	33680	454	121189
数据集2	333	38	105	726	3590	8009	12760	16831	17965	402	60759
数据集3	897	118	125	140	832	2057	3050	3800	4033	69	15121
数据集4	20	6	4	18	177	432	677	845	892	9	3080
数据集C	8	1	1	14	152	393	633	715	428	0	2345

初始期(2011年以前):有关文献主要以“大数据集”“海量数据”等形式出现,文献累积量

少,增长缓慢。

爆发期(2012~2013年):有关文献以“大数

据”形式出现,文献量剧增,环比增速在5~10倍之间。

稳定增长期(2014年至今):大数据在更多领域得到应用,文献累积量不断增加,增速放缓,环比增速逐步回落到1~2倍。2014年环比增速大于2,2015年开始环比增速低于2。

相关度高、质量高的文献增速低于全部文献的增速,且爆发年份略晚于全部文献。对比同期WOS平台数据,国内会议、报纸等文献在大数据文献方面的爆发时间与WOS文献一致,而期刊论文滞后一年。国内经管类和社会科学领域的大数据文献在2013年显著增加。

(二) 文献来源和类型分布

有9种文献排在数据集1和2的文献来源前十强:贵阳日报、贵州日报、人民邮电、电脑知识与

技术、中国计算机报、中国电脑教育报、科技日报、通信世界。文献来源最多的是贵州省的两种报纸以及计算机相关领域的报刊。WOS平台数据也表明,大数据相关文献超过一半来自计算机领域,且会议论文占比高。贵州省近年来在大数据建设方面取得巨大成就在文献方面得到很清晰的体现。大数据相关文献来源多样,报纸和会议成为主要的文献产出途径,研究论文也占据了重要的地位;报纸是政府信息沟通的主要路径。

(三) 研究领域分布

大数据已经渗透到各个学科,大数据相关文献涉及到大数据技术、大数据应用、大数据管理等问题,表现在相关文献主要来自计算机科学研究领域,同时分布在经济、管理等更广泛的领域。表2是数据集1和2中相关文献研究领域分布。

表 2 大数据研究文献的学科分布

	计算机软件及 计算机应用	信息经济与 邮政经济	工业 经济	企业 经济	贸易 经济	金融	新闻 与媒体	高等 教育	宏观经济管理与 可持续发展	教育理论与 教育管理	图书情报与 数字图书馆	行政学及国家 行政管理
1	19.04	11.74	5.58	5.57	5.47	5.18	4.58	3.44	3.33	3.04	-	-
2	24.14	4.71	-	7.22	3.70	2.89	5.79	6.95	-	3.58	6.88	3.13

三、研究机构、基金支持与著者分析

(一) 机构与基金资助

在数据集1和2中,主要发文机构是:吉林大学、电子科技大学、武汉大学、北京邮电大学、山东大学、南京大学、清华大学、华中科技大学、中国人民大学和浙江大学。数据集4中发文最多的机构分别是:重庆理工大学会计学院、重庆理工大学云会计研究所、中国人民大学财政金融学院、中国人民大学新闻学院和辽宁大学商学院。通过CiteSpace对机构间合作进行分析,发现国内经管界在该研究领域合作不明显。

数据集1和2表明我国大数据文献主要基金资助来自:国家自然科学基金、国家社会科学基金、国家高技术研究发展计划、国家科技支撑计划、国家重点基础研究发展和中国博士后科学基金,其中国家自然科学基金和社科基金是最主要的资助机构。

(二) 高产作者与高被引作者

数据集1和2中,高产作者包括:程平、喻国明、杜创、朱恒鹏、王泽宇、高秋明、黄欣荣、邬贺铨、甄峰、廖震松等。其中重庆理工大学程平和中国人民大学喻国明发文最多。通过CiteSpace作者聚类分析,数据集4仅突显出以程平为核心的一个研究团队。整体上,经管领域大数据文献表现出的合作团队较少。

数据集C中,来自中文社科领域的高被引作者包括:《大数据时代》作者维克托·迈尔-舍恩伯

格,《大数据》作者涂子沛,中科院计算所李国杰,人大信息学院孟小峰,中国工程院邬贺铨,国家自然科学基金委管理学部冯芷艳,人大新闻学院彭兰以及桂林理工大学图书馆樊伟红。

四、基于高被引与共被引分析的知识基础识别

文献引用和被引用反映出知识继承、传递和发展脉络。1973年Small H提出共被引,认为两篇文献同时被其他文献引用形成知识网络,这两篇共被引文献在主题上具有相似性或相关性^[14]。引用关系和共被引网络可以反映出学科的结构、特征和演化过程。

(一) 高被引文献分析

数据集1中输出被引次数最高的12篇文献,见表3。有3篇文献发表在2011年之前,其中文献[2]和[8]是大数据技术研究,文献[12]是地理信息系统大数据集的研究。其余9篇文献分布在2011~2014年,与大数据研究密切相关。除文献[2],[8],[10],[12]外,其余8篇文献在CSSCI中属于被引次数大于50的高被引文献。CSSCI期刊论文对于技术性强的文献引用相对较少。高被引论文主要集中在2011~2014年,可以认为目前国内大数据研究依然沿着爆发期形成的方向发展,没有出现明显转折点,这些高被引文献构成了大数据研究的主

表 3 高被引文献列表

	作者	题名	刊名	发表年	被引次数
1	孟小峰、慈祥	大数据管理：概念、技术与挑战	计算机研究与发展	2013	2252
2	陈全、邓倩妮	云计算及其关键技术	计算机应用	2009	1540
3	李国杰、程学旗	大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考	中国科学院院刊	2012	1382
4	王珊、王会举、覃雄派、周烜	架构大数据：挑战、现状与展望	计算机学报	2011	824
5	王元卓、靳小龙、程学旗	网络大数据：现状与展望	计算机学报	2013	730
6	覃雄派、王会举、杜小勇、王珊	大数据分析——RDBMS与MapReduce的竞争与共生	软件学报	2012	602
7	冯登国、张敏、李昊	大数据安全与隐私保护	计算机学报	2014	537
8	马华东、陶丹	多媒体传感器网络及其研究进展	软件学报	2006	581
9	冯芷艳、郭迅华、曾大军、陈煜波、陈国青	大数据背景下商务管理研究若干前沿课题	管理科学学报	2013	467
10	程学旗、靳小龙、王元卓、郭嘉丰、张铁赢	大数据系统和分析技术综述	软件学报	2014	464
11	邬贺铨	大数据时代的机遇与挑战	求是	2013	456
12	肖乐斌、钟耳顺、刘纪远、宋关福	三维GIS的基本问题探讨	中国图像图形学报	2001	455

要知识基础。

高被引文献通常具有更高的影响力。在WOS平台获得的12篇高被引文献中，有两篇文献的作者机构是合肥工大、华中科大以及清华大学；一篇文献的作者机构是澳门大学。比较检索结果的机构分布和高产作者，在经管大数据领域，我国的文献产出数量多且增长快，但在影响力方面还需要加强。中国科学技术信息研究所发布的中国科技论文统计结果也表明，我国的高影响力论文多分布在农学、

化学、计算机、工程等学科，经济管理学科乃至整个社会科学研究的国际影响力还有待提高。

（二）基于共被引的大数据研究知识基础辨识

通过VOSviewer对数据集C进行共被引分析以发现社会科学领域大数据研究的知识基础。本文选择VOSviewer的Fractional counting计数方法，对2345篇文献的参考文献进行统计，得到24268条引用关系，对被引次数不少于50次的175篇文献进行共被引聚类分析（其中1篇属于孤立点没有被显示）。聚

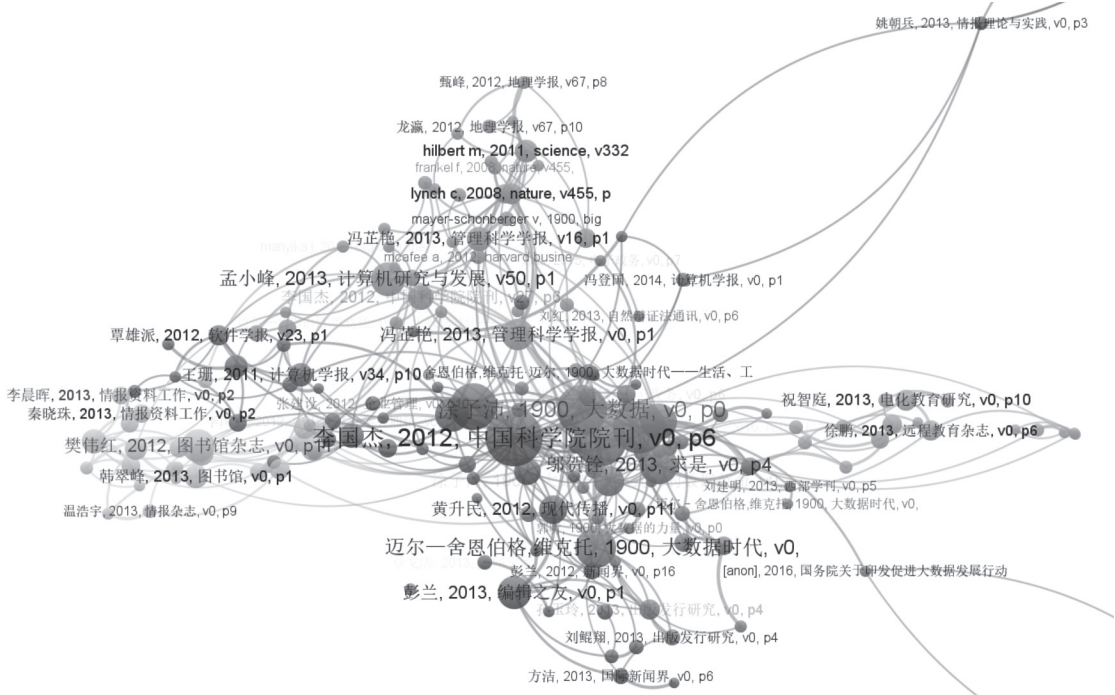


图 1 数据集C的共被引聚类图谱

（右上方缺失节点是“蒋洁，2014情报科学”，右下方缺失节点是“张兴旺，2014中国图书馆学报”）

类参数选择如下: Normalization Method=LinLog/Modularity且最小聚类规模不小于15个节点,得到6个聚类。即数据集C中通过共被引关系识别出中文社科领域大数据研究的6类知识基础,见图1。

聚类1#: 红色区域包括42个节点,该区域高被引文献多,大部分节点关系紧密。突出的作者包括维克托·迈尔-舍恩伯格、李国杰、涂子沛、郭贺铨、郭晓科等。该区域的学科界限模糊,属于多学科研究,时间跨度长。

聚类2#: 绿色区域包括35个节点,以管理学领域为主,主要与计算机领域交叉。突出的作者包括孟小峰、冯芷艳、李国杰、俞立平等,这个区域出现较多WOS平台高被引英文文献。主要分布在2012~2014年,外文参考文献年代略早于中文参考文献。

聚类3#: 蓝色区域包括34个节点,以传播学领域为主,与计算机、图书情报领域交叉。主要作者有孟小峰、李国杰、彭兰、黄升民等。主要分布2012~2014年,该类别界限较模糊。

聚类4#: 黄色区域包括26个节点,以图书情报领域为主,突出的作者包括樊伟红、张兴旺、黄晓斌等,主要分布在2012~2013年。

聚类5#: 紫色区域包括20个节点,以计算机类期刊为主,夹杂图书情报学期刊,主要作者有李国杰、王珊、覃雄派、李广建等,主要分布在2012~2014年,该类别界限较模糊。

聚类6#: 青色区域包括17篇文献,集中在教育研究领域,年代在2012~2016年。

通过共被引聚类分析,中文社科领域大数据研究的知识基础由计算机科学、图书情报学、管理学、新闻传播学和教育学构成;英文文献主要与管

理类和计算机类文献共被引;英文文献大多属于WOS领域高被引文献,如文献“Chen HC, 2012”“Mcafee A, 2012”“Lavalley S, 2011”“Manyika J, 2011”“Davenport TH, 2014”“Mayer-Schonberger, 2013”“Boyd D, 2012”“Lazer D, 2014”“Dean J, 2008”等,国内外经管领域大数据研究的知识基础有共性;被引用文献主要分布在2012~2014年,与WOS平台数据一致。

五、基于关键词共现分析的研究热点及其演变过程

张勤等指出共词分析法是一种内容分析方法,主要通过对能够表达某一学科领域研究主题或研究方向的专业术语共同出现在一篇文献中的现象进行分析,判断学科领域中主题间的关系,从而展现该学科的研究结构^[16]。本文主要分析经济与管理领域大数据文献研究热点及其演变过程。为了更准确地进行分析,对数据集4进行进一步精炼,剔除会议通知、征稿、总目录等非学术论文,剔除2011年及以前文献,得到2801篇发表在2012~2018年来自CNKI较高质量期刊中经管领域的文献。具体数量分布为(15, 151, 389, 621, 781, 837, 7),合计占数据集4的90.94%,命名为数据集4'。

(一) TimeLine图谱

经济与管理学科发展有其自律性,也有他律性,既体现学科自身发展,也顺应时代环境变化,CiteSpace的关键词图谱清楚地显示出该特征。基于关键词共现关系的TimeLine图谱将所有关键词聚为11类,每类由若干关键词组成,节点最大的词出现

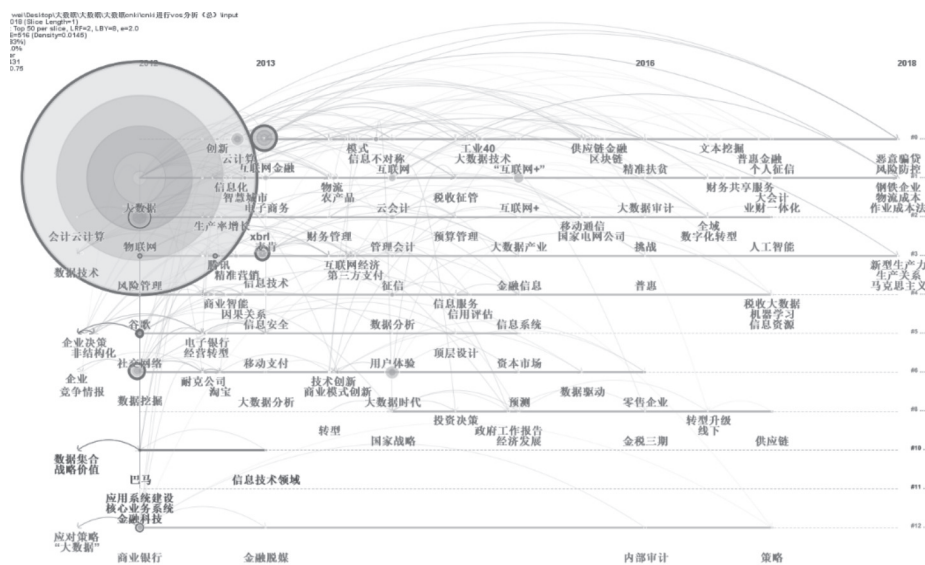


图2 基于关键词共现关系的TimeLine图谱

频次最高,紫色外圈的节点具有高中心性,与其他词的结合能力强,词被标识在首次出现的年代,连线表示关键词之间的共现关系,如图2所示。

通过图谱可以直观地看出2012年以来大数据如何被应用到各个领域。“互联网+”在文献中的出现与我国政府2015年的系列政策密切相关。“精准扶贫”是习总书记自2013年以来多次做出的重要指示,2016年文献中开始出现基于大数据的精准扶贫研究,既顺应国家政策,也是大数据不断渗透到各个领域的必然结果;2017年提出基于大数据的个人征信;2018年最新的主题是大数据与风险防控、马克思主义等的结合。

(二) Density Visualization图谱

VOSviewer绘制的密度图(density visualization)可以更加直观地观察到研究热点的分布和演变。2012年和2018年数据量小,无明显聚类。图3系列(为了更好地显示聚类效果,没有显示“大数据”节点)分别显示了2013~2017年每年的研究热点以及多年叠加后的研究热点。

2013年开始大量出现“大数据”“社交网络”和“物联网”成为重点主题,“云计算”“互联网金融”“数据挖掘”“麦肯锡”等属于较重要的主题。2014年“互联网金融”成为重点主题,“大数据时代”“物联网”“云计算”“国家战略”“社交媒体”等是较重要的主题。2015年“物联网”和

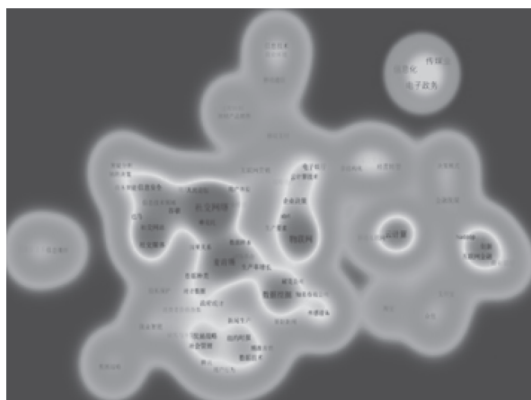


图3-1密度图(2013)

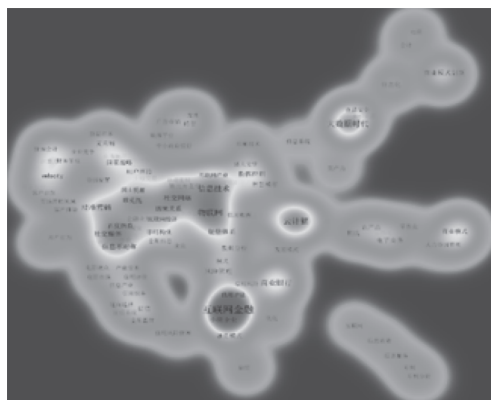


图3-2密度图(2014)

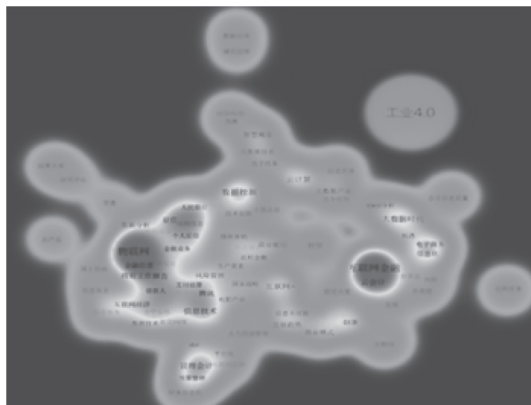


图3-3密度图(2015)

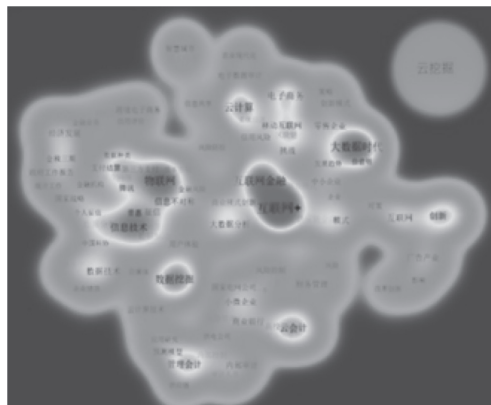


图3-4密度图(2016)

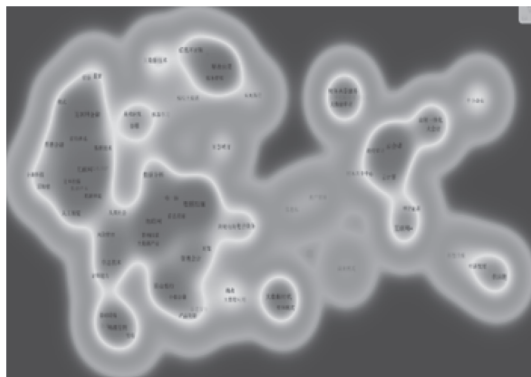


图3-5密度图(2017)

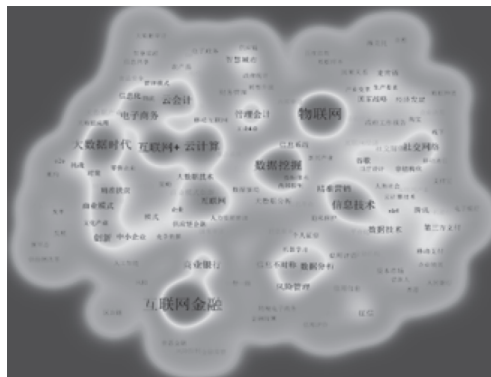


图3-6密度图(2012-2018)

图3 密度图(2013)

“互联网金融”成为重点主题,“数据挖掘”“征信”和“云会计”等是较重要的主题。2016年“互联网+”成为最重要的主题,“物联网”“互联网金融”“大数据时代”“信息技术”“数据挖掘”等依然是重点主题,更多主题涌现出来。2017年由于数据不全,话题密度较2016有所下降,“精准扶贫”“大数据审计”“精准营销”等主题成为新重点,“互联网金融”“数据挖掘”“物联网”“云会计”等主题持续重要。2012~2018年全部文献叠加后的关键词共现图谱显示出研究热点主要包括:互联网金融、物联网、数据挖掘、互联网+、云计算、大数据时代、电子商务、社交网络、信息技术、云会计等。

结合图2中高频词出现年代,通常热点词出现后第二年会文献中反映出来,如出现在2013年的“互联网金融”在2014年以后持续成为热点主题,出现在2015年的“互联网+”成为2016年最大主题。“物联网”“云计算”“数据挖掘”作为与“大数据”密切相关的主题,成为文献持续焦点。

国内外经管领域大数据的热点主题与演变过程有显著差异。英文文献有明显的技术导向,云计算、大数据分析、Hadoop、数据挖掘等是持续热点话题,大数据与经管领域结合路径明显,如知识管理、用户分析、供应链管理、商务智能、创新、电子商务等。中文文献使用的词汇更丰富,新概念能更快速地显现出来。“物联网”“互联网金融”不仅出现较早,而且是持续热点话题;“互联网+”“精准扶贫”等词汇与国家政策密切相关;技术导向不明显,技术类词汇被淹没在众多的新词汇中。

六、结论

本文采用文献计量方法对国内大数据领域的相关文献进行分析。根据文献数量变化,可以将国内大数据研究分为三个阶段:初始期、爆发期和稳定增长期。大数据领域文献主要来自计算机领域,贵州省大数据产业发展态势通过文献被凸显出来。国家级研究基金是文献产出的最重要支持力量,高校是文献产出的主力军。经管领域大数据研究尚无明显的高影响力团队,经管领域大数据研究热点和演变过程既体现客观规律,也体现出国家政策的影响。我国大数据研究与国际大数据研究有共性,也有特性,互联网金融、云会计等术语在WOS数据中没有显著出现。我国文献存量巨大、增长快速,但经管和社会学科研究的国际影响力还较低。

本文存在以下不足:首先,2017年数据不完整导致分析不完整;其次,经管领域文献内容对比研究中,为平衡文献数量采用的检索策略不完全相同,中文检索更强调查准率而英文检索强调查全率,WOS平台学科划分与CNKI不同,本文通过人工筛选将12个WOS学科对应于CNKI的经济与管理学科,可能存在一定偏差。

参考文献

- [1] 赵强,单炜.大数据政府创新:基于数据流的公共价值创造[J].中国科技论坛,2014(12):23-27.
- [2] Big Data, Big Impact: New Possibilities for International Development[DB/OL]. [2018-1-20]. <https://www.weforum.org>.
- [3] HUSER V, CIMINO J J. Impending Challenges for the Use of Big Data[J]. International Journal of Radiation Oncology Biology Physics, 2015, 95(3): 890-894.
- [4] ASHTON K. That “Internet of Things” Thing-RFID Journal[DB/OL]. [2018-01-20]. <https://www.mendeley.com/research-papers/internet-things-thing-rfid-journalpdf/>.
- [5] 张姣姣,刘云,程旖婕.基于文献计量学定律的大数据应用领域发展规律研究[J].知识管理论坛,2016(5):384-392.
- [6] 王雪.基于文献计量的国内外大数据研究对比分析[J].高校图书馆工作,2017(4):49-54.
- [7] 赵丹,王晰巍,李嘉兴,等.国内外大数据工具学术论文比较研究——基于文献计量方法[J].情报科学,2016,34(6):133-137.
- [8] 陈军,谢卫红,陈扬森,等.国内外大数据可视化学术论文比较研究——基于文献计量与SNA方法[J].科技管理研究,2017,37(8):44-53.
- [9] 李贺,袁翠敏,李亚峰.基于文献计量的大数据研究综述[J].情报科学,2014(6):148-155.
- [10] 王岑岚,尤建新.大数据文献评述:基于软件Citespace的可视化研究[J].科技管理研究,2017,37(21):180-189.
- [11] 王曰芬,宋爽,熊铭辉.基于共现分析的文本知识挖掘方法研究[J].图书情报工作,2007,51(4):66-70.
- [12] SYNNESTVEDT M B, CHEN C, HOLMES J H. CiteSpace II: visualization and knowledge discovery in bibliographic databases [C]//Annual Symposium proceedings, 2005:724-731.
- [13] VAN ECK N J, WALTMAN L. VOSviewer: A Computer Program for Bibliometric Mapping[J]. Social Science Electronic Publishing, 2009, 84(2): 523-538.
- [14] 宋秀芳,迟培娟. VOSviewer与Citespace应用比较研究[J].情报科学,2016,34(7):108-112+146.
- [15] SMALL H. Cocitation in scientific literature-new measure onf relationship between 2 documents[J]. Journal of the American Society for Information Science, 1973,24(4):265-269.
- [16] 张勤,马费成.国外知识管理研究范式——以共词分析为方法[J].管理科学学报,2007,10(6):65-75.

编辑 何婧