·专 题·

基于 LDA 模型的国内大数据研究热点主题分析

邱均平[1,2] 沈 超[1,2]
（1. 杭州电子科技大学中国科教评价研究院，浙江 杭州 310018；
2. 杭州电子科技大学管理学院，浙江 杭州 310018）

摘 要： [目的/意义] 揭示国内在大数据领域研究中的热点主题。[方法/过程] 收集 CNKI 收录的 2008—2020 年关于大数据的论文，通过 LDA 主题模型抽取研究主题，并识别热点主题。选取具有代表性的主题，对其再次进行 LDA 主题聚类，并运用 LDAvis 对主题进行可视化。[结果/结论] LDA 模型能够较为准确地提取大数据领域文献的研究主题，这有利于研究人员了解该领域的发展状态，把握未来的研究方向，探寻新兴主题。

关键词： LDA；大数据；LDAvis；热点主题
DOI: 10.3969/j.issn.1008-0821.2021.09.003
〔中图分类号〕G250.2 〔文献标识码〕A 〔文章编号〕1008-0821（2021）09-0022-10

Analysis of Hot Topics in Domestic Big Data Research Based on LDA Model

Qiu Junping[1,2] Shen Chao[1,2]
（1. Chinese Academy of Science and Education Evaluation，Hangzhou Dianzi University，Hangzhou 310018，China；
2. School of Management，Hangzhou Dianzi University，Hangzhou 310018，China）

Abstract: [Purpose/Significance] The paper reveales the hot topics in domestic research in the field of big data. [Method/Process] The paper collected papers on big data from 2008 to 2020 included in CNKI，extracted research topics through the LDA topic model，and identified hot topics. The paper selected representative topics，performed LDA topic clustering on them again，and used LDAvis to visualize the topics. [Result/Conclusion] The LDA model can more accurately extract the research topics of the literature in the big data field，which helps researchers understand the development status of the field，grasp the future research direction，and explore emerging topics.

Key words: LDA；big data；LDAvis；hot topics

“大数据”一词在《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》（以下简称《建议》）中一共出现了 3 次[1]。《建议》指出，要推动大数据同各产业深度融合，加快大数据中心建设，以及加强宏观经济治理数据库等建设，提升大数据等现代技术手段辅助治理能力，可见国家对于大数据的重视。大数据一词于 2008 年 9 月首次在《Nature》杂志被提出[2]，此后大数据迅速引起了学术界的广泛关注，不同学科的学者纷纷开始从自己学科的角度去探索大数据的含义概念、大数据的分析方法和大数据在本学科的应用。在以文献、信息和数据为研究对象的图书情报学科内对大数据进行的研究也迅速增长。科技文献作为学术成果的重要载体，是科研人员大量智慧

收稿日期：2021-04-08
基金项目：国家社会科学基金重大项目“基于大数据的科教评价信息云平台构建和智能服务研究”（项目编号：19ZDA348）。
作者简介：邱均平（1947-），男，院长，教授，博士生导师，研究方向：信息计量与科学评价。沈超（1997-），男，硕士研究生，研究方向：信息计量与科学评价。

汗水的结晶，是进行知识传播和学术交流的重要途径。因此，对科技文献进行计量分析，能够了解大数据的发展现状，预测其发展趋势。

对大数据领域相关文献进行计量分析的研究，国内一些学者多是通过文献计量的相关方法进行的。赵悦阳等使用 Web of Science 的分析功能和可视化软件 UCINET、gCLUTO，对 Web of Science 数据库中的相关论文进行发文分析、词频分析、共现分析和双聚类分析。得出高产国家(地区)、高产机构、高产作者和研究热点的结论[3]。童子颐以中国学术期刊网络出版总库中我国大数据研究相关论文的高频关键词作为统计数据，对其进行共词分析，并在此基础上对其进行聚类分析和多维尺度分析，得出我国大数据研究的热点主要有：大数据技术、新闻传媒中的大数据、大数据应用等[4]。代芯瑜等对我国 2003—2012 年发表的有关大数据研究论文进行统计，利用文献计量法和可视化分析方法，从宏观上把握近年来我国的大数据研究现状及研究重点，并得到相关研究结论[5]。夏军辉等对图书情报领域关于大数据的研究现状、热点、主要研究方法和发展趋势进行了分析和整理[6]。王春华等使用 CNKI 数据库，将学科定为图书情报与数字图书馆，将主题定为大数据，将得到的文献数据运用词频统计方法和共词分析方法，借助 SPSS 软件和 UCINET 软件，对这些文献数据进行聚类分析、战略坐标图分析和核心—边缘结构分析，得出了国内图书情报领域大数据研究的 8 个热点研究主题，并对热点研究主题进行了详细的解析[7]。虞秋雨等以近 5 年图书情报领域在中国知网数据库中核心期刊收录的有关“大数据”主题的文献为研究对象，建立了一种以 g 指数为主要基础的划分高频词的方法，利用 Excel 软件进行数据统计并构造共词矩阵。同时借助 SPSS、Pajek 软件对矩阵进行可视化分析、K-core 分析以及聚类分析，研究文献中各关键词间的关系，探讨了近 5 年我国图书情报领域关于大数据主题的研究热点[8]。黄鹂等基于 Web of Science 数据库，从发文年代、国家(地区)和机构、核心作者及主要期刊分布几方面分析了医学信息学科大数据研究的现状和进展，借助软件对关键词进行聚类分析，发现研究热点主要集中在临床决策支持系统、临床研究数据管理、电子健康档案、转化生物信息学和遗传流行病学等方面。范婷等运用双聚类法对医学大数据的研究热点进行了分析，得出发文量分布、期刊分布、高频主题词及共词聚类结果[9]。

上述研究中，对研究热点进行分析多用文献计量的方法，借助 LDA 主题模型分析大数据领域文献的热点研究主题的文章极少。为适应当前文献数量剧增的现状，本文试图借助 LDA 模型，对 CNKI 数据库中标题带有“大数据”的中文期刊论文进行文本建模，通过困惑度确定模型的最优主题数，并对文档—主题矩阵和主题—词矩阵进行分析，以期了解大数据领域研究现状和研究热点，为研究人员提供参考。

1 相关技术

1.1 LDA

LDA (Latent Dirichlet Allocation) 模型，是由 Blei D M 等于 2003 年提出的一种基于概率模型的主题模型算法，LDA 是一种非监督机器学习的文本挖掘技术，可以用来识别大规模文档集或语料库中的潜在隐藏的主题信息[10]。

LDA 模型算法中一篇文章的每个词都是通过以一定概率选择了某个主题，并从该主题中以一定概率选择某个词语这样的过程得到的。该方法假设文本中的每个词是从一个潜在隐藏的主题中抽取出来的。对于语料库中的每篇文档，LDA 定义了如下生成过程：首先，对每一篇文档，从主题分布中抽取一个主题；其次，从上述被抽到的主题中所对应的单词分布中抽取一个单词；最后，重复上述过程直至遍历文档中的每一个单词。

本文主要采用 Gibbs 采样算法求解得到全局的主题 Z 的分布和词语 W 的分布。作为无监督机器学习，需要事先确定 3 个超参数 α、β、k (最优主题数)，α、β 选取一般为默认值，最优主题数 k 通过困惑度计算确定。困惑度的计算公式如式 (1) 所示。其中，D 为测试集；M 为文本数量；d_i 为文档 d 中的单词序列；N_i 为文档 d 的单词数目。

$$Perlexity(D) = \exp\left\{-\frac{\sum_{i=1}^{M}\ln p(d_i)}{\sum_{i=1}^{M}N_i}\right\} \quad (1)$$

困惑度能够衡量 LDA 主题模型预测样本的精确程度，因此，理论上困惑度越小说明模型预测精准度越高，困惑度最低或是拐点处对应的 k 就为最佳主题数。

1.2 LDAvis

LDAvis 是一种主题可视化方法，于 2014 年由 Sievert C 等[11]提出。LDAvis 以特征词和主题的关联程度选择表示主题的特征词，并且 LDAvis 可视化图可以帮助人们从整体的视角观察各个主题之间的关系[12]。简单来说，就是 LDAvis 探究了主题—主题、主题—词语之间的关联。主题—主题用多维标度的方式，将两者投影在低维空间，从而进行比较分析；主题与词语之间的关联综合了词频和词语的独特性两种属性。其中 λ 就是调节两种属性哪个占比更大的重要参数。λ 的取值在 0~1 之间，λ 的最优取值需要根据具体问题进行具体分析。

2 研究设计

2.1 数据来源

实验领域为大数据。中文数据采集时间段为 2008—2020 年，检索日期为 2021 年 3 月 10 日。

中国知网(CNKI)为文献的数据来源。限定文献为 SCI 期刊、EI 期刊、核心期刊、CSSCI/CSCD，文献类型设定为期刊、中文。设定使用专业检索，设定检索式为“TI = ‘大数据’”，其中“TI”表示标题。下载全记录文献信息并以 Excel 格式批量导出，筛去重复以及不完整的文献记录，最后得到总计80 797篇文献。

按年份将样本中的文献进行分类，并统计每年的文献数量，绘制折线图，其变化情况如图 1 所示。可以看出，2008—2011 年发文量较少，从 2012 年开始发文量呈稳定增长的趋势。总的来说，可以将时间分为两段，2008—2011 年为初步探索期，2012—2020 年为快速增长期。

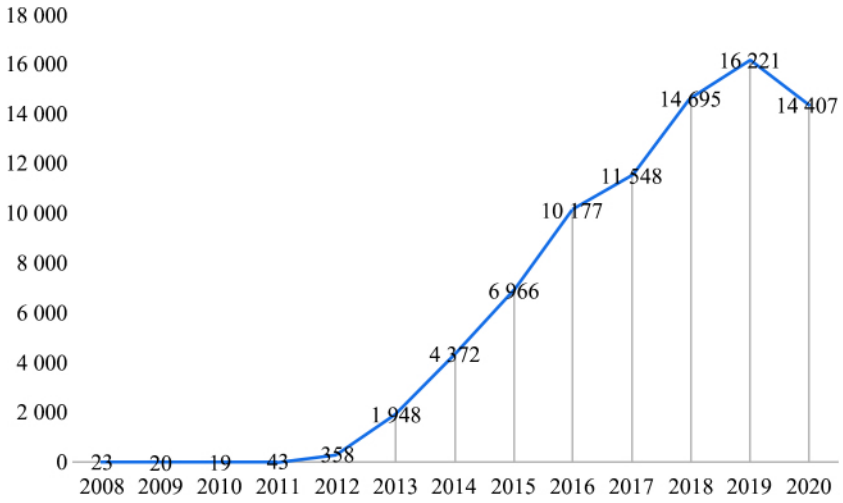

图 1 2008—2020 年样本文献年度总量变化趋势

2.2 文本预处理和 LDA 参数设置

将下载好的文献按年份进行分类，形成各年份的文献信息文件，并将文件格式转换为 CSV 以便后续处理，从文献信息中提取标题、关键词、摘要信息形成 LDA 模型的语料来源。对语料来源文件用 Python 中的 Jieba 分词组件进行分词操作，从而得到文档—词矩阵。然后，借助 Python 软件中 Sklearn 包构建 LDA 模型。在构建模型前需要确定模型的最优主题数，本文结合模型困惑度来确定模型的最优主题数。首先，将主题数区间设为 [0，40]，步长设为 2，α、β 设为默认值；其次，根据各个主题数的困惑度，选取困惑度最低的主题数作为最优主题数。运行程序得到的结果主要有两个：文献—主题分布和主题—词分布。

2.3 实验结果与分析

2.3.1 主题内容分析

结合文献—主题分布和主题—词分布，绘制出各主题及各主题下所对应的文献数量按年份排列的表格，如表 1、表 2 所示。

表 1 2008—2016 年各年度主题数、主题词及对应文献数量

年 份	主题数	主题(对应文献数量)
2008—2012	5	海量文件存储策略（32）、优化数据库系统性能（78）、各国大数据技术的发展（80）、大数据分析技术（55）、企业信息计算分析（218）
2013	22	计算平台（47）、信息管理平台（19）、图书馆服务（0）、传媒产业（49）、商业金融（28）、数据分析（309）、城市交通建设（2）、研究会议（126）、教育业（17）、智慧城市（33）、数据仓库（7）、电商企业（84）、通信运营（53）、智能化（14）、市场营销（325）、数字基础设施（8）、网络安全（1）、信息管理（434）、信息安全（38）、海量数据管理（54）、媒体（171）、互联网（130）
2014	25	通信运营商（159）、数字档案（52）、社会国家治理（52）、企业管理（265）、组织数字化（44）、智能交通（82）、主题论坛（131）、资源管理（14）、互联网营销（418）、产业转型发展（73）、传统媒体（178）、图书馆服务（258）、实时信息处理（211）、经济发展战略（514）、数据挖掘（341）、互联网金融（143）、风险管理（9）、智慧医疗电网（24）、信息安全（474）、智慧城市建设（94）、客户信息分析（201）、教育（233）、信息化建设（134）、模型算法研究（225）、电子政务（43）
2015	24	医疗信息数字化（80）、工业信息化（118）、网络安全（71）、舆情分析（239）、交通建设（1）、档案管理（272）、算法模型（482）、信息安全（248）、企业管理（503）、传媒（276）、市场营销（414）、教育（223）、教育改革（141）、图书馆服务（371）、技术发展（743）、互联网金融（323）、政府治理（374）、产业发展战略（667）、电力系统（161）、数据分析（95）、文旅数字化（33）、大数据时代面临的机遇挑战（613）、用户定位（48）、智慧平台建设（470）
2016	21	大数据时代面临的机遇挑战（1859）、传媒（67）、信息安全（329）、数据分析（852）、智慧交通（116）、智能电网（263）、组织管理（179）、档案管理（112）、电子商务（402）、市场营销（912）、知识情报挖掘（171）、社会治理（366）、企业创新发展（584）、产业发展（870）、教育（629）、智慧城市（325）、医疗信息化（257）、信息统计（454）、财务管理（242）、图书馆服务（591）、算法模型（599）

表 2 2017—2020 年各年度主题数、主题词及对应文献数量

年份	主题数	主题（对应文献数量）
2017	32	国家战略规划（398）、信息安全（472）、资源数字化（77）、产业转型（399）、档案管理（177）、算法模型（555）、智慧城市（191）、数据信息统计（86）、战略规划（66）、网络医疗（175）、系统平台设计（1002）、教育（357）、社会治理（633）、研究方法（584）、电子政务（22）、数据分析（415）、信息化建设（403）、市场营销（0）、大数据对社会的影响（1634）、大数据时代的机遇与挑战（370）、发展模式路径选择（34）、构建评价体系（245）、传媒（260）、社区管理（13）、图书馆服务（337）、人工智能（658）、精准扶贫（90）、教学模式（407）、企业管理（975）、互联网金融（256）、警务侦查（39）、高校创新管理（214）
2018	27	医疗信息化（170）、企业管理（1575）、国家战略规划（664）、人才培养模式（543）、犯罪侦查（173）、大数据时代的机遇与挑战（1787）、城市交通（79）、图书馆服务（503）、人工智能思维（346）、教育扶贫（365）、信息技术（1289）、数据分析（237）、数字经济（196）、信息安全（794）、智能电网监测系统（491）、互联网金融（395）、智慧城市（653）、电力资源（176）、企业管理（416）、网络质量评估（329）、统计分析（31）、研究分析方法（789）、算法模型（373）、政府信息化建设（290）、公共管理（791）、教育（833）、数据挖掘（409）
2019	30	数据挖掘技术（759）、犯罪侦查模式创新（266）、农业产业（226）、管理工作（679）、智能平台系统（1213）、方法模型评价（756）、图书馆服务（418）、信息化建设立法（414）、传媒（144）、企业管理（350）、信息资源管理（112）、教学改革（1059）、档案管理（274）、精准扶贫/市场营销（304）、数据处理方法（705）、政府治理（638）、信息安全（936）、企业管理（1078）、教育（706）、大数据对社会带来的影响（2108）、智慧城市（387）、审计（471）、医疗信息化（347）、会计信息化（404）、金融数据安全（265）、数据分析（211）、网络安全（221）、实践（33）、互联网经济（283）、教育信息化（454）

表 2（续）

年份	主题数	主题（对应文献数量）
2020	29	评价体系构建（254）、算法模型设计（1336）、产业转型（464）、地区规划（102）、企业管理（1808）、质量标准化（28）、疫情防控/政府治理（744）、大数据时代对社会生活的影响（2742）、市场营销（383）、审计（283）、信息安全（143）、数据安全（365）、教育（77）、精准扶贫（171）、网络安全（308）、人工智能（275）、医疗信息化（198）、智能电力系统（435）、数据分析（380）、智慧城市建设（750）、在线学习（146）、思政教育（594）、管理（139）、互联网金融（228）、教育（940）、传媒（81）、框架（29）、信息化建设/档案管理（641）、图书馆服务（363）

通过对表 1、表 2 中的内容进行分析可以看出，大数据技术在各个时间段的研究中应用型研究较多，且在许多领域中的研究具有连续性和一贯性。下面挑选出具有代表性的 9 个主题，并以主题标签为关键词在原始数据中的标题列筛选包含该主题标签的论文，对这些论文进行 LDA 主题聚类和 pyLDAvis 可视化，并进行分析。

1）图书馆服务。这一主题在 2013—2020 年都有体现，并且其所对应的文献数量较多，说明国内图书馆领域的大数据研究较多。以“图书馆”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是图书馆和大数据主题，所以在停用词表中加入“大数据’‘数据”和“图书馆”3 个词，避免这 3 个词多次出现，影响主题聚类效果。从结果可以看出，有关图书馆这个主题的研究有以下几个分支：对高校图书馆的研究；对图书馆员的研究，如对图书馆员的信息素养进行研究；对图书馆内的文献进行研究；对图书馆用户读者行为进行研究，如对读者的隐私保护的研究等。图 2 为“图书馆服务”主题下各个主题的 LDAvis 可视图。

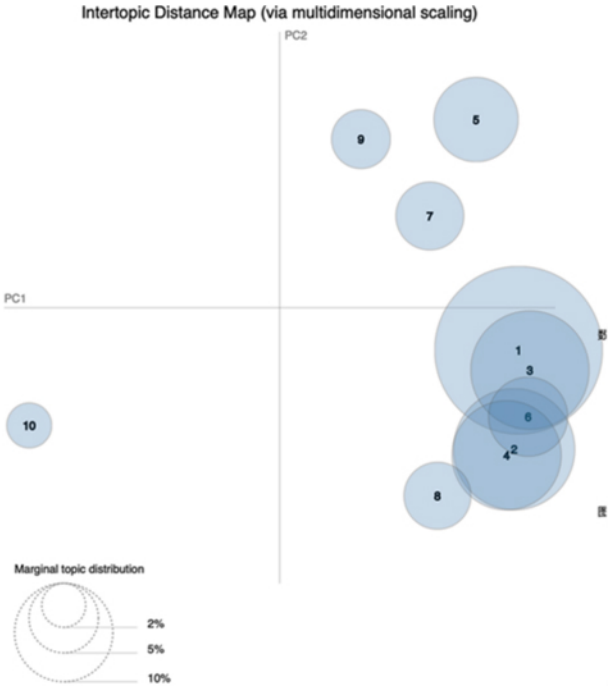

图 2 LDAvis 主题可视化图——‘图书馆”

2）智慧城市建设、城市智慧交通等有关地区规划建设的主题也在许多年份中有所体现，说明大数据技术在城市建设、交通建设等方面应用较多。以“智慧城市”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是智慧城市和大数据主题，所以在停用词表中加入“大数据’‘数据’‘智慧’‘城市”4 个词，避免这 4 个词多次出现，影响主题聚类效果。从结果可以看出，有关智慧城市主题的研究有以下几个分支：有关城乡规划、城市规划的研究；有关城乡治理的研究；有关数据及数据中心平台的研究；有关交通的研究等。图 3 为“智慧城市”主题下各个主题的 LDAvis 可视图。

3）大数据在教育领域的应用研究也较多。主要集中在教学模式改革等方面。以“教育”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是教育和大数据主题，所以在停用词表中加入“大数据’‘数据’‘教育”3 个词，避免这 3 个词多次出现，影响主题聚类效果。从结果可以看出，有关教育主题的研究有以下几个分支：对大学生创新创业的研究；对网络教育资源的研究；对远程教育的研究；对教师教学评价的研究；对高校思政教育的研究等。图 4 为“教育”主题下各个主题的 LDAvis 可视图。

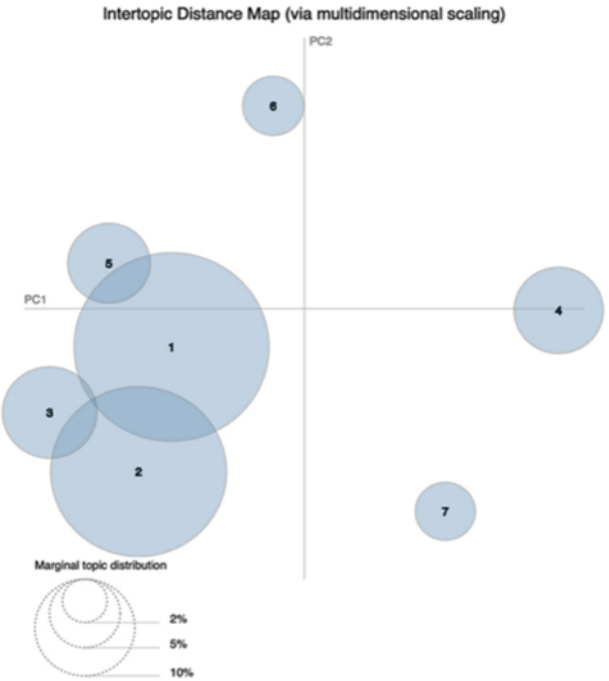

图 3 LDAvis 主题可视化图——“智慧城市”

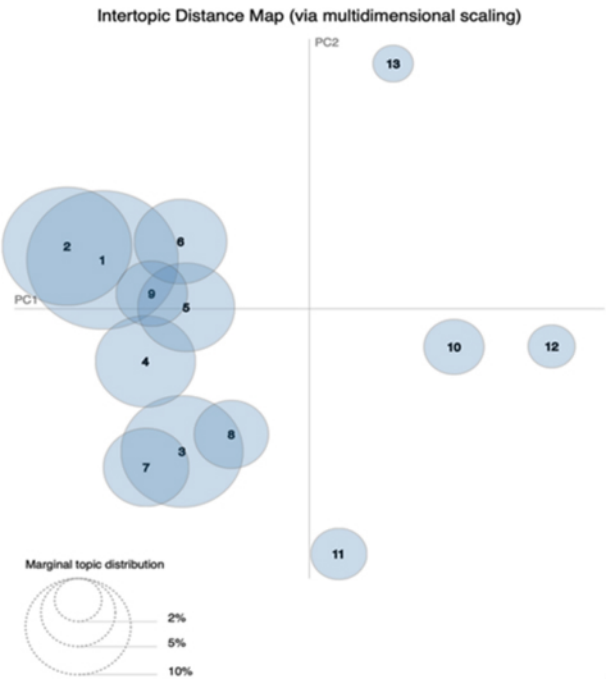

图 4 LDAvis 主题可视化图——“教育”

4) 大数据与商务、金融相结合，对电子商务、互联网金融、网络营销等新兴产业起到了极大的促进作用，例如利用用户大数据，电子商务平台能够精准地投放广告和商品，促进商品的销量。以“商务”“金融”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是商务、金融和大数据主题，所以在停用词表中加入“大数据”“数据”“商务”“金融”4 个词，避免这 4 个词多次出现，影响主题聚类效果。从结果可以看出，金融商务的研究有以下几个分支：对小微企业的研究；对金融风险的研究；对物流供应链的研究；对信息智能分析的研究；对个性化服务的研究；对电子商务的研究等。图 5 为“金融商务”主题下各个主题的 LDAvis 可视图。

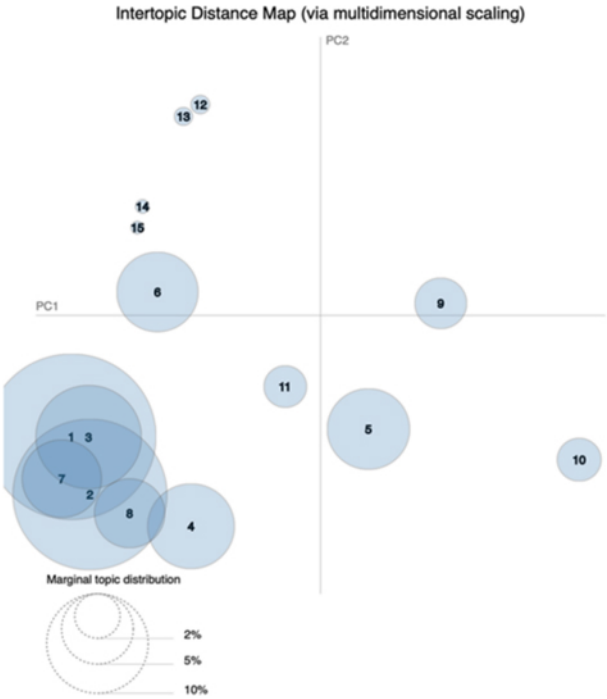

图 5 LDAvis 主题可视化图——“金融商务”

5) 大数据在政务工作和国家治理方面也发挥着极大的作用。如浙江省推出的“最多跑一次”便民服务，便是依靠着大数据技术。极大地方便了办事群众，同时也精简了政府机构人员。以“政务”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是政务和大数据主题，所以在停用词表中加入“大数据”“数据”“政务”3 个词，避免这 3 个词多次出现，影响主题聚类效果。从结果可以看出，政务主题的研究有以下几个分支：信息公开、资源共享；对电子政务的研究等。图 6 为“政务”主题下各个主题的 LDAvis 可视图。

6) 大数据技术与传播媒体的结合，使得信息的传播更为快速和准确。以“传媒”为关键词在

图 6 LDAvis 主题可视化图——“政务”

各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是传媒和大数据主题，所以在停用词表中加入“大数据”“数据”“传媒”3 个词，避免这 3 个词多次出现，影响主题聚类效果。从结果可以看出，传媒主题的研究有以下几个分支：传播媒介创新和传统媒体面临的挑战等。图 7 为“传媒”主题下各个主题的 LDAvis 可视图。

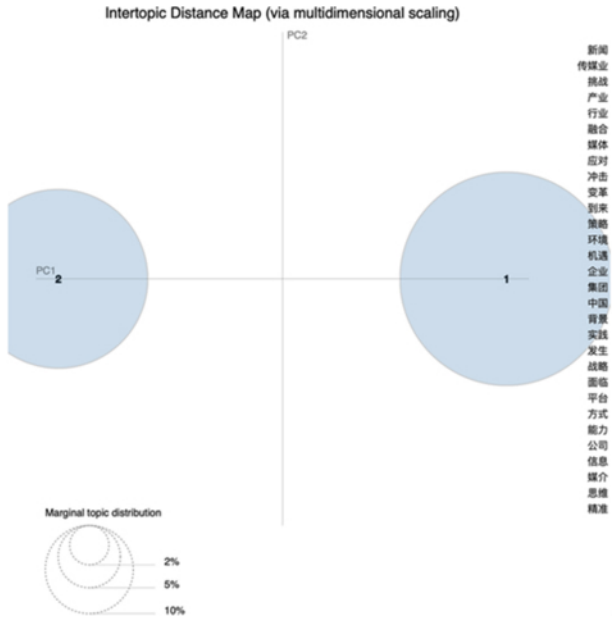

图 7 LDAvis 主题可视化图——“传媒”

7) 数字医疗、医疗信息数字化则是大数据技术在医疗领域中的应用，这样的应用能够在一定程度上解决“看病难、看病贵”的问题。以“医疗”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是医疗和大数据主题，所以在停用词表中加入“大数据”“数据”“医疗”3 个词，避免这 3 个词多次出现，影响主题聚类效果。从结果可以看出，医疗主题的研究有以下几个分支：医疗信息化；对患者隐私保护的研究；对智能监测的研究等。图 8 为“医疗”主题下各个主题的 LDAvis 可视图。

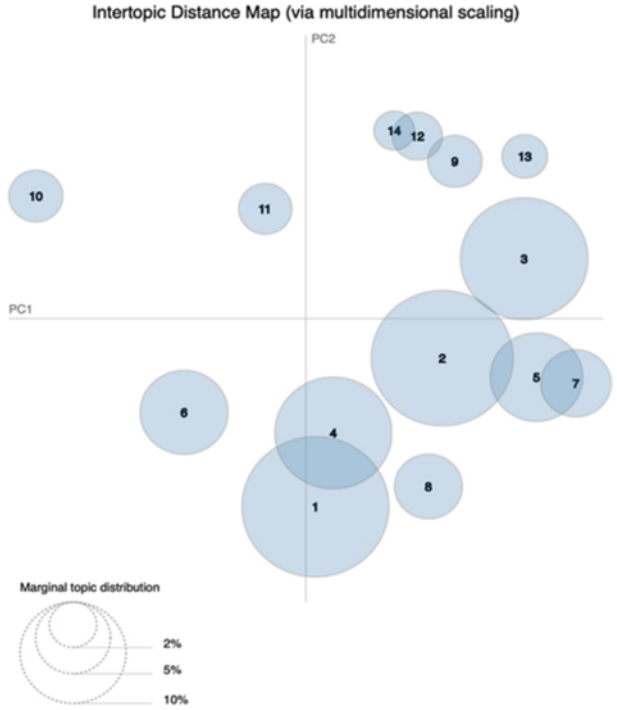

图 8 LDAvis 主题可视化图——“医疗”

8) 网络安全、信息安全两大主题在绝大多数年份中都有体现，说明人们十分重视网络安全和信息安全。大数据时代，人们的各种信息都以数据的形式存在于网络上，如何保障这些数据信息的安全就成了全民所关注的热点。以“网络安全”和“信息安全”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是网络安全和信息安全以及大数据主题，所以在停用词表中加入“大数据”“数据”“网络安全”“信息安全”4 个词，避免这 4 个词多次出现，影响主题聚类效果。从结果可以看出，网络

信息安全的研究有以下几个分支：个人隐私安全的研究；信息管理的研究；防御系统的研究；网络环境的研究等。图 9 为“网络、信息安全”主题下各个主题的 LDAvis 可视图。

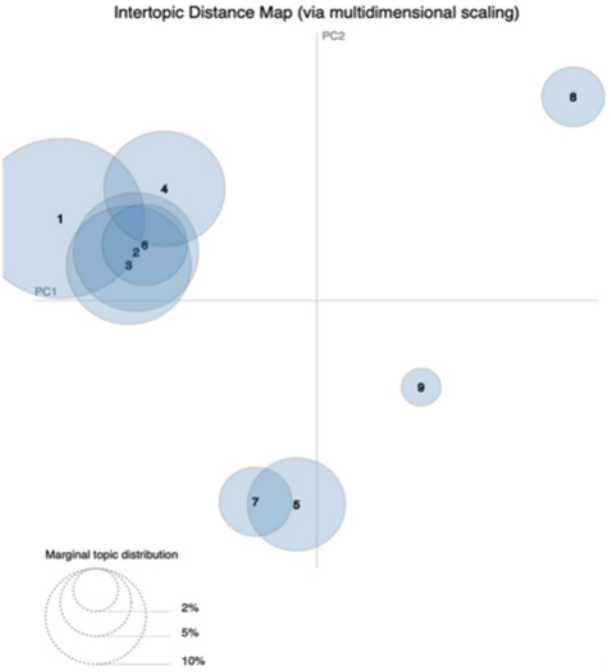

图 9 LDAvis 主题可视化图——“网络、信息安全”

9) 大数据技术还应用在企业组织管理、产业转型升级等方面，大数据技术的应用能够更好地促进社会经济发展。以“企业”和“产业”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是企业、产业以及大数据主题，所以在停用词表中加入“大数据’‘数据’‘企业’‘产业”4 个词，避免这 4 个词多次出现，影响主题聚类效果。从结果可以看出，企业产业主题的研究有以下几个分支：对财务管理的研究；对管理模式创新的研究；对工业产业链的研究；对中小企业的研究；对人力资源的研究等。图 10 为“企业、产业”主题下各个主题的 LDAvis 可视图。

大数据的研究也体现了与时俱进性，说明大数据技术与各个领域的研究都具有可融合性。

1) 精准扶贫这一主题在 2017 年首次出现，在随后的几年中也有体现。以“扶贫”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是扶贫以及

图 10 LDAvis 主题可视化图——“企业、产业”

大数据主题，所以在停用词表中加入“大数据”“数据’‘扶贫”3 个词，避免这 3 个词多次出现，影响主题聚类效果。从结果可以看出，这些主题词代表了扶贫的各个方面，比如教育、政府治理；也可以看出对农村进行扶贫的重要性等。图 11 为“精准扶贫”主题下各个主题的 LDAvis 可视图。

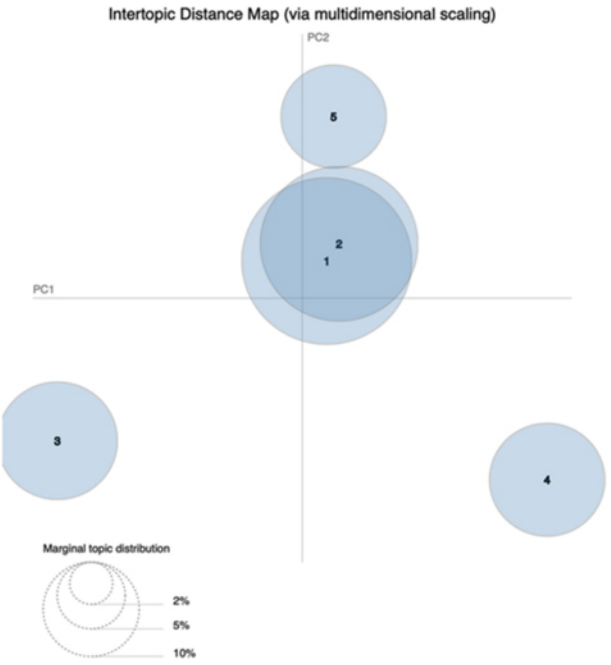

图 11 LDAvis 主题可视化图——“扶贫”

2) 在 2020 年的研究主题中出现了疫情防控，2020 年新冠疫情肆虐，而利用新技术防控疫情能

够更有效率地对疫情进行遏制。以“疫情”为关键词在各年度原始数据中的标题列进行筛选，将得到的论文进行 LDA 主题聚类，由于研究的就是疫情以及大数据主题，所以在停用词表中加入“大数据”“数据”“疫情”3 个词，避免这 3 个词多次出现，影响主题聚类效果。从结果可以看出，疫情防控主题的研究有以下几个分支：对人口流动进行研究；企业复工复产的研究；传染病预测预警的研究；对政府治理的研究等。图 12 为“疫情”主题下各个主题的 LDAvis 可视图。

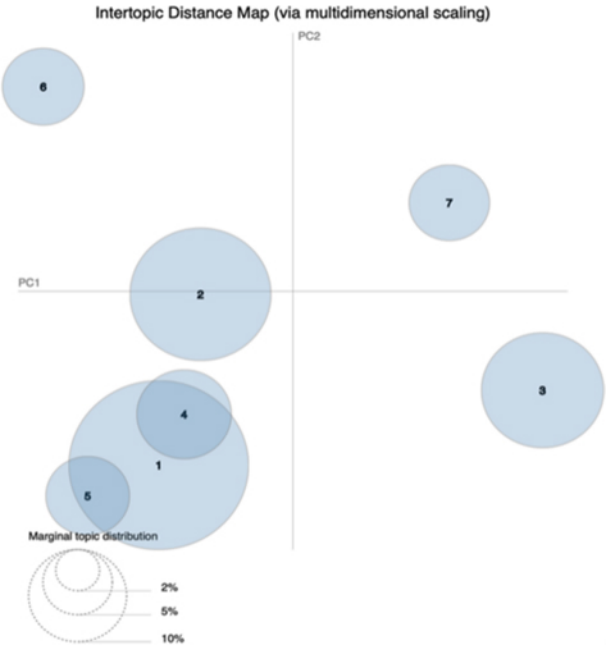

图 12 LDAvis 主题可视化图——“疫情”

2.3.2 文献聚类及主题强度分析

为了更好地了解不同主题的研究状态，根据文档—主题分布，对每个主题下的文献数量进行分析，同时结合主题强度的定义，即主题强度主要描述了主题在某一时期的热门程度。在某一时期关于某个主题的文献数量越多，说明该主题的强度越高，可以被认为是热点主题。对表 1、表 2 进行分析可以得出：

1）2008—2012 年时间窗口内，“企业信息计算分析”主题所包含的文献数量最多，所以“企业信息计算分析”主题为该时间窗口的热点主题。同理，可以得出 2013 年的热点主题为“信息管理”；2014 年的热点主题为“经济发展战略”；2015 年的热点主题为“技术发展”；2016 年的热点主题为“大数据时代面临的机遇与挑战”；2017 年的热点主题为“大数据对社会的影响”；2018 年的热点主题为“大数据时代的机遇与挑战”；2019 年的热点主题为“大数据对社会的影响”；2020 年的热点主题为“大数据对社会生活的影响”。从各个时间窗口的热点主题可以看出，在前期关于大数据的研究主要集中在对数据信息的管理和分析上，后期研究热点逐渐转变为大数据产生的影响。

2）不难发现，除了包含文献数量最多的主题外，还有一些主题在时间窗口中也占据较大的比例。并且有许多主题在多个时间窗口中出现。所以本文挑选 4 个占据比例较大的且在多个时间窗口出现的主题，对其进行主题强度随时间变化的分析。结果如图 13 所示。可以看出“图书馆服务”这一主题在 2013—2016 年呈现稳定上升的趋势，在 2017 年有所下降，但 2018 年又有所回升，之后呈下降趋势；“智慧城市”这一主题总体呈上升趋势，但在个别年份有下降的波动；“市场营销”主题在 2016 年之前呈上升趋势，并在 2016 年文献数量达到最高，接着在 2017 年、2018 年、2019 年下降至 0 篇，2020 年又上升至 383 篇；“信息安全”主题在 2008—2014 年呈上升趋势，到 2015 年下降至 0 篇，接着又呈现上升趋势，直到 2019 年，之后呈现下降趋势。

3 结论与分析

本文借助 LDA 主题模型，结合模型困惑度判断确定模型的最优主题数，同时考虑文献发表时间，以年为单位划分时间窗口（由于 2008—2012 年的论文数较少，所以将这 4 年合并成一个时间窗口），共分为 9 个时间窗口。对这 9 个时间窗口中的文献进行主题挖掘，对挖掘到的主题的内容进行研究，并选取 11 个具有代表性的主题对这些主题内的论文再次进行 LDA 主题聚类和 LDAvis 主题可视化，进一步分析主题内的研究热点；对挖掘到的主题进行强度分析，按照主题包含的文献数量确定每个时间窗口的热点主题，笔者还挑选了 4 个主题对其主题强度随时间的变化趋势进行展示分析。结

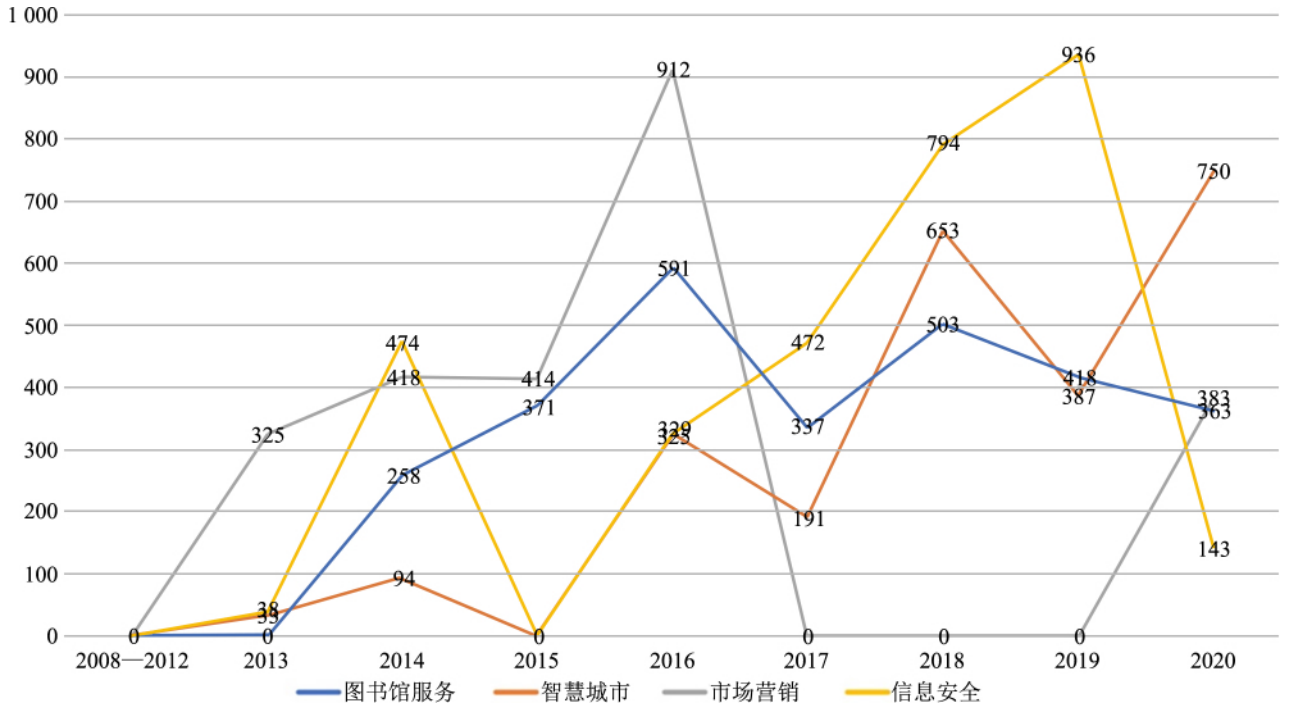

图 13 主题强度随时间变化趋势

果表明，LDA 模型能够较为准确地提取大数据领域文献的研究主题，这有利于研究人员了解该领域的发展状态，把握未来的研究方向，探寻新兴主题。

当然，本文亦存在不足之处：①本文选取的样本数量较大(篇)，具有一定的实践意义，但仅仅考虑到了中文期刊文献，未考虑到外文文献，未来研究可以考虑扩大样本容量，以充分了解大数据领域的发展状态；②各个主题的标签是笔者根据关键词和自己的主观判断总结的，具有一定的主观性。

参 考 文 献

[1] 新华社. 中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议 [EB/OL]. http://www.gov.cn/zhengce/2020-11/03/content_5556991.htm，2021-03-25.

[2] 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，2013，31(9)：10-14，136.

[3] 赵悦阳，曹霞，方丽. 大数据领域理论研究的发文分析与热点分析 [J]. 预防医学情报杂志，2018，34(8)：1121-1126.

[4] 童子颐. 国内大数据研究热点分析 [J]. 情报探索，2015，(7)：38-41，46.

[5] 代芯瑜，张文曦. 我国大数据研究现状与热点分析 [J]. 思想战线，2013，39(S2)：149-154.

[6] 夏军辉，扈桂红. 图书情报领域关于大数据研究的热点分析 [J]. 资源信息与工程，2021，36(1)：144-147.

[7] 王春华，李维，文庭孝. 我国图书情报领域大数据研究热点分析 [J]. 图书情报知识，2015，(4)：82-89.

[8] 虞秋雨，徐跃权. 近 5 年我国图书情报领域大数据研究热点分析 [J]. 图书馆学研究，2020，(8)：10-18.

[9] 黄鹂，曹东维. 国际医学信息学领域大数据研究热点分析 [J]. 医学信息学杂志，2018，39(4)：2-7.

[10] Blei D M，Ng A Y，Jordan M I. Latentdirichlet Allocation [J]. The Journal of Machine Learning Research，2003，(3)：993-1022.

[11] Sievert C，Shirley K E. LDAvis：A Method for Visualizing and Interpreting Topics [J]. Procedings of the Workshop on Interactive Language Learning，Visualization，and Interfaces，2014：63-70.

[12] 赵公民，吕京芹，武勇杰. 基于 LDA 模型的新能源汽车政策文本量化分析 [J]. 科技和产业，2021，21(1)：49-55.

(责任编辑：陈 媛)