arXiv:2303.10239v1 [cond-mat.other] 16 Feb 2023

凝聚态物质电子结构软件包引文的密度泛函理论主题建模

Marie Dumaz [*1,3], Camila Romero-Bohórquez[2], Donald Adjeroh[3],和 Aldo H. Romero[1]

[1] *Department of Physics and Astronomy, West Virginia University, Morgantown,WV 26506-6315, USA*[2] *Neuroscience Department, West Virginia University, Morgantown, WV26506-6315, USA*
[3] *Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506-6315, USA*

摘要

随着新科学论文不断涌现，研究人员越来越难以了解自己研究领域的最新文章。准确地对论文进行分类，是实现个性化推荐、便捷获取感兴趣的研究成果的第一步。密度泛函理论（DFT）领域尤其能体现这一点：它是一种应用于各种不同研究和相互关联学科的方法，拥有一个规模庞大、发表了大量研究文章的活跃学术群体。我们提出了一种基于主题建模的全新无监督出版物分类方法，并以一组与 DFT 相关的文献为案例。首先，通过对出版物摘要进行词语分析和聚类来创建主题，然后根据词语相似度，将每篇出版物／论文归入相应主题。接着，我们通过分析主题与出版商、期刊、国家或出版年份之间的关联，得出了一些有意思的发现。所提出的方法具有通用性，可用于分析其他研究领域的出版物和引文趋势，而不仅限于密度泛函理论领域。

1 引言

多年来，基础科学领域的科学文献一直采用物理与天文学分类方案（PACS）进行分类。该方法将一系列代码与关键词对应起来，以涵盖出版物的主要主题。这

*通讯作者：mcd0029@mix.wvu.edu

通常由作者完成，并为实际分类提供了相当大的自由度。尽管一些期刊仍在采用这种方法，但这一环节正逐渐转向更聚焦的主题分类[13]。基本而言，这一变化源于引入该方法的机构（美国物理学会，AIP）决定停止提供支持，主要原因是稿件提交过程中行政管理和论文处理成本高昂。因此，有必要采用一种更具动态性和灵活性的方法。通过这种方法，不同主题能够反映科学发展和科学兴趣的持续变化，并能在这一领域中为审稿人乃至期刊分类。鉴于此，机器学习中的主题建模正受到广泛关注 [14, 2]。本文提出一种无监督学习技术，旨在找出文本之间的关联，从而依据一小组主题对文献进行分类。这些主题通过一种无监督机器学习技术得出：该技术在潜在空间中寻找文献之间的相似性，进而在所选文献中聚类出有意义的词语。尽管过去已有研究使用主题建模对研究主题进行分类 [10, 15], ，但在本研究中，我们从一个定义明确的数据库出发，将这一方法应用于与电子结构软件包使用相关的出版物。

尽管有一种能够对基础科学中的主题进行分类的通用方法将是理想的，但在本文中，我们决定聚焦于一个特定的研究领域，在这里我们能够获取一个控制得非常完善的数据库 [6]。密度泛函理论（DFT）是目前用于表征原子尺度材料的最常用方法 [9, 8]。它属于所谓的电子结构方法，其优势在于大量计算代码和软件都采用了这种方法，由此形成了相关社群，推动研究工作。我们根据引用这些计算代码程序库的研究论文来选择文献，从而将研究范围缩小到仅与 DFT 相关的论文。

已有一些工作尝试分析不同子领域中与 DFT 相关的论文，但这些工作是依据 PACS 编号 [1, 5]对文章进行分类的。然而，我们从科学网数据库下载的数据集并未保留 PACS 编号记录，只保留了关键词。此外，由于 DFT 是一种具有诸多不同应用的方法，且应用领域并不限于物理学，并非所有论文都遵循 PACS 结构。这再次表明，有必要建立一种全新的分类体系，完全基于文献的内容。我们的方法不仅能够根据内容对文献进行分类，还可用于发现新的主题以及主题之间的关系，从而识别新兴的研究领域。

2 方法论

研究论文及其摘要采用与以往研究相同的方法 [6], ，通过 Web of Science 数据库（WoS）获取，并将范围扩展至包括 2020年发表的文章。

从 WoS 数据库中提取摘要后，我们进行了若干预处理步骤，以从文本中筛选出最有意义的词语。首先，我们删除了某些词组，例如“左箭头”或“竖线”，因为它们仅用于创建数学表示。出于同样的原因，我们删除了两个美元符号之间的所有字符，因为这是生成方程的标准命令，而这些字符只会干扰算法。不过，这条规则有一个例外：如果两个美元符号之间出现下划线，则会保留该内容，因为这种模式用于表示化学式，而化学式是密度泛函理论各子领域中不可或缺且有意义的概念。我们还删除了出版和版权标签、DOI 编号以及所有非字母数字字符。

然后，对剩余句子进行分词，也就是说，将文本拆分为多个单元，并从中生成词元列表。接着，我们从这些列表中剔除所有停用词，即那些过于常见、无法帮助区分不同子领域的词语。我们使用了标准的 NLTK [3] 停用词列表，并根据多轮试验和反馈中积累的数据经验，添加了其他停用词。例如，“理论”和“方法”并非磁性或气体吸附等某些物理学领域所特有的词语。我们还会删除所有单个字母和数字，它们常用于数学表达式或表示数值，本身并无实际含义。随后，我们创建二元词组和三元词组，也就是经常连续出现的两个或三个词元，并用下划线将它们连接成一个词。例如，“从头算”和“密度泛函理论”就是这样的词组。最后，对词元进行词形还原，将词语还原为词根；然后筛选所得词元列表，只保留至少出现在两篇文献中、且出现在不到 20% 的论文中的词元。我们根据最终列表创建词典和词袋（BoW），并将它们作为算法的输入。

潜在狄利克雷分配（LDA） [4] 是一种主题建模算法，它使用贝叶斯统计方法，根据以词袋形式传入的文献生成文本。这意味着 LDA 不考虑词语的顺序。事实上，潜在狄利克雷分配模型会创建一组主题，其中每个主题都是从分布β_k ∼狄利克雷(η)中抽取的语料库中词语的概率分布，而每篇文档都是从 θ_k ∼狄利克雷(α)中抽取的主题概率分布。

这里，我们将 M 定义为文献总数，将 N 定义为一篇文档 m中的词语总数，并将 k 定义为主题数。如图 1 所示，LDA 跨越三个层次。参数 α 和 η 是语料库级参数，这意味着它们在每个语料库中仅采样一次。 θ 表示文档级变量，每篇文档采样一次。最后，变量 z_n 和w_n 处于词语级：每篇文档中的每个词语都要采样一次。w_n 也是唯一可观测的变量，从作为输入提供的词袋中选取。

进行主题建模时，我们使用 Python 库 Gensim 实现了潜在狄利克雷分配（LDA）算法 [11] ，该库已提供一套便捷的函数，可用于创建、评估和可视化潜在狄利克雷分配模型。我们使用了两个词袋：一个包含数据集中 1990年 至 2019年 的所有文献，另一个仅包含 2020年 发布的文献。2020年 的语料库被用作测试

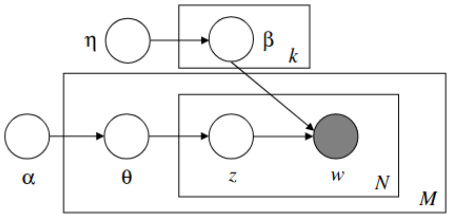

图 1：潜在狄利克雷分配模型的图示 [4]

set.

我们创建了 16 个模型，主题数各不相同（k），范围从 5 到 80，同时将所有其他参数设为 null。在此分析中，我们选择对主题数为 15、25、35 和 45 时的其余参数进行调优。所有模型均使用C_v 一致性得分 [12], 进行评估；该得分与人工排名的相关性最强。随后，我们运行了 480 个模型，对超参数进行调优，包括 α 和 η，并分别考察 k的 5 个不同取值。对于每个 k取值，我们都选用得分最高的模型 C_v，并在本文后续部分中沿用该模型。

为了加快训练速度，我们在多个核心上进行参数调优。需要注意的是，onlineLDA 中使用的随机优化过程包含数据分区的随机化，因此如果与并行化结合使用，就无法确保结果完全可复现。

为了发现趋势和兴趣，我们将主题分配给出版物的文献计量属性。分配主题时，我们将潜在狄利克雷分配模型返回的概率视为权重。我们创建一个矩阵 A，其中每一行代表一篇文献，每一列代表一个主题。矩阵中的元素表示将某个主题分配给某篇文献的概率。

该矩阵对于分析不同国家、期刊和年份的主题分布尤其有用。例如，在计算每个国家各主题的出版物数量时，我们可以利用这些权重进行更精确的分析。为此，我们创建一个矩阵，其中每一行代表一个主题，每一列代表一个国家。例如，如果一篇文献是在美国创建的，那么每个主题都会将其在该文献中对应的权重加到“美国”列中。这些分析也都经过归一化处理，即矩阵中的每个值都除以A 矩阵中所有值的总和，该总和等于文献总数。

3 结果

在此，我们介绍数据集、实验和结果。

为了创建一个专注于密度泛函理论的研究论文数据集，我们使用了“科学网”（WOS）的图形界面，并下载了所有引用若干最常用的计算软件包来研究晶体系统的文章。

我们收集了从首次引文到2019年期间的所有论文作为训练集；并收集了2020年的所有引文作为测试集。 [6]

3.1 文档与主题的分布

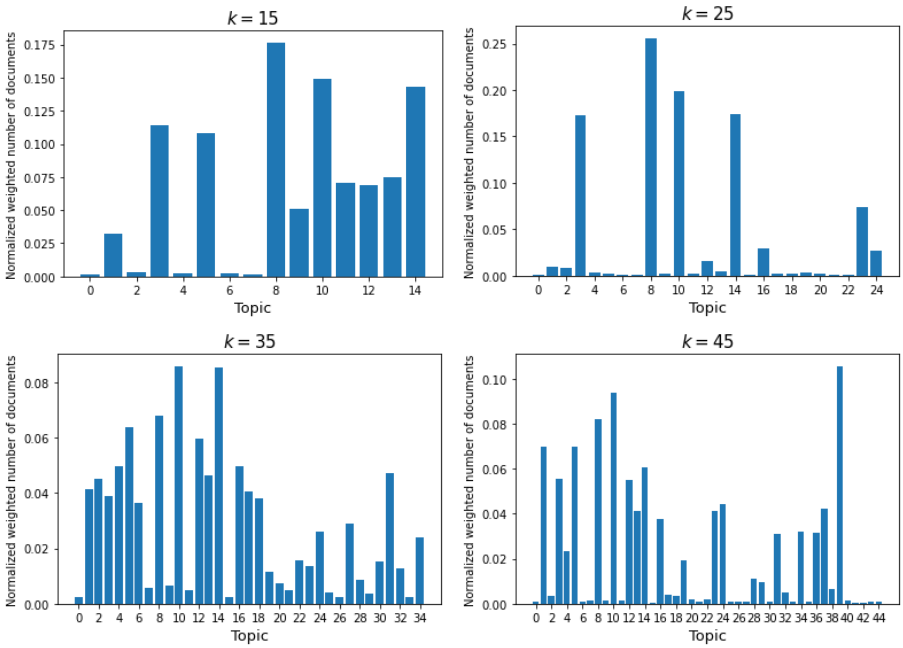

图2：各主题的文档分布（针对1990—2019年的出版物）

图2展示了数据集中各主题的文档分布，也就是说，它显示了包含每个主题的文档数量。请注意，我们使用某个主题出现在文档中的概率作为权重，具体方法见方法论部分。我们将每个主题在所有文档中的权重相加，再除以文档总数。该图让我们大致了解各模型识别出的主题分布得有多均匀，以及哪些主题在文献中更常见。

可以看出，包含 25 个主题的模型极不均衡，仅 4 个主题就占到了文献的 80%。接下来，针对 $k = 15$ and $k = 45$ 的潜在狄利克雷分配模型似乎分布得比较均匀，因为没有任何主题包含超过 18% 和 11% 的文献。通过直方图的纵轴可以轻松看出这一点；纵轴也是快速了解分布概况的方法：如果纵轴的范围更大，就意味着至少有一个主题在文献中占比更高，也就是说分布不那么均匀。同样，我们可以研究所归属文档数归一化值的范围。包含 25 个主题的模型范围最大，为 0.255，

其次是包含 15 个主题的模型，为 0.175；包含 45 个主题的模型为 0.105。范围最小、因而分布最均匀的是 $k = 35$ ，范围为 0.086。事实上，要覆盖 90% 的文献，对于 $k = 15$，需要取占比最高的前 8 个主题（即主题总数的 53%）；对于 $k = 25$，需要取 24% 的主题；对于 $k = 35$ ，需要取 57%；对于 $k = 45$，需要取 37%。此外，包含 35 个主题的模型也是文献占比最低值最高的模型。

$k = 35$ 和 $k = 45$ 模型的多样性似乎也表明，对于我们的语料库而言，15个主题并不足够。事实上，我们发现了主题专门化的证据，进一步证实了增加主题数量的好处。例如，在 $k = 15$模型中，主题 3包含以下词语：“形成”“活性”“氢”“金属”“反应”“机理”“过程”“催化剂”“氧化”和“锂”。我们可以将这一主题对应到 $k = 35$ 模型中的两个不同主题：它与主题6共享“活性”“金属”“反应”“催化剂”“氧化”这些词语，与主题27共享“反应”“形成”“机理”“过程”这些词语。我们可以清楚地看到，它分成了两类：一类更侧重于特定过程和应用（主题6），另一类则更为宽泛（主题27）。

从这些分析中我们可以得出结论：最均衡的模型对应于 k= 35。基于这两个原因，本节余下部分将主要聚焦于这个特定的潜在狄利克雷分配模型及其结果。

3.2 深入了解主题

图 3 展示了潜在狄利克雷分配模型中 35 个主题的词云。每个词云呈现了该主题中概率最高的 10 个词语。每个词语的大小与其在对应主题中的概率成正比。词语越大，说明其概率越高，在该主题中的影响也越大。我们还为主题添加了名称，以便快速了解各个词云所代表的内容。尽管有些主题有所重叠，但大多数主题彼此相对独立，界定清晰。例如，主题13中可以找到与磁性问题相关的词语，而主题 7 则更多涉及氢和氢化物。氢并非磁性原子，因此这两个学科不太可能被归为同一组；事实也确实如此：它们相当不同。不过，主题 1 涉及合金，而合金与磁性有关，因为某些合金金属具有磁性；然而，它还是被归入了另一个独立的组（主题13），这表明潜在狄利克雷分配模型善于识别不同的词语关系。主题14也是一个很好的例子，其中的词语可能适用于许多不同领域，例如“数据”“模拟”“模型”或“方法”，但它们彼此相关联，构成了一个独特的主题。

我们还可以使用 t-SNE 降维方法 [7] 将模型的主题可视化，从而分析这些主题。图 4 展示了将数据降至二维后的结果，其中每个点代表一篇文献，其位置取决于它与其他文献的相似度，每种颜色代表一个主导主题。需要注意的是，有些主题没有显示出来，因为它们不是任何文献中的主导主题，例如主题 0 和 29。由于这些主题在文献中的分布较为均匀，我们可以看到许多主题都有所呈现，而不是只有少数主题



图 3：构成 $k = 35$ 潜在狄利克雷分配模型中各主题的前 10 个词语

主导整个图。我们还可以注意到，有些主题的分布更为分散，例如新方法这一主题（如主题31）。这是意料之中的，因为研究人员正在许多不同领域开发新过程并进行数据模拟。相比之下，其他主题则聚集在一起，界定清晰，例如第一个关于合金化和晶体形成的主题。这幅图还有助于直观呈现主题之间的相似性。两个主题越接近，它们之间的联系就越紧密。

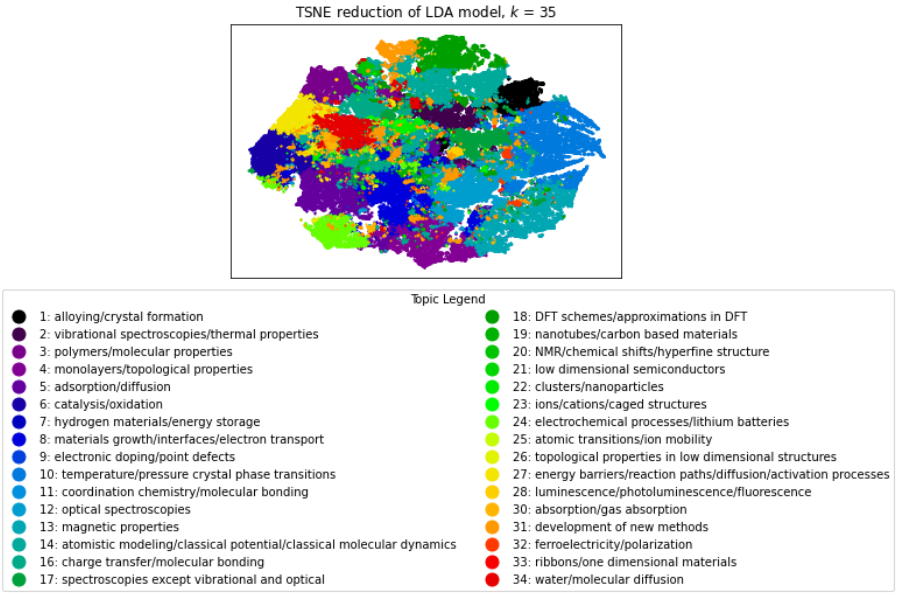

图 4： $k = 35$的潜在狄利克雷分配模型的 t-SNE 降维

例如，主题31（能垒/反应路径/扩散/活化过程）与主题6（催化/氧化）和主题3（聚合物/分子性质）较为接近，但与主题1（合金化/晶体形成）和主题13（磁性）位于相对的另一侧。事实上，分子性质以及催化等活化过程与化学相关，而磁性和合金化则更多是在物理学中研究。

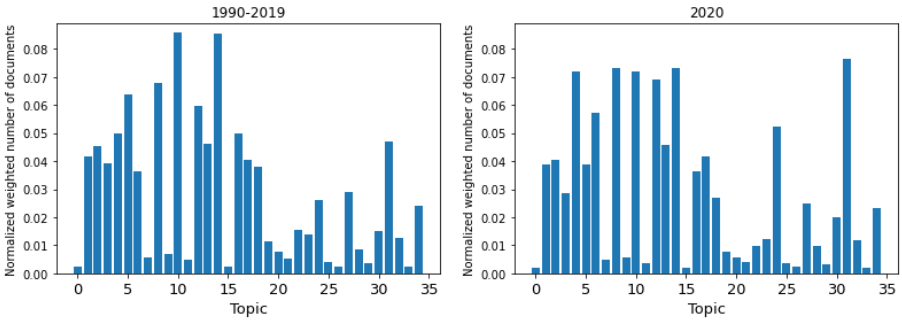

图5：文档在 k=35 个主题上的分布。左：1990—2019年语料库。右：2020年语料库

4 主题的时间演变

图5比较了各主题下文献的分布：左侧为1990—2019年语料库（已在图2中展示），右侧为2020年语料库。两幅图相似，且都显示文献分布良好。直到2019年，主题10（温度/压力/晶相转变）和主题14（原子模拟/经典势/经典分子动力学）是最常见的学科；到了2020年，其他一些主题也赶了上来，其中包括主题31（新方法开发），它如今被分配的文献数量最多。这也可能意味着，虽然主题8（材料生长/界面/电子输运）和主题10直到2019年都非常重要，而且如今依然如此，但主题4（单层材料/拓扑性质）、主题24（电化学过程/锂电池）和主题31（新方法开发）近来正越来越受关注。两幅图的相似性也表明，我们的模型具有良好的泛化能力，能够高效地为新文献分配主题。

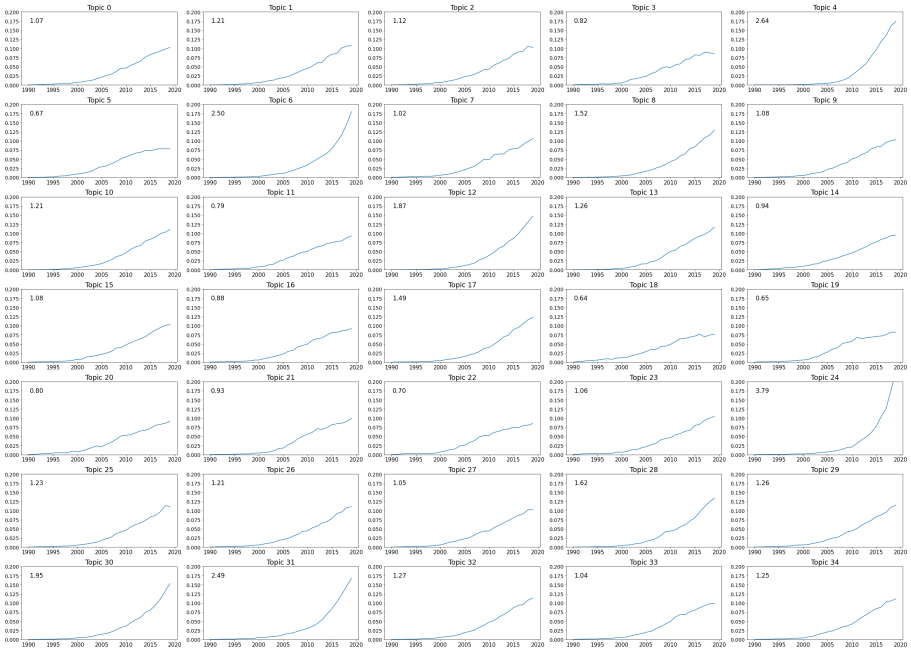

图6： $k = 35$各主题随年份变化的归一化加权文献数量。每个图（左上角）中的数字表示最佳拟合功率因子。

我们针对每个主题分别计算了逐年归一化加权文献数量。每年发表的DFT 相关文章数量都呈指数级增长，因此，几乎所有主题也

呈指数级增长。为便于比较，我们计算了各条曲线自 2000年以来的幂律因子。尽管有些主题的演变更符合指数曲线，但幂律函数在所有主题中的平均值最高 r^2。结果如图6所示，每个子图左上角的数字表示最适合该主题的功率因子。我们只考虑 2000年之后的出版物，因为 DFT软件正是在这一时期开始真正得到全球范围的应用。

共有九个主题的幂律因子小于 1。尽管自 2000年以来，这些主题的重要性仍有所提升，但其中一些已开始趋于稳定，甚至逐渐失去 DFT社群的关注。例如，DFT方案相关的主题 18 和纳米管相关的主题19，其记录到的幂律因子最低，近年来都未能获得新的发展势头；此外，主题 18 的出版物数量还在2017年首次出现下降。相反，与电化学过程和锂电池相关的主题24，以及主题4（单层材料/拓扑性质）、主题6（催化/氧化）和主题31（新方法开发）的增长率显著较高，似乎已成为当今 DFT领域最重要的学科之一。其他学科的增长率均处于中等水平，体现了 DFT的多样性，也说明这一方法对于不同主题都十分重要。

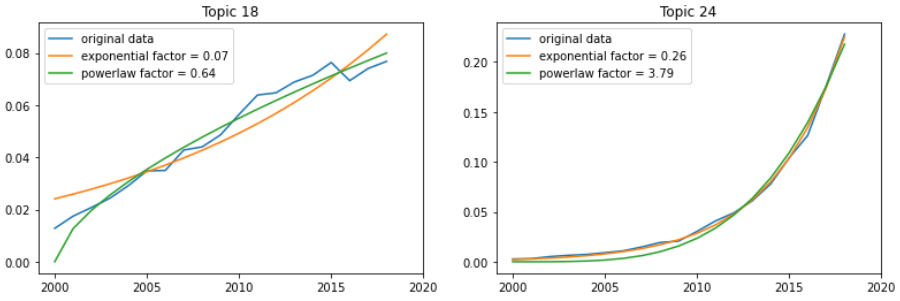

图7：比较增长率最低和最高的两个主题，并分别拟合最佳指数曲线和幂律曲线。左：幂律因子最低的主题 18。右：幂律因子最高的主题24。

图7展示了我们的模型中35个主题里增长最慢和最快的两个主题。我们分别用幂律函数和指数曲线拟合这两条曲线，并将拟合系数列于左上角。主题24（电化学过程/锂电池）就是一个更适合用指数函数拟合的增长案例，其 r^2 达到0.998，尽管幂律函数也取得了很高的 r^2 分数0.988。除了增长速度的显著差异外，我们还可以注意到主题 18（DFT方案）的演变过程有多么“起伏不定”。与主题24平滑的曲线相比，这些不规则波动也可能表明人们的兴趣正在减退（或至少不太稳定）。

图8还展示了各主题随年份的增长情况，形式为一幅

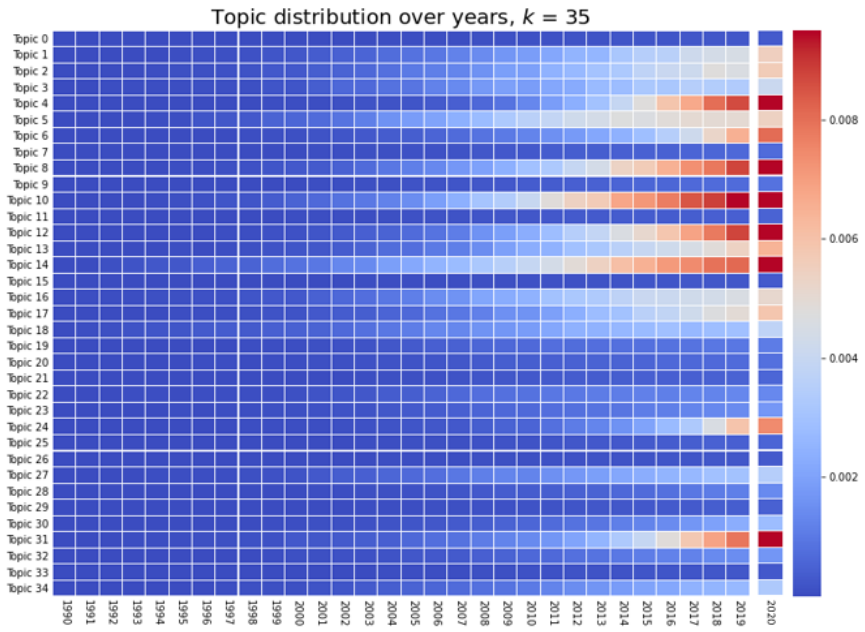

图8： $k = 35$各主题按年份统计的归一化加权文献数量

热力图的形式展示了各主题多年来的增长情况。我们注意到，2000年之前没有出现显著增长，这印证了我们仅研究2000年之后指数因子的选择。此外，大多数主题至少到2005年之后才开始获得发展势头。值得注意的是，指数因子最高的三个主题（主题4、6和24）在短短几年内迅速增长，这正是其因子较高的原因。尤其是主题24，直到2018年才开始显现出与其他主题的差异。有趣的是，在这三个主题中，有两个（主题6和24）并不属于文献占比最高的主题。不过，鉴于其较高的功率因子，我们只能预测它们在未来几年会变得更加重要。同样，通过观察一些学科的发展轨迹，例如主题1或17，我们可以有把握地预测这些子领域在未来几年会受到更多关注。图中的2020年这一列似乎印证了这些预测，因为主题1、4、6、17和24在过去一年中的占比均有所上升。另一方面，我们也很容易判断出哪些主题趋于稳定、势头正在减弱。例如，主题3（聚合物/分子性质）和主题18（DFT方案/密度泛函理论近似）至少十年来一直受到科学界的关注，但其增长速度仍不及其他主题。

比较主题14和主题31的发展变化很有意思，因为它们彼此相关。主题14包含“数据”“模型”

以及“模拟”等词语，体现了学界将代码、机器学习和新的计算方法融入研究工作的努力。随着这类研究在当今实验中所占比重越来越大，该主题多年来一直稳步增长。其次，与代码和过程改进及优化相关的主题31直到2014年才开始受到关注，比主题14晚了约10年。不过，它增长得更快，因此到2020年，这两个主题都跻身密度泛函理论领域发表论文最多的学科之列。我们还可以提及与石墨烯和单层材料相关的主题4的发展变化。它在2010年之后几年开始增长；2010年，安德烈·海姆和康斯坦丁·诺沃肖洛夫因在二维材料石墨烯方面的工作而获得诺贝尔物理学奖。

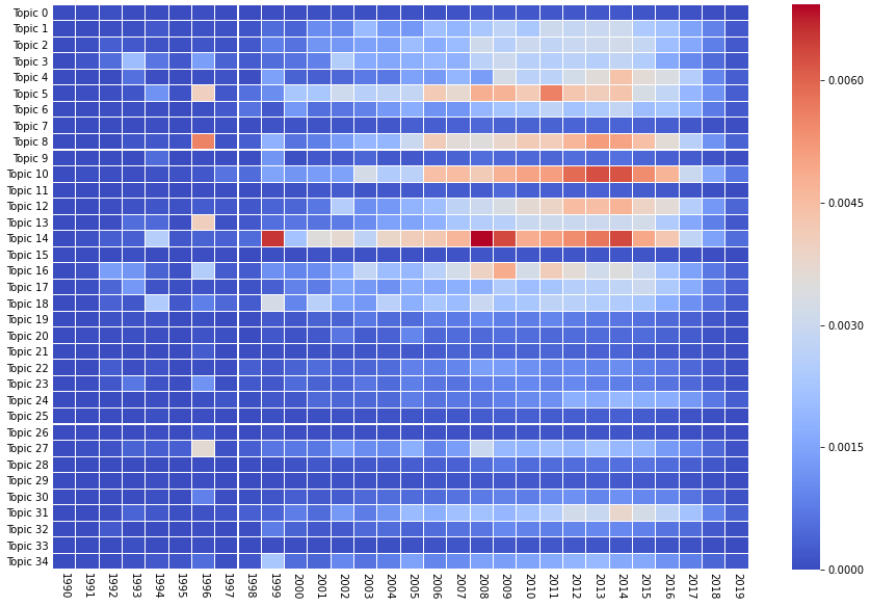

图9： $k = 35$各主题按年份统计的引文数，经归一化和加权处理

5 引文趋势

观察图9中各主题每年的引文数量，我们会发现与上一幅图中相同的趋势，只不过时间提前了几年。事实上，如今发表的论文引用的是较早的研究文章，因此，2019年出版物数量的增加意味着对较早论文的引文数量也会增加，这些论文通常至少早于当前3年。这一时间差无疑源于研究过程：阅读文献、形成新想法和假设、获取数据并积累经验，最后得出结论并发表。

因此，观察引文数量随年份变化的趋势，可能比仅观察已发表论文的数量更能帮助我们深入了解情况，也更具预测价值。对于刚开始吸引更多关注的主题，例如主题6和24，我们可以追溯其趋势的起点至2000年代末或2010年代初；而对于存在时间较长的主题，例如主题8、10或14，则可以将其归因于1990年代末的发现。

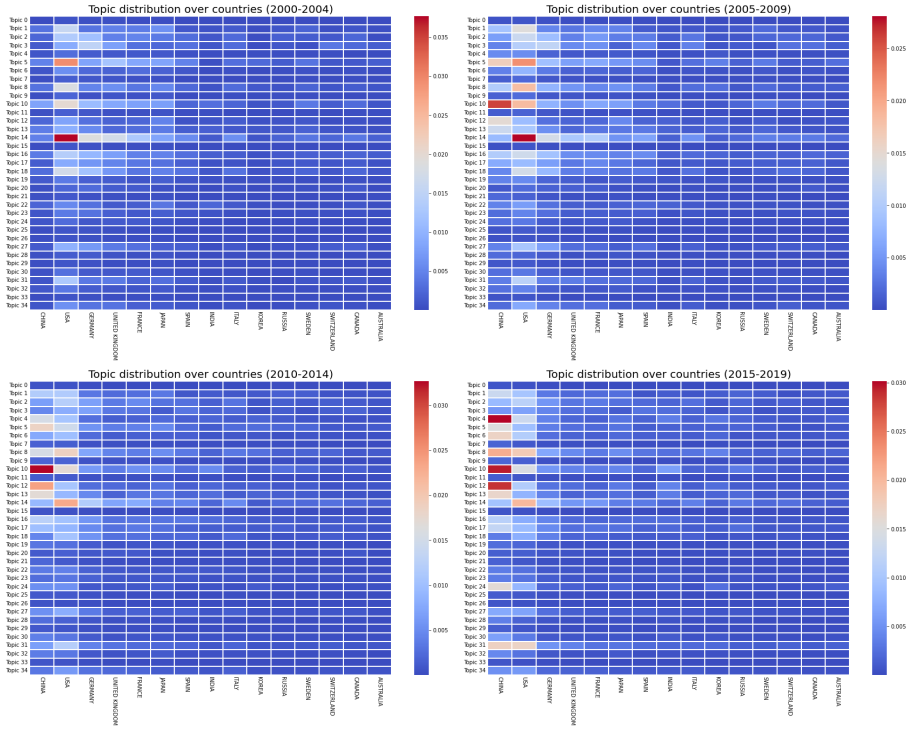

图10：不同时间段内各主题按国家统计的归一化加权文献数量（仅显示排名前 15 的国家），适用于 $k = 35$

5.1 主题的空间演化

我们还研究了密度泛函理论中各主题多年来的空间演化，尤其是按国家进行的演化。为此，我们根据所有作者的所属机构识别国家。不过，每篇论文中每个国家只计一次。

图10将各国的主题热图按四个不同的时间段划分。2005年之前，美国在密度泛函理论领域占据主导地位，在更多主题中发表的出版物数量也最多。随后，这一比例逐渐转向中国；如今，中国发表的文章更多，研究兴趣也更加多元。2005年之后，中国开始涉足密度泛函理论，首先关注的是

晶相变这一学科（主题10）。压力诱导的晶相变是密度泛函理论的主要优势之一，也是最活跃的子领域之一，这解释了它为何最先吸引关注。直到现在，主题10仍是中国研究工作的重要组成部分；2015年至2019年间，中国作者发表的主要学科中仍包括这一主题，以及主题4和主题12。如今，美国仍在密度泛函理论的许多不同主题上有所贡献，对主题14的兴趣有所增加，但在主题5（吸附/扩散）上的出版物数量有所减少。

我们可以准确预测，中国很快将在密度泛函理论领域占据主导地位，因为其出版物数量只会持续增长。图11展示了2020年各国家的文献在不同主题上的归一化分布，证实了中国在密度泛函理论领域的影响力。此外，尽管本研究统计的大多数软件都在欧洲开发，但它们在中国和美国的使用最为广泛。德国的出版物数量位居第三，Turbomole也起源于德国；然而，作为我们数据集中使用量排名第六的代码，Turbomole在2005年之后几乎无法与中国和美国竞争。同样，英国和法国的影响力多年来持续下降，尽管在2005年之前，它们在若干不同主题上都十分活跃。

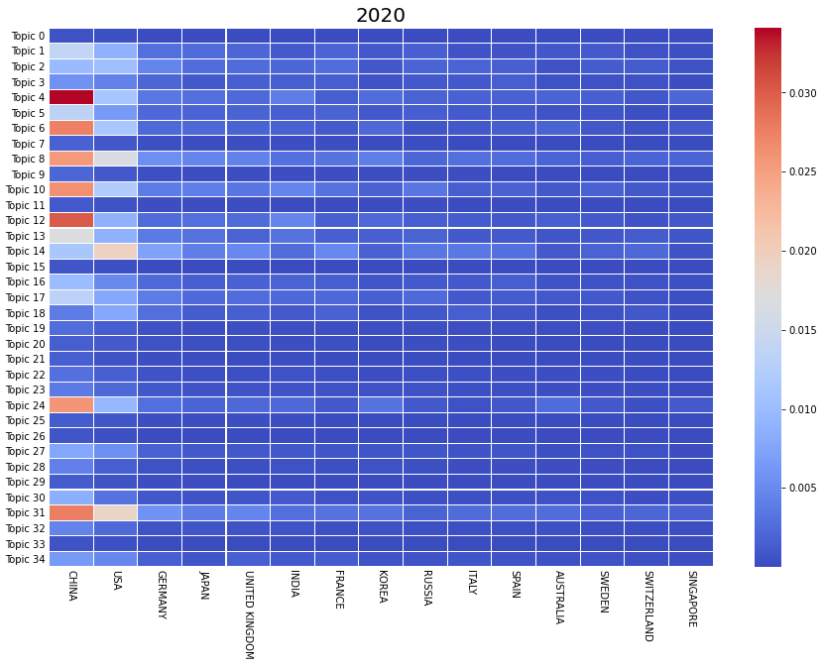

图11： $k = 35$ 中 2020年发表的文章按主题划分的各国家归一化加权文献数量（仅显示前 15 个国家）

6 期刊和出版商的主题分布

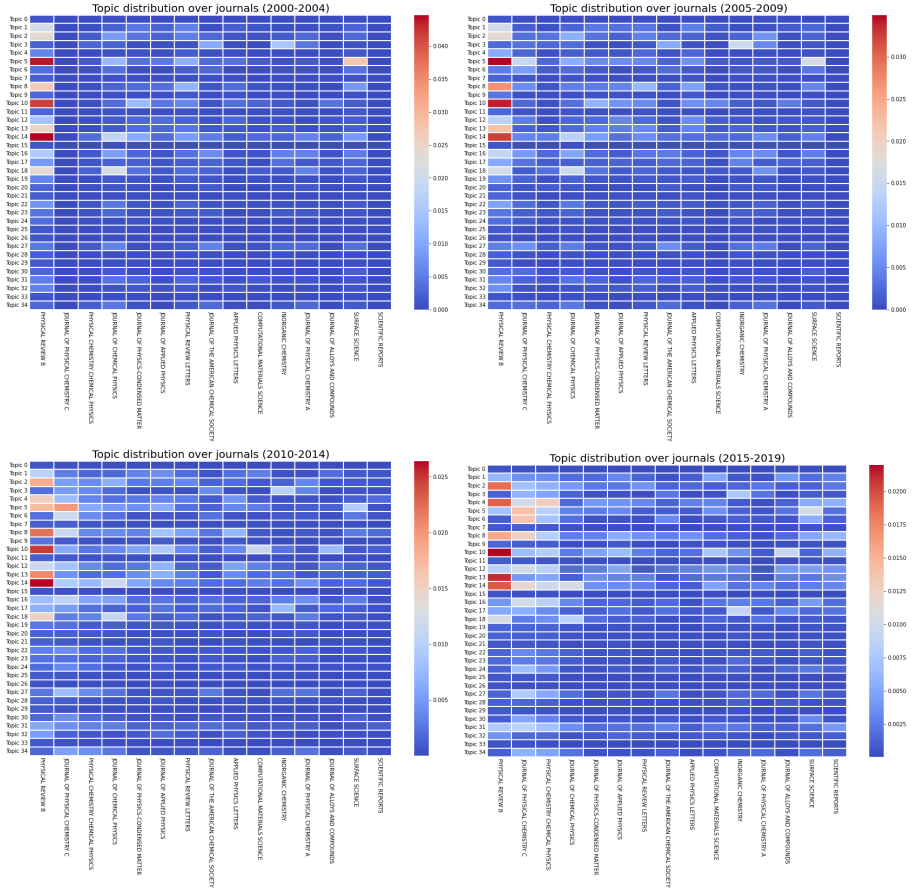

图12： $k = 35$在不同时间段内各主题的每种期刊文献归一化加权数量（仅显示排名前 15 的期刊）

与图10类似，图12展示了按期刊划分、各主题上的出版物归一化数量（以概率为权重计算），并根据论文发表年份将其分为四个五年期。与之相同，这些图表呈现了密度泛函理论出版物在不同期刊间的多样化趋势，同时突显了少数几个重要期刊的地位。然而，图10显示，各国家自2005年起便呈现出多样化趋势；而直到更近些年，大多数论文仍发表在Physical Review B上，且似乎没有其他期刊能够在出版物数量上取而代之。

所展示的 15 种期刊发表的出版物数量最多，并且

按从左到右的顺序排列。尽管有些期刊（如 Surface Science）专注于某个主题并长期坚持这一方向，但我们可以看到，大多数期刊都涉及多个不同学科。Physical Review B 和 Journal of Physical Chemistry C 是两种重要期刊，在密度泛函理论领域都颇具影响力，因此预计它们会是该领域最大的贡献者。此外，由于密度泛函理论主要应用于物理学和化学，专注于凝聚态分析的这两种期刊分别代表这两个领域，也就不足为奇了。

事实上，排名第 2、第 3 和第 4 的三本期刊都与密度泛函理论的化学方向相关。尽管它们的目标相似，却均分了这些出版物。《Journal of Chemical Physics》侧重于主题14和18，《Journal of Physical Chemistry C》对主题5、6、8及其他少数主题的贡献更大，而《Physical Chemistry Chemical Physics》涵盖的学科范围则更广。因此，从主题14和18中出现的词语（建模/近似）来看，《Journal of Chemical Physics》似乎更多地关注方法，而非材料特性。物理学研究人员大多只在某一本特定期刊上发表文章，而化学家则可以根据兴趣选择更多期刊。

图13似乎证实了我们从上一幅图中得出的推论。然而，尽管期刊在许多不同主题上发表论文，它们似乎只因少数几个主题而被引用，这或许会影响它们未来的编辑方针。事实上，《Physical Review B》中被引用最多的四个主题，也是该期刊发表论文最多的四个主题。基于这一假设，我们可以预测，截至2019年，主题16和27发表的论文数量将会增加，因为这两个主题的被引次数更多，发表量也更大。需要指出的是，尽管数值已经过归一化处理，少数被引次数极高的文章仍可能使结果产生偏差。

图14与图12相似，但横轴显示的是出版商（而非期刊）。另一个区别是，这幅图仅根据发表于开放获取期刊的文章绘制。值得注意的是，这两幅图所呈现的关注重点各不相同。例如，开放获取期刊似乎特别关注主题31，而图12中的订阅期刊仅发表了少量与这一主题相关的文章。开放获取期刊中另外三个最重要的主题是主题4、8和10；这些主题在付费期刊中也有大量发表，但程度各异。主题4和8的占比有所上升，而在开放获取出版商发表的文献中，主题10、13和14的占比均有所下降，尽管这些主题仍然具有相关性。此外，与付费期刊覆盖许多不同学科不同，开放获取期刊似乎更加聚焦于少数几个主题。我们还可以注意到，主题随时间演变的趋势与此存在相当好的相关性：在开放获取期刊中发表最多的论文，也是在2020年发表最多的论文（见图8）。

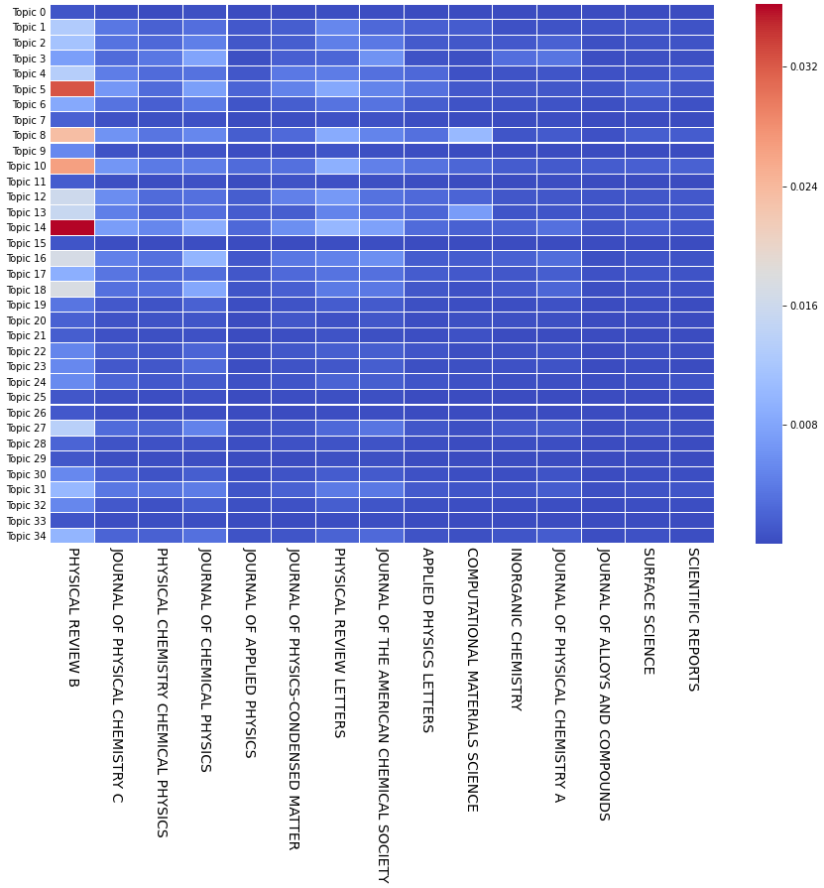

图13：仅考虑开放获取期刊时，每个主题及每种期刊的标准化加权引文数（仅显示前 15 名），适用于 k =35

7 讨论

文献计量学和主题建模是强大的工具，可用于研究和评估任何特定领域的发展及其影响。在本文中，我们围绕密度泛函理论这一特定领域，开展了若干关于主题演变的分析。这里所说的演变，既包括时间维度（跨年份）和空间维度（跨国家），也涉及期刊、出版商、机构和学科；分析依据的是密度泛函理论领域超过12万篇论文的信息。我们发现，根据摘要自动对论文进行分类，是一种有望提升分类效率、加快分类速度并提高客观性的方案。事实上，我们的潜在狄利克雷分配模型得出的35个主题中，大多数

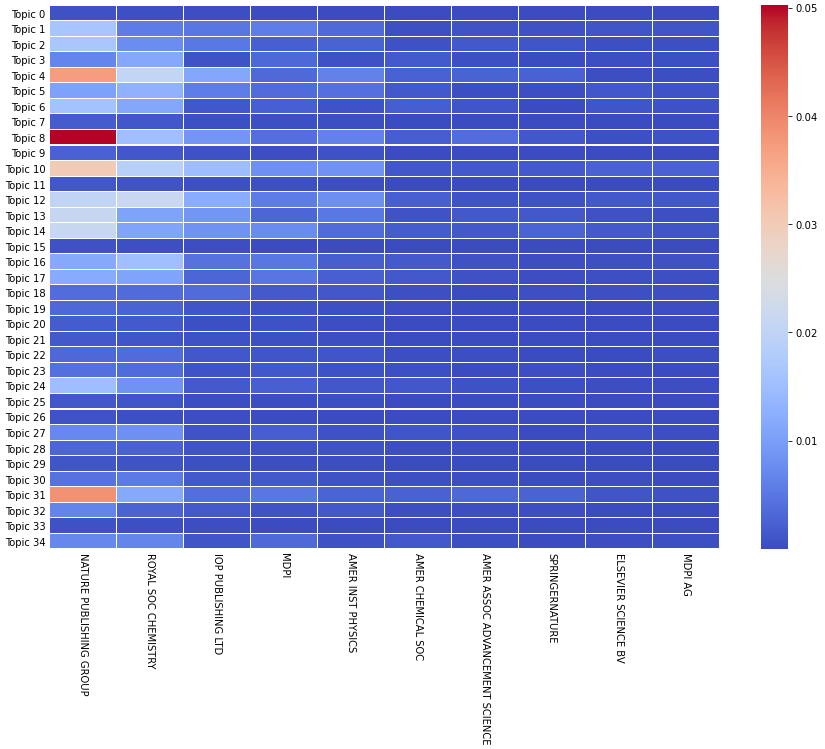

图14：针对 $k = 35$，按主题和出版商统计的归一化加权文献数量（仅显示前10名），仅考虑开放获取期刊。

与我们的文献语料库一致，因此可以为其命名。

尽管大多数软件包是在欧洲开发的，但它们主要在美国和中国使用，这两个国家在密度泛函理论领域的出版物数量上占据主导地位。两国对许多不同学科都抱有共同兴趣，但美国科学家使用许多最重要的密度泛函理论软件包，而中国的研究人员主要使用 VASP，这是一款需授权的软件包。这种差异可能源于各软件内置工具集的不同，以及它们主要服务的子领域不同。事实上，中国似乎将大部分研究重点放在材料科学和应用上，聚焦于二维材料（主题4）、化学反应（主题6）或电池（主题24）；而美国的研究人员发表的研究则大多涉及模拟（主题14）和新方法（主题31）。

此外，我们的分析中出现频率最高的两个主题，也是最受有影响力的期刊和国家关注的主题，都与代码开发、模拟和建模有关。由于我们的研究聚焦于计算软件包，因此一些论文关注相关方法也就不足为奇了

和软件过程。然而，这些学科受到的广泛关注表明，物理学和化学领域近期正大力开发和优化代码包与程序库。

本文的主题建模基于论文摘要完成，并已得出连贯清晰的主题。未来的工作可以将我们的方法应用于数据集中的全部 120,000 篇论文全文。此外，所采用的方法具有很强的通用性，易于调整后应用于密度泛函理论以外的其他研究领域。

补充信息

作者声明不存在利益冲突。

数据可用性

支持本研究结果的数据集可在 Figshare 上获取，标识符为 https://doi.org/10.6084/m9.figshare.12494654.v2。

参考文献

[1] Alberto Aleta 等人：《谨慎探索：绘制物理学科学兴趣的演变图谱》。载于：*EPJ Data Science* 8.1（2019年），第 27 页。[2] Bhagyashree Vyankatrao Barde 和 Anant Madhavrao Bainwad：《主题建模方法与工具概述》。载于：*2017*年国际智能计算与控制系统会议。IEEE。2017年，第 745–750 页。[3] Steven Bird 、 Ewan Klein 和 Edward Loper：《使用 Python 进行自然语言处理：利用自然语言工具包分析文本》。”O’Reilly Media, Inc.”，2009年。[4] David M Blei、Andrew Y Ng 和 Michael I Jordan：《潜在狄利克雷分配》。载于：机器学习研究杂志 3.Jan（2003年），第 993–1022 页。[5] Matteo Chinazzi 等人：《绘制物理学研究空间图谱：一种机器学习方法》。载于：*EPJ Data Science* 8.1（2019年），第 33 页。[6] Marie Dumaz 等人：《固态计算软件包中密度泛函理论的作者身份与引文文化特征》。载于：计量学 126.8（2021年），第 6681–6695 页。[7] Geoffrey EHintonand Sam Roweis：《随机邻居嵌入》。载于：神经信息处理系统进展 15（2002年）。[8] Pierre Hohenberg 和 Walter Kohn：《非均匀电子气》。载于：《物理评论》 136.3B（1964年），B864。

[9] Walter Kohn 和 Lu Jeu Sham：《包含交换和关联效应的自洽方程》。载于：《物理评论》 140.4A（1965年），A1133。[10] Michael Paul 和 Roxana Girju：《研究领域的主题建模：跨学科视角》。载于：*RANLP-2009* 国际会议论文集。2009年，第 337–342 页。[11] Radim ˇReh ◦ uˇrek 和 Petr Sojka：《面向大型语料库的主题建模软件框架》。英文。载于：*2010*年 *LREC* 自然语言处理框架新挑战研讨会论文集。瓦莱塔，马耳他：ELRA，2010年5月，第 45–50 页。[12] Michael R¨oder、Andreas Both 和 Alexander Hinneburg：《探索主题连贯性度量的空间》。载于：第八届美国计算机协会网络搜索与数据挖掘国际会议论文集。2015年，第 399–408 页。[13] Arthur Smith：《从 PACS 到 PhySH》。载于：《自然评论：物理学》 1.1（2019年），第 8–11 页。[14] Ike Vayansky 和 Sathish AP Kumar：《主题建模方法综述》。载于：《信息系统》 94（2020年），第 101582 页。[15] Weizhong Zhao、Wen Zou 和 James J Chen：《用于大型生物医学数据集聚类分析的主题建模》。载于：《*BMC*生物信息学》。第 15 卷。第 11 期。BioMed Central。2014年，第 1–11 页。

作者贡献

M.D. 和 C.R.-B. 从科学网数据库下载了数据，并进行了数据清洗和数据预处理。M.D. 执行了模型训练和测试。M.D.、D.A. 和 A.H.R. 进行了数据分析和解读，制作了数据图表并撰写了正文。所有作者都审阅了手稿。

图

图 1 潜在狄利克雷分配模型的图示 [4]

图 2 （1990—2019年出版物的）文档主题分布

图 3 $k = 35$潜在狄利克雷分配模型中构成各主题的前 10 个词语

图 4 $k = 35$潜在狄利克雷分配模型的 t-SNE 降维结果

图 5 k=35个主题的文档分布。左：1990—2019年语料库。右：2020年语料库

图 6 $k = 35$各主题按年份归一化加权的文献数量。每个图（左上角）中的数字表示拟合效果最佳的功率因子。

图 7 增长率最低和最高的两个主题与最佳拟合指数曲线和幂律曲线的比较。左：幂律因子最低的主题 18。右：幂律因子最高的主题24。

图 8 $k = 35$ 各主题逐年归一化加权文献数量

图 9 $k = 35$ 各主题逐年归一化加权引文数量

图 10 $k = 35$ 各主题在不同时间段的各国归一化加权文献数量（仅显示排名前 15 的国家）

图 11 $k = 35$ 2020年发表的文章中各主题的各国归一化加权文献数量（仅显示排名前 15 的国家）

图 12 $k = 35$ 各主题在不同时间段的各期刊归一化加权文献数量（仅显示排名前 15 的期刊）

图 13 仅考虑开放获取期刊时，每个主题和每种期刊的归一化加权引文数量（仅显示前 15 名），适用于 k =35

图 14 仅考虑开放获取期刊时，每个主题和每家出版商的归一化加权文献数量（仅显示前 10 名），适用于 k= 35。