基于 LDA 模型的专利文本主题分析

——以国内元宇宙领域为例

陆振昇[1,2]，马 超[1]

(1. 深圳信息职业技术学院 素质赋能中心，广东 深圳 518172；2. 湘潭大学 公共管理学院，湖南 湘潭 411105)

摘要：为了探究元宇宙作为新兴产业的热点，解决国内元宇宙领域研究重点不明确的问题，提出使用 LDA 主题模型的专利文本分析方法。将 LDA 主题模型运用到国内元宇宙领域相关专利文本分析中，结合人为判断和主题困惑度的方法，实现了对专利技术主题的识别和划分。通过实验分析得出结论：人工智能、区块链、云计算等是当前中国元宇宙产业应用专利的热点技术；通过 LDA 主题模型分析国内元宇宙的专利文本，可以实现其技术热点主题的分类和细分判别，可以为未来的行业发展提供建议。

关键词：LDA 主题模型；元宇宙；专利文本分析

中图分类号：G255.53 **文献标志码**：A **文章编号**：1671－1807(2023)11－0085－04

2021 年是元宇宙概念火爆全球的一年，被称为“元宇宙元年”。2021 年 3 月 10 日，号称为元宇宙第一股的 Roblox 在美国纽约上市，当日暴涨 50%以上；同年 10 月 28 日，国外社交媒体巨头 Facebook 更名为 Meta，“meta”一词代表元宇宙的“元”；随后微软也宣布进军元宇宙。国内互联网头部公司诸如阿里巴巴、腾讯、字节跳动等也纷纷开始布局元宇宙，金融界开始大量注资拥有元宇宙概念的相关企业，产业界各团体展开了元宇宙这个新赛道的竞争[1]。2022 年，南京信息工程大学、安徽大学和香港理工大学分别开设了元宇宙相关专业，元宇宙自此成为当前社会、政府、产业、学界等争相关注的焦点。

元宇宙是一个大的由虚拟世界和现实世界高度融合的数字空间，包括所有虚拟世界、增强现实和互联网的总和[2]。2021 年底以来，北京、上海、武汉、合肥等多地政府出台了元宇宙和虚拟现实的相关政策文件。2021 年底，上海市政府年度经济会议上便指出“引导企业 研究虚拟世界与现实世界相交互的平台”；2022 年，北京市政府宣布要把通州区打造成元宇宙示范应用区；深圳市也提出在前海建立元宇宙应用试验区。2021 年 3 月发布的《十四五规划和 2035 年远景目标纲要》中提出“加快建设数字经济、数字社会、数字政府，以数字化转型整体驱动生产方式、生活方式和治理方式变革”。2022 年 10 月 28 日，国务院五部门(工业和信息化部、教育部、文化和旅游部、国家广电总局、国家体育总局)联合发布《虚拟现实与行业应用融合发展行动计划(2022—2026 年)》强调应用场景落地[3]。应用落地需要相关的专利技术作为支撑。

1 专利分析相关研究

丁鹏斐[4] 提出了一种基于 LDA (latent dirichlet allocation)模型的中药专利内容热点领域分析方法，并以中药材三七为例，实现了中药专利领域主题细分和热点子领域判断。张世玉等[5]提出在传统技术层面专利组合分析方法的基础上，采用文本挖掘技术，通过技术领域标签抽取、专利文本特征表示、采用文本聚类等流程来对专利文本所属技术领域进行划分。张素娟等[6]使用 LDA 主题模型和聚类标签的方法实现了对西洋参领域专利的主题热度分析。艾楚涵等[7]提出了 LDA 模型与 Kmeans聚类算法结合的方法，对我国转基因玉米育种领域的专利文本进行了分析。伊惠芳等[8]用了融合时间标签的 LDA 主题模型和战略坐标法相结合，将石墨烯领域专利分析以二维的形式展现出来。

收稿日期：2022-12-22

基金项目：广东省普通高校创新团队及特色创新项目(2020KCXTD040、2020KTSCX302)；广东省普通高校特色创新项目(KJ2021C006)；2020 年深圳市教育科学规划课题(SK2020C018)。

作者简介：陆振昇(1996—)，男，湖南长沙人，湘潭大学公共管理学院，硕士研究生，研究方向为信息分析与评价和网络舆情分析；马超(1983—)，男，辽宁葫芦岛人，深圳信息职业技术学院，讲师，博士，研究方向为人工智能与大数据和机器学习。

2 LDA主题模型理论基础

2.1 LDA主题模型

LDA主题模型由David Blei于2003年提出[9]，是一种文档主题生成模型，它包含了三层结构，分别是主题、文档、词，是一个贝叶斯概率模型。LDA模型是一个无监督的机器学习方法，可以用来识别大规模文档集或语料集中的潜在主题信息[10]。同时，LDA采用了词袋模型，通过将每一篇文档视为一个词频向量，文档直接用这些向量集合来表示，并且这个词袋方法没有考虑词与词之间的顺序，降低了计算的复杂度。在LDA模型中每一篇文档代表一些主题所构成的概率分布，在每一个主题中主题又代表了很多单词所构成的一个概率分布[11]。LDA模型的核心是Dirichlet分布，在贝叶斯概率理论中被称为共轭先验分布[12]。

LDA模型的大体思想为：运用先验分布的理念（即先设定一个猜想值去计算）通过不断迭代调整每个文档中每个词汇对应主题的概率分布和每个主题对应文档的概率分布，使最终结果符合实际的文档集中单词对应文档的分布。用数学公式表示为

$$P(w \mid d) = P(w \mid t)P(t \mid d) \tag{1}$$

式中：w为词汇；d为文档；t为主题。

2.2 主题困惑度

由于LDA模型在训练时需要事先设定好主题分类的个数，困惑度的概念是一种用于评价语言模型好坏的指标[13]。使用主题困惑度作为确定最佳主题数的指标，其在LDA模型中计算公式为

$$\text{perplexity}(D) = e^{\frac{-\sum p(d)}{\sum Nd}} \tag{2}$$

$$p(d) = \Sigma \ln\ p(w) \tag{3}$$

$$p(w) = [\Sigma z p(z \mid d)] p(w \mid z) \tag{4}$$

式中：D为整个文档集；$p(w)$为测试集每一个词汇出现的概率；N为测试集所有词集合；z为训练过的主题；d为测试集的每篇文档。

最终计算出来的困惑度代表文档主题的不确定性，因此理论上来说困惑度越小模型性能越好，在困惑度曲线上显示为最低点或拐点处的主题数是最佳主题数[14]。

3 实验与分析

3.1 数据获取与处理

通过对CNKI中国知网的中国专利数据库中的关键词“元宇宙”进行检索，得到364条匹配结果，经过筛选，去除相似重复项和“多元宇宙算法”干扰项，选取其中234条专利的摘要构建文档集。数据获取截止时间为2022年12月19日。进行数据清洗，如“元宇宙”“专利”“申请”“发明”等词出现频率极高但对分析目标没有作用；使用jieba分词对文档集进行中文分词，并建立去停用词库。

3.2 实验结果

基于Python3.8软件的Sklearn库中LDA模型包对处理好的文档集进行主题划分，需要事先人为设置主题划分数，考虑到目前“元宇宙”专利数据集规模比较小，所以将主题数设置成3～5个。经对比发现，区分为3个主题时，主题区分度还不错，但细分技术领域和细分应用领域区分度不高；区分为4个主题时，出现极少数相似主题词分到相邻主题的情况；区分为5个主题时，出现较多相似主题词被分到不同主题的情况。主题数为3、4、5时的主题-主题词分布如表1～表3所示。

将超参数α设置为0.1，β设置为0.01，最大迭代次数设置为50次，得出不同主题数下主题困惑度的变化曲线（图1）。

由图1可知困惑度在主题数为7时出现拐点，最终确定最佳主题数为7。

将主题数设置为7后，得到的主题-主题词的分布情况如表4所示。可以将这7类分别对应其所在的技术领域，分别是“人工智能技术”“区块链技术”“物联网技术”“人机交互技术”“3D建模技术”“扩展现实技术”“云计算技术”。

表1 主题数为3时的主题分布

主题	主题词（部分）
Topic 0	用户 节点 区块链 交易 资产 身份
Topic 1	虚拟现实 场景 模块 设备 平台 终端
Topic 2	传感器 动作 图像 空间 装置 人体

表2 主题数为4时的主题分布

主题	主题词（部分）
Topic 0	交易 数据 区块链 装置 业务
Topic 1	节点 用户 云端 区块链 虚拟现实
Topic 2	场景 模块 空间 传感器 图像
Topic 3	设备 动作 环境 人物 场景

表3 主题数为5时的主题分布

主题	主题词（部分）
Topic 0	交易 资产 区块链 数字化 用户
Topic 1	节点 虚拟现实 云端 身份 体验
Topic 2	传感器 图像 空间 显示屏 客户
Topic 3	设备 目标 人体 轨迹 移动
Topic 4	场景 用户 虚拟空间 环境 预测

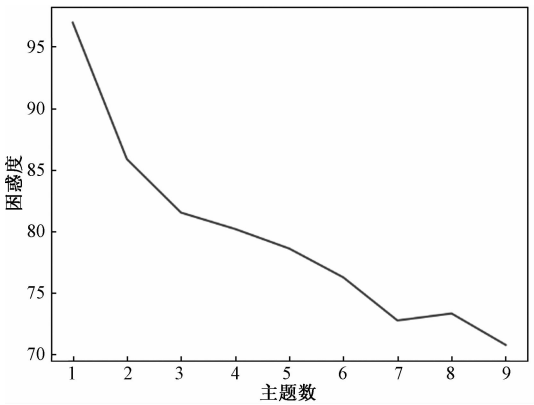

图 1　不同主题数下主题困惑度变化曲线

表 4　主题数为 7 时的主题分布

序号	主题	主题词(部分)
1	人工智能技术	优化 预测 内容 图片 数据
2	云计算技术	数据 平台 服务器 云端 联网
3	物联网技术	传感器 物体 监测 控制器 装置
4	人机交互技术	设备 动作 人体 运动 装置
5	扩展现实技术	场景 虚拟空间 参数 元素 现实
6	3D 建模技术	空间 图像 区域 环境 物理
7	区块链技术	用户 节点 区块链 资产 交易

3.3　实验结果分析

结合 LDA 主题模型训练输出最后的主题分类结果，使用 Excel 软件进行专利主题分类数据统计分析，得到图 2 所示结果。

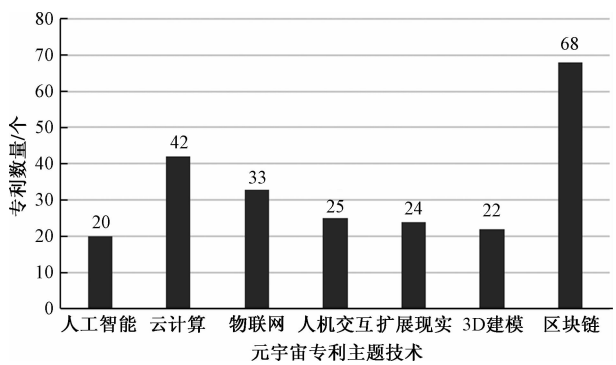

图 2　专利数量分类统计柱状图

由图 2 可得，目前基于区块链技术在元宇宙环境中开发的专利数量占比最大，云计算技术次之。而在 2022 年国务院五部门联合发布的有关促进加强虚拟现实技术与行业产业界融合应用发展计划中重点关注的虚拟现实技术方面的专利数量排名靠后。通过分析可以发现这七类关键技术在元宇宙产业应用中存在发展不平衡的现象，在《“十四五”规划和 2035 远景规划》中“加强数字化发展 建设数字中国”篇章里明确提到了人工智能和扩展现实这两个数字经济重点产业技术，有关部门应当在元宇宙产业的布局上着重发展这两项技术的创新开发和落地应用。截至 2022 年 7 月 20 日，根据全球专利数据供应商 IFI CLAIMS 的情报，过去 5 年拥有元宇宙相关专利的前 10 位公司有微软 158 件、三星 122 件、Magic Leap 109 件、IBM71 件、迪士尼 40 件、Facebook 38 件、Adobe 31 件、Verizon 30 件、英特尔 27 件、Snap 27 件[15]。这些公司的元宇宙相关专利数量总和是国内专利数据库中元宇宙专利的两倍以上，国内的元宇宙产业尚处于起步阶段，在未来的数字化进程中，国内元宇宙产业的应用专利和技术专利具有相当大的发展空间。

4　结语

LDA 主题模型可以应用到元宇宙专利文本数据的主题分类中，实现对元宇宙专利主题领域、技术领域的现状的分析和判断，揭示了热门技术领域和热门产业发展的紧密关联性，为后续元宇宙产业的研究提供了参考意见。经过实证环节总结出以下结论：通过实验分析基于 LDA 主题模型得出专利-主题的具体分布，将中国元宇宙领域相关专利细分成七大技术类别，填补了当前国内元宇宙领域内专利文本分析的空白。

1)通过对当前中国专利数据库中元宇宙相关专利的分析研究，发现以下局限：获取的专利文本数据规模小，无法更深层次地、更广维度地对国内元宇宙专利数据进行挖掘。

2)国内目前在元宇宙产业还处于初期发展阶段，产业界和高校已经陆续着手扩大元宇宙方向的布局，相信在不久的将来，随着生成式 AI(artificial intelligence)技术的蓬勃发展，在此项技术的加持下，中国元宇宙相关专利会在虚拟现实技术领域出现井喷式的增长。

参考文献

[1]　赵星，乔利利，张家榕，等. 元宇宙研究的理论原则与实用场景探讨[J]. 中国图书馆学报，2022，48(6)：6-15.

[2]　郑诚慧. 元宇宙关键技术及与数字孪生的异同[J]. 网络安全技术与应用，2022(9)：124-126.

[3]　周鑫，王海英，柯平，等. 国内外元宇宙研究综述[J]. 现代情报，2022，42(12)：147-159.

[4]　丁鹏斐. 基于主题模型的中药材专利文本挖掘方法研究及应用[D]. 昆明：昆明理工大学，2019.

[5]　张世玉，王伟，于跃，等. 基于文本挖掘技术的技术层面专利组合分析方法优化[J]. 情报理论与实践，2015，38(10)：127-129，144.

[6]　张素娟，宸铁梅，张云倩，等. 基于 LDA 模型的西洋参专

利热点内容及创新趋势分析方法研究[J].情报探索,2019(10):57-62.
[7] 艾楚涵,熊新,吴建德.基于LDA主题模型的专利文本分析应用研究[J].科技和产业,2019,19(3):77-82.
[8] 伊惠芳,吴红,马永新,等.基于LDA和战略坐标的专利技术主题分析——以石墨烯领域为例[J].情报杂志,2018,37(5):97-102.
[9] 范宇,符红光,文奕.基于LDA模型的专利信息聚类技术[J].计算机应用,2013,33(S1):87-89,93.
[10] 王鹏,高铖,陈晓美.基于LDA模型的文本聚类研究[J].情报科学,2015,33(1):63-68.
[11] 李湘东,张娇,袁满.基于LDA模型的科技期刊主题演化研究[J].情报杂志,2014,33(7):115-121.
[12] 孙宁宁.基于主题模型的专利文本分析及应用研究[D].北京:北京工业大学,2017.
[13] 谭春辉,熊梦媛.基于LDA模型的国内外数据挖掘研究热点主题演化对比分析[J].情报科学,2021,39(4):174-185.
[14] 王国睿,张亚飞,尚有为,等.基于LDA主题模型的电子病历热点主题发现[J].中华医学图书情报杂志,2021,30(2):33-39.
[15] 赵颖会.IFI Claim发布元宇宙专利分析报告[J].世界科技研究与发展,2022,44(4):503.

Technical Topic Analysis in Patents Based on LDA:

Taking metaverse in China as an example

LU Zhensheng[1,2], MA Chao[1]

(1. COME Center, Shenzhen Institute of Information Technology, Shenzhen 518172, Guangdong, China;
2. School of Public Administration, Xiangtan University, Xiangtan 411105, Hunan, China)

Abstract: In order to explore the metaverse as a hot spot for emerging industries, and to solve the problem of unclear research focus in the field of domestic metaverse, a patent textanalysis method based on LDA(latent dirichlet allocation) topic model is proposed. By applying LDA topic model to do the patent text analysis of domestic metaverse industry, combined with human judgement and topic perplexity methods, the identification and classification of patent technology topic had been achieved. Through empirical research, some conclusions are as follows: artificial intelligence, blockchain, cloud computing, etc are hotspot technologies of pratical patent of domestic metaverse industry; Analyzing the patent text of domestic metaverse by LDA topic model, which can classify and segment technology hottopics on domestic metaverse industry; also, it can provide suggestions for the future of metaverse industry development in China.

Keywords: LDA(latent dirichlet allocation) topic model; metaverse; analysis in patents