基于机器学习的中国区块链专利技术主题识别与自动分类研究*

胡泽文 王梦雅 韩雅蓉
（南京信息工程大学气象灾害预报预警与评估协同创新中心，南京 210044）

摘要：区块链领域技术主题的自动识别与技术主题范畴的自动分类研究，为拓展领域研发主题和推动领域发展提供情报支持。以德温特专利数据库中的中国区块链技术专利为样本，设计和实现基于机器学习的区块链技术主题识别与自动分类模型，实现基于LDA主题模型的区块链技术主题识别。基于专利文献特征向量空间，形成技术主题范畴的分类体系，最终实现基于传统机器学习和深度学习模型的区块链技术主题自动分类。研究发现：LDA主题模型能够有效识别出区块链技术领域的主题类别，并构建出技术主题类别的特征向量空间，共识别出18个技术主题，按照研究方向归纳为区块链架构研究、区块链行业应用研究、数据存储和数据安全保护研究、高新技术应用研究4类主题范畴；通过交叉融合LDA主题模型、传统机器学习与深度学习等机器学习方法，能够有效实现领域技术主题范畴的自动分类。分类结果显示，支持向量机、LightGBM、LSTM、BP神经网络、逻辑回归模型等分类模型的性能较优，准确率为84%~87%，精确率为79%~83%，其中逻辑回归模型的自动分类效果更显著。

关键词：LDA主题模型；机器学习；区块链；主题识别；自动分类
中图分类号：G250 DOI：10.3772/j.issn.1673-2286.2023.12.004
引文格式：胡泽文，王梦雅，韩雅蓉. 基于机器学习的中国区块链专利技术主题识别与自动分类研究[J]. 数字图书馆论坛，2023（12）：32-43.

区块链技术早期（2008年）作为比特币底层技术框架由中本聪提出[1]。区块链技术又称分布式共享账本技术[2]，是分布式数据存储、点对点传输、共识机制、加密算法等技术集成应用形成的颠覆性创新技术[3]。2019年10月，中共中央政治局就区块链技术发展现状和趋势进行第十八次集体学习，习近平总书记在主持学习时强调，区块链技术的集成应用在新的技术革新和产业变革中起着重要作用，要把区块链作为核心技术自主创新的重要突破口，明确主攻方向，加大投入力度，着力攻克一批关键核心技术，加快推动区块链技术和产业创新发展[4]。2021年，区块链被写入《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》。2023年，工业和信息化部等部门联合印发《关于促进数据安全产业发展的指导意见》，要求布局新兴领域融合创新，加快数据安全技术与区块链等新兴技术的交叉融合创新。国家各个部门纷纷出台相关政策，强调要将区块链技术与各个领域结合起来，以加快推动区块链技术和产业的创新发展，从而不断改善区块链产业的政策环境。

随着区块链技术的不断进步，社会各界逐渐认识到区块链技术的重要性。专利文献是记录科学技术发展的重要成果，专利文献的数量可以反映一个国家、机

收稿日期：2023-10-25
*本研究得到国家社会科学基金项目“面向海量科技文献的潜在‘精品’识别方法与应用研究”（编号：20CTQ031）、江苏高校“青蓝工程”资助。

构、企业和技术领域的发展水平，而专利文献所涉及的内容和主题则能够揭示一个技术领域的发展方向和未来科技发展的趋势。随着大数据和物联网时代的到来，区块链技术得到了快速的发展，同时区块链技术专利文献的数量也大幅增加。在这种情况下，如何科学准确地从大量区块链专利文献中挖掘出领域的技术主题范畴和代表性主题方向，以及如何实现对这些主题范畴的自动分类，成为了全面分析和掌握区块链领域技术主题分布以及发现重点研发主题的关键问题。

1 技术主题识别与自动分类相关研究

有效的专利技术挖掘既可以帮助企业发现技术创新主题和形成技术创新成果，也可以帮助企业识别预测核心专利和增加有价值专利资源储备，从而有效应对技术竞争，同时提高企业自身的专利风险警告和应对能力。目前国内外学者对专利技术挖掘进行了广泛的研究[5]，具体研究主题涵盖：基于专利文献的技术发展机会与技术发展趋势挖掘研究、专利文献领域技术现状与发展趋势的挖掘研究、技术竞争趋势与技术发展路径的专利地图挖掘研究、专利价值的挖掘与评估研究。

技术主题识别与自动分类研究是众多专利技术挖掘研究中的一个方向，国内外学者已经对主题识别与自动分类进行了诸多研究，主题识别方法主要分为基于关键词汇的主题识别、基于传统引文分析的主题识别、基于文本挖掘的主题识别等[6]。①基于关键词汇的主题识别研究主要融合共词分析与社会网络分析方法，实现关键词汇的聚类和主题识别。Liu等[7]采用共词分析、双聚类算法和战略坐标分析等，对金融领域论文进行挖掘，探讨该领域的热点主题。张蕴娣等[8]通过提取图情领域论文的关键词，由杰卡德系数计算文献相似度，形成主题聚类，并计算出关键词在聚类中的重要程度，确定聚类主题内容，进行主题识别。②基于传统引文分析的主题识别研究主要以科技期刊、专利文献等为分析对象，对其引证和被引证情况进行分析，通过构建的引文网络，对文献进行聚类主题识别[9-10]。③基于文本挖掘的主题识别是指从大量的半结构或非结构化文本信息中提取未知、潜在、可理解的知识或数据模式的过程，主要分为主题模型与文本聚类两种方法[11]。吴俊等[12]以36氪网站的区块链新闻为样本，通过结合结构化主题模型与深度学习情感分析技术的新兴产业新闻文本监测方法，发现各个时段区块链技术热点主题。周健等[13]通过主题模型方法直观展示了区块链技术研究主题的演化过程，识别出热点主题。陈虹枢等[14]将主题模型与词嵌入算法等结合起来构建区块链领域主题网络，识别出区块链技术中具有突破性创新特征且最显著的主题是神经网络和边缘计算。Kim等[15]融合文档-词向量构建模型和主题聚类方法，对区块链相关论文的摘要进行主题建模并识别主题。张长鲁等[16]以聚类的方法对中国知网中区块链相关文献进行主题聚类分析，并对主题内涵进行解读。郝雯柯等[17]通过BERT模型和UMAP算法对文本进行语义表示和向量降维，使用HDBSCAN聚类算法进行文本聚类，将新颖度、成长性、影响力达标的主题视为新兴主题。

随着知识产权意识的不断提升，专利数量剧增，如何对不同技术领域的海量专利文献进行细粒度的技术主题分类成为专利检索领域面临的一个问题。专利文献自动分类研究按照分类方法可以分成基于特定规则的分类、基于引证关系的分类、基于文本内容挖掘的分类等[18]。①基于特定规则的分类是指基于文献的特定规则对科技文献进行分类。He等[19]基于关联规则挖掘方法识别类目规则，进而构建自动分类器。②基于引证关系的分类主要以科技文献为分析对象，基于文献的引证与被引证关系构建引文网络关系，进而形成分类体系。彭爱东[20]通过对专利文献进行同被引聚类，对专利进行分类；Chang等[21]基于专利引证关系对专利进行聚类并对类簇涉及技术进行解读，进而构建分类体系。③基于文本内容挖掘的分类隶属于自然语言处理范畴，主要采用机器学习与特征工程相结合的方法，对专利文献进行特征提取，构建专利文献的特征向量空间，进而采用朴素贝叶斯、逻辑回归、支持向量机等机器学习算法对领域文档进行自动分类[22]。贾杉杉等[23]提出多特征多分类器集成的专利自动分类方法，通过提取专利文献中的特征，构建特征向量空间，训练朴素贝叶斯、支持向量机、AdaBoost分类器，实现对专利的分类；Li等[24]提出基于卷积神经网络和词嵌入技术的方法，对专利文献的IPC小类进行分类；吕璐成等[18]针对传统机器学习存在的缺陷，引入CNN和RNN等深度学习模型，构建中文专利文献的特征向量空间和类别体系，实现基于深度学习模型的专利文献分类。

综上所述，已有的主题识别研究主要基于科技文献样本，分别运用主题词共现网络、聚类图谱、共词分析等方法识别领域研究主题。然而专利技术主题的机

器学习分类研究以专利IPC分类号的自动分类研究为主，很少涉及具体技术领域的细粒度主题的自动分类研究。本研究主要基于中国的区块链领域专利文献，融合LDA（Latent Dirichlet Allocation）主题模型、传统机器学习和深度学习模型自动识别出区块链领域的细粒度技术主题，并实现技术主题的自动分类。

本研究的创新之处在于：通过将LDA主题模型与机器学习模型融合，以区块链技术领域的专利文献为样本，解决区块链领域的技术主题识别与自动分类问题。全面展示区块链领域的技术主题范畴，为从业人员选择从业方向、科研人员选择研究方向、技术人员选择技术攻关方向和管理人员制定投资决策提供知识和情报支撑。同时LDA主题识别与区块链技术主题范畴自动分类的融合式研究成果能够为前沿技术领域的主题识别、自动分类、知识元检索、推荐与利用提供新的研究思路和方法。

2 研究思路与方法

2.1 研究思路

基于文本挖掘的方法，提出将LDA主题模型、传统机器学习与深度学习结合，利用TF-IDF（Term Frequency-Inverse Document Frequency）方法表示领域专利文本的主题特征向量空间，形成领域技术主题的分类体系，并通过机器学习模型进行技术主题的自动分类实验，训练出领域技术主题的分类模型，并通过测试集对分类模型进行验证与评价，得出技术主题分类模型的评价结果，以实现区块链领域的技术主题识别与自动分类。研究思路如图1所示。

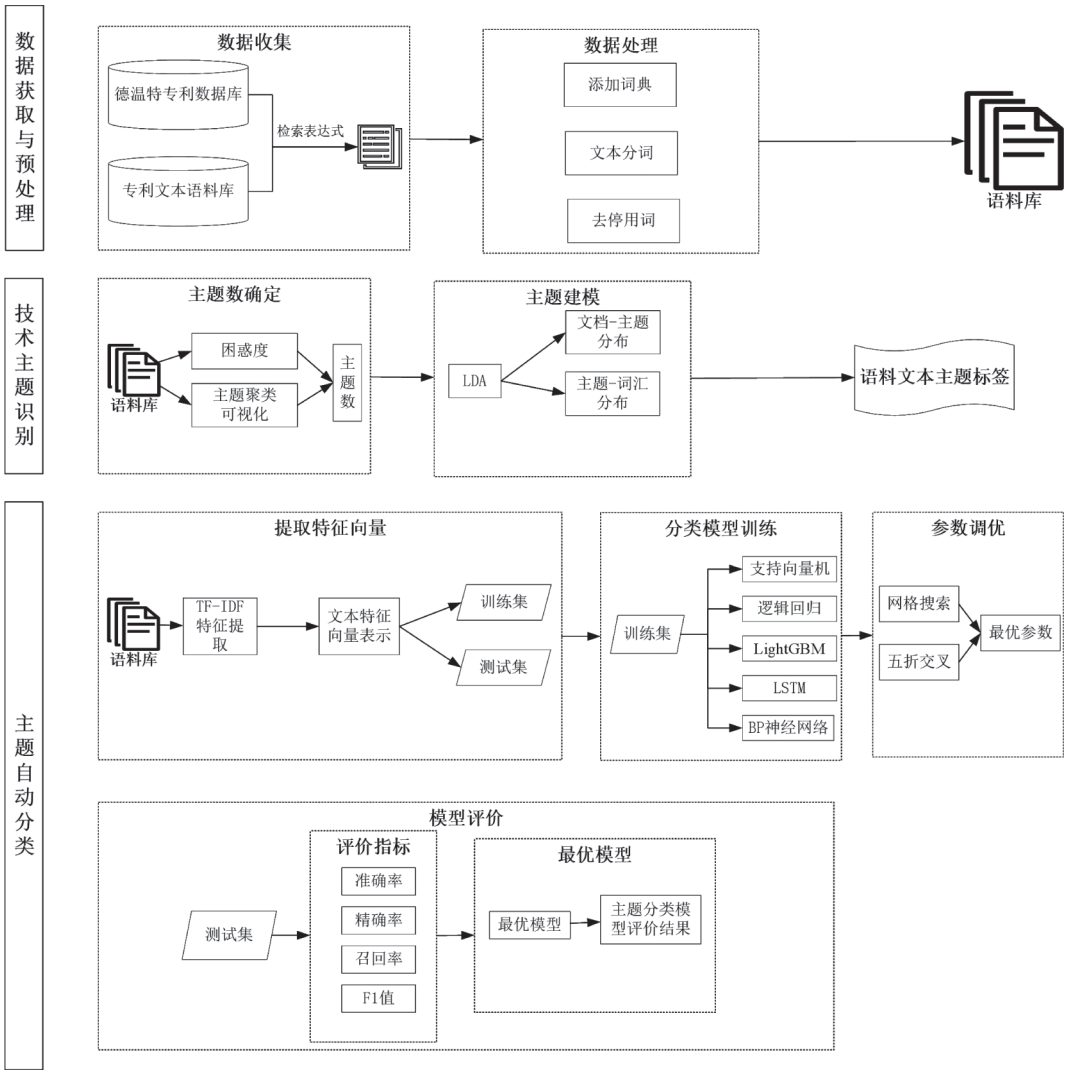

图1 区块链领域的技术主题识别与自动分类流程

首先，进行数据收集，从专利数据库中获取数据，形成区块链专利文献初始语料库，对专利文献标题和摘要进行分词，添加专业词汇形成词典，对分词结果进行停用词处理，形成语料库。其次，在语料库中按照主题困惑度和主题聚类结果，确定区块链技术领域主题数，利用LDA主题模型进行区块链技术主题建模，并对各技术主题的专利文献进行标签标记。最后，通过TF-IDF对区块链专利主题特征进行权重测度和向量化表示，划分训练集和测试集，在训练集中构建技术主题的机器学习分类模型，通过机器学习模型对训练集进行训练，由网格搜索和五折交叉的参数调优方法确定最优参数，通过测试集进行训练结果评优，确定最优模型。

2.2 研究方法

2.2.1 LDA主题模型

LDA是用于文本主题建模的三层贝叶斯概率模型，在2003年由Blei等提出[25]。在LDA主题分析模型中，假设文档主题的先验分布是Dirichlet分布，即每篇文档的主题服从θ_m=Dirichlet（$\boldsymbol{\alpha}$）（$\boldsymbol{\alpha}$是一个k维向量），通过对该分布的计算可得文档中第n个词的主题编号 $z_{m,n}$=multi（θ_m），利用公式$w_{m,n}$=multi（$\varphi_{z_{m,n}}$）得到该词的出现概率，φ表示主题词分布。以上$\boldsymbol{\alpha}\rightarrow\theta_m\rightarrow z_{m,n}$的过程构成了Dirichlet-multi共轭。同时，假设任何一个主题词的先验分布也服从Dirichlet分布，即主题k的主题词服从φ_k=Dirichlet（$\boldsymbol{\beta}$）（$\boldsymbol{\beta}$是一个v维向量）。两个先验分布同时存在，最终得出每个主题下的主题词，过程如图2所示。

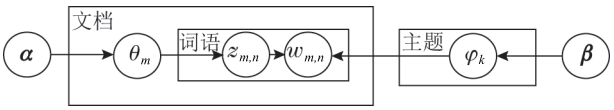

图2 LDA模型运行机制

在使用LDA主题模型进行主题建模时，需要预先确定最终的主题数。为获取最合适的主题数，采用困惑度指标。困惑度常用于评估主题模型的拟合优度，当困惑度较低时，模型的主题结构相对稳定，预期误差值较小。一般情况下，当困惑度下降趋势不再明显或困惑度处于拐点时，对应的数值为最优主题数。困惑度的计算公式如式（1）所示。

$$I_{\text{Perplexity}}=\exp\left\{\frac{-\sum\log[p(w)]}{N_{\text{test}}}\right\} \tag{1}$$

式中：p（w）表示测试集中每个词语的出现概率；N_{test}表示测试集中词语的数目。

2.2.2 文档特征向量权重

TF-IDF[26]是一种广泛使用的算法，用于将文本转化为特征向量，并衡量词语在语料库中的重要性。其中：TF表示词语在当前文档中的频率，即词频；IDF表示逆文档频率，用来衡量一个词语在整个语料库中的重要性。通过计算TF-IDF，可以得出一个词语在当前文档中的重要程度，从而实现对文档的区分。TF-IDF具体计算公式如式（2）~（4）所示。

$$\text{TF}\ (w)=\frac{F_w}{N_{\text{total}}} \tag{2}$$

$$\text{IDF}\ (w)=\log\left(\frac{N_{\text{D}}}{N_{\text{D},\,w}+1}\right) \tag{3}$$

$$\text{TF-IDF}\ (w)=\text{TF}\ (w)\cdot\text{IDF}\ (w) \tag{4}$$

式中：F_w表示单词w在文档中的出现频次；N_{total}表示文档的单词总数；N_{D}表示文档集中的文档总数；$N_{\text{D},\,w}$表示包含单词w的文档数。

2.2.3 机器学习分类模型

选用支持向量机、逻辑回归、LightGBM、LSTM、BP神经网络5种机器学习模型进行多分类实验分析。首先利用已有的分类数据对模型进行训练，并多次迭代优化，得到最佳模型。接着，利用训练后的模型对未知类别的测试数据进行判断、识别和标记。

（1）支持向量机。支持向量机通过开发一个最优超平面来执行分类决策，该超平面需满足两侧最近向量的点间隔最大的条件[27]。该分类器由核函数的方法将输入的样本数据映射到高维特征空间，在该空间寻找使得样本线性分开的最优超平面。函数ϕ实现从低维特征空间到高维特征空间的映射。如果存在函数K，对于任意的低维特征向量$\boldsymbol{\gamma}_1$和$\boldsymbol{\gamma}_2$，都有K（$\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$）=ϕ（$\boldsymbol{\gamma}_1$）·ϕ（$\boldsymbol{\gamma}_2$），则称函数K为核函数，常用的核函数包括线性核、多项式核、高斯核、径向基函数等。径向基函数的核心思想是以某个中心点为基准，根据样本点到该中心点的距离来确定样本点的权重。

（2）逻辑回归。由于传统的逻辑回归算法无法解决多分类问题，采用多元逻辑回归算法。假设数据集$\boldsymbol{D}=\{\boldsymbol{x}_i, y_i\}$（$i$=1，2，…，$M$），$\boldsymbol{x}_i$为一个特征向量，$y_i\in\{1, 2, \cdots, K\}$（$K$>2），其中$M$为训练样本数[28]。该算法模型如式（5）所示，代价函数如式（6）所示。

$$\boldsymbol{h}(\boldsymbol{x}_i)=\begin{bmatrix} P(y_i=1|\boldsymbol{x}_i;\ \boldsymbol{\theta}) \\ P(y_i=2|\boldsymbol{x}_i;\ \boldsymbol{\theta}) \\ \vdots \\ P(y_i=K|\boldsymbol{x}_i;\ \boldsymbol{\theta}) \end{bmatrix}=\frac{1}{\sum_{j=1}^{K}\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}_i)}\begin{bmatrix} \exp(\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{x}_i) \\ \exp(\boldsymbol{\theta}_2^{\mathrm{T}}\boldsymbol{x}_i) \\ \vdots \\ \exp(\boldsymbol{\theta}_K^{\mathrm{T}}\boldsymbol{x}_i) \end{bmatrix} \tag{5}$$

$$J(\boldsymbol{\theta})=-\frac{1}{M}\sum_{i=1}^{K}\sum_{j=1}^{K}\{y_i=j\}\log\frac{\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}_i)}{\sum_{i=1}^{K}\exp(\boldsymbol{\theta}_i^{\mathrm{T}}\boldsymbol{x}_i)} \tag{6}$$

式中：P表示模型的输出概率；$\boldsymbol{\theta}$表示模型的参数向量。

（3）LightGBM。LightGBM是一个基于梯度提升决策树算法的机器学习模型，相较于传统的梯度提升决策树算法，具有更快的训练速度和更高的精度[29]。它的原理是使用损失函数的负梯度来近似当前决策树的残差值，从而拟合新的决策树。为了减少训练数据量，LightGBM在建立决策树时采用按叶生长（Leaf-wise）策略代替按层生长（Level-wise）策略（见图3）。此外，还增加了最大深度的限制，以确保高效且避免过拟合。LightGBM的损失函数如式（7）所示。

$$L(\xi)=\sum_{q=1}^{Q}l[y_q,\ f_{\text{prediction}}(x_q;\ \xi)]+\sum_{k'=1}^{K'}\Omega(f_{\text{prediction},\ k'}) \tag{7}$$

式中：Q表示输入的样本量；l表示样本的损失函数；$f_{\text{prediction}}$表示模型的预测函数；ξ表示模型参数；K'表示树的数量；Ω表示正则化项。

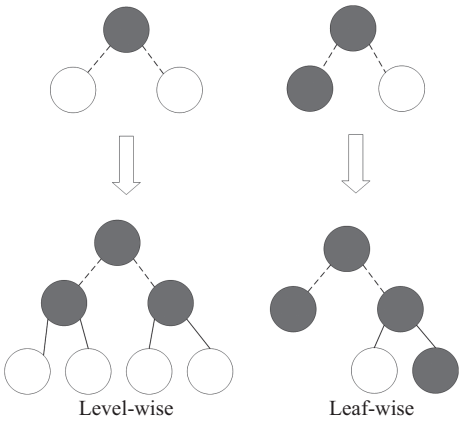

图3　Level-wise策略和Leaf-wise策略

（4）LSTM。LSTM是一种常见的RNN模型，其通过调整信息的输入、输出、遗忘，有效地解决了传统RNN中的梯度消失问题，有助于更好地处理长序列数据。LSTM主要由四部分要素组成：记忆单元、输入门、遗忘门和输出门[30]。具体计算过程如式（8）~（13）所示。

$$f_t=\sigma(W_f[h_{t-1}, x_t]+b_f) \tag{8}$$

$$i_t=\sigma(W_i[h_{t-1}, x_t]+b_i) \tag{9}$$

$$\tilde{c}_t=\tanh(W_c[h_{t-1}, x_t]+b_c) \tag{10}$$

$$c_t=f_t*c_{t-1}+i_t*\tilde{c}_t \tag{11}$$

$$o_t=\sigma(W_o[h_{t-1}, x_t]+b_o) \tag{12}$$

$$h_t=o_t*\tanh(c_t) \tag{13}$$

式中：f_t表示遗忘门的输出；i_t表示输入门的输出；$\tilde{c}_t$表示新的记忆单元状态；c_t表示当前时间的记忆单元；o_t表示输出门的输出；t表示时间；h_t表示记忆单元的状态；x_t表示信息输入；W表示矩阵乘法操作；b表示函数的偏置项；σ表示sigmoid函数；*表示点乘操作；tanh表示tanh函数。

（5）BP神经网络。BP神经网络是一种常用的前馈神经网络模型，由多层感知机和BP算法组成，采用误差逆传播算法进行训练，属于包括隐藏层的多层前馈式网络[31]。BP神经网络有3个层次，分别是输入层、隐藏层和输出层。它能够通过迭代学习来建立并保存大量的输入与输出之间的映射关系，具体网络架构如图4所示。

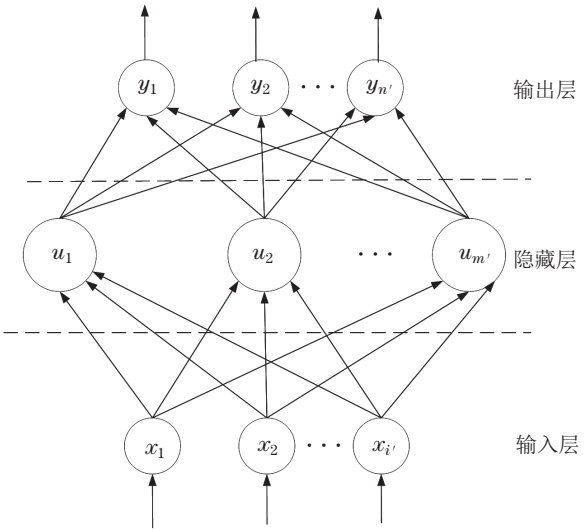

图4　BP神经网络算法

输入层信号可以形式化为$\boldsymbol{X}=(x_1, x_2, \cdots, x_{i'})^{\mathrm{T}}$，其中$i'$为神经元数量；从输入层到隐藏层的信号可以形式化为$\boldsymbol{U}=(u_1, u_2, \cdots, u_{m'})^{\mathrm{T}}$，$m'$为隐藏层的神经元数量，该信号也是输出层的输入信号；输出层的神经元数量为n'，输出结果可以用$\boldsymbol{Y}=(y_1, y_2, \cdots, y_{n'})^{\mathrm{T}}$表示。

2.2.4 实验评价指标

在各二分类评价指标的基础上进一步计算整体综合评价指标，进行机器学习二分类和多分类效果评价。二分类和多分类效果评价指标的具体公式如表1所示。

表1 机器学习的二分类和多分类评价指标

指 标	二分类评价公式	多分类评价公式
准确率	$R_{ACC}=\frac{C_{TP}+C_{TN}}{C_{TP}+C_{FP}+C_{FN}+C_{TN}}$	$R_{macro\_ACC}=\frac{1}{n}\sum_{i=1}^{n}R_{ACC}$
精确率	$R_{precision}=\frac{C_{TP}}{C_{TP}+C_{FP}}$	$R_{macro\_precision}=\frac{1}{n}\sum_{i=1}^{n}R_{precision}$
召回率	$R_{recall}=\frac{C_{TP}}{C_{TP}+C_{FN}}$	$R_{macro\_recall}=\frac{1}{n}\sum_{i=1}^{n}R_{recall}$
F1值	$S_{F1}=\frac{2R_{precision}\cdot R_{recall}}{R_{precision}+R_{recall}}$	$S_{macro\_F1}=\frac{2R_{macro\_precision}\cdot R_{macro\_recall}}{R_{macro\_precision}+R_{macro\_recall}}$

注：C_{TP}表示预测为正向、实际上预测正确的样本量；C_{TN}表示预测为负向、实际上预测正确的样本量；C_{FP}表示预测为正向、实际上预测错误的样本量；C_{FN}表示预测为负向、实际上预测错误的样本量。多分类指标即对二分类指标之和求平均值。

3 实证分析

通过在PyCharm 2020.1软件中配置Python 3.8环境，编写并运行实验程序，实验过程中主要使用的库包括Pandas、Numpy、jieba、Sklearn、Matplotlib、pyLDAvis等。

3.1 数据采集与处理

区块链的英文翻译为blockchain，不同类型区块链的英文翻译全部含有blockchain，因此无需对blockchain进行同义词扩展。为兼顾查全率，需要考虑blockchain的一些变形，最终确定检索式为TAB=（“blockchain”OR“block chain”OR“block? chain”）AND AC=（CN）AND KI=（A* OR B*）AND PY>=（2008）AND PY<=（2022），在德温特专利数据库中进行检索，检索日期为2023年1月，共检索出47 117篇中国专利文献。由于医药和化学领域高分子聚合物的链式结构通常用blockchain表示，与区块链技术并无关联，在专利分类中应予以去除，即排除IPC分类号中的A、C两个部分，并且去除无效记录和数据缺失的专利文献，最终获得中国区块链领域的43 582篇专利文献。由于中国的专利文献包含中文标题和摘要，采用中文分词组件jieba对区块链领域中国专利文献的中文标题和摘要进行分词处理和数据清洗。

在数据处理阶段需要将自定义的区块链相关词汇加入特征词词典，对文档进行自动分词，并通过停用词词典过滤掉分词结果中的“噪声”，从而提升文档分词的准确性。针对区块链领域专利，基于专利文献题名、摘要数据项构建特征词与停用词词典，具体说明如下：由于专利文献没有关键词，通过在万方和中国知网数据库中搜索主题“区块链”，确定相关领域关键词并纳入区块链领域特征词词典。对于停用词词典，最初仅采用哈工大停用词词库，但效果不佳。分析发现，产生这种情况是因为专利文献的形式化描述词汇较多，如“发明”“方法”“包含”“公开”等句首词，这些词汇容易误导机器学习。基于此，停用词词典除包含哈工大停用词词库外，还加入了非相关句首词。经过分词处理将专利文献划分成单个词语，以便提取专利文献词语特征[32]。

3.2 区块链技术主题的LDA模型识别

在LDA主题模型中需要预先确定主题数，通过主题困惑度确定区块链领域的技术主题数。基于式（1）对区块链领域专利文献主题进行困惑度分析，结果如图5所示。可以看出，困惑度在主题数为20、24、28个时达到低值或拐点。

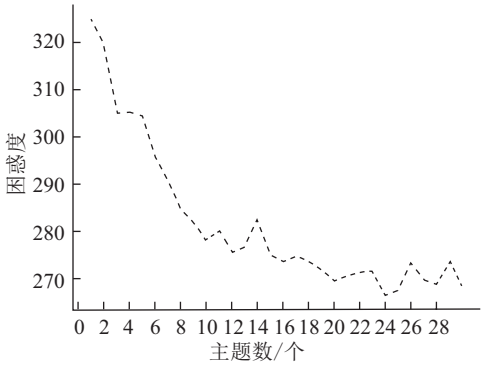

图5 主题困惑度

对主题识别结果使用pyLDAvis库进行聚类和可视化，并比较主题数为20、24、28个时的主题聚类可视化结果，确定最佳主题数。主题聚类结果中，每个圆代表一个主题，圆的大小表示主题相对应的文档数量多少。两圆的距离越远，两个主题的相似性越低，主题区分效果越好。受篇幅限制，只展示当主题数为20个时的主题聚类可视化结果（见图6）。当主题数为20个时，主题重叠度较小，并且主题均匀分布在4个象限，因此选取20作为LDA模型的最优主题数。通过对20个主题、主题词以及主题文献进行人工判读，去除与区块链相关性不强的主题，将具有包含关系的主题进行合并，将相似性较大的主题进行合并，最终识别出18个技术主题，如表2所示。

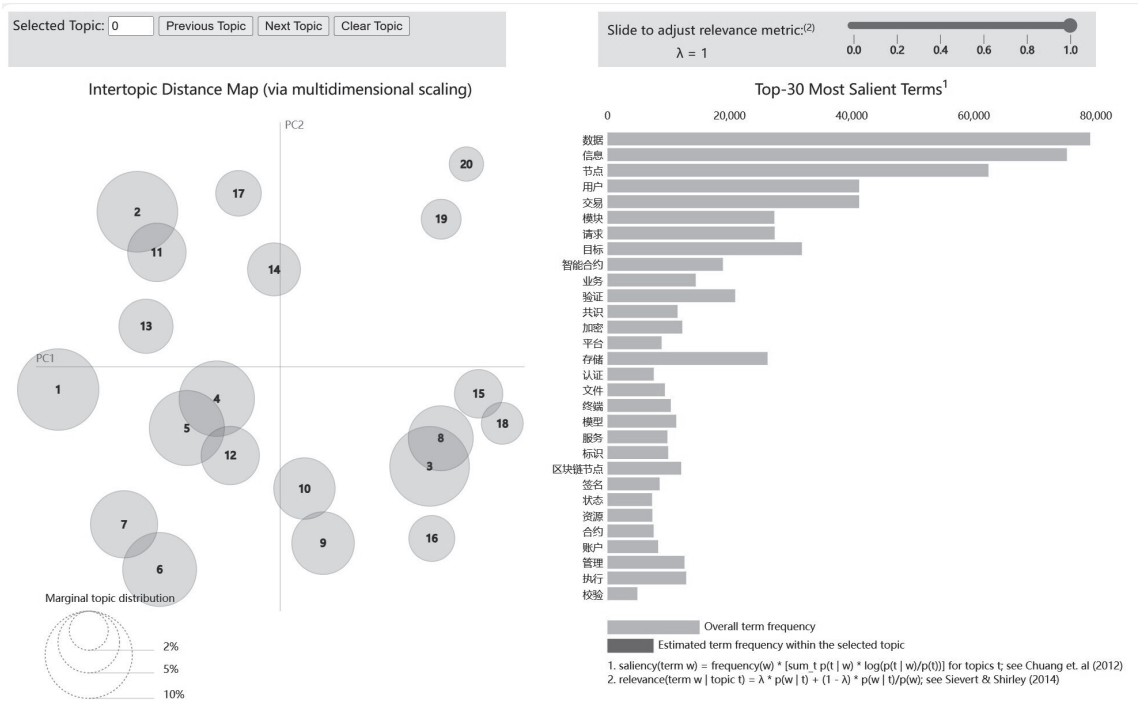

图6 20个主题的聚类可视化结果

表2 区块链领域技术主题及部分主题词

主 题	主题词
Topic1：数据溯源	认证 溯源 区域 结构 合同 接入 身份认证 防伪 中心 二维码
Topic2：车辆互联网	服务 管理 车辆 区块链技术 技术 步骤 联盟链 共享 利用 可信
Topic3：审计	校验 数据共享 金融 审计 费用数据源 捐赠 对账 货币 目录 物资
Topic4：网络交易	交易 签名 账户 地址 验证 交易数据 接收 区块链交易 金额 转账
Topic5：分布式账本	节点 接收 发送 记账 账本 主链 主节点 记账节点 背书 同步
Topic6：信息存储与分发	数据 存储 数据存储 数据处理 物联网 获取 上链 数据库 日志 上传
Topic7：共识机制	共识 资产 连接 监控 固定 设置 终端设备 区块链平台 安装 队列
Topic8：物联网	模块 采集 控制 监测 服务器 通信 物流 管理 终端 传输
Topic9：产品推荐	产品 配置 操作 组件 推荐 类型 页面 执行 展示 功能
Topic10：人工智能	模型 图像 特征 识别 文本 分类 标签 预测 训练 人工智能
Topic11：供应链	商品 客户 生产 钱包 购买 登记 流通 消费者 销售 供应商
Topic12：电子档案	信息 验证 存证 证书 记录 内容 文档 身份验证 数字证书 数字签名
Topic13：业务流程	业务数据 审核 风险 指标 需求 票据 用户端 理赔 电子票据 需求方
Topic14：金融科技	平台 企业 测试 监管 数字资产 银行 资金 发行 数值金融机构 交易平台
Topic15：智能合约	智能合约 支付 订单 协议 商家 确权 收款 以太坊 运行 付款
Topic16：数字身份认证	用户 终端 身份 消息 机构 服务器 数字身份 用户身份 核验 账号
Topic17：数据安全	加密 文件 密钥 私钥 公钥 密文 解密 存储 共享 安全性
Topic18：访问控制	请求 目标 标识 接收 客户端 查询 发送 获取 访问 授权

基于LDA主题模型挖掘出的各个主题，将研究方向归纳如下：①区块链架构研究，具体涵盖分布式账本、共识机制、智能合约、访问控制等；②区块链行业应用研究，具体涵盖审计、网络交易、产品推荐、供应链、电子档案、金融科技、数字身份认证等；③数据存储和数据安全保护研究，具体涵盖数据溯源、信息存储与分发、电子档案、数据安全等；④高新技术应用研究，具体涵盖车辆互联网、物联网、人工智能等。

3.3　区块链技术主题类别的专利文献数量分布

通过LDA主题模型，将区块链技术主题分为数据溯源、车辆互联网、审计、网络交易、分布式账本、信息存储与分发、共识机制、物联网、产品推荐、人工智能、供应链、电子档案、业务流程、金融科技、智能合约、数字身份认证、数据安全、访问控制。区块链技术主题的专利文献数量分布如表3所示。

表3　主题文献分布

主　题	文献数量/篇	主　题	文献数量/篇
数据溯源	518	人工智能	4 270
车辆互联网	5 324	供应链	838
审计	159	电子档案	2 624
网络交易	3 191	业务流程	730
分布式账本	4 092	金融科技	405
信息存储与分发	3 369	智能合约	1 912
共识机制	1 217	数字身份认证	1 455
物联网	3 185	数据安全	2 315
产品推荐	1 606	访问控制	4 190

3.4　区块链技术主题类别的机器学习分类

3.4.1　区块链技术主题类别的特征向量空间构建

每个区块链技术主题类别涵盖一定数量的专利文献，通过对专利文献进行分词和停用词处理，开展LDA主题词提取和TF-IDF权重计算，形成如表4所示的区块链领域技术主题类别的特征向量空间。特征向量空间涵盖每个主题类别、主题词。

表4　主题类别的特征向量空间（样例）

主 题	t_1	t_2	t_3	t_4
s_1	–2.471 4	–2.995 7	–2.995 7	–2.995 7
s_2	6.682 7	–2.995 7	6.868 8	5.370 3
s_3	–2.995 7	–2.995 7	–2.995 7	–2.995 7
s_4	5.679 9	–2.995 7	–2.995 7	–2.995 7

注：s表示主题类别，t表示主题词，数值表示TF-IDF值。

3.4.2　区块链技术主题类别的机器学习训练与测试

基于区块链技术主题类别特征向量空间，采用支持向量机、逻辑回归、LightGBM、LSTM、BP神经网络等机器学习模型来识别区块链技术主题类别。在这一阶段，对样本进行分割，随机抽取80%的样本作为训练集的数据进行学习训练和参数调整，剩下的20%作为测试集的数据进行验证。使用一组标记有技术主题类别的样本集寻找最佳的机器学习模型。在训练集中，使用网格搜索方法在指定范围内调整参数，并使用修正的参数训练模型，寻找所有参数中最精确的参数。对于神经网络模型，经过多次实验，选择最佳神经元数量等参数，参数的最终选择结果见表5。利用基于训练集的机器学习方法参数设定，采用五折交叉验证对各机器学习方法进行评估，计算各评价指标。表6显示了各机器学习算法的准确率、精确率、召回率和F1值。图7展示了各机器学习算法的准确率、精确率、召回率和F1值对比情况。

结合表6和图7的评价结果发现，在训练样本相同的情况下，逻辑回归模型的表现最优，准确率、精确率、召回率、F1值分别为0.87、0.79、0.88和0.82。从准确率来看，逻辑回归模型表现最好，达到0.87，其次是达到0.86的支持向量机模型；从精确率来看，

表5　模型实验参数设置

支持向量机		逻辑回归		LightGBM		LSTM		BP神经网络	
参　数	设　置	参　数	设　置	参　数	设　置	参　数	设　置	参　数	设　置
C	15	class_weight	balanced	boosting_type	gbdt	batch_size	40	learning_rate	0.01
degree	2	max_iter	100	num_leaves	10	num_epoch	50	batch_size	40
gamma	0.1	multi_class	ovr	learning_rate	0.1	optimizer	adam	num_epoch	50
kernel	rbf	solver	newton-cg	n_estimators	500	learning_rate	0.01		

LightGBM模型表现最好，达到0.83，其次为支持向量机模型，达到0.82；从召回率来看，逻辑回归模型表现最好，达到0.88，支持向量机模型仍为表现次优的模型，达到0.82；从F1值来看，逻辑回归模型、支持向量机模型表现较好，达到0.82。尽管LightGBM、支持向量机模型的识别精确率达到83%和82%，但是其分类准确率、召回率和F1值表现相对较差。综合各项评价指标结果，在样本相同的情况下，表现最好的模型是逻辑回归模型，其次是支持向量机模型。因此，设计的多元逻辑回归模型较其他4种模型更适用于区块链技术领域的主题识别与自动分类。

表6 机器学习模型对测试集的识别效果

模 型	准确率	精确率	召回率	F1值
支持向量机	0.86	0.82	0.82	0.82
逻辑回归	0.87	0.79	0.88	0.82
LightGBM	0.85	0.83	0.77	0.80
LSTM	0.84	0.79	0.79	0.79
BP神经网络	0.84	0.81	0.78	0.79

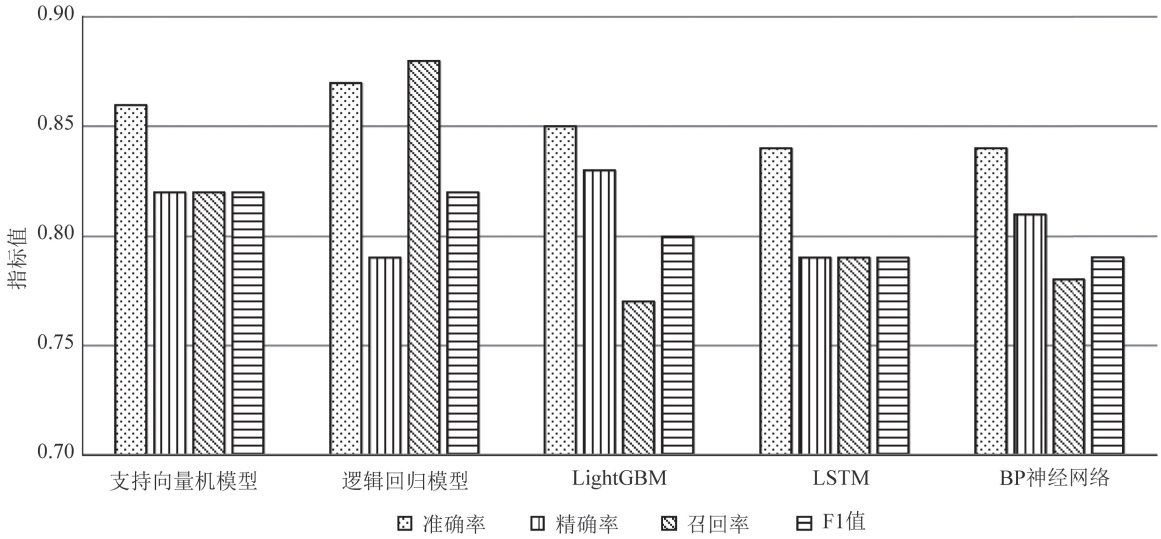

图7 5种模型的分类准确率、精确率、召回率和F1值对比

3.4.3 区块链技术主题类别的机器学习分类结果

采用主题识别和分类效果较好的多元逻辑回归模型，计算参与模型检验的技术主题的出现频次和各评价指标的值，如表7所示。

从多元逻辑回归模型主题分类结果能够看出：①模型的分类效果优，多元逻辑回归模型基于测试集识别各技术主题的效果指标总体达到80%以上。其中网络交易、分布式账本、物联网和人工智能的评价结果显示，各评价指标均高于85%。这表明该模型能够准确识别与这些主题相关的信息，这可能与训练样本中包含大量该主题的数据有关。②从主题覆盖率来看，多元逻辑回归模型将专利文献数据自动分类为不同的主题，分类主题覆盖LDA模型识别的所有主题类别。③不同主题的出现频次不同，因为每个主题的专利文献数量不同，分层抽样后测试集包含的主题数量也不同。为了更好地展示逻辑回归模型在主题识别中的分类效果，

表7 多元逻辑回归模型主题分类精度评价结果

主 题	精确率	召回率	F1值	出现频次/次
数据溯源	0.61	0.91	0.73	97
车辆互联网	0.92	0.80	0.86	1 019
审计	0.31	0.92	0.46	25
网络交易	0.90	0.88	0.89	653
分布式账本	0.91	0.89	0.90	797
信息存储与分发	0.89	0.89	0.89	711
共识机制	0.82	0.89	0.85	239
物联网	0.91	0.89	0.90	660
产品推荐	0.75	0.86	0.80	314
人工智能	0.97	0.86	0.91	881
供应链	0.73	0.91	0.81	190
电子档案	0.87	0.83	0.85	514
业务流程	0.73	0.93	0.82	157
金融科技	0.59	0.91	0.71	77
智能合约	0.80	0.90	0.85	361
数字身份认证	0.78	0.87	0.82	283
数据安全	0.86	0.89	0.88	479
访问控制	0.93	0.82	0.87	823

生成了混淆矩阵热力图（见图8），以获得关于实际正确预测主题的详细信息。热力图的颜色越深表示对应主题的实际出现频次越多。

混淆矩阵热力图显示，车辆互联网和人工智能的专利文献在正确识别数量方面表现出了显著优势。其中，车辆互联网专利文献的正确识别数量为820，而人工智能专利文献的正确识别数量为760，这一结果与测试集中该类主题的文档数量较多有一定关系。具体而言，车辆互联网专利文献主题的正确识别率达到了0.92，而人工智能专利文献主题的正确识别率更高，达到了0.97。尽管分布式账本和访问控制专利文献主题的正确识别率分别达到了0.91和0.93，但主题的表达方式截然不同。从被错误分类的角度来看，物联网专利文献主题主要被错误地归类到车辆互联网中。专利文献主题的误分类主要是模型训练过程中的一些因素导致的，比如参数和训练样本的数量。为了解决模型的过拟合问题，在实验中采用了网格搜索法和五折交叉验证来寻找最佳参数，并将其应用于模型的测试阶段。此外，由于不同技术主题类别的专利文献的特征向量概念语义之间存在一定的交叉重叠情况，同时一篇专利文献可能包含多个主题，比如物联网主题的专利文献中可能包含车辆互联网的内容，专利文献的主题信息复杂多样，分类的难度增加，导致对某些主题的分类精度降低。

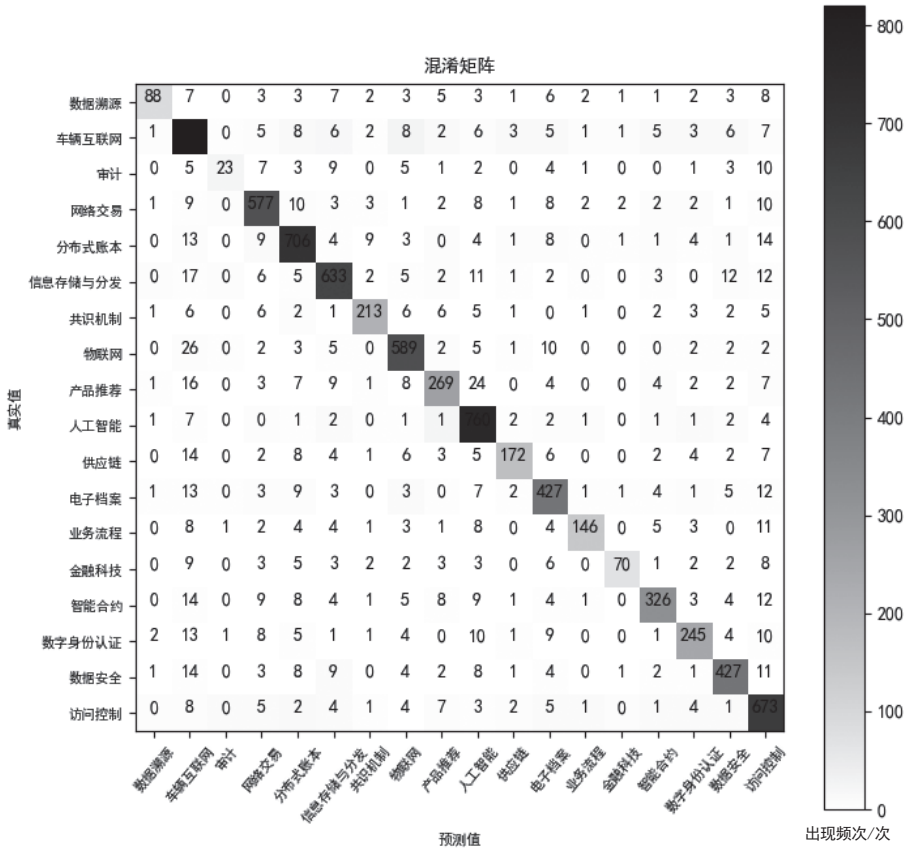

图8　多元逻辑回归模型主题识别混淆矩阵热力图

4　结论

面对领域交叉、技术体系繁杂且体量庞大的专利文献数据，传统的研究主要通过人工方式对专利文献进行分类，存在分类耗时且分类效率较低、容易因理解差异出现错分和误分等问题，并且无法对领域细分技术主题进行挖掘和分类。而通过融合LDA主题模型、传统机器学习和深度学习模型，能够实现技术领域中细分技术主题类别的挖掘和技术主题类别专利文献的自动分类，极大地提升主题识别和自动分类的效率和准确性。

本研究通过对区块链技术领域专利文献的标题和摘要进行中文分词、LDA主题建模和TF-IDF权重测算，有效识别出区块链技术领域的18个技术主题类

别，主题类别的区分度较高，为区块链领域的主题自动分类奠定基础。同时归纳出4类主题范畴（区块链架构研究、区块链行业应用研究、数据存储和数据安全保护研究、高新技术应用研究），为后续研究人员和从业者选择研究或攻关方向提供借鉴。为实现区块链领域细粒度主题的自动分类，本研究厘清了区块链领域技术主题类别的专利文献数量范围，进而构建区块链技术主题类别的特征向量空间。将机器学习算法与LDA主题模型交叉融合，构建了区块链技术主题类别的识别与自动分类模型，实现区块链领域技术主题类别专利文献的自动化分类。分类结果显示，支持向量机、LightGBM、LSTM、BP神经网络、逻辑回归模型等分类模型的表现性能较优，准确率处于84%~87%范围内，精确率处于79%~83%范围内，其中逻辑回归模型的自动分类效果更显著。

参考文献

[1] NAKAMOTO S. Bitcoin：a peer-to-peer electronic cash system[J]. Decentralized Business Review，2008：21260.

[2] SWAN M. Blockchain：Blueprint for a New Economy[M]. New York：O’Reilly Media，2015.

[3] 中国区块链技术和产业发展论坛. 中国区块链技术和应用发展白皮书2016[R]. 北京：工业和信息化部，2016：1-10.

[4] 叶蓁蓁. 我国必须走在区块链发展前列[EB/OL]. [2023-12-10]. http://politics.people.com.cn/n1/2019/1026/c1001-31421642.html.

[5] 林静，连晓振，侯亮. 面向主控式创新的高价值专利挖掘研究[J]. 情报杂志，2022，41（6）：164-172，163.

[6] 柴文越，刘小平，梁爽. 新兴主题识别方法研究综述[J/OL]. 现代情报：1-14[2023-11-14]. http://kns.cnki.net/kcms/detail/22.1182.G3.20231110.0908.002.html.

[7] LIU Y M，ZHANG S，CHEN M，et al. The sustainable development of financial topic detection and trend prediction by data mining[J]. Sustainability，2021，13（14）：7585.

[8] 张蕴娣，于宁，赵闯. 国内图情领域区块链研究热点与展望[J]. 情报科学，2022，40（10）：187-192.

[9] 罗双玲，张文琪，夏昊翔. 基于半积累引文网络社区发现的学科领域主题演化分析：以“合作演化”领域为例[J]. 情报学报，2017，36（1）：100-110.

[10] SHIBATA N，KAJIKAWA Y，TAKEDA Y，et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications[J]. Technovation，2008，28（11）：758-775.

[11] 李尚昊，朝乐门. 文本挖掘在中文信息分析中的应用研究述评[J]. 情报科学，2016，34（8）：153-159.

[12] 吴俊，邵丹睿，姜尚杨帆. 融合语义与情感分析的区块链产业新闻监测研究[J]. 现代情报，2020，40（11）：22-33.

[13] 周健，张杰，屈冉，等. 基于LDA的国内外区块链主题挖掘与演化分析[J]. 情报杂志，2021，40（9）：161-169.

[14] 陈虹枢，宋亚慧，金茜茜，等. 动态主题网络视角下的突破性创新主题识别：以区块链领域为例[J]. 图书情报工作，2022，66（10）：45-58.

[15] KIM S，PARK H，LEE J. Word2vec-based latent semantic analysis （W2V-LSA） for topic modeling：a study on blockchain technology trend analysis[J]. Expert Systems with Applications，2020，152：113401.

[16] 张长鲁，张健. 国内区块链研究主题挖掘、热点分析及趋势探究[J]. 统计与信息论坛，2021，36（2）：119-128.

[17] 郝雯柯，杨建林. 基于语义表示和动态主题模型的社科领域新兴主题预测研究[J]. 情报理论与实践，2023，46（2）：184-193.

[18] 吕璐成，韩涛，周健，等. 基于深度学习的中文专利自动分类方法研究[J]. 图书情报工作，2020，64（10）：75-85.

[19] HE C，LOH H T. Pattern-oriented associative rule-based patent classification[J]. Expert Systems with Applications，2010，37（3）：2395-2404.

[20] 彭爱东. 基于同被引分析的专利分类方法及相关问题探讨[J]. 情报科学，2008，26（11）：1676-1679，1684.

[21] CHANG S B，LAI K K，CHANG S M. Exploring technology diffusion and classification of business methods：using the patent citation network[J]. Technological Forecasting and Social Change，2009，76（1）：107-117.

[22] 李程雄，丁月华，文贵华. SVM-KNN组合改进算法在专利文本分类中的应用[J]. 计算机工程与应用，2006，42（20）：193-195，212.

[23] 贾杉杉，刘畅，孙连英，等. 基于多特征多分类器集成的专利自动分类研究[J]. 数据分析与知识发现，2017，1（8）：76-84.

[24] LI S B，HU J，CUI Y X，et al. DeepPatent：patent classification with convolutional neural networks and word embedding[J]. Scientometrics，2018，117（2）：721-744.

[25] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003，3（1）：993-1022.

[26] 施聪莺，徐朝军，杨晓江. TFIDF算法研究综述[J]. 计算机应

用，2009，29（S1）：167-170，180.

[27] HAPSARI D P，UTOYO I，PURNAMI S W. Text categorization with fractional gradient descent support vector machine[J]. Journal of Physics：Conference Series，2020，1477（2）：022038.

[28] BRADLEY R，MACLAREN W M. Ordinal logistic regression analysis of flight task ratings[J]. The Aeronautical Journal，2006，110（1109）：447-456.

[29] KE G，MENG Q，FINLEY T，et al. Lightgbm：a highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems，2017：3149-3157.

[30] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9（8）：1735-1780.

[31] MCCLELLAND J L，RUMELHART D E，PDP Research Group. Parallel Distributed Processing，Volume 2：Psychological and Biological Models[M]. Cambridge：The MIT Press，1987.

[32] 雷兵，刘小，钟镇. 基于题录信息的领域学术文献细粒度分类方法研究[J]. 图书情报工作，2021，65（14）：128-137.

作者简介

胡泽文，男，博士，副教授，博士生导师，研究方向：数据科学与知识工程，E-mail：huzewen915@163.com。
王梦雅，女，硕士研究生，研究方向：数据科学与知识工程。
韩雅蓉，女，硕士研究生，研究方向：数据科学与知识工程。

Topic Recognition and Automatic Classification of Chinese Blockchain Patent Technology Based on Machine Learning

HU ZeWen　WANG MengYa　HAN YaRong
(Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science & Technology, Nanjing 210044, P. R. China)

Abstract: The automatic recognition of technology topics in the field of blockchain and the automatic classification of technology topic categories provide intelligence support for expanding research and development topics in the field and promoting the development of the field. This paper takes the Chinese blockchain technology patents in the Derwent patent database as samples, designs and implements the blockchain technology topic recognition and automatic classification model based on machine learning, and realizes the blockchain technology topic recognition based on the LDA topic model. Based on the characteristic vector space of patent literature, a classification system for technology topic categories is formed, ultimately achieving automatic classification of blockchain technology topics based on traditional machine learning and deep learning models. The results show that the LDA topic model can effectively identify the topic categories in the blockchain technology field, and construct the characteristic vector space of the technology topic categories. 18 technology topics are identified, which can be summarized as four topic categories according to the research direction: blockchain architecture research, blockchain industry application research, data storage and data security protection research, and high-tech application research. Through the cross-fusion of LDA topic model, traditional machine learning and deep learning, and other machine learning methods, we can effectively realize the automatic classification of technology topic categories in the domain. The classification results show that the performance of classification models such as support vector machine, LightGBM, LSTM, BP neural network, and logistic regression model is better. The accuracy rate is 84%−87%, and the precision rate is 79%−83%, among which the automatic classification effect of logistic regression model is more significant.

Keywords: LDA Topic Model; Machine Learning; Blockchain; Topic Recognition; Automatic Classification

（责任编辑：王玮）