基于数据挖掘的图书馆信息分类检索算法与系统性能评估

黄海涛[1]
HUANG Haitao

摘 要 针对图书馆信息资源结构复杂、来源多样，传统检索方法难以应对语义异构与排序精准性不足等问题，文章提出一种基于数据挖掘的分类检索算法，构建双通道结构，实现结构字段与文本语义的联合建模；引入动态聚类机制，解决主题边界模糊与冷启动样本难聚类问题；设计多因子排序函数，结合语义扩展、行为偏好与置信度重构排序路径。系统基于 310 742 条图书管理与用户行为数据进行验证，评估指标包括 Accuracy、Macro-F_1、NDCG@10 等。研究构建“分类 - 聚类 - 排序”一体化技术框架，为图书馆智能检索系统提供方法支撑。

关键词 图书馆信息检索；数据挖掘；双通道模型；动态聚类；多因子排序；系统性能评估

doi：10.3969/j.issn.1672-9528.2026.02.037

0 引言

在数字资源快速增长与用户需求多样化背景下，传统图书馆检索系统在数据异构、分类粒度与排序机制方面存在明显局限。尽管图书馆管理系统积累了大量流通数据，但受结构标准、时间粒度与语义表达不一致影响，难以在自动分类与智能检索中高效利用，亟需借助数据挖掘重构信息组织与访问逻辑。

吴小凤[1]融合图卷积与注意力机制，将图书内容与用户行为深度耦合，拓展了异构特征融合思路；杨月[2]构建多通道语义卷积模型，兼顾分类精度与响应效率；康丽丽等[3]引入冷启动与时间窗口机制，缓解用户兴趣波动影响；王欣玥[4]从分类体系出发，揭示标签规范性与语义对齐不足，启示了结构优化路径。

本文针对图书馆数据的结构异构与语义模糊问题，提出融合双通道分类、动态聚类与多因子排序的复合检索框架，通过主键对齐、语义向量聚合与行为权重调控机制，实现信息统一表达与检索加速，助力分类检索系统的智能化演进。

1 数据预处理与特征工程

1.1 多源异构数据融合

图书馆信息系统中，多源数据在结构、语义与时间维度存在异构性，需统一映射为模型输入。结构化数据基于 CN-MARC 编码（如“200$a”表题名、“210$c”表出版者），经 YAML 转为扁平 JSON，压缩为 24 个主键与 12 个辅助字段，经 Apache NiFi 每 5 s 写入 HBase，主键为“UserID+BookID”。非结构化文本（如书评、摘录）由 Tesseract OCR 提取，置信度阈值 0.88，误差 ≤3 字，清洗后输入 fastText CBOW 模型，生成 128 维向量，词频阈值 10，训练 15 轮，单文件 ≤120 MB。日志与借阅记录时间粒度分别为 ms 和 s，采用滑动窗口对齐（宽度 20 s，步长 5 s），匹配规则用公式表示为：

$$\Delta T = \min\left\{\left|t_i^{(A)} - t_j^{(B)}\right| \left|t_i^{(A)} - t_j^{(B)}\right| \le \delta\right\} \tag{1}$$

式中：$t_i^{(A)}$ 为操作日志的时间戳；$t_j^{(B)}$ 为借阅记录的时间戳；δ=20 s 为滑动窗口容差阈值；ΔT 为匹配差值。

匹配结果生成主键“UID_OPID_TID”，以 Parquet 格式存储，UTF-8 编码，空值以“NULL”表示，结果包含结构向量、语义向量与时间索引。

1.2 特征增强与降维

多源融合后形成的 286 维高维向量涵盖结构、语义与行为特征，存在冗余、量纲不统一与偏态分布问题。为提升表征能力，采用归一化、语义压缩与线性降维三阶段处理。对借阅频率与出版年限等数值特征，使用 *Z*-score 标准化统一尺度：

$$\boldsymbol{x}_i^* = \frac{\boldsymbol{x}_i - \mu}{\sigma} \tag{2}$$

式中：$\boldsymbol{x}_i$ 为原始特征；μ 与 σ 分别为训练集中该特征的均值与标准差。

语义特征由 fastText 提取 128 维嵌入，经堆叠自编码器压缩至 32 维，网络结构为 128-64-32，激活函数为 LeakyReLU，损失函数为均方误差。模型运行于 NVIDIA RTX A4000

1. 安徽省图书馆 安徽合肥 230001

（16 GB）平台，batch size 为 256，训练耗时 17.6 min。

主成分分析（PCA）保留前 37 个主成分，累计方差解释率超过 95%。协方差矩阵在 cuML 环境中以 tile 模式并行分解，块尺寸为 32×32。输出特征矩阵维度为 N×37，满足压缩性、抗噪性与可分性要求。特征降维前后分布如图 1 所示。

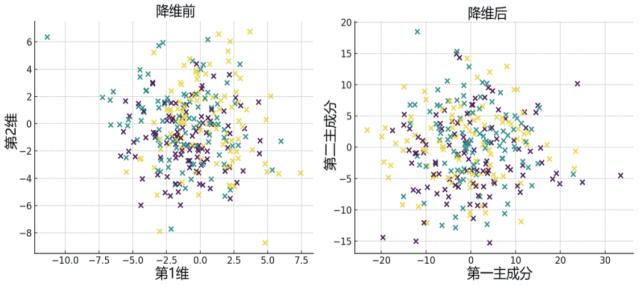

图 1 特征降维前后分布对比图

2 分类检索算法设计

2.1 双通道分类模型

双通道模型分别处理结构特征与语义向量，提升对多源异构数据的判别能力。结构通道输入为 37 维降维向量，包含出版年、借阅频率、分类号等字段 [5]。采用两层全连接网络，第一层 64 节点，激活函数为 LeakyReLU，引入 BatchNorm 与 Dropout（p=0.3）以增强泛化能力，输出 32 维表示 h_s，计算公式为：

$$\boldsymbol{h}_s = \mathrm{Dropout}(\mathrm{BN}(\mathrm{ReLU}(\boldsymbol{W}_s \boldsymbol{x}_s + \boldsymbol{b}_s))) \tag{3}$$

式中：$\boldsymbol{x}_s \in \mathbb{R}^{37}$ 为结构输入；$\boldsymbol{W}_s \in \mathbb{R}^{64\times37}$，$\boldsymbol{b}_s \in \mathbb{R}^{64}$，最终输出 $\boldsymbol{h}_s \in \mathbb{R}^{32}$ 通过 FC 压缩得到。

语义通道输入为自编码器压缩后的32维向量，采用卷积–池化–全连接结构提取上下文特征。卷积核大小为 3，通道数 16，最大池化核为 2，Flatten 后连接至 128 维 FC，再压缩至 32 维输出 $\boldsymbol{h}_t$：

$$\boldsymbol{h}_t = \mathrm{FC}(\mathrm{Flatten}(\mathrm{MaxPool}(\mathrm{Conv1D}(\boldsymbol{x}_t)))) \tag{4}$$

式中：$\boldsymbol{x}_t \in \mathbb{R}^{32\times1}$ 为输入语义矩阵；$\boldsymbol{h}_t \in \mathbb{R}^{32}$ 为文本通道最终输出向量。拼接 $\boldsymbol{h}_s$ 与 $\boldsymbol{h}_t$，生成 64 维联合向量 $\boldsymbol{h}$=[$\boldsymbol{h}_s$; $\boldsymbol{h}_t$]。输入分类器（结构为 64-32-C）进行预测，损失函数为交叉熵：

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{5}$$

式中：y_i 为真实标签的独热编码；$\hat{y}_i$ 为 Softmax 输出预测值 [6-7]。模型运行于 NVIDIA RTX A4000，参数量约 4.3×10^6，batch size 为 64，最大训练轮数 100，优化器为 AdamW，初始学习率 0.001。

2.2 动态主题聚类增强

为提升相似性分辨能力，引入结构特征与文本语义向量的加权联合度量函数 $S(i,j)$，用公式表示为：

$$S(i,j) = \alpha \cdot \cos(\boldsymbol{v}_i^t, \boldsymbol{v}_j^t) + (1-\alpha)\cdot\left(1 - \frac{\|\boldsymbol{v}_i^s - \boldsymbol{v}_j^s\|_2}{\max_k \|\boldsymbol{v}_i^s - \boldsymbol{v}_k^s\|_2}\right) \tag{6}$$

式中：$\boldsymbol{v}_i^t$表示样本 i 的文本语义向量；$\boldsymbol{v}_i^s$表示结构特征向量；$\cos(\cdot)$ 表示余弦相似度；$\|\cdot\|_2$ 表示欧氏距离；α 表示融合系数，设为 0.65。

归一化项用于统一不同特征域的量纲与尺度，提升融合稳定性。聚类阶段采用凝聚型 AHC 算法，初始每条样本独立成簇，基于 $S(i,j)$ 构造相似度矩阵，按最大相似度依次合并 [8]。设最小类簇容量为 20、最大层级为 5、相似度阈值为 0.68，输出嵌套聚类树结构。高密度区域保留子簇细分，低密度区域则聚合简化，支持多粒度索引。为解决新入库图书或低频行为数据冷启动难聚类问题，引入语义锚点映射生成软标签。设未聚类样本 $\boldsymbol{x}_u$ 的词向量为 $\boldsymbol{v}_u$，从已聚类样本中筛选满足 $\cos(\boldsymbol{v}_u, \boldsymbol{v}_k) > 0.85$ 的锚点集合$\mathcal{N}(\boldsymbol{x}_u)$，其所属类别分布 $\{c_k\}$ 构成初始概率标签分布：

$$P(c/\boldsymbol{x}_u) = \frac{1}{|\mathcal{N}(\boldsymbol{x}_u)|}\sum_{\boldsymbol{x}_k\in\mathcal{N}(\boldsymbol{x}_u)} \mathbb{I}(c_k = c) \tag{7}$$

式中：$\mathbb{I}(\cdot)$为示性函数，表示锚点样本对类别的投票。该机制融合结构与语义双域信息，通过层次聚类与软标签估计应对冷启动，适配语义动态演化环境。

2.3 关联规则驱动的排序

基于 Apriori 算法构建关键词–主题共现矩阵，从历史查询记录中提取频繁项集。设最小支持度 σ=0.012，置信度阈值 θ=0.75。查询项 q 与候选主题词集合 T={t_1,t_2,⋯,t_n} 建立扩展关系：

$$\mathrm{conf}(q \Rightarrow t_k) = \frac{\mathrm{freq}(q,t_k)}{\mathrm{freq}(q)} > \theta \tag{8}$$

式中：freq(q, t_k) 为查询项与主题词的共现频次；freq(q) 为查询项出现次数。满足条件的 t_k 作为扩展词纳入排序空间，提升语义覆盖能力 [9]。排序得分函数融合行为评分 S_b、语义匹配度 S_s、元数据置信度 S_m，加权用公式表示为：

$$S(u,d) = \lambda_b \cdot S_b(u,d) + \lambda_s \cdot S_s(q,d) + \lambda_m \cdot S_m(d) \tag{9}$$

式中：u 为用户；d 为目标文献。$S_b(u,d)$ 基于邻域协同过滤模型提取的评分值，历史行为矩阵通过 Spark MLlib 实现，截取近 1 000 条行为样本构建相似度向量；$S_s(q,d)$ 采用双向编码器表示的 BERT 模型提取查询 q 与文献 d 摘要的向量，使用余弦相似度计算；$S_m(d)$ 聚合出版年、分类等级与图书馆馆藏量，规范化至 [0,1] 区间内，反映文献可靠性 [10]。权重 λ 通过交叉验证设定为 [0.4, 0.35, 0.25]。后端集成 Faiss，配置 IVF-PQ 索引结构，聚类中心 512，压缩维度 16，文献向量为 128 维 float32，索引容量控制在 600 MB 内。系统支持 8 线程并行检索，Top-K=10，单次查询平均延迟低于 25 ms。该机制实现查询扩展与多因子排序向量融合的端到端优化，显著提升多特征场景下的响应效率与检索准确性。

3 系统性能评估

3.1 数据集与预处理

构建统一格式的异构图书信息数据集，涵盖结构化借阅记录、用户日志及非结构化文本与图像，共计 310 742 条样本。结构字段由 Interlib 系统导出，按 YAML 规则映射为 JSON，压缩为 24 个主键字段，通过 Apache NiFi 以 5 s 周期写入 HBase。文本信息（书评、摘录）经 Tesseract OCR（v5.3.1）识别，置信度阈值设为 0.88，字符误差控制在 3 字内，归一化后输入 fastText 模型，嵌入维度 128，词频阈值 10，训练 15 轮。日志时间粒度为 ms，借阅记录为 s，采用滑动窗口机制（Δt=20s，步长 5 s）完成对齐。全部数据封装为 Parquet 格式，UTF-8 编码，包含结构向量、语义向量及时间索引，作为分类检索模型的标准输入。

3.2 评估指标

为全面评估分类检索系统的性能，设置五项核心指标，涵盖分类精度、排序结构与响应效率三个维度。分类性能采用“准确率（Accuracy）”和“宏平均 F_1 值”衡量，其中准确率反映模型整体判断的正确性，宏平均 F_1 综合各类精度与召回率，适用于多分类场景。排序效果通过“NDCG@k”指标度量，反映系统推荐结果与理想排序的一致程度。检索效率则以“平均延迟（Latency）”衡量，计算多轮查询的平均响应时间。另设“召回率（Recall）”评估模型发现正样本的能力，兼顾样本不均衡问题。实验在 Python 3.10 环境下进行，使用 Scikit-learn 与 RankEval 工具，硬件平台为 NVIDIA RTX A4000（16 GB）与 Intel Xeon Gold 5218（2.3 GHz，16 核），统一采用五折交叉验证，确保评估过程一致性与结果可靠性。

3.3 对比实验设计

为验证各功能模块的结构有效性与设计合理性，设置三组控制变量实验，分别考察分类结构、聚类策略与排序机制对系统性能的影响。实验基于 310 742 条样本，按 8∶2 划分训练与测试集，采用五折交叉验证取平均结果，确保样本分布一致。测试平台配置为 NVIDIA RTX A4000（16 GB）与 Intel Xeon Gold 5218，运行环境为 Ubuntu 22.04，训练框架为 PyTorch 2.1.1，batch size 为 64，优化器采用 AdamW，学习率 0.001，最大轮次 100。为确保对比有效性，实验统一数据编码、输入向量与评估指标，仅调整目标变量。各组实验设置如表 1 所示。

3.4 结果分析

为验证分类检索系统在真实应用场景下的性能表现，设计三组对比实验，分别考察分类准确性、排序质量与系统效率。测试集包含 2.4 万条结构化元数据和 3.6 万条文本描述，采用五折交叉验证获取平均值。对比模型包括单通道分类器、K-means 聚类结构及 TF-IDF 排序算法，实验平台配置为 RTX 3090（24 GB）、Intel Xeon W-2255（3.7 GHz）与 64 GB 内存。各组关键指标统计如表 2 所示。

实验结果表明，双通道结构在 Accuracy 与 Macro-F_1 上表现突出，验证了结构与语义特征融合的有效性。动态聚类增强显著提升主题排序指标，改善候选覆盖。多因子排序结合 Faiss 索引大幅降低延迟，提升系统吞吐与稳定性。整体上，本文方法在分类、聚类与排序三个环节实现协同优化。

4 结论

本文提出面向图书馆异构信息资源的分类检索系统，构建融合双通道特征建模、动态主题聚类与多因子排序的算法框架。系统通过结构字段与文本语义联合建模，解决了语义割裂与特征冗余问题；引入混合相似度与软标签引导的聚类机制，提升了低频样本的聚类精度；排序阶段结合语义相关度与行为因子优化多因子函数。实验结果表明本系统在 Accuracy、NDCG@10 等指标上具有显著优势。后续研究可探索因果推理与跨语种多模态检索的集成路径。

表 1 分类检索系统对比实验设计方案

编号	比较维度	对照方案	本文方案	控制变量	训练设置
A1	分类结构	单通道 FC	双通道结构 + 文本语义解耦	向量结构	轮次 100，lr=0.001
A2	聚类策略	K-means（k=20）	层次聚类 + 软标签引导	聚类方式	轮次 50，lr=0.002
A3	排序机制	TF-IDF+ 余弦相似度	规则挖掘 + 多因子打分 +Faiss 索引	排序路径	—

表 2 分类检索系统对比性能指标统计表

实验组别	Accuracy	Macro-F_1	NDCG@10	MRR@10	Recall@10	Precision@10	Latency/ms	CPU Util/%	GPU Util/%
单通道分类器	0.843	0.821	0.756	0.693	0.726	0.701	34.7	46.3	18.7
静态 K-means 聚类	0.856	0.836	0.782	0.711	0.741	0.723	31.4	51.5	20.2
TF-IDF 排序方案	0.861	0.847	0.803	0.735	0.755	0.738	29.1	54.8	21
本文所提方法	0.918	0.894	0.882	0.793	0.821	0.794	22.5	61.4	33.6

（下转第 161 页）

算法实现高效数据加密，双重加密确保传输安全；传输优化处理通过 PCA 降维减少数据量，多维身份验证确保合法性；资源动态分配采用 OFDM 技术实现多链路调度，基于带宽计算与权重分配。实验结果表明，该方法实现了终端数据包的实时加密传输，在降低解码失败率的同时提升了传输效率。随着网络技术发展，需持续优化加密技术以应对日益复杂的安全挑战。

参考文献：

[1] 张伟 , 李泽亚 , 张充 , 等 . 基于 WSN 的流程行业安全生产数据多优先级多路径传输方法 [J]. 中国安全生产科学技术 , 2023, 19(9):5-11.

[2] 蔺海荣 , 段晨星 , 邓晓衡 , 等 . 双忆阻类脑混沌神经网络及其在 IoMT 数据隐私保护中应用 [J]. 电子与信息学报 , 2025, 47:1-17.

[3] 张杰 , 杜金华 , 刘立 , 等 . 水下通信网络中基于公钥加密体制的安全数据传输方法 [J]. 小型微型计算机系统 , 2023, 44(8): 1805-1811.

[4] 涂剑峰 , 李芳丽 . 冒充性攻击下无线传感节点信息数据安全推荐算法 [J]. 传感技术学报 , 2024, 37(8):1434-1440.

[5] 欧家祥 , 胡厚鹏 , 何沛林 , 等 . 基于国密算法的电网 AMI 数据安全防护技术 [J]. 西南师范大学学报 : 自然科学版 , 2023, 48(3):17-24.

[6] 陈佟 , 黄文雯 , 夏小萌 , 等 . 基于 RSA 算法的电力通信工程环境安全监测方法 [J]. 微电子学与计算机 , 2023, 40(4):63-71.

[7] 王惠明 , 刘志明 , 何娜 , 等 . 复杂环境山地灾害监测智能感知与数据传输关键技术 [J]. 科学技术与工程 ,2025, 25(2):640-648.

[8] 张晓烁 , 倪唯一 , 肖海林 . 基于 RIS-UAV 协作的车载通信系统安全传输研究 [J]. 电子科技大学学报 , 2023, 52(4): 539-548.

[9] 王怀乐 , 刘然 , 马明 , 等 . 基于云资源的青藏高原科考数据收集与传输平台设计与实现 [J]. 气象科技 , 2023, 51(5): 648-657.

[10] 许子恒 , 李王睿 , 张立东 , 等 . 基于信息备份技术提升城市轨道交通数据单向传输可靠性的设计方法 [J]. 城市轨道交通研究 , 2024, 27(9):227-230.

[11] 覃润楠 , 彭晓东 , 谢文明 , 等 . 面向航天器大数据安全传输的发布 / 订阅系统设计 [J]. 系统工程与电子技术 , 2024, 46(3): 963-971.

[12] 刘炜 , 郭灵贝 , 夏玉洁 , 等 . 基于门限聚合签名的区块链预言机数据传输模型 [J]. 郑州大学学报 : 理学版 ,2023, 55(4): 23-29.

【作者简介】

孙思雨（1991—），女，黑龙江大庆人，本科，助理工程师，研究方向：数字运维。

（收稿日期：2025-05-14　修回日期：2026-02-12）

（上接第 156 页）

参考文献：

[1] 吴小凤 . 基于数据挖掘的智慧图书馆信息自动化检索系统设计 [J]. 自动化技术与应用 ,2024,43(4):155-158.

[2] 杨月 . 数字图书馆交互式信息分类检索模型设计 [J]. 科技通报 , 2021,37(12):112-116.

[3] 康丽丽 , 钱婧 . 网络信息资源分类检索方法的应用 [J]. 集成电路应用 , 2023,40(5):158-159.

[4] 王欣玥 . 信息检索领域 IPC 分类号与 CPC 分类号的对比分析 [J]. 海峡科技与产业 ,2022,35(12):77-83.

[5] 孟怡悦 , 彭蓉 , 吕其标 . 一种结合标签分类和语义查询扩展的文本素材推荐方法 [J]. 计算机科学 ,2023,50(1):76-86.

[6] 代佳洋 , 周栋 . 基于多任务学习的跨语言信息检索方法研究 [J]. 广西师范大学学报 (自然科学版),2022,40(6):69-81.

[7] 江雪姣 . 基于大数据技术的网络信息资源分类检索方法 [J]. 信息与电脑 (理论版),2022,34(13):10-12.

[8] 林广朋 . 基于贝叶斯算法的网络信息安全过滤系统设计 [J]. 长江信息通信 , 2022,35(6):54-56.

[9]ZHANG J .Data-driven development of entrepreneurial ecosystem based on the apriori algorithm[J].Journal of computational methods in sciences and engineering,2025,25(4):3224-3238.

[10]LI T T, LIU F, CHEN X B, et al.Publisher correction: web log mining techniques to optimize apriori association rule algorithm in sports data information management[J].Scientific reports, 2024,14(1):30568.

【作者简介】

黄海涛（1992—），男，安徽霍邱人，本科，高级工程师，研究方向：大数据、人工智能、智慧图书馆。

（收稿日期：2025-07-08　修回日期：2026-01-14）