doi:10.3772/j.issn.2095-915x.2023.03.006

基于多源信息融合的学位论文自动分类标引

开放科学（资源服务）标识码（OSID）

谢庆恒

国家图书馆　北京　100081

摘要：[目的 / 意义] 学位论文是图书馆的特色馆藏文献，实现学位论文的自动分类标引对智慧图书馆建设具有积极意义。[方法 / 过程] 首先基于 BERT 分别获取题名和摘要的词向量表示，然后将二者进行加权代数和计算得到融合向量，最后将其输入到基于 Pytorch 框架构建的 Softmax 经典分类器进行学位论文的自动分类标引实践探讨。 [局限] 在数据信息源和学科内容的多样性方面尚需进一步加强。[结果 / 结论] 模型分类 F1 值达到了 79.55%，优于基于单一信息的题名或摘要的分类效果，能较好满足实际应用要求。

关键词：学位论文；自动分类；信息融合；BERT

中图分类号：G254　TP391

Automatic Classification and Indexing of Dissertations Based on Multi-source Information Fusion

XIE Qingheng

National Library of China, Beijing 100081, China

Abstract: [Objective/Significance] Dissertation is the distinguishing collection of the library, and it is of positive significance to realize the automatic classification and indexing of dissertations for the construction of a Smart Library. [Methods/Processes] Firstly, based on BERT, the word vector representations of the title and abstract are obtained, and then the weighted algebraic sum of them is calculated to obtain the fusion vector. Finally, it is inputted into the Softmax classic classifier constructed based on the Python framework for practical exploration of automatic classification and indexing of dissertations. [Limitations] Further exploration is needed in diversity of data information sources and subject content. [Results/Conclusions] The results show that the F1 value of this model reaches 79.55%, which is better than that of title or abstract based on single information, and can fairly meet the requirements of practical application.

Keywords: Dissertation; automatic classification; information fusion; BERT

基金项目　中国图书馆学会青年项目“智慧图书馆中学位论文自动分类标引研究”（2022LSCKYXM-ZZ-QN003）。

作者简介　谢庆恒（1988-），硕士，馆员，研究方向为文献编目，E-mail: xqheng2012@163.com。

引用格式　谢庆恒 . 基于多源信息融合的学位论文自动分类标引 [J]. 情报工程 , 2023, 9(3): 70-80.

引言

文献分类是图书馆赖以生存的重要基础，分类标引是文献分类的重要环节，如何拥抱新一代信息技术实现文献分类标引的自动化是图书馆面临的一大挑战，也是图书馆智慧化升级的重要方面。学位论文是图书馆的特色馆藏文献，实现学位论文的自动分类标引对智慧图书馆建设具有积极意义。与一般文献不同，学位论文特别是博士学位论文大都反映的是某一学科专业最前沿的研究成果，内容专深，学科交叉渗透，其内容的揭示要求实现深层次细粒度的文献分类标引。当前人工分类标引一般要细分到中图法四级甚至五级类目才能科学定位论文类别，保证分类质量。比如《白洋淀沉积物污染特征与源—汇过程研究》一文，主要研究白洋淀湖泊污染物的分布特征及其来源，若分入 X52（水体污染）则只反映了水污染问题，未反应具体的湖泊污染问题，专指性不够，应分入 X524（湖泊、水库污染）；又如《影响南京地区 NO2 污染的气象因子及典型个例数值模拟研究》一文，研究南京地区的氮氧化物污染问题，若只分到 X51（大气污染），则同样专指性不够，应划分到更具体的下位类 X511（气相污染物）才能准确揭示论文类别；再如《ssGSEA 在膀胱癌中的运用及其预后模型的构建》主要讲述单样本基因富集分析方法在膀胱癌免疫相关基因分析中的运用，应划入 R737.14（膀胱肿瘤），如若划入其上位类 R737.1（泌尿器肿瘤）则专指性不强，导致用户检索时不能精准定位，降低查准率。学位论文分类标引的高专指性决定了其自动分类标引的深层次性，自然增加了其自动分类标引的难度，同时也对语料信息质量提出了更高要求，单一信息源语料显然难以满足。因此，开发一套融合多源信息的分类模型是实现学位论文自动分类标引的关键。

1 相关研究

学位论文自动分类标引是一项典型的文本分类任务，相关研究并不多见，赵国荣 [1] 基于支持向量机方法研究了山西大学图书馆学位论文的自动分类问题；Lighton[2] 基于有监督机器学习技术探讨了机构知识库中的电子论文和学位论文自动分类的可行性，取得较好效果。相近研究主要集中于期刊论文 [3]、专利文本 [4]、新闻分类 [5] 等，早期主要采用朴素贝叶斯 [6]、支持向量机 [1] 和 K- 近邻算法 [7] 等基于统计的分类方法，由于这些方法存在依赖人工选择特征、建模过程复杂费力、文本特征抽取不足等问题，分类效果不甚理想。随着神经网络模型的兴起，特别是以词向量为代表的文本表征技术的进步，文本分类效果得到显著提升。早期的 Word2Vec[8]、Glove[9] 等静态词向量模型由于不能解决一词多义问题使得分类效果提升有限。后来业界相继提出 Elmo[10]、GPT[11]、BERT[12] 等动态词向量模型，逐步解决了一词多义、特征提取不足、上下文信息融合不充分等缺陷，促使文本分类效果稳步上升，尤其是 BERT 刷新了多达 11 项 NLP 任务的记录。目前，凭借突出的文本表征能力和便捷的下游任务接口，BERT 模型广泛承担了网络舆情 [13]、情感分析 [14]、问答系统 [15]、新闻传播 [5] 等领域的文本分类任务，

doi:10.3772/j.issn.2095-915x.2023.03.006

效果较为理想。

已有相关研究数据大多来自单一信息源，有的基于文本较短的标题数据[5]，有的基于文本较长的摘要数据[3-4,16]，均取得一定效果，但仍有较大提升空间，一个可能原因就是数据信息源单一，难以充分揭示所表达主题的主要内容。研究表明，多源信息融合是解决单一信息源问题的有效方案。这里的多源不仅包括不同模态的信息，如文本、图像、音频、视频等，也包括同一模态下不同位点的信息，如文本中的标题、摘要、关键词、引（前）言、结束语等，还包括同一模态不同属性信息，如图像中的分辨率、色度、亮度等。要将这些多源信息进行融合就需要采用一定的融合方法。数据融合一般可划分为数据层、特征层与决策层融合三种形式[17]，其中特征融合的常用方法有直接叠加、串行或并行连接以及加权求和等。一方面，学者运用特征融合方法基于不同模态数据进行了探讨，均取得理想效果，如康丽萍等[18]分别抽取图像与文本分类标签特征后进行图文数据融合，马超等[19]基于 BERT 和 SqueezeNet 实现了对图文结合的多模态产品评论有用性的分类，黄欢等[20]基于 CNN 与 ResNet 提取短视频的视觉与音频两种模态特征并通过融合方法进行短视频情感分类；另一方面，有学者探讨了同模态不同位点信息或同模态不同角度信息的特征融合问题，效果同样较为理想，如赵国荣[1]融合了同一文本信息中的标题、摘要和关键词探讨了学位论文的自动分类问题，杨桄等[21]通过提取图像的边缘特征、平均梯度特征、相关信号强度比特征实现红外和可见光图像的融合，谢将剑等[22]基于 VGG16 的特征迁移模型从三个不同侧面提取图像特征并通过线性加权进行特征融合。这些彰显了信息融合方法的应用广泛性和效果显著性，同时也凸显了其在学位论文等科技文献分类领域应用的相对匮乏性。鉴于此，本文拟结合学位论文题名和摘要两个位点信息，基于 BERT 模型抽取信息特征，通过线性加权求和方法进行特征融合，以探讨信息融合方法在学位论文自动分类标引方面的适用性并希望以此优化提升学位论文自动分类标引的效果。

2 研究方法与数据处理

2.1 研究方法

2.1.1 分类架构

进行多源信息融合需要确定信息来源点和信息融合方法。就学位论文分类标引而言，信息来源点主要包括题名和摘要。题名是学位论文内容的直观表达，是学位论文最直观的标签。摘要能较全面的反映学位论文的内容，是学位论文的精简版[1]。二者基于不同角度在不同程度上揭示了学位论文主要内容，是进行自动分类标引的重要信息点。题名简短直观、摘要详细全面，二者各有侧重，按照一定方式进行融合可以充分发掘利用二者信息，实现互通有无，相互补充，再通过多次试验找到一个最优融合比以达到理想的分类标引效果。因此，本文以相当规模的包含题名和摘要内容的高质量人工分类标引学位论文数据为语料，通过构建融合分类模型，并多次实验找到题名和摘要表示向量的最优融合比，探讨基于多源信息融合方法

的学位论文自动分类标引效果（图 1a）。为了验证融合方法的效果，本文还设计了基于单独题名或摘要内容的分类对比试验（图 1b，图 1c）。

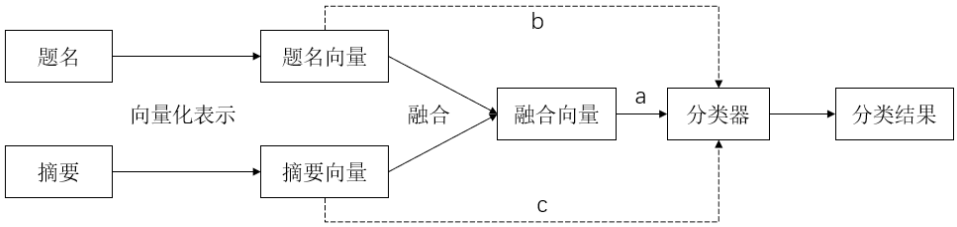

图 1 基于信息融合方法的分类框架

2.1.2 模型结构

BERT[12] 是 Google 团队于 2018 年提出的基于 Transformer 衍生出的双向自编码表征模型，在 11 种经典自然语言处理测试任务中取得 SOTA 表现，具有里程碑意义。BERT 采用预训练 - 微调的上下游模式，基于自注意力机制融合上下文信息，通过完成遮盖语言模型（masked language model，MLM）和下一句预测（next sentence prediction，NSP）两个预训练任务，模型分别获取单词和句子级别的语义信息，从而获得较强的语义表征能力。BERT 下游任务主要是使用预训练阶段的词向量完成诸如情感分析、句子关系判断、问答任务、命名实体识别等任务，即所谓的微调（fine-tunning）。BERT 输出已完成封装，对于单句（段）分类任务，可以直接输入单个句子（段落），将输出结果（CLS 向量）直接输入到分类器进行分类，得到分类结果。

将数据预处理之后输入模型，基于 BERT 分别获取题名和摘要文本特征向量，之后分别通过 Softmax 分类器得到基于题名的分类结果、基于摘要的分类结果以及基于二者融合的分类结果。融合方法选用简单的矩阵代数和公式，即将题名表示向量和摘要表示向量按照一定权重比例进行相加来表示融合向量，具体结构如图 2 所示。

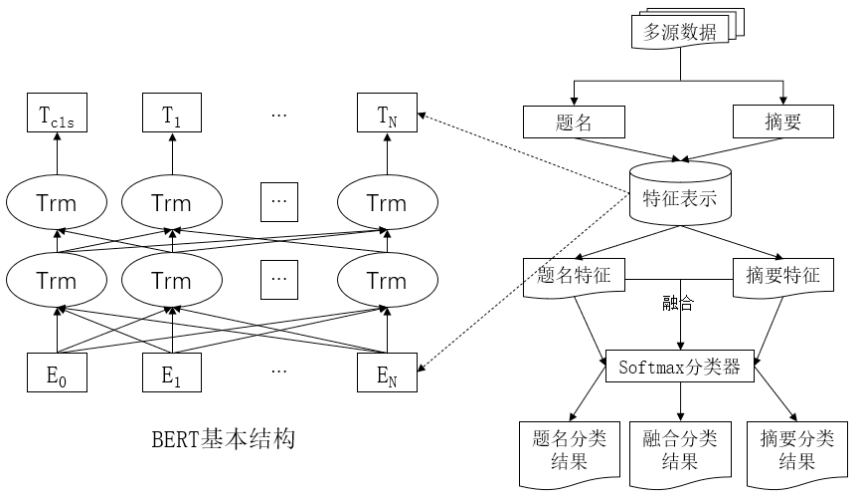

图 2 基于 BERT 的分类模型结构

doi:10.3772/j.issn.2095-915x.2023.03.006

2.2 数据处理

本文数据采集自万方中国学位论文全文数据库[23]。以中图分类号 R 类下的 1123 个四至五级类目为检索项，共采集到近些年来论文数据 260995 条，每条数据包含题名、摘要和类号，类号细分到四至五级，如四级类号 R737.1、R649.3，五级类号 R811.1、R332.1，数据样例如图 3 所示。

序号	题名	摘要	分类号
1	rBlag2-PLGA纳米疫苗对rBlag2诱导的小鼠气道变态反应炎症的影响及	是一种新型药物载体用作抗原载体可以控制抗原释放接种机体后能诱导持	R562.2
2	血根碱对胃癌细胞增殖和侵袭的影响及其机制研究	胃癌是消化系统常见恶性肿瘤其致死率在所有肿瘤中排第三位早期胃癌术	R735.2
3	68Ga-PSMA PET/CT与Nomogram诊断前列腺癌的回顾性临床研究	背景前列腺癌是最常见男性恶性肿瘤之一在欧美所有男性恶性肿瘤中其发	R737.2
4	乳粘素与膜联蛋白在口腔癌细胞凋亡检测中的应用及比较	细胞凋亡时细胞膜会发生一系列变化在早期阶段磷脂层不对称性分布特性	R739.8
5	在胰腺癌靶区勾画中磁共振弥散加权成像(MR-DWI)与增强电子计算机	旨在进一步研究探讨胰腺癌肝脏与区域淋巴转移瘤在增强电子计算机增强	R730.4
6	COPD呼吸衰竭机械通气患者两种不同通气模式的比较	机械通气主用于各种原因引起急慢性呼吸衰竭和呼吸功能不全为患者提供	R563.8
7	血清脂质运载蛋白2与急性心肌梗死后左室重构的相关性研究	本文旨在探讨血清水平与急性心肌梗死后患者左室重构相关性收集急性段	R542.2
8	基于社会公平视角的重庆市基本公共卫生服务区域均等化研究	十九大报告中提出当前社会主矛盾转化为发展不平衡和不充分这种不平衡	R197.1
9	小鼠骨髓间充质干细胞体外诱导成肝样细胞的实验研究	体外分离培养扩增鉴定小鼠骨髓间充质干细胞体外诱导小鼠定向分化为肝	R329.2
10	动脉粥样硬化斑块处内皮间质转化的变化及硫化氢的调控作用与机制研	动脉粥样硬化引起心脑血管疾病已成为全球人口死亡最主原因血管内皮功	R543.5
11	新化合物ECPIRM及ATRA对角质形成细胞炎症因子的影响和相关机制研	银屑病是一种常见免疫介导慢性复发性炎症性皮肤病以表皮角质形成细胞	R758.6
12	超声内镜、CT对胃癌诊断及TNM分期判断的对比研究	为探讨超声内镜联合诊断胃癌准确率灵敏度及分期价值了解二者对胃癌诊	R730.4

图 3 数据样例

初步分析发现，数据集中存在一些多个类号的样本，该样本会同时出现在多个类号中，为保持训练数据样本的唯一性，选取这些样本中的一个类号作为该样本类号，其他类号中则将该样本删除。还有一些明显分类错误的数据也进行了删除处理。为了增加样本数据的随机性，随机选取各类样本的 80% 作为备选样本，经过清洗、去重、删除等操作得到无重复的高质量数据共计 1123 类 199451 条数据，其中相当数量的类号下文献数量不足 10 条。为消除训练数据的分布不均带来的影响，本文制定了分类训练样本的类目筛选规则，具体如下：

（1）以 1123 个深层级类目为起点，自下而上进行各类目下文献的统计筛选工作；

（2）若某个类目下的文献数量大于等于 300，则将该类目直接列入最终类目列表中；

（3）若某个类目下的文献数量小于 300，则将该类目下的文献合并到其上级类目中；

（4）若合并后的类目层级小于四级，则将该类目舍弃，否则，统计合并后的类目下文献数量，若小于 300，则重复步骤（3）；

（5）若合并后类目下的文献数量不小于 300，则将该类目列入最终列表。

按照此规则最终筛选得到 129 类共计 169495 条数据，再按照 8:1:1 分别将各类随机分成训练集（135643 条）、验证集（16897 条）和测试集（16955 条），组成本文模型数据集。

3 实验过程与结果分析

3.1 实验环境

本文实验处理器采用 AMD EPYC 7742 核处理器，内存 122GB，GPU 为单张 RTX A6000，显存 48GB，程序语言为 Python，运行环境是 Python3.9.12，深度学习框架采用 PyTorch1.11.0。

3.2 评价标准

采用经典的精确率（P）、召回率（R）、F1 值作为模型分类结果评价依据。其中 P 表示预测为正样本中实际为正样本的比率，反应

的是模型查准的能力；R 表示实际为正样本中预测为正样本的比率，反应的是模型查全的能力；F1 值综合平衡了二者信息，计算公式为：$F1=(2\times P\times R)/(P+R)$。

3.3 结果分析

3.3.1 基线模型对比

为了检验模型的分类效果，本文基线模型采用经典模型 Word2Vec 和 FastText。为提高检验效率，训练语料采用数据集中的题名数据部分，模型的基本设置如下：

（1）Word2Vec：使用 Jieba 分词工具将每个句子进行切词并转换为二维 list 形式，基于文献 [24] 已有的 Word2Vec 词向量生成词向量再取各词的平均向量作为句子的向量表示。模型参数设置为学习率 0.01，训练轮次 epoch 为 3，批次大小为 160，dropout 率为 0.5，采用交叉熵计算损失并通过随机梯度下降法更新模型权重。

（2）FastText：使用文献 [25] 已有的 FastText 中文词向量生成词向量，其他与（1）一致。

（3）BERT：采用 Google 发布的预训练好的中文模型“BERT-Base，Chinese”生成词向量，取模型输出结果中的“pooler_output”层作为句子向量表示，模型学习率设置为 5e-5，epoch 为 3，批次大小 160，最大序列长度为 32，dropout 率为 0.5，损失函数仍采用交叉熵计算方法，通过随机梯度下降法更新模型权重计算出分类结果。

表 1 基于题名的模型分类结果

模型	P	R	加权F1值
Wrod2Vec	0.6426	0.6178	0.6172
FastText	0.5921	0.5693	0.5640
BERT	0.7528	0.7548	0.7504

从表 1 可以看出，BERT 模型的 F1 值均高于传统的 Word2Vec 和 FastText 模型，分别高出 13.32 和 18.64 个百分点，可见 BERT 模型凭借良好的特征抽取能力保证了分类效果好于传统模型。因此，本文选用 BERT 作为分类模型。

3.3.2 BERT模型结果

模型词向量采用 Google 发布的预训练好的中文模型“BERT-Base，Chinese”，该模型采用 12 层 Transformer，隐藏层的维度为 768，Multi-Head-Attention 的参数为 12，模型总参数大小为 110 MB。模型语料数据包括两部分，即题名数据和摘要数据。基于题名训练前文 3.3.1 部分已讲述；基于摘要进行训练时，模型训练参数批次大小（batch_size）设置为 80，学习率为 5e-5，最大序列长度（max_seq_len）为 400，优化器为 Adam，dropout 率为 0.5，训练轮次 epoch 为 3，采用随机梯度下降法对模型权重进行更新；基于融合方法进行训练时，模型训练参数批次大小（batch_size）设置为 40，学习率为 5e-5，最大序列长度（max_seq_len）为 400，优化器为 Adam，dropout 率为 0.5，训练轮次 epoch 为 2，题名、摘要融合比分别尝试 3:7,4:6,5:5 和 6:4，损失函数仍采用交叉熵计算方法，采用随机梯度下降法对模型权重进行更新，运行结果如表 2 所示。

从表 2 可以看出，基于摘要的分类效果在 P、R、F1 三个指标上全面优于基于题名的分类效果，分别高出 2.69、2.93 和 2.74 个百分点，表明详细全面的摘要数据能较好揭示论文主要内容，相对而言，简短直观的题名对论文内容的揭示则不够充分。一般地，题名是学位论文内

doi:10.3772/j.issn.2095-915x.2023.03.006

表 2　BERT 模型运行结果

语料		P	R	F1	类别F1提升占比
题名		75.28%	75.48%	75.04%	-
摘要		**77.97%**	**78.41%**	**77.78%**	**23.26%**
题名＋摘要	3:7	79.25%	79.57%	78.96%	78.29%
	4:6	79.51%	79.86%	79.25%	87.60%
	5:5	79.78%	80.07%	79.55%	92.25%
	6:4	79.63%	79.99%	79.45%	91.47%

容的直观表达，是学位论文最直观的标签。摘要则能较全面的反映学位论文的内容，是学位论文的精简版。摘要对论文内容的表达更为详尽，对论文主题的揭示更加充分。同时也表明 BERT 能够提取出摘要中的文本特征，能够较好识别摘要中的详细信息。129 个类别中，基于摘要的 F1 值提升的占比只有 23.26%，表明摘要数据相对于题名数据效果虽有提升但仍有较大改善空间。基于题名和摘要融合数据的分类结果 F1 值均高于基于摘要的 F1 值，其中融合比 5:5 时 F1 值达到最高的 79.55%，模型分类效果最佳，表明题名和摘要对模型分类结果的影响同等重要，二者从不同角度和不同程度上揭示论文内容，相比于各自单独揭示，二者融合能够起到信息互补的作用，对论文主题的揭示更为充分、更为全面。具体来看，129 个类别中，基于融合比为 3:7、4:6、5:5、6:4 融合数据的分类 F1 值提升的分别有 101、113、119 和 118 个，占比分别为 78.29%、87.60%、91.47% 和 92.25%（表 2），表明融合数据的分类效果好于基于摘要的分类效果，尤其是融合比为 5:5 时效果最佳。

模型 epoch 为 2，batch_size 为 40，故共有约 6800 个 batch，取每第 1000 个 batch 的值（损失值、准确率）为纵坐标绘制本模型训练后的 loss 损失曲线和 accuracy 准确率曲线，如图 4 和图 5 所示。从图 4 可以看出，训练 loss 值在前 3000 个 batch 呈大幅下降趋势，表明深度学习网络模型的学习能力较强。模型在第 4000 个 batch 后 loss 值稳定在 0.5 左右，趋于收敛。从图 5 可以看出，在前 3000 个 batch 模型准确率显著上升，第 4000 个 batch 后趋于稳定，保持在 0.79 左右，模型未出现过拟合现象。

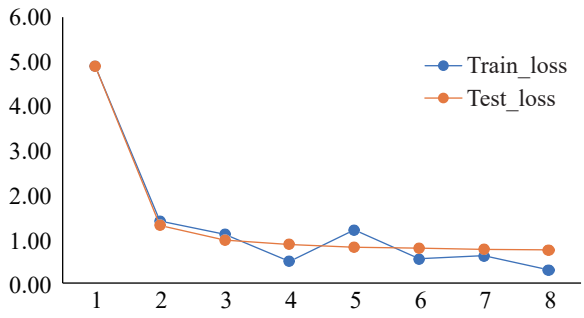

图 4　loss 曲线

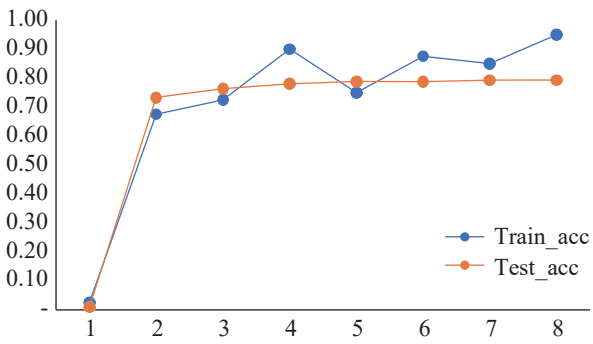

图 5　accuracy 曲线

3.3.3　结果讨论

通过分析比较基于题名、摘要和二者融合的分类结果 F1 值（分别记为 $F1_t$、$F1_a$、$F1_{ta}$），按照三者的大小关系可以把样本分为 3 大类型，即双主导型、摘要主导型和题名主导型（表 3）。

表 3 不同语料下 F1 值比较及样本类型

$F1_a>F1_t$	$F1_{ta}>F1_a$	$F1_{ta}>F1_t$	类型
√	√	√	（1）双主导
√	×	√	（2）摘要主导
×	√	√	（3）题名主导

注：“√”表示符合条件，“×”表示不符合条件。

（1）双主导型。指的是 $F1_a>F1_t$、$F1_{ta}>F1_a$、$F1_{ta}>F1_t$。$F1_a>F1_t$ 表明详细的摘要信息相比简洁的题名信息能更好表达论文主题内容。$F1_{ta}>F1_a$ 表明融合进题名信息后能更进一步的揭示论文主题，说明二者能够相互补充、相互完善。$F1_{ta}>F1_t$ 表明融合后的信息比单纯的题名信息表达能力要强，同样也体现了二者信息的相互促进作用。

例 1：

2001#$a 结缔组织移植术腭部供区黏膜厚度的横断面研究

330##$a 本研究旨在测量分析成年健康人群上腭部不同位置黏膜厚度及上腭粘膜厚度与年龄、性别的关系，以期为临床医生在上腭部获取游离移植物时选择适当的位置、方法、大小及厚度提供理论参考。方法：……结果：上颌腭侧黏膜厚度随着距离龈缘越远呈逐渐增厚的趋势，……结论：中老年组腭部黏膜厚度比青年组更厚，中老年组中的男性腭部黏膜厚度要比女性的厚；上颌腭侧黏膜厚度随着距离龈缘越远呈逐渐增厚的趋势；从上颌第二前磨牙到第二磨牙各牙位腭侧黏膜厚度平均值均大于 3mm，在此区域均可切取厚度≧ 1mm 的上皮下结缔组织移植物，是临床上获取游离移植物的理想供区部位。

解析：基于题名、摘要和二者融合的分类结果分别为 R593.2（结缔组织疾病）、R781.4（牙周病）和 R782.2（口腔颌面部疾病）。论文主要介绍的是口腔美学外科手术的自体结缔组织供区部位黏膜厚度问题，属于口腔颌面部整形外科学范畴，正确分类为 R782.2。由于题名中含有“结缔组织移植术”等关键字使得模型误认为论文主题为“结缔组织疾病”，而摘要中“磨牙”等关键字也使得模型误认为研究主题为“牙周病”，二者对主题的把握均比较片面，而将二者融合，既能提炼出结缔组织、牙周疾病等重要信息，又能捕捉到移植物、口腔上颚等口腔整形外科学等关键信息，从而结合二者信息得到正确的分类结果。

（2）摘要主导型。指的是 $F1_a>F1_t$、$F1_{ta}<F1_a$、$F1_{ta}>F1_t$。$F1_a>F1_t$ 表明摘要信息相比题名信息能更好表达论文主题内容。$F1_{ta}>F1_t$ 表明融合后的信息比单纯的题名信息表达能力要强。$F1_{ta}<F1_a$ 表明融合进题名信息后对揭示论文主题的能力反而降低，说明题名信息对论文主题的揭示造成了一定的干扰。

例 2：

2001#$a 行人年龄和碰撞速度对行人下肢损伤的影响研究

330##$a 在复杂的人 - 车 - 道路交通环境中，随着机动车辆保有量迅速提升，交通事故频繁发生，行人在道路环境中的伤亡率较高，其中行人下肢是最比较容易受到损伤的部位之一，……找出影响行人下肢损伤风险的相关参数，有利于降低行人在交通事故中的损伤风险……本文从 CIDAS 数据库中整理了 2012-2017 年发生的行人案例……结果显示，当速度大于 44.3km/h 时，AIS3+ 重伤损伤风险增加，随着行人年龄逐渐变大，重伤损伤风险随之变

doi:10.3772/j.issn.2095-915x.2023.03.006

大，当下肢重伤损伤风险为 50%时，行人年龄为 33 岁。通过损伤风险预测发现，在相同碰撞速度下，年龄越大越容易发生骨折。因此，年龄和速度是影响行人下肢长骨骨折和重伤风险的两个重要参数……本文的研究结果可为具有年龄差异的行人下肢损伤评估指标的设定提供参考。

解析：基于题名、摘要和二者融合的分类结果分别为 R654（心脏血管和淋巴系外科学）、R683.4（四肢骨折）和 R687.3（骨骼手术）。该论文主要论述交通事故中影响行人下肢长骨骨折和重伤风险的两个重要参数是年龄和碰撞速度，应该分类到 R683.4（四肢骨折），这些内容在摘要中表达的较为充分。而基于题名的模型将其分类为 R654，出现了错误。这一错误同时也影响到了融合信息模型的分类结果，将融合结果带偏到了外科手术类，从而将论文误分到了 R687.3（骨骼手术）。

（3）题名主导型。指的是 $F1_a<F1_t$、$F1_{ta}>F1_a$、$F1_{ta}>F1_t$。这种情况主要是题名中能较为直观的提取到论文主题信息，大都是题名中出现了反应主题信息的关键字眼，如“肝癌”“肺癌”等。而摘要中虽也表达了主题信息，但往往在进行研究背景介绍时会出现干扰信息产生“掩盖”现象，反而从论文主题的揭示造成一定困扰，导致模型分类出现偏差。二者融合后，分类效果则取决于摘要的“掩盖”效应的大小。效应小，则融合模型的分类效果尚可，效应大，则可能出现融合模型效果不如题名。

例 3：

2001#$a ETV 治疗 CHB 患者肝细胞癌发生率及预测因子的研究

330##$a 目的：对接受恩替卡韦（Entecavir，ETV）治疗的慢性乙型肝炎（Chronic Hepatitis B viral hepatitis，CHB）患者原发性肝细胞癌（Hepatocellular Carcinoma，HCC）的发生率及预测因子的研究，探讨 ETV 治疗慢性乙型肝炎患者对 HCC 发生率的影响。方法：……结论：1.CHB 患者长期口服 ETV 仍可发生 HCC。2. 在长期 ETV 治疗下仍有 HCC 发生，除了已知的 HCC 预测因素，如高龄、肝硬化等，HBeAg 阴性、低血小板也是较高的 HCC 风险预测因素。3. 对于年龄大于 50 岁的男性患者，尤其是 HBeAg 阴性慢性乙型肝炎患者，应定期加强监测，尽早发现肝癌，从而改善患者预后。

解析：基于题名、摘要和二者融合的分类结果分别为 R735.7（肝癌）、R512.6（病毒性肝炎）和 R735.7（肝癌）。本例中，题名中含有关键词“肝细胞癌”，能较为明显地揭示论文主题，模型分类效果较好。而摘要中开篇介绍乙型肝炎的一些背景知识，“掩盖”了肝癌这一主题的信息，导致模型分类为 R512.6（病毒性肝炎）类。但这种“掩盖”作用效果有限，因此二者融合的分类效果仍较为准确。

4 实际应用

为了验证融合模型的实际应用效果，随机采集中国知网 [26] 未出现在上述训练样本中的论文数据若干条进行实验，结果如表 4 所示。结果显示，在 5 个类别上，融合模型的预测结果准确率平均达到 85% 以上，效果较为理想，表明本文提出的融合模型在 R 类别的学位论文的自动分类标引实际应用中能达到预期效果，具有较大的实用价值。

表 4　融合模型预测结果

类别	样本量	准确率
R373.1	40	87.50%
R541.6	40	85.00%
R575.2	40	72.50%
R681.5	40	92.50%
R782.2	40	90.00%

例 1：

2001#$a Shox2/Alx1 信号轴调控小鼠上颌骨的形成

330##$a 腭裂是最常见的出生缺陷之一，在全世界的发病概率大约为 1‰～ 2‰，在我国的发生率大约为 1.82‰，表现为上腭的软组织畸形，还常伴有不同程度的骨组织缺损和畸形。根据缺损发生部位，腭裂可以分为前端的硬腭裂（Hard Cleft Palate）和后端的软腭裂（Soft Cleft Palate），……本研究通过组织形态学和生物信息学手段，从基因组非编码 DNA 角度，揭示了 Shox2 作为上游转录因子通过 Alx1En 结合到 Alx1 启动子区域，调控 Alx1 的转录激活；而 Alx1 调控着上腭板间充质细胞成骨相关基因的表达，完整的上颌骨保证了小鼠上腭板正常的形态发生。该研究对深入理解上腭板发育过程基因调控的精细分子机制及腭裂小鼠病理发生的机制，具有重要的科学意义。

解析：融合模型基于题名和摘要信息准确判断出论文主题为“腭裂”的病因分析，因此将其分类为 R782.2（口腔颌面部疾病）。

例 2：

2001#$a 综合医院消化内科肝硬化患者的流行病学和临床特征分析

330##$a 背景：肝硬化是一种病因多样的慢性疾病，临床表现复杂，并发症凶猛，死亡率高，深入的基础和临床研究对于改善肝硬化至关重要……目的：了解综合医院消化内科肝硬化患者临床流行病学情况及临床特征，为消化内科医师临床诊治肝硬化患者提供病因学、临床特点、病情程度和疾病负担等参考依据。方法……结论：消化内科病毒性肝炎肝硬化占比可能呈下降趋势，隐源性肝硬化占比增多。且隐源性肝硬化多合并银屑病，应慎重诊断。消化内科肝硬化患者肝功能衰竭少见，住院天数与 Child-Pugh 评分和终末期肝病模型评分呈正相关。病毒性肝炎肝硬化患者病情以轻中度为主，而非病毒性肝炎肝硬化患者相对病情较重，并发症多见；相对于丙肝肝硬化，乙肝肝硬化更倾向于有同种疾病家族史。

解析：论文主要论述“肝硬化”的临床特征和流行病学调查。融合模型基于题名和摘要两个信息点进行特征提取，准确无误的将其分类到了 R575.2（肝硬变）。

5　结束语

学位论文的自动分类标引是智慧图书馆的重要建设内容。学位论文因其内容专深、学科交叉渗透的特点实现其自动分类标引需达到深层级类别，而单一信息源难以满足这一要求。本文提出的融合题名和摘要二者信息的融合模型能够较好解决这一难题，通过对医学类学位论文的分类实践得到融合模型分类精度能够达到 79.55%，优于基于单一的题名或者摘要信息的分类效果。在实际应用中也取得了理想效果，分类准确率平均值达到了 85% 以上，表明了本

doi:10.3772/j.issn.2095-915x.2023.03.006

文模型的有效性和实用性。本文仅基于标题和摘要两个信息点对医学类学位论文进行探讨，信息点较少，学科内容较为单一，信息源和学科内容的多样性尚需进一步加强。今后将继续优化模型，增加关键词、引言、结束语等信息点，以期达到更好的分类效果，并将模型推广到更多的学科类别，拓宽模型应用场景。

参 考 文 献

[1] 赵国荣 . 基于支持向量机的学位论文自动分类研究 [J]. 晋图学刊 , 2016(4): 11-15.

[2] Lighton Phiri. Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories[J]. International Journal of Metadata, Semantics and Ontologies, 2021, 14(3).

[3] 赵旸 , 张智雄 , 刘欢 . 基于层次分类法的中文医学文献分类研究 [J]. 图书馆学研究 , 2021(21): 49-55+61.

[4] 陆晓蕾 , 倪斌 . 基于预训练语言模型的 BERT-CNN 多层级专利分类研究 [J]. 中文信息学报 , 2021, 35(11): 70-79.

[5] 苗将 , 张仰森 , 李剑龙 . 基于 BERT 的中文新闻标题分类 [J]. 计算机工程与设计 , 2022, 43(8): 2311-2316.

[6] 邸鹏 , 段利国 . 一种新型朴素贝叶斯文本分类算法 [J]. 数据采集与处理 , 2014, 29(1): 71-75.

[7] 黄贤英 , 熊李媛 , 刘英涛 , 等 . 基于类别特征改进的 KNN 短文本分类算法 [J]. 计算机工程与科学 , 2018, 40(1): 148-154.

[8] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv:1301.3781, 2013.

[9] Jeffrey Pennington, Richard Socher, Christopher D. Manning. GloVe: Global Vectors for Word Representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014.

[10] Matthew E. Peters, Mark Neumann, Mohit Iyyer, et al. Deep contextualized word representations[J]. arXiv:1802.05365, 2018.

[11] Alec Radford, Karthik Narasimhan, Tim Salimans, et al. Improving Language Understanding by Generative Pre-Training[EB/OL]. [2022-11-20]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[12] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. Arxiv:1810.04805, 2018.

[13] 庄穆妮 , 李勇 , 谭旭 , 等 . 基于 BERT-LDA 模型的新冠肺炎疫情网络舆情演化仿真 [J]. 系统仿真学报 , 2021, 33(1): 24-36.

[14] 刘继 , 顾凤云 . 基于 BERT 与 BiLSTM 混合方法的网络舆情非平衡文本情感分析 [J]. 情报杂志 , 2022, 41(4): 104-110.

[15] 杨陟卓 , 韩晖 , 张虎 , 等 . 融合 BERT 语义表示的高考语文阅读理解问答研究 [J]. 中文信息学报 , 2022, 36(5): 59-66.

[16] 张智雄 , 赵旸 , 刘欢 . 构建面向实际应用的科技文献自动分类引擎 [J]. 中国图书馆学报 , 2022, 48(4): 104-115.

[17] 蒋雨肖 , 丁晟春 , 吴鹏 . 基于 BiLSTM-VGG16 的多模态信息特征分类研究 [J]. 情报理论与实践 , 2021, 44(11): 180-186+179.

[18] 康丽萍 , 孙显 , 许光銮 . 加权 KNN 的图文数据融合分类 [J]. 中国图象图形学报 , 2016, 21(7): 854-864.

[19] 马超 , 李纲 , 陈思菁 , 等 . 基于多模态数据语义融合的旅游在线评论有用性识别研究 [J]. 情报学报 , 2020, 39(2): 199-207.

[20] 黄欢 , 孙力娟 , 曹莹 , 等 . 基于注意力的短视频多模态情感分析 [J]. 图学学报 , 2021, 42(1): 8-14.

[21] 杨桄 , 童涛 , 陆松岩 , 等 . 基于多特征的红外与可见光图像融合 [J]. 光学精密工程 , 2014, 22(2): 489-496.

[22] 谢将剑 , 杨俊 , 邢照亮 , 等 . 多特征融合的鸟类物种识别方法 [J]. 应用声学 , 2020, 39(2): 199-206.

[23] 万方中国学位论文全文数据库 [EB/OL]. [2022-03-20]. https://c.wanfangdata.com.cn/thesis.

[24] Shen Li, Zhe Zhao, Renfen Hu, et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[C]// ACL 2018.

[25] Facebook Open Source[EB/OL]. [2022-08-20]. https://fasttext.cc/docs/en/crawl-vectors.html

[26] 中国知网学位论文库 [EB/OL]. [2022-12-20]. https://kns.cnki.net/kns8?dbcode=CDMD.