情报学报　2020 年 10 月　第 39 卷　第 10 期
Journal of the China Society for Scientific and Technical Information, Oct. 2020, 39(10): 1046-1059

DOI: 10.3772/j.issn.1000-0135.2020.10.004

基于深度预训练语言模型的文献学科自动分类研究

罗鹏程[1,2]，王一博[2]，王继民[1]

（1. 北京大学信息管理系，北京　100871；2. 北京大学图书馆，北京　100871）

摘　要　为了支撑“一流学科”相关的情报和文献服务，本文探索利用深度预训练语言模型实现文献的教育部一级学科自动分类。通过构建基于BERT和ERNIE的文献学科分类模型，在21个人文社科一级学科近10万条期刊文献数据集上进行实验验证，并与传统机器学习方法（朴素贝叶斯、支持向量机等）、典型深度学习方法（卷积神经网络、循环神经网络）进行对比分析。结果显示，基于深度预训练语言模型的方法效果最好，其中ERNIE在测试集上的Top 1和Top 2准确率分别可达到75.56%、89.35%；同时使用标题、关键词和摘要作为输入的分类模型效果最优；一些学科的学科独立性强，分类效果好，如体育学F1值高达0.98；另一些学科间交叉性高，分类效果欠佳，如理论经济学和应用经济学的F1值在0.6左右。此外，本文还对学科交叉融合、模型应用场景、分类效果优化做了进一步的探讨。

关键词　文献学科分类；深度学习；文本分类；预训练语言模型

Automatic Discipline Classification for Scientific Papers Based on a Deep Pre-training Language Model

Luo Pengcheng[1,2], Wang Yibo[2] and Wang Jimin[1]

(1. *Department of Information Management, Peking University, Beijing*　100871; 2. *Peking University Library, Beijing*　100871)

Abstract: In order to support discipline-related intelligence and literature services, this paper explores the use of a deep pre-training language model to automatically classify scientific papers for the Ministry of Education. Based on BERT and ERNIE, we constructed a literature classification model. The model was verified using a dataset that consisted of about 100,000 journal papers from 21 first-level disciplines belonging to the humanities and social sciences. We compared our model with traditional machine learning methods (such as, Naïve Bayes, Support Vector Machines) and typical deep learning methods (i.e., Convolution Neural Network and Recurrent Neural Network). The results showed that the method based on the deep pre-training language model works best, and the top-1 and top-2 accuracy of ERNIE could reach 75.56% and 89.35%, respectively. The classifier that simultaneously used the title, keyword, and abstract of the papers as the input achieved the best result. Relatively independent disciplines achieved good classification accuracy. For example, the F1 score of Sports Science was 0.98. Other disciplines demonstrated poor accuracy owing to their relatively high intersection with other disciplines. For example, the F1 score of Theoretical Economics and Applied Economics was around 0.6. In addition, this paper further discusses the topics of disciplinary intersection, model application, and optimization.

Key words: literature discipline classification; deep learning; text classification; pre-training language model

收稿日期：2019-09-02；**修回日期**：2019-11-05

基金项目：国家社会科学基金重点项目“开放科学数据集统一发现的关键问题与平台构建研究”（20ATQ007）。

作者简介：罗鹏程，男，1989年生，博士研究生，馆员，主要研究领域为学术数据挖掘、科学数据管理、开放获取等；王一博，男，1992年生，硕士，助理馆员，主要研究领域为数据挖掘、科学评价等；王继民，男，1966年生，博士，教授，博士生导师，主要研究领域为Web数据挖掘、信息可视化、科学计量学、科学评价等，E-mail：wjm@pku.edu.cn。

1 引　言

随着2015年国家发布《统筹推进世界一流大学和一流学科建设总体方案》，以及2017年发布的《统筹推进世界一流大学和一流学科建设实施办法（暂行）》，“一流学科”的建设成为国内各大高校关注的重点。“一流学科”通常指教育部《学位授予和人才培养学科目录》中的学科[1]（简称“教育部学科”），其最新版包含13个学科门类，111个一级学科。它是教育部学科评估的依据，因而也是各高校管理部门以及图书馆学科服务关注的重点。“一流学科”的建设需要一流的学科服务，当前学科服务正进入知识服务阶段[2]。在学科服务中，如学科信息门户、学科情报分析等服务内容均需要收集学科相关文献。而各种来源的文献分类体系（如《中国图书馆分类法》、中国知网和Web of Science文献分类目录）与教育部学科无法直接对应，这为许多学科相关服务带来了困难。

在学科情报服务中，如北京大学学科竞争力分析[3]和学科前沿分析[4]，均需要先将文献映射到教育部学科，然后基于映射结果展开文献计量等分析。然而，由于数据库来源众多，国内外数据库分类体系差异较大，仅部分学科能够通过类目映射（即各数据库类目映射到教育部学科）获取数据。在学科文献服务中，如学科信息门户，需要收集学科相关的文献信息，随着学科的交叉融合，一些学科的外延变得越来越大，相关文献分散在许多其他领域期刊。例如，过去海洋学科仅涵盖海洋生物、海洋化学等自然科学，而现在还包括海洋经济、海洋历史等人文社会科学，因而很难从现有的数据库文献分类体系直接获取学科相关文献[5]。在学科评估中，如教育部学科评估[6]，通常有学术论文等科研成果评价指标，然而一所学校关于某一学科的所有成果可能分布在不同的二级单位，仅仅通过学科所在院系报送材料会遗漏大量成果。例如，笔者依据来源期刊对北京大学2015年图情档所涉及二级单位分析发现，有多达20余个二级单位在图情档相关的期刊上发文。

由此可见，在涉及文献的学科服务中，基于教育部学科的文献分类成为一项重要的基础性工作。针对这一问题，在实践中通常采用数据库分类到教育部学科分类映射、期刊到教育部学科分类映射。由于数据库类目与教育部学科无法一一对应，同一期刊可能涉及多个学科。因此，这些方法都不能完全地解决文献教育部学科分类问题。在本研究中，笔者探索利用深度学习预训练语言模型对文献进行分类。与基于类目和期刊的映射方法相比，该方法在单篇文献的粒度上识别文献学科，能够在一定程度上解决映射方法粒度过粗导致的问题。

2 相关研究

文献学科分类方法总体上可以分为3类：基于映射的方法、基于引用信息的方法、基于机器学习的方法。下面将详细介绍。

2.1 基于映射的文献学科分类

在该类方法中，按照映射粒度的不同可以划分为：类目映射、期刊映射。类目映射，即先将数据库文献分类类目映射到学科，再将数据库类目下的相关文献归类到学科。学科分析和评价中常使用的Web of Science（WoS）便提供教育部学科与WoS类目的映射表[7]，很多学科相关的工作都基于该映射[8-9]。国内的主流文献数据库分类体系均参考《中国图书馆分类法》（简称《中图法》）设计[10]，因此一些研究者总结了《中图法》文献分类到学科分类的一些经验映射方法，可适用于国内数据库类目映射[11]。此外，也有研究者针对专业文献分类体系进行映射研究。例如，单连慧等[12]采用编辑距离计算医学大类学科与《医学专业分类表》分类体系类目词汇的相似度，并结合人工判断建立了两个类目体系的映射关系。期刊映射，即先将期刊分类到学科，再将期刊下所有文献归类到学科。梁瑛等[13]通过利用CALIS联合目录公共检索系统、全国期刊联合目录数据库、JCR期刊主题与期刊名结合的3种归类方式，获取ESI工程许可排名期刊与教育部一级学科的对应关系。基于映射的方式具有成本低、易操作的优点，在实践中使用的较多；然而由于映射粒度较大，通常会导致一些文献无法进行学科归属。

2.2 基于引用信息的文献学科分类

由于文献之间存在广泛的引用关系，一些研究者探索利用这一信息来分类文献。基于期刊映射的方式无法解决综合性期刊文献的学科分类，Glänzel等[14]探索利用参考文献的学科分类来识别综合性期刊文献学科的可能性，Fang[15]则以《美国国家科学院院刊》（Proceedings of the National Academy of Sciences of the United States of America，PNAS）为例，

验证了这一方法的有效性。Taheriyan[16]也尝试利用引文、共引、共同作者等信息来识别文献学科，结果显示当关系图稠密时，该方法具有较好的效果。除了分类文献外，一些研究者也利用引用信息来优化期刊学科分类，例如，Gómez-Núñez等[17]利用引文分析提高Scimago期刊和国家排名（Scimago Journal & Country Rank，SJR）中的期刊学科分类效果。基于引用信息的分类方法需要预先知道参考文献的学科类别，因此这一方法仅适用于特定场景，如综合性期刊文献分类，或者优化已有的分类。

2.3 基于机器学习的文献学科分类

机器学习方法能够从训练数据中学习特征，然后在单篇论文的粒度上识别文献学科。目前利用机器学习方法，按照我国教育部学科类别进行文献分类的研究较少，绝大多数相关研究集中在按照传统文献分类体系（如《中图法》）进行分类。王昊等[18]、杨敏等[19]、李湘东等[20]使用*K*最近邻（*K* nearest neighbor，*K*NN）、朴素贝叶斯（naive Bayes，NB）、支持向量机（support vector machine，SVM）等传统机器学习方法对来自中图分类的书目记录进行自动分类研究。王昊等[21]还使用SVM和BP（back propagation）神经网络对来自中图分类的期刊论文进行自动分类研究。除了利用传统机器学习方法，近年来也有一些研究利用最新的深度学习相关方法实现文献分类。例如，郭利敏[22]使用卷积神经网络（convolutional neural networks，CNN）对全国报刊索引文献进行分类，傅余洋子[23]使用长短期记忆网络（long short-term memory，LSTM）对中文图书进行分类。以上基于机器学习的文献分类依据均是依据《中图法》，目前仅少量研究对文献学科分类进行探索。董微等[24]基于来自计算机、航天、医学、农业、生物学5个领域的3000余篇文献数据，使用SVM对交叉学科的科技文献分类进行研究；王效岳等[25]基于人工标注的1000篇网络学术文献，使用SVM按照教育部12个学科大类对其进行分类。

从前述相关研究分析可以看出，目前利用机器学习方法实现文献教育部学科分类研究的数量不足，涉及的学科和实验数据都较少。在研究较多的文献中图法分类中，主要使用的是传统机器学习方法（如NB、*K*NN、SVM等），也有一些研究利用了典型的深度学习方法（如卷积神经网络、循环神经网络）。最近几年来，深度学习在自然语言处理中取得了较大的进展，特别是基于预训练语言模型的深度学习方法在很多自然语言处理任务中获得较大突破。本研究将以教育部21个人文社科一级学科为分类依据，通过收集近10万期刊文献数据，基于预训练语言模型的深度学习方法实现文献学科分类。与传统机器学习方法、典型深度学习方法相比，该方法的分类效果有显著的提升，能够更好地支撑学科相关情报和文献服务。

3 研究方法

3.1 问题定义

（1）文献：在本研究中，一篇文献d=(Ti, Ke, Ab)，其中Ti表示标题，Ke表示关键词，Ab表示摘要，它们均由中文字及标点符号序列组成。

（2）分类目标：教育部一级学科集合（记为$S=\{s_1,s_2,\cdots,s_k\}$），即《学位授予和人才培养学科目录》。该目录中的一级学科是教育部学科评估的依据，也是“一流学科”建设中所指的学科，因此是高校科研管理部门和图书馆学科服务关注的重点。

（3）分类模型：它是一个函数f，将文档d映射为教育部一级学科s，即$s=f(d)$。

本研究的目标在于构造一个分类函数f，实现对未知学科类别的文档d的学科类别预测。在具体实现中，将基于BERT[26]和ERNIE[27]深度预训练语言模型来构造分类函数f。ERNIE与BERT几乎完全一样，仅在预训练过程中有所不同，如无特殊说明，下文对BERT模型的描述也适用于ERNIE。

3.2 深度预训练语言模型

深度预训练语言模型是目前自然语言处理中最为前沿的研究领域，在许多自然语言处理任务中都取得了非常好的效果。同时它也是一种迁移学习（transfer learning）算法，能够在一个容易获取的大数据集上进行模型预训练，之后将学习得到的模型迁移到其他任务中，以获取更好的效果。

与常见的卷积神经网络和循环神经网络结构不同，BERT和ERNIE基于注意力机制（attention mechanism）对语言数据进行建模。在自然语言处理中，注意力机制首先在神经网络机器翻译中被引入，早期主要用来配合循环神经网络（recurrent neural network，RNN），解决序列数据建模中容易出现“遗忘”、解码器“焦点分散”的问题[28]。神经网络机器翻译采用“编码器-解码器”（seq2seq）架构，通常编码器、解码器均采用循环神经网络来

实现。2017 年，谷歌提出变换器（Transformer）模型[29]，使用自注意力机制（self-attention mechanism）在编码器和解码器中完全取代循环神经网络。该模型不仅取得了更好的翻译效果，而且与 RNN 相比，能更容易地支持并行计算。2018 年，谷歌基于 Transformer 模型的编码器，提出了 BERT 模型，取得了当时 11 项自然语言处理任务中的最好效果。下面将对 BERT 模型架构（如图 1 所示）进行介绍。

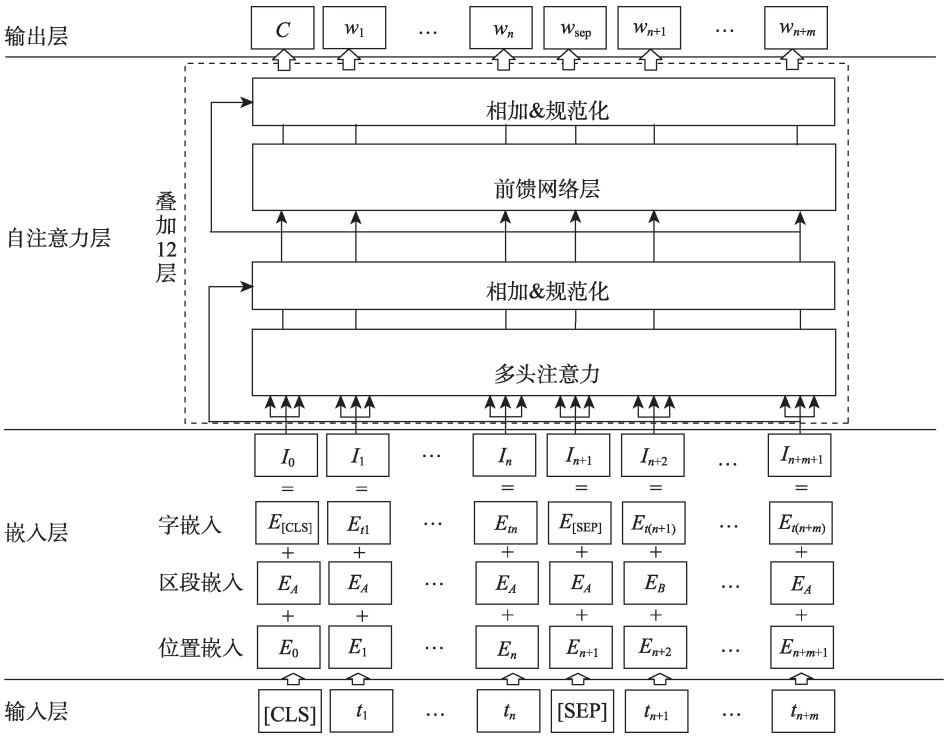

图 1　BERT 神经网络结构

1）模型输入

BERT 具有强大的特征抽取功能，在对中文进行建模时，不需要分词，而是采用以“字”为单位的输入特征。与动辄数十万，甚至上百万的词汇空间相比，以“字”为单位的优势在于特征空间大大缩减。事实上，如果模型能力足够强，完全不需要进行分词，如同人在理解文本时也没有必要先切分词汇。BERT 的输入为一个句子对（这里的句子指连续的文本片段，而非真正的句子），分别用两个特殊的标记[CLS]、[SEP]表示句子对的开始和句子对之间的分割。需要注意的是，虽然 BERT 支持句子对输入，但并不要求一定要使用两个句子，可以只输入一个句子。

2）嵌入层

在嵌入层（embedding layer）中，BERT 将每个中文字表示成一个实数向量。与一般的神经网络模型仅使用字嵌入（word embedding）不同，BERT 还包含了区段嵌入（segment embedding）和位置嵌入（position embedding）。

由于输入为句子对，因此为了区分两个句子，引入了区段嵌入：属于第一个句子的字使用 E_A 向量表示，属于第二个句子的字使用 E_B 向量表示。由于自注意力层无法区分文本中字的顺序，因此在输入自注意力层之前，需要将字的位置信息加入进去。BERT 使用位置嵌入表示词的位置信息，常将文本最大长度设置为 512。因此，从位置 0~511，分别对应 $E_0 \cdots E_{511}$ 共 512 个向量。字嵌入与通常神经网络嵌入层一样，即将每个字表示成一个单独的向量。假设不同的中文字以及一些特殊的标记字符（即[CLS]、[SEP]、[MASK]）的总数为 L，最终共有 L 个不同的字嵌入向量。

通过将每一个字对应的字嵌入、区段嵌入、位置嵌入 3 个向量相加，最终得到输入文本对应的实数矩阵 I，I 将作为自注意力层的输入。需要注意的是，嵌入层的这些不同的向量是模型的参数，可在模型训练时学习得到。

3）自注意力层

在自注意力层中，首先需要对输入矩阵 I 做如

下矩阵运算得到查询（query）Q、键（key）K、值（value）V，

$$Q=W^{q}I,\ K=W^{k}I,\ V=W^{v}I \tag{1}$$

式中，W^{q}、W^{k}、W^{v}均为参数，在模型训练时学习得到。有了Q、K、V之后，再做如下注意力运算，

$$\text{Attention}(Q、K、V)=\text{softmax}(QK^{T}/d^{1/2})V \tag{2}$$

式中，d为Q中向量的维数；softmax为神经网络中常用归一化函数。以上注意力模型本质是上对输入I进行线性变换得到Q和K，进而通过softmax运算得到一个权重矩阵，再对V加权求和。softmax得到权重矩阵中各个取值的大小，即反映模型对各个“字”分配的注意力大小。

为了使得模型能够学习到多种不同的注意力，BERT使用了多头注意力（multi-head attention）模型，即对Q、K、I分别做多个不同的线性变换（乘以不同的参数矩阵），得到不同的Q_i、K_i、V_i，再分别进行Attention函数运算，最后把不同头（head）得到结果进行拼接，并做线性变换（乘以一个参数矩阵）得到多头注意力层的最终输出。

在“相加&规范化”（Add&Norm）层，BERT将多头注意力层的输入和输出相加，然后做规范化。规范化的结果再进入到全连接的前馈网络层，这一层的输出和输入再做一次相加，然后做规范化。规范化的结果作为整个自注意力层的输出。在BERT中，会对自注意力层进行多次堆叠。在谷歌的中文BERT预训练模型中，堆叠了12层。

4）模型输出

最后一个自注意力层的输出将作为整个BERT模型的最终输出。其中输入[CLS]对应的第一个输出C用于模型预训练中的分类任务，其余各个字对应的输出用于模型预训练中的字预测。

3.3 文献学科分类模型

本研究使用图2所示的神经网络结构进行文献学科分类。在输入层中，可将文档d=(Ti, Ke, Ab)的标题、关键词、摘要单独或者组合（拼接）在一起作为输入；在输入层之后接BERT（或ERNIE）模型；然后加入一层包含|S|个节点的全连接层，最后再接softmax层计算各类别的概率分布。该模型的训练分为两个步骤：①基于大量无标注中文文本数据，对图1所示的网络结构进行“预训练”；②以预训练学习得到的参数初始化图2中的BERT（或ERNIE）层网络结构，然后基于文献学科分类任务数据集，对图2中整个网络结构中的参数进行调优。

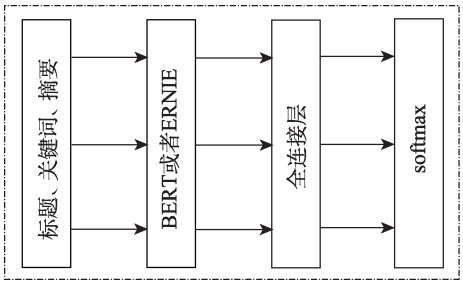

图2 基于深度预训练语言模型的文献学科分类神经网络结构

3.3.1 模型预训练

基于BERT模型进行文献学科分类的一大优势在于，可以使用大量无标注的文本数据进行模型参数的学习。预训练过程在图1所示的神经网络结构上进行，包含两个部分：掩码语言模型（masked language model，Masked LM）和下一个句子预测（next sentence prediction，NSP）。

1）掩码语言模型

掩码语言模型的训练类似于完形填空。在输入层，BERT随机掩盖掉15%的字。对于掩盖的文字，BERT分别以80%、10%、10%的概率将其表示为特殊标记[MASK]、随机选择一个字、保持原来的字不变；在输出层，BERT要求模型在对应位置预测被掩盖掉的字。通过这种方式，BERT能够学会文字之间隐含的语义关系。

ERNIE对BERT的改进仅在掩码语言模型训练阶段。在BERT中，掩码是完全随机的，因此对于一个短语或者命名实体，如“内蒙古”，BERT掩码可能只是把“古”字给掩盖掉，这使得预测变得相对容易。在ERNIE中，将掩码训练分成3个阶段：基本级掩码、短语级掩码、实体级掩码。基本级掩码与前述BERT掩码方式一样，在短语级和实体级掩码训练阶段，ERNIE会将一个短语或命名实体中的所有字掩盖掉，如“内蒙古”3个字会一起被掩盖掉。这使得ERNIE需要结合上下文预测整个短语或者命名实体。

2）下一个句子预测

在很多自然语言处理任务（如问答任务）都需要理解两个句子之间的语义关系，而在掩码语言模型训练中并不能直接捕获这种关系。因此，BERT构造句子对(A,B)作为输入。在句子对中，B以50%的概率为A的下一个句子，以50%的概率从语料中随机选择一个句子。在输出层，C为分类预测结果（是否为下一个句子）。通过加入该预训练步骤，问答任务的效果有明显提升。

3.3.2　模型调优

BERT 模型可以在大量无标注中文文本数据上进行预训练，这一过程耗时耗力。一旦预训练步骤完成，便可以针对不同的任务做调优，这一过程相对轻松。在本研究的文献分类问题中，将使用构造的文献学科数据集，基于图 2 所示的神经网络结构进行模型调优。

BERT 具有非常强的特征抽取能力，有研究表明 BERT 各层特征抽取功能类似一个传统自然语言处理的流水线。越低层输出的特征越适合于初级的自然语言处理任务，如词性标注；越高层输出的特征越适用于高级的自然语言处理任务，如语义角色、指代分析[30]。因此，通过在 BERT 最高层输出后面接入简单的全连接层和 softmax 层，便可以实现很好的分类效果。

4　实验设计

本实验包括 3 个步骤：实验数据的构建、分类模型的实验设置、分类效果的比较与分析。为了对比分析基于深度预训练语言模型的文献学科分类效果，本研究还选择了传统机器学习方法和两种典型的深度学习方法做比较。

4.1　实验数据构建

本实验的目的在于自动实现文献的教育部一级学科分类（即《学位授予和人才培养学科目录》中的一级学科）。因此，本实验选择中国社会科学引文索引（Chinese Social Sciences Citation Index，CSSCI），并按照高级检索中的“学位分类”获取数据。CSSCI 构建得比较早，所以其“学位分类”中的学科与 2018 年最新版教育部一级学科分类略有差别（本实验使用的学科中，有 2 个原先的一级学科分裂为多个一级学科）。由于按照最新版教育部学科分类构造训练数据集工作量会较大，所以本研究仍旧按照 CSSCI 中所使用的老版本学科分类来进行实验。实验数据详细构造步骤如下。

（1）使用“中文社会科学引文索引”获取人文社科一级学科对应的文献列表。编写爬虫程序，在“中文社会科学引文索引”高级检索页面，按照一级“学位分类”（遍历所有一级学位分类下拉选项）逐年（选择 2014—2018 年，共 5 年）检索，获取每个检索结果返回的所有文献列表（最多返回 1000 条），并提取文献的标题。

（2）利用百度学术补充文献信息。编写爬虫程序，以文献标题作为关键词，在百度学术中进行检索。当百度学术能够完全匹配时，将直接返回文献的详情页面，从其中可以提取文章的关键词和摘要。对于不能够完全匹配的文献，则直接丢弃，不进入数据集。

（3）训练数据提取与准备。提取每篇文献的标题、关键词、摘要以及对应的一级学科，删除摘要为空、非中文描述的文献，共得到 90514 篇文献，各一级学科对应的文献样本数量如表 1 所示。对于文献数量较少的军事相关一级学科（军事思想及军事历史、军队政治工作学、军事后勤学与军事装备学、军队指挥学、战略学、战术学）直接删除，不进入训练集。使用随机数按照 4∶1 将数据集划分为训练集和测试集，最终训练集文献数量为 72106，测试集文献数量为 18082，表 1 给出了各一级学科对应的训练集和测试集样本数量。

此外，表 1 还给出了每个一级学科文献分布期刊的数量。可以看出，各学科文献分布的期刊数量较多，对该学科具有较好的代表性。以图书馆、情报与档案管理为例，图 3 给出了文献分布数量最多的前 20 个期刊，可以看出这些期刊几乎都是图情档学科常见的核心期刊。由于 CSSCI 数据库中针对单篇论文按照一级学位分类，因而图情档学科文献除了主要分布在本学科相关的核心期刊中，还有少部分文献分布在其他领域核心期刊中。

4.2　分类模型实验设置

为了对比本实验使用的基于预训练语言模型的深度学习分类效果，笔者还选择了传统的机器学习方法（SVM、NB 等）、典型深度学习方法（CNN、RNN）进行比较。各个模型首先在训练数据集上进行训练，然后在测试集上进行模型效果评价。下面分别对 3 类模型训练时相关的设置进行描述。

1）基于深度预训练语言模型学科分类实验设置

模型预训练。BERT 和 ERNIE 的预训练需要大量的 GPU 计算资源，训练时间长，成本高昂。因此，本实验直接使用 BERT 和 ERNIE 已经发布的预训练模型，作为学科分类模型中 BERT 和 ERNIE 层初始的参数。本实验使用包含 1.1 亿个参数的“BERT-Base, Chinese”预训练模型，该模型在中文维基百科上训练得到，并使用同样模型规模的“ERNIE 1.0 中文 Base 模型（max_len=512）”预训练模型，该模型在百度百科、百度知道、百度贴吧

表 1 数据集构成情况

一级学科	训练集样例数	测试集样例数	总数样例数	分布期刊数	一级学科	训练集样例数	测试集样例数	总数样例数	分布期刊数
应用经济学	3879	943	4822	278	中国语言文学	3660	931	4591	200
图书馆、情报与档案管理	3853	957	4810	107	心理学	3611	964	4575	180
工商管理	3841	963	4804	254	哲学	3405	902	4307	221
体育学	3840	960	4800	66	政治学	3468	832	4300	251
法学	3886	890	4776	275	民族学	2412	586	2998	129
社会学	3762	1006	4768	284	环境科学与工程	1571	416	1987	169
新闻传播学	3796	966	4762	194	农林经济管理	1054	269	1323	171
公共管理	3814	925	4739	310	军事思想及军事历史			261	
管理科学与工程	3721	933	4654	225	军队政治工作学			25	
历史学	3761	893	4654	230	军事后勤学与军事装备学			16	
理论经济学	3721	925	4646	300	军队指挥学			13	
教育学	3728	910	4638	181	战略学			9	
外国语言文学	3696	940	4636	176	战术学			2	
艺术学	3627	971	4598	149					

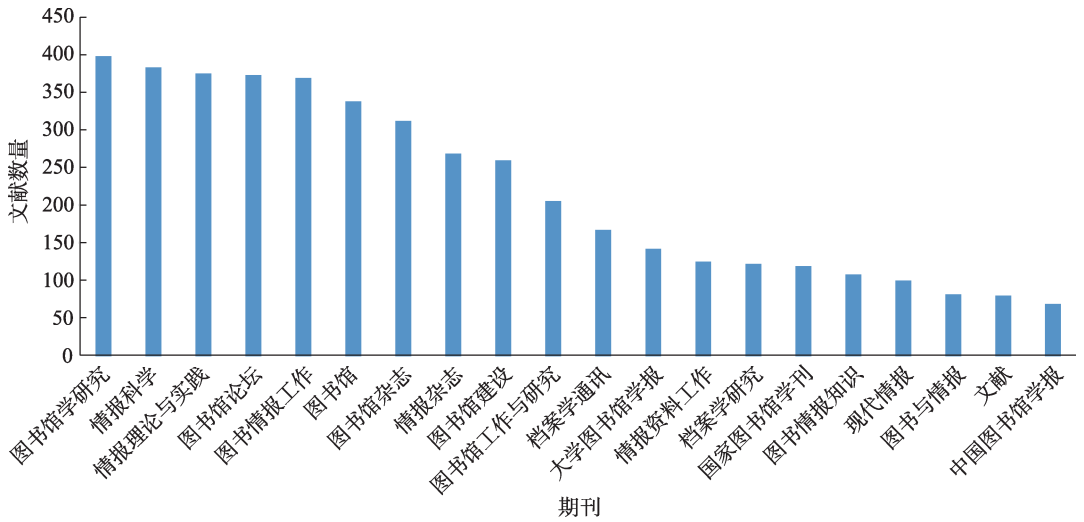

图 3 "图书馆、情报与档案管理"文献分布数量最多的前 20 个期刊

数据集上训练得到。

模型调优。在训练和预测的过程中，由于标题和关键词比较短，所以实验中采用“摘要”“摘要+关键词”“摘要+标题”“摘要+标题+关键词”4 种输入方式。在训练数据集上进行调优时，允许 BERT 和 ERNIE 模型进行参数调整。本实验使用的 BERT 模型基于 TensorFlow 实现[31]，ERNIE 模型基于 PaddlePaddle 的实现[32]。

2）传统机器学习模型学科分类实验设置

本实验选用 scikit-learn[33] 中提供的多种传统机器学习方法进行文献学科分类，包括朴素贝叶斯（BernoulliNB、MultinomialNB、ComplementNB）、决策树（DecisionTreeClassifier）、*K* 最近邻（*K*NeighborsClassifier）、逻辑回归（LogisticRegression、LogisticRegressionCV）、支持向量机（LinearSVC、SVC），括号里为 scikit-learn 对相应算法实现的类名（含不同变体）。传统机器学习方法主要使用向量空间模型表示文本。在本研究中首先使用 jieba[34] 对标题、关键词和摘要进行分词，然后基于文档频率筛选特征词，之后将标题、关键词、摘要文本分别表示为 TF-IDF 向量。使用“摘要”“摘要+关键词”“摘要+标题”“摘要+关键词+标题”4 种不同方式获取组合特征：对摘要、关键词、标题做加权，加权后的向量再规范化为单位向量。

3）典型深度学习模型学科分类实验设置

在深度学习中，典型的神经网络结构为卷积神

经网络（CNN）和循环神经网络（RNN）。CNN 主要应用于计算机视觉领域，在自然语言处理中，Kim[35]提出了基于 CNN 的文本分类方法 TextCNN，本实验将使用 TextCNN 进行文献学科分类。RNN 适用于处理序列数据，在自然语言处理中有着广泛应用。RNN 有多种变形，本实验尝试了简单形式的 RNN、长短期记忆网络（LSTM）、门控循环单元网络（gated recurrent unit，GRU），最终选择效果较好的 GRU 报告文献学科分类效果。

在神经网络输入层需要将词表示为向量，通常使用词嵌入（word embedding）方式将词映射到低维、稠密的向量空间。词嵌入层在模型训练开始前可以随机初始化，也可以使用已有的预训练词向量进行初始化。由于监督学习的训练数据有限，不太容易获得较好的词向量。在实际中通常使用非监督学习方法在较大的语料库中预先训练词向量，然后将预训练好的词向量应用于下游任务。本实验将使用已有的预训练中文词向量，包括腾讯公司 Song 等[36]、脸书 Grave 等[37]、北京师范大学 Li 等[38]提供的预训练中文词向量。

4.3　效果比较与分析

下面对各种模型的分类效果做对比分析。在效果评价中，使用测试集上学科分类的准确率（accuracy）来衡量，即正确预测学科类别的文献数量占总文献数量的比重。然后，选择效果最优的 ERNIE 模型，对其分类结果做进一步分析。

4.3.1　模型分类效果分析

图 4~图 6 分别给出了基于深度预训练语言模型、典型深度学习方法、传统机器学习方法在文献学科分类上的效果。从图 6 可以看出，效果最好的传统机器学习分类方法为非线性的支持向量机（SVC），当同时使用“摘要+标题+关键词”时，能够达到 71.54% 的准确率。从图 5 可以看出，在典型深度学习方法中 CNN 效果较好，当同时使用“摘要+标题+关键词”以及北师大词向量时效果最优，准确率达到 72.73%。从图 4 可以看出，基于深度预训练语言模型的最好分类效果为 75.56%，显著高于传统机器学习方法和典型的深度学习方法。可见，该类方法在文献学科分类中具有明显的优势。

从图 4 还可以看出，ERNIE 的分类效果较 BERT 高出约 1%，可见增加了短语和实体级掩码预训练的分类效果有显著提升。同时，随着输入信息的增加，模型的分类效果也在提升。仅使用“摘要”的

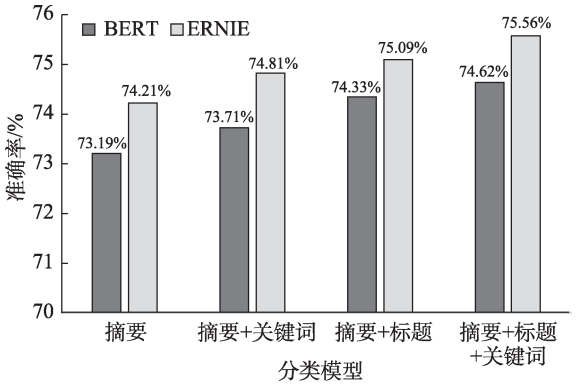

图 4　基于深度预训练语言模型的学科分类准确率

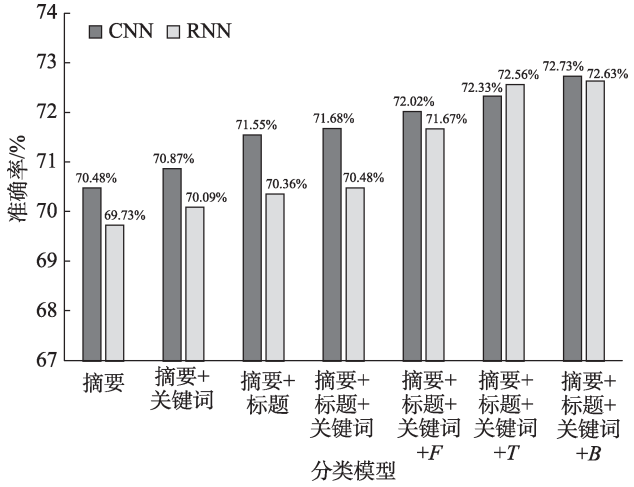

图 5　典型深度学习分类方法的学科分类准确率

F 表示脸书词向量，*T* 表示腾讯词向量，*B* 表示北师大词向量。

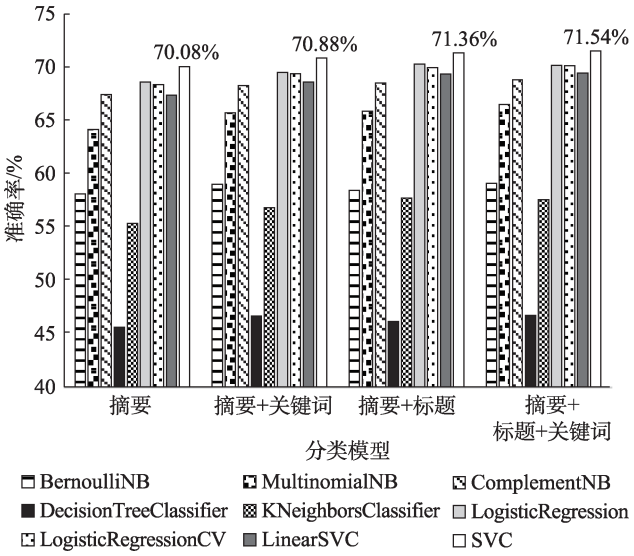

图 6　传统机器学习方法的学科分类准确率

ERNIE 模型分类准确率为 74.21%；当同时使用“摘要+标题+关键词”作为输入时，分类效果最优，达到 75.56%。对比“摘要+关键词”“摘要+标题”的

分类效果可以发现，标题相对于关键词而言，更能提升分类效果。

4.3.2 ERNIE 分类结果分析

从前面的分析可以看出，同时使用“摘要+标题+关键词”的 ERNIE 分类效果最优。本小节将对 ERNIE 在测试集上的分类结果进行详细分析，以深入了解其文献学科分类效果。

1）各学科分类效果

表 2 给出了 ERNIE 在测试集上各个学科分类预测的情况。其中，第一行的学科表示测试集中文献的正确学科类别，第一列表示 ERNIE 预测的学科类别。表格中，单元格表示预测类别在各个正确类别上的文献数量分布。例如，单元格（法学,法学）=754、（法学,环科）=3 表示 ERNIE 预测为法学的文献中，有 754 个正确预测为法学，有 3 个错误预测为环科。从表 2 可以看出，对角线上的数值较大，这些数值相加除以测试集文献总数即为准确率 75.56%。

为了更加清晰地看出各个学科的分类效果，表 3 给出了各个学科分类的精确率（precision）、召回率（recall）和 F1 值（F1 score）。从表 3 可以看出，体育学分类效果最佳，F1 值高达 0.9802；外国语言

表 2 ERNIE 在测试集上的分类预测情况

预测	正确																				
	法学	工商	公共	管科	环科	教育	理经	历史	民族	农经	社会	体育	图情	外语	心理	新传	艺术	应经	哲学	政治	中文
法学	754	4	31	5	3	0	11	12	9	3	11	0	1	0	2	8	1	12	5	11	0
工商	6	687	2	110	10	2	52	1	6	1	12	0	4	0	7	1	1	93	1	2	0
公共	35	3	576	37	15	12	31	1	14	5	88	0	7	0	2	14	0	50	1	75	0
管科	5	134	15	598	7	7	19	3	2	7	14	1	51	0	29	13	2	42	16	2	1
环科	6	6	23	10	301	0	25	1	5	9	12	0	2	0	2	0	0	21	0	8	0
教育	5	2	49	12	1	794	2	6	6	0	17	1	4	7	19	7	4	3	7	23	3
理经	5	26	18	27	36	3	562	17	5	3	23	0	0	1	5	0	0	134	12	30	1
历史	8	2	3	2	0	4	15	665	35	0	26	0	13	6	1	9	21	3	19	31	21
民族	3	4	41	1	2	3	8	29	396	3	87	0	0	2	3	2	4	7	14	18	10
农经	1	7	7	4	8	0	17	0	1	212	13	0	2	0	0	0	0	54	0	1	0
社会	4	8	63	10	10	8	20	11	40	6	557	3	1	1	22	12	8	24	11	33	4
体育	2	2	0	1	1	3	0	0	2	0	3	944	0	0	2	3	0	1	2	0	0
图情	7	3	3	34	1	11	0	8	3	0	4	0	816	2	1	73	2	0	3	1	9
外语	0	0	1	0	0	14	0	5	4	0	5	0	3	862	2	2	14	0	7	3	59
心理	0	8	4	19	0	11	1	0	2	0	21	6	0	3	827	4	1	4	13	2	4
新传	9	4	16	14	1	5	2	8	3	1	13	1	39	6	11	768	20	8	5	6	11
艺术	1	0	1	2	0	5	1	15	19	0	15	1	1	7	0	21	863	0	12	4	29
应经	7	58	20	34	10	0	135	2	4	18	22	2	1	1	2	3	1	482	1	6	1
哲学	5	1	5	9	2	3	10	48	15	1	17	1	5	6	21	7	13	1	724	49	26
政治	27	4	46	4	8	19	14	35	7	0	35	0	1	0	5	7	0	3	32	526	3
中文	0	0	1	0	0	6	0	26	8	0	11	0	6	36	1	12	16	1	17	1	749

注：为便于显示，对学科名称进行了简化。各简化名称对应的全称分别为：体育-体育学；外语-外国语言文学；艺术-艺术学；心理-心理学；法学-法学；教育-教育学；图情-图书馆、情报与档案管理；中文-中国语言文学；新传-新闻传播学；哲学-哲学；历史-历史学；农经-农林经济管理；环科-环境科学与工程；工商-工商管理；政治-政治学；民族-民族学；管科-管理科学与工程；理经-理论经济学；公共-公共管理；社会-社会学；应经-应用经济学。

表 3 ERNIE 在测试集上的各个学科分类效果

	体育	外语	艺术	心理	法学	教育	图情	中文	新传	哲学	历史
精确率	0.9772	0.8787	0.8656	0.8892	0.8539	0.8169	0.8318	0.8406	0.8076	0.7472	0.7523
召回率	0.9833	0.9170	0.8888	0.8579	0.8472	0.8725	0.8527	0.8045	0.7950	0.8027	0.7447
F1 值	0.9802	0.8974	0.8770	0.8733	0.8505	0.8438	0.8421	0.8222	0.8013	0.7740	0.7485
	农经	环科	工商	政治	民族	管科	理经	公共	社会	应经	
精确率	0.6483	0.6984	0.6884	0.6778	0.6217	0.6178	0.6189	0.5963	0.6507	0.5951	
召回率	0.7881	0.7236	0.7134	0.6322	0.6758	0.6409	0.6076	0.6227	0.5537	0.5111	
F1 值	0.7114	0.7108	0.7007	0.6542	0.6476	0.6291	0.6132	0.6092	0.5983	0.5499	

注：学科名称简化信息同表 2。

文学、艺术学、心理学、法学也有较好的预测效果，F1 值均在 0.85 以上。相对而言，应用经济学、社会学分类效果较差，F1 值低于 0.6。从表 2 可以看出，应用经济学和理论经济学混淆程度很高，有大量相互错误分类的文献，ERNIE 预测的应用经济文献中有 0.1667 的比例为理论经济学；社会学和公共管理的混淆度也较高，社会学的文献有 0.0875 的比例被错分为公共管理。

从以上分析可以看出，一些学科由于相互之间的相似度较高、容易混淆，造成分类效果较差。为了展示测试集中各个学科之间相似程度，笔者依据测试集中摘要数据，对学科进行层次聚类。具体步骤为：对测试集中各文献摘要使用规范化的 TF-IDF 表示，将各个学科下的文献向量求均值得到该学科的向量表示，然后进行全连接的凝聚层次聚类，聚类结果如图 7 所示。从图 7 可以看出，理论经济学、应用经济学相似度最高，最先被聚类在一起；其次是公共管理和社会学，工商管理和管理科学与工程，政治学也较早地被聚类，反映出这些学科的相似度较高，从表 3 中也观察出它们的分类效果确实不佳。此外，从图 7 可以看出，基于 TF-IDF 表示的文、史、哲 3 个学科也比较相似，但是 ERNIE 模型较好地表示了这 3 个学科的特征，能够获得稍好的分类效果。

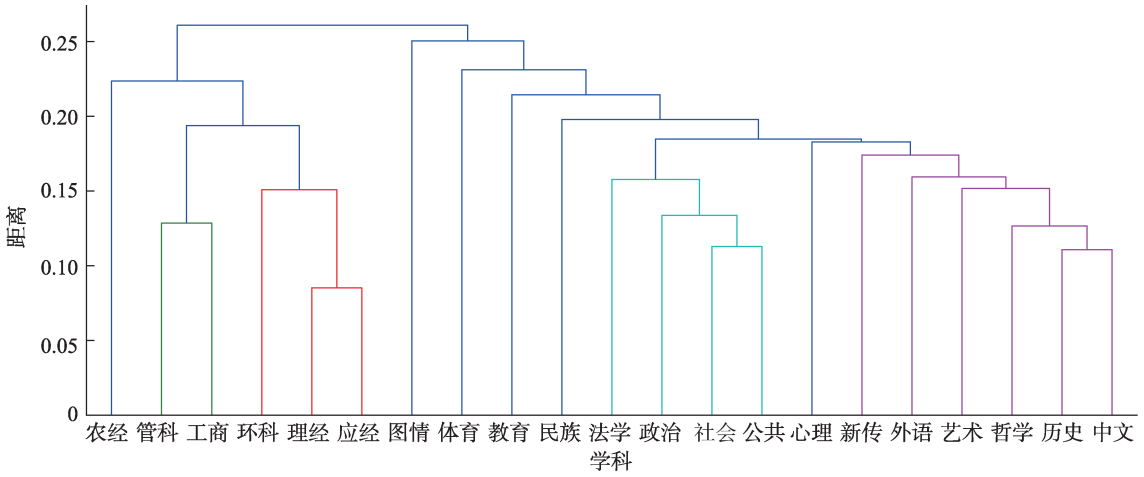

图 7　测试集上的学科层次聚类

学科名称简化信息同表 2。

2）Top *N* 准确率

加入了 softmax 层的神经网络可以输出各个类别的预测概率，依据预测概率的高低可以获得分类结果。前文的分析仅给出了预测概率最高的学科的准确率（即 Top 1 准确率），事实上，可以通过统计 Top *N* 个预测学科中是否包含真实学科的比例来进一步分析 ERNIE 的分类效果。图 8 给出了 ERNIE 在测试集上 Top 1~Top 5 的分类准确率。从图 8 可以看出，Top 2 的准确率到达 89.35%，而 Top 5 的准确率更是高达 97.81%。也就是说，对于每篇待分类的文献，如果分类器给出概率最高的 2 个学科，那么这 2 个学科中包含正确学科的概率可达到约 90%。

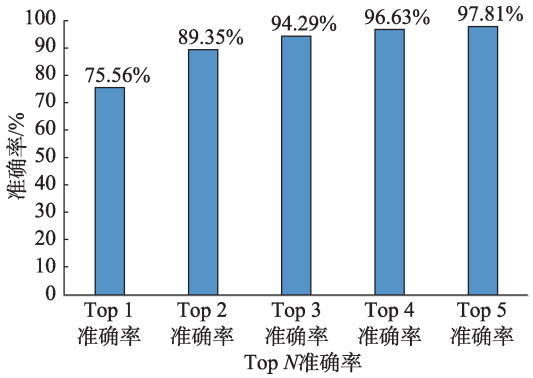

图 8　ERNIE 在测试集上 Top *N* 分类准确率

由于测试集中文献仅给出了一个正确的学科类别，事实上一些文献确实可以属于多个学科。表 4 给出了从测试集中随机选取的 5 个 Top 1 预测学科错误的文献，并给出了这些文献的标题、真实学科（*T*）、预测的前 5 个学科（*P*）。从这些样例可以看出，分类器预测的第一个学科不一定就是完全错误。第 2 条、第 4 条样例分别来自图书情报、教育学的期刊，因此预测的第一个学科事实上是正确的，只不过在 CSSCI 中也为它们赋予了其他学科，而本实验收集的数据集中恰好使用了这些其他学

表 4 随机选取的错分类样例

(1)标题:管理者任期、内部控制与战略差异 [T:管理科学与工程][P:工商管理, 管理科学与工程, 应用经济学, 理论经济学, 公共管理]
(2)标题:社交网络学术活动特征分析及应对策略 [T:新闻传播学][P:图书馆、情报与档案管理, 新闻传播学, 管理科学与工程, 社会学, 工商管理]
(3)标题:基于多目标遗传算法的水资源优化配置研究——以徐州市为例 [T:应用经济学][P:理论经济学, 环境科学与工程, 应用经济学, 管理科学与工程, 农林经济管理]
(4)标题:高校创业教育学院化现象分析与治理对策 [T:公共管理][P:教育学, 公共管理, 社会学, 政治学, 管理科学与工程]
(5)标题:维吾尔族流动人口研究综述 [T:社会学][P:民族学, 社会学, 公共管理, 历史学, 心理学]

科。第 1、第 3、第 5 条样例来自相对综合性的期刊，ERNIE 预测的第一个学科或多或少与其主题有一定的关系，例如，第 5 条样例作者给出的中图分类号为“C924.24”，其与民族学分类“C95”很接近。由此可见，ERNIE 在测试集上真实的 Top 1 准确率应该高于 75.56%。

5 讨 论

下面将从学科交叉对分类效果的影响、学科分类应用场景以及分类模型效果优化 3 个方面对基于深度预训练语言模型的文献学科分类进行讨论。

5.1 学科交叉对分类效果的影响

学科之间不是泾渭分明，而是既有相互独立，又有交叉融合。当前，学科间的交叉融合已经成为科学发展创新的动力源泉之一。对于文献学科分类来说，由于学科之间存在交叉融合，很难只将一篇文献归属到一个学科。为了更好地理解机器分类的效果，需要对不同学科间的交叉和融合有更多的了解。为此，笔者基于不同学科在期刊上的发文数量分布的相似性来表示它们的交叉性。假设学科 s_i 在期刊 p_k 上的发文量为 sp_{ik}，定义如下指标表示学科 s_i 和 s_j 的交叉性，也即学科在期刊上发文数量分布向量的余弦相似度，

$$\mathrm{sim}(s_i,s_j)=\frac{\sum_k(\mathrm{sp}_{ik}\times\mathrm{sp}_{jk})}{\sqrt{\sum_k(\mathrm{sp}_{ik})^2}\times\sqrt{\sum_k(\mathrm{sp}_{jk})^2}} \tag{3}$$

如果两个学科在不同期刊上的发文分布相同，则它们的交叉性为 1；如果两个学科完全不在相同期刊上发文，则它们的交叉性为 0。图 9 给出了在整个数据集上 21 个科学间交叉性计算结果，图中小方格的颜色表示学科交叉性的高低。

图 9 中学科与表 3 分类效果高低顺序一致。从图 9 可以看出，从左上角向右下角方向（除对角线外），整体来看小方格的颜色越来越明亮，这反映出，按照这一学科顺序，不同学科之间更多地在相同期刊上发文。在图 9 中，体育学、外国语言文学、艺术学、心理学对应的几列颜色都很深，反映出它们很少与其他学科在相同的期刊上发文，学科独立性较高。在所有方格中，理论经济学与应用经济学的交叉性最高，达到 0.73，反映出它们常常在相同的期刊上发文。观察图 9 中“图情”一列可以发现，它整体颜色都比较暗，但是与“新传”“管科”交叉的方格颜色较浅，取值分别为 0.17、0.13，这一结果与我们直观感受也很接近，在一定程度上验证了以上指标的有效性。通过对图 9 的分析可以看出，学科的交叉性对分类效果有很大影响，越是相对独立的学科，机器分类的效果越好；越是交叉度高的学科，机器分类的效果越差。

5.2 基于深度预训练语言模型的文献学科分类应用

文献分类本身具有很强的主观性，特别是对于综合、交叉的文献，分类更是难以达成一致。笔者从测试集上随机筛选出来自“综合社科期刊”“高校综合性学报”且 ERNIE 分类错误的 100 篇文献，并从知网、万方、维普、论文作者处获取他们给出的中图分类号。分析发现，在这 100 篇论文中，有 47 篇论文知网、万方、维普和作者给出的一级大类不一致，仅有 16 篇论文的分类号完全一样。这说明，即便是人工分类，对于同一篇论文给出的分类结果也会存在非常大的差异。造成这一现象的原因有多方面。从客观角度来看，论文本身可能涉及多个学科，因而确实也可以赋予不同的分类类目；从主观角度来看，每个人知识结构不同，理解文献内容的视角也不同，因此选择类目时也会存在偏颇和局限。将 ERNIE 预测的第一个学科类别与知网、万

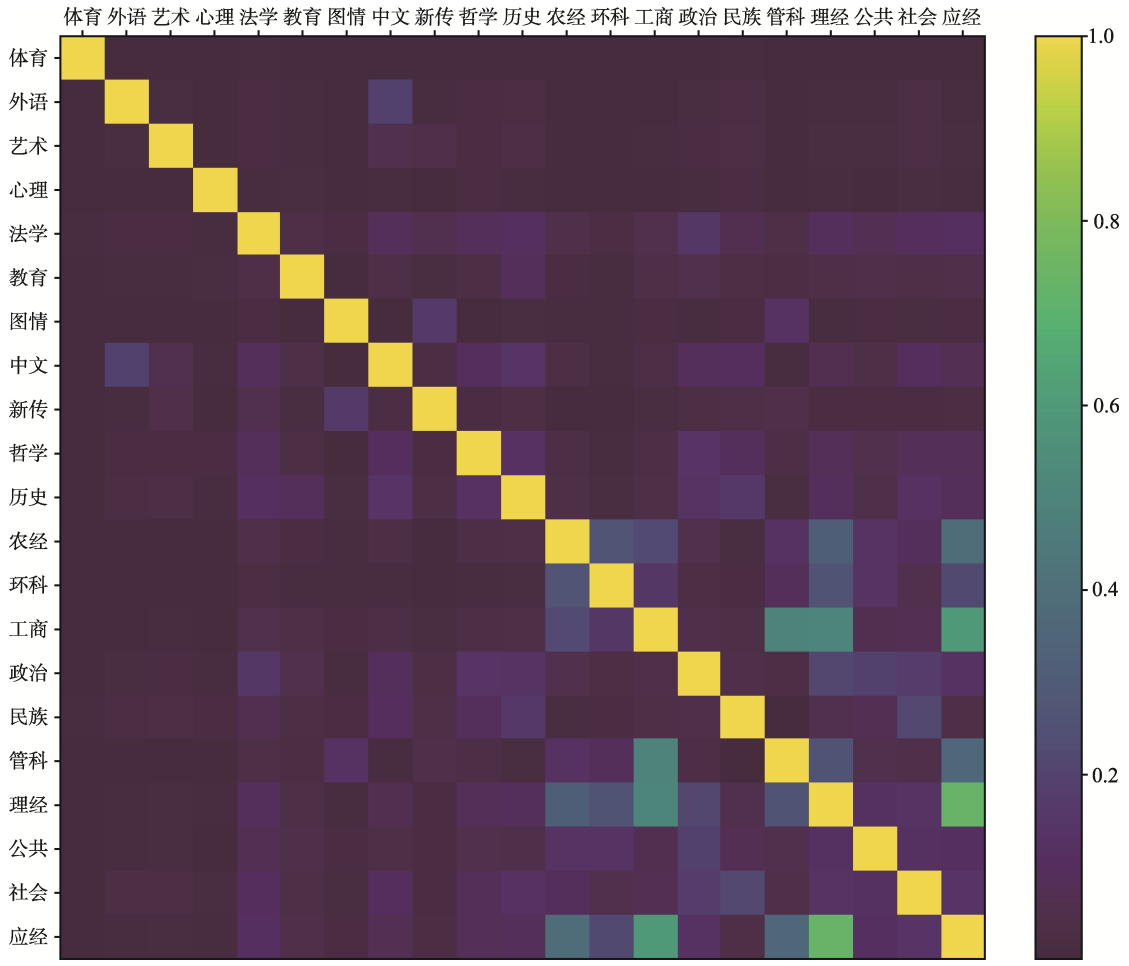

图 9　学科之间的交叉性(彩图请见 http://qbxb.istic.ac.cn/CN/volumn/home.shtml)

学科名称简化信息同表 2。

方、维普和作者给出的分类号相比较，有 46 篇文献的第一个预测学科与 4 个来源的分类号中至少 1 个有直接的较强联系（如 G641 与教育学，D61 与政治学），其他大多数也或多或少有一定关系。这反映出，即使是目前所谓“错误”预测学科类别的文献，实际上它的分类也有较大可能性是正确的或至少是具有参考价值的。因此，使用深度预训练语言模型给出的学科分类预测列表，具有很大的参考价值，可以辅助人工分类，克服人工分类的局限性。

与基于映射的方法相比，基于深度预训练语言模型的文献学科分类结果具有不可解释性，因此该类方法适用于对分类精度有一定容忍性的应用场景，或者可作为一种辅助人工的手段进行应用。对于如学科评价等对学科分类要求高的应用场景，可以将机器学习分类作为一种辅助手段。对于能够完全属于某一学科的期刊，则可将该期刊下的文献全部归类到该学科；而对于不属于该学科的期刊，则可以利用机器分类，尽可能筛选出与该学科相关的文献，之后再做人工审核。对于学科信息门户，也可以通过将期刊映射和机器分类相结合的方式，来尽可能多地获取学科相关信息。除了以上学科相关应用场景，通常的文献信息检索系统也可以应用该类方法。当前，图书情报机构都在不断积累自己的数字馆藏，例如，各高校和科研院所都在建立机构知识库，一些机构还建立了学术发现系统，如中国科学院文献情报中心的“科技大数据知识发现平台”[39]。目前这些系统平台缺乏以学科为中心的检索浏览机制，通过利用基于深度预训练语言模型的文献学科分类方法，可以为这类检索平台提供学科过滤的功能。

5.3　基于深度预训练语言模型的文献学科分类效果优化

通过前面的研究分析可以发现，当引入更多信息字段时，模型的分类效果更佳（本研究中同时使

用标题、关键词和摘要的效果最佳）。对于一篇文献的元数据信息，除了标题、关键词和摘要之外，还可以利用参考文献中论文的标题及其学科分类。从第 2 节的相关研究可以发现，已有研究者利用非监督的方式来分类文献[14-17]。通过引入参考文献论文标题，有可能提高机器分类效果。机器分类方法可以为每篇文章赋予一个初始的学科类别，之后也可以利用引文网络信息来优化分类结果。

深度学习方法的优势在训练数据集较大的情况下才能更容易地发挥出来，模型规模越大，在小数据集上越容易过拟合，因此效果可能欠佳。目前本研究使用的训练数据仅为数万篇文献，每个学科类别仅有不超过 4000 篇的训练样本，因此本研究基于深度预训练语言模型方法的效果可能还未完全释放出来。未来，通过采集更多的学科训练数据，模型的分类效果将有可能得到进一步的提升。在数据采集时，可以先筛选出每个学科具有代表性的核心期刊，再通过这些期刊去选择文献，这样可以收集到各个学科的训练数据，而不是仅限于本研究的 20 多个一级学科，当然这一工作量会更大。

除了增加训练数据集的规模，还可以通过在领域文本上做预训练来提高分类效果。本研究使用的 ERNIE 模型是在由中文维基百科、百度百科、百度新闻和百度贴吧组成的数据集上训练得到的，而中文 BERT 模型也是在维基百科上训练得到的。显然这些数据集与科学文献还是存在一定的差异，目前已经有研究者在英文科学文献上进行语言模型的预训练，并在相应的下游任务中取得了更好的效果，如 SciBERT[40]。未来，可以在大规模中文科学文献上进行语言模型的预训练，以进一步提高中文文献学科分类效果。

参 考 文 献

[1] 教育部. 学位授予和人才培养学科目录(2018 年 4 月更新)[EB/OL]. [2019-10-22]. http://www.moe.gov.cn/s78/A22/xwb_left/moe_833/201804/t20180419_333655.html.

[2] 肖琨. 支持“双一流”建设的高校图书馆服务创新趋势研究[J]. 大学图书馆学报, 2018, 36(5): 43-51.

[3] 吴爱芝, 肖琨, 张春红, 等. 基于文献计量的高校学科竞争力评估方法与体系[J]. 大学图书馆学报, 2018, 36(1): 62-67, 26.

[4] 北京大学图书馆. 北京大学科学研究前沿(2018 年版)[EB/OL]. [2019-08-18]. https://www.lib.pku.edu.cn/portal/cn/fw/kyzc/zhishichanquan.

[5] 马芳珍, 李峰, 肖琨. 基于知识服务的海洋学科门户建设[J]. 大学图书馆学报, 2018, 36(3): 46-51.

[6] 学位中心关于第四轮学科评估成果及人员归属说明[EB/OL]. [2019-08-18]. http://yjs.jlict.edu.cn/show.aspx?id=476&cid=50.

[7] CSSC category to Web of Science category mapping 2012[EB/OL]. [2019-08-18]. http://help.incites.clarivate.com/inCites2Live/filterValuesGroup/researchAreaSchema/chinaSCADCSubjCat.html.

[8] 蔺梅芳, 刘静. 基于 InCites 学科映射的一级学科文献计量分析——以电子科技大学为例[J]. 四川图书馆学报, 2015(3): 71-73.

[9] 刘虹, 徐嘉莹. 上海市高校学科国际影响力评价——基于 InCites 数据库学科映射的文献计量分析[J]. 复旦教育论坛, 2014, 12(4): 29-34.

[10] 刘文娟. 国内三种期刊数据库学科分类之比较[J]. 中国信息化, 2019(1): 83-84.

[11] 詹萌. 学科(专业)分类与文献分类之间的映射关系研究[J]. 情报理论与实践, 2013, 36(10): 40-43, 35.

[12] 单连慧, 赵迎光, 钱庆. 基于词汇相似度的医学分类体系映射研究与实现[J]. 医学信息学杂志, 2016, 37(11): 46-50.

[13] 梁瑛, 邹小筑. ESI 工程类与中国教育部学科分类的对比研究[J]. 农业图书情报学刊, 2016, 28(1): 76-81.

[14] Glänzel W, Schubert A, Czerwon H J. An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis[J]. Scientometrics, 1999, 44(3): 427-439.

[15] Fang H. Classifying research articles in multidisciplinary sciences journals into subject categories[J]. Knowledge Organization, 2015, 42(3): 139-153.

[16] Taheriyan M. Subject classification of research papers based on interrelationships analysis[C]// Proceedings of the 2011 Workshop on Knowledge Discovery, Modeling and Simulation, New York: ACM Press, 2011: 39-44.

[17] Gómez-Núñez A J, Vargas-Quesada B, de Moya-Anegón F, et al. Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis[J]. Scientometrics, 2011, 89(3): 741-758.

[18] 王昊, 严明, 苏新宁. 基于机器学习的中文书目自动分类研究[J]. 中国图书馆学报, 2010, 36(6): 28-39.

[19] 杨敏, 谷俊. 基于 SVM 的中文书目自动分类及应用研究[J]. 图书情报工作, 2012, 56(9): 114-119.

[20] 李湘东, 阮涛. 内容相近类目实现自动分类时相关分类技术的比较研究——以《中图法》E271 和 E712.51 为例[J]. 图书馆杂志, 2018, 37(6): 11-21, 30.

[21] 王昊, 叶鹏, 邓三鸿. 机器学习在中文期刊论文自动分类研究中的应用[J]. 现代图书情报技术, 2014(3): 80-87.

[22] 郭利敏. 基于卷积神经网络的文献自动分类研究[J]. 图书与情报, 2017(6): 96-103.

[23] 傅余洋子. 基于 LSTM 模型的中文图书分类研究[D]. 南京: 南京大学, 2017.

[24] 董微, 赵捷. 基于密度分布单类支持向量机的科技文献分类研究[J]. 情报工程, 2018, 4(3): 67-72.

[25] 王效岳, 白如江, 王晓笛, 等. 海量网络学术文献自动分类系统[J]. 图书情报工作, 2013, 57(16): 117-122.

[26] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[27] Sun Y, Wang S H, Li Y K, et al. ERNIE: Enhanced representation through knowledge integration[OL]. [2019-08-18]. https://arxiv.org/abs/1904.09223v1.

[28] Hu D C. An introductory survey on attention mechanisms in NLP problems[C]// Proceedings of SAI Intelligent Systems Conference. Cham: Springer, 2020: 432-448.

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook: Curran Associates Inc., 2017: 6000-6010.

[30] Tenney I, Das D, Pavlick E. BERT rediscovers the classical NLP pipeline[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 4593-4601.

[31] TensorFlow code and pre-trained models for BERT[EB/OL]. [2019-08-18]. https://github.com/google-research/bert.

[32] An implementation of ERNIE for language understanding[EB/OL]. [2019-08-18]. https://github.com/PaddlePaddle/ERNIE.

[33] scikit-learn: Machine learning in python[EB/OL]. [2019-08-18]. https://scikit-learn.org/.

[34] 结巴中文分词[EB/OL]. [2019-08-18]. https://github.com/fxsjy/jieba.

[35] Kim Y. Convolutional neural networks for sentence classification [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1746-1751.

[36] Song Y, Shi S M, Li J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 175-180.

[37] Grave E, Bojanowski P, Gupta P, et al. Learning word vectors for 157 languages[C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation, European Language Resources Association, 2018: 3483-3487.

[38] Li S, Zhao Z, Hu R F, et al. Analogical reasoning on Chinese morphological and semantic relations[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Stroudsburg: Association for Computational Linguistics, 2018: 138-143.

[39] 谢靖, 钱力, 师洪波, 等. 科研学术大数据的精准服务架构设计[J]. 数据分析与知识发现, 2019, 3(1): 63-71.

[40] Beltagy I, Cohan A, Lo K. SciBERT: Pretrained contextualized embeddings for scientific text[OL]. [2019-08-18]. https://arxiv.org/abs/1903.10676.

（责任编辑　王克平）