基于神经网络的文献主题国别标引方法研究

王新
（中国农业科学院农业信息研究所，北京 100081）

摘要：为解决海量文献的主题国别标引问题，探讨“互联网+大数据”时代背景下深度学习技术在知识组织领域的应用方法，本文提出基于深度卷积神经网络的文献主题国别标引方法。该方法在探讨主题国别标引任务转换为多标签分类任务的可行性基础上，首先利用自然语言处理方法将文献全文向量化，然后使用预训练的词嵌入将文献向量转换为富含词汇间语义关系的张量，再利用深层卷积神经网络对文本特征由词汇、句子、段落、篇章逐层学习自动提取，生成富含全文语义的张量，最后由全连接层学习分类权重后输出各个国别的概率，实现文献主题国别的自动标引。实验结果表明，该方法达到预期效果，具有高度精确的分类性能和良好的泛化能力，为深度学习算法在知识组织领域的应用提供了有价值的参考。

关键词：知识组织；主题标引；深度学习；深度卷积神经网络
中图分类号：G254.3
DOI：10.3772/j.issn.1673-2286.2019.07.006

在过去几十年中，已发表的科学论文数量每年攀升8%~9%。仅在生物医学领域，每年有超过100万篇论文进入PubMed数据库，每分钟约有2篇论文[1]。然而，研究显示20世纪70年代以来科研人员的人均每年论文阅读数量趋于平稳[2]。芝加哥大学的社会学家Evans[3]认为，大量的论文和相关的在线访问方式导致了“科学和学术的缩小”。浩如烟海的信息远超出了个人有效利用的范围，但信息的序化和组织手段没有跟上时代发展的需要。大量文献由于缺乏有效的组织和揭示而游离于用户视野之外。如何从文献信息中精准挖掘主题信息，从而有效实现对文献内容进行组织和揭示，是当前资源建设工作亟待解决的问题。

在XML主题地图（XTM）格式规范中，文献主题国别（country topic）被定义为一种主题类型，即国别本身是一种文献主题。与传统书目记录格式如MARC中反映文献外在特征的“出版国别”字段不同，它从语义层面对文献内容的国家主题进行揭示，是指文献内容所讨论的空间范畴，它既可以是研究对象（如国别报告），也可以是揭示研究对象的主题之一（如特定国家的农产品报告）。主题国别是揭示宏观研究报告类文献主题内容的重要属性，对于文献发现和缩小搜索范围有不可替代的作用。随着我国“一带一路”倡议的推进，企业“走出去”的进程不断加快，对相关报告类文献的客观需求量激增，对文献主题识别与标引实践提出了更高的要求。

在实际标引操作中，主题国别同时涉及主题和空间范畴，识别更加复杂。一方面，主题国别与文献出版国别、作者机构等来源信息不一致，无法通过文献来源信息判断文献主题国别。如经济学人智库提供193个国家和地区的报告，依靠出版地或作者机构信息已经无法判断其内容的主题国别。另一方面，通过命名实体识别（NER）出的国别信息未必是文献主题国别，报告中经常出现引用相关国家的部分指标数据与文献主题国别的经济指标进行对比的情况，这种少量引用的国家信息不能作为该报告的主题国别。仅仅依靠命名实体识别已经无法准确捕捉文献所讨论的国别对象。针对这一问题，本文从机器学习视角，将主题标引问题转换为多标签分类问题，在已有人工标注数据基础上结合自然语言处理和深度学习技术，提出一种利用深度卷积神经网络，通过监督学习的方式，实现主题国别字段自动识别标引的方法，从而为“互联网+大数据”时代背景下知识组织领域的相关研究提供有价值的参考，同时为探索

深度学习技术在知识领域的相关研究提供借鉴。

1 相关研究

文献主题识别、抽取及标引一直是信息组织领域的研究热点，针对这一问题国内外学者分别从不同角度、运用不同方法开展研究，目前所采用的方法主要有以下几种。

（1）基于文献计量指标的技术，如关键词的词频分析和共现分析的方法。Carthy[4]介绍了一种使用简单的自然语言处理技术将词汇相关术语分组到词汇链中的方法，利用词汇链跟踪新闻并发现主题；Michael-Schultz等[5]提出利用关键词idf加权的余弦相似度检测新闻主题的方法；王曰芬等[6]介绍了共现分析在文本知识挖掘中的应用；Hu等[7]利用共现分析方法对我国图书情报学研究主题的演变进行了分析；叶春蕾等[8]提出了一种共词分析的改进方法，并用于学科主题演化研究；丁晟春等[9]提出了基于关键词共现网络对网络舆情潜在主题的抽取方法。

（2）基于主题概率模型LDA、潜在语义分析PLSA及其改进模型的方法。Wang等[10]改进了LDA模型并用于检测随时间变化的主题发展趋势；Glynn等[11]利用贝叶斯分析和动态线性主题模型，对带有时间戳的文本主题时间演变趋势进行了研究；Jagarlamudi等[12]提出JointLDA生成模型，用于将不同语言的相关主题合并为一个多语言主题；王曰芬等[13-14]利用LDA主题模型识别科学文献主题并进行了多视角探讨；关鹏等[15]对LDA模型主题抽取数量的确定方法、结合生命周期理论对科学文献主题挖掘方法[16]以及不同语料下的主题抽取效果进行了探索[17]。

（3）其他方法。夏火松等[18]提出一种基于改进K-means聚类的主题抽取方法，并对新闻评论主题进行了实验；Zhang等[19]分析了词嵌入结合K-means聚类算法对文献主题抽取效果，并通过实验证明词嵌入配合K-means方法更适用于大规模集群主题提取任务；祝清松等[20]发现在引文内容分析基础上识别的主题具有更好的主题代表性，提出基于引文分析的高被引论文主题识别方法；孟令恩等[21]提出一种引入语义角色标注信息外加辅助规则提取专利的方法。

以上研究利用不同方法从多个角度对主题识别抽取进行研究，无论是基于LDA主题模型及其改进方法的研究、基于文献计量的方法，还是基于K-means聚类、引文分析及语义角色标注的方法，都以关键词和摘要乃至引文作为数据分析的对象，限于数据质量，关键词等信息对全文的代表性和概括性相对有限，这从根本上限制了以上方法的效果。此外，以上方法所采用的模型和算法往往基于词袋，受限于算法本身，较少考虑词序和歧义等问题，这也是限制主题发现效果的重要原因。

2 研究思路及关键技术

为进一步提高主题发现和抽取能力，同时能够很好地抽象表示并处理全文中所含词汇、词序、句子、句序乃至段落语义信息，本文提出基于深度卷积神经网络（Deep Convolutional Neural Network，DCNN）的文献主题国别标引方法。

2.1 深度卷积神经网络

DCNN是以张量为输入，通过卷积层和池化层的多层连接拓扑结构，能够定义足够充分的假设空间，对文本的语义进行深层次的学习和表示。DCNN继承了卷积神经网络的特点，同时具有更多的参数和表示能力，它具有以下特征。

（1）卷积神经网络善于学习层次结构。通过卷积层、池化层的反复交替，卷积神经网络可以逐层自动提取文本中的深层次信息。如第一层卷积学习每个前后相邻3个单词之间的关系，位于其后的池化层可以将第一层卷积学习到的特征进行“蒸馏”和“组合”，从而获取6个单词之间的关系的特征表示，然后再传递给下一层卷积层。每层学习粒度大小可以通过卷积核和池化层的超参数控制。通过多层卷积即可实现对全文内容的抽象特征表示。

（2）局部特征的平移不变性（translation invariant）。局部特征一经学习，无论它再次出现在输入中的任何位置，都不影响对其再次识别。如卷积神经网络学习到某个特定短语的搭配特征后，该短语再次出现在文首和文末都不影响对其识别。这一特征使卷积神经网络可以通过较少的样本学到具有泛化能力的数据表示。

2.2 整体模型架构及思路

文献主题国别标引问题，可以转换为多标签分类问题。在标引问题转换为分类问题的过程中，其可行性

主要基于以下考虑。

（1）国别数量有限，而且国别名称在一定时间范围内变动较小。从机器学习的视角看，主题国别标引是一个平稳问题（nonstationary problem），这是衡量机器学习能否胜任该问题的前提。

（2）输入数据为全文，全文包含用于发现主题国别以及文献研究对象等高层次语义的全部信息，足够回答一篇文献的主题国别。而且，深度学习算法可以通过增加隐藏层来扩展假设空间的容量，能够完全承载一篇文献所含的语义特征表示及运算。

（3）输出为国别，一般来说，一篇文献只有一个主题国别，但不排除存在一篇文献有多个主题国别的可能。国别名称可以作为标签来衡量机器学习的输出是否正确。

基于以上分析，使用DCNN解决文献主题国别标引问题具备了基础条件。由此文献主题国别标引问题可以定义如下。

令X代表文献样本空间，$L=\{\lambda_1, \lambda_2 \cdots \lambda_n\}$为有限个国别名称标签集合，假设X中的文献样本实例$\chi \in X$和国别标签集合L的一个子集$l \in L$相关，则该子集l为文献实例χ的相关标签集。同时l的补集L\l被认为与χ不相关。χ的相关国别标签集l可以表示为向量$y=\{y_1, y_2 \cdots y_n\}$，其中$y_i$与$\lambda_i$一一对应，$y_i=1$等价于$\lambda_i \in L$，则相关标签集可用$y=\{0, 1\}^n$表示。多标签分类器h是一个映射$\chi \rightarrow y$，对每个文献实例$\chi \in X$，h为其分配一个标签子集，其输出为$h(\chi)=(h_1(\chi), h_2(\chi) \cdots h_n(\chi))$。

从实现角度对基于DCNN的文献主题国别标引架构安排如下：首先，对文献全文文本进行预处理，转换为机器可以计算的形式；其次，引入经过预训练的词嵌入（word embedding）作为词汇语义的表征，经过词嵌入处理，全文由单词向量序列构成的1阶张量转变为2阶张量；再次，将全文张量输入至DCNN，由神经网络对文本特征从局部到整体进行逐层学习，自动提取全文特征；最后，将每篇全文的特征压平为1阶向量传入全连接层，通过训练后分类后输出每个类别的概率。整体实验流程见图1。

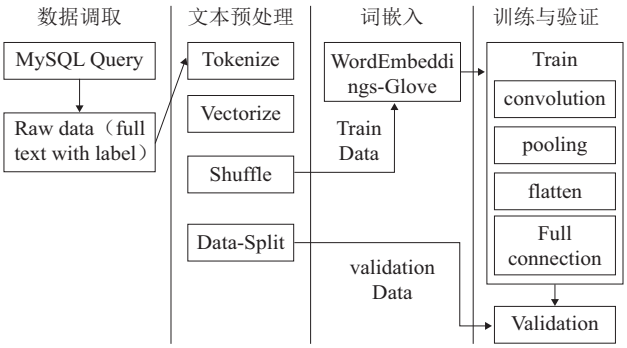


图1　整体实验流程

3　数据实证

3.1　数据获取及分析

本文以“农业”“农业经济”“农产品”为主题，从中国农业科学院农业对外合作公共信息服务平台接入的40个不同来源平台数据集中选取英文文献2.8万篇，该平台发布的数据由专业人员人工标引并通过审核，可以确保数据标注的精度要求。各来源数据分布及语料文本结构分布情况见图2。

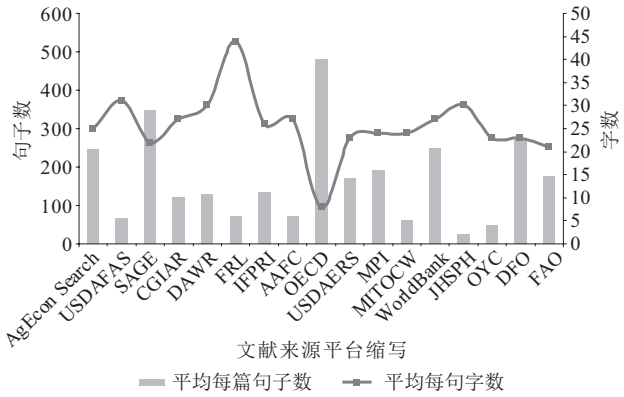


图2　各来源数据分布及语料文本结构分布情况

对各来源文献数量分布和语料文本结构分布情况，本文利用自然语言处理工具包NLTK[22]做了初步处理和分析。分析语料文本结构的目的在于为DCNN超参数的配置提供参考。从数据语料文本结构的统计来看，各来源文献数据集中的句子长度大多在30个单词左右，句子数量反应了文本的篇幅长短不一，但不影响DCNN对句子内部微观结构及句子之间关系的学习。

国别标签的分布情况存在极大的差异，将会在数据清洗和训练过程中进行处理，处理方法为：保留样本数量超过300的国别，对于国别数量分布不均采用加权方法进行均衡处理。国别标签包括316个国家和地区，部分原始数据分布情况见图3。

3.2　标签及文本预处理

（1）数据清洗和规范。将数据集中不符合要求的

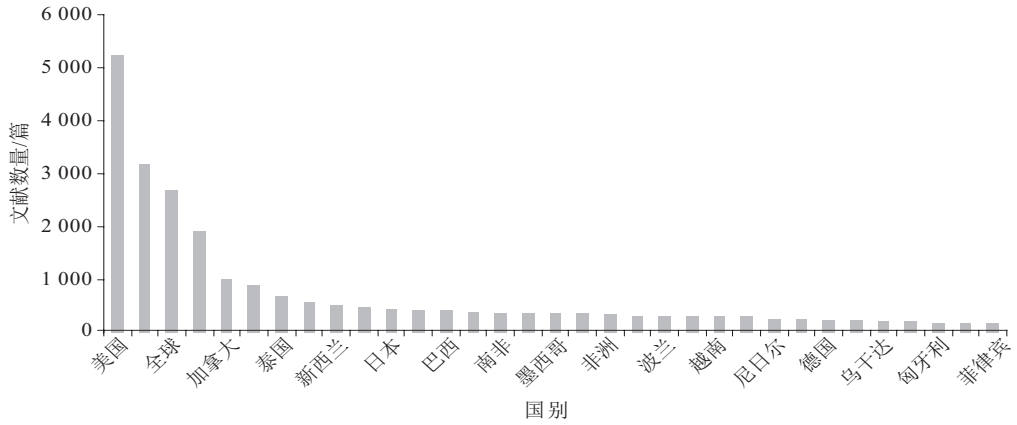


图3 部分主题国别的文献数量分布情况

数据剔除，如缺失全文或国别字段、国别标注不规范等情况。将样本数量低于300的国别数据作为噪声数据滤掉，保留样本数量大于300的24个国家和地区，不属于这24个国家和地区的区域用other标识。

（2）数据类别平衡处理。由于国别标签的数量分布严重不平衡，会对机器学习任务的效果产生不利影响，所以需要对其类别进行平衡处理。常用的数据平衡处理方法有数据级方法和算法级方法。其中，数据级方法分为过度抽样和欠采样，过度抽样方法导致模型易于过拟合，欠采样会造成样本量过小，二者都不适合本任务的数据平衡处理。为此，本文使用算法级加权平衡方法，为每个类别分配一个表示其样本重要性的权重，通过权重和样本数量相乘，从而解决各类别样本数量不平衡的问题。

（3）国别标签向量化。使用独热编码（one-hot）方法对标签进行编码，每篇文献的主题国别被编码为25维的向量，每个维度对应一个国别，每篇文献只在其主题国别的位置为1，其他为0。如只有中国、美国、日本、韩国4个国别，按此顺序生成独热编码，则国别主题为美国的文献国别标签为（[0 1 0 0]），而以中美两国为主题国别的文献标签为（[1 1 0 0]）。

（4）特征工程和文本向量化。由于机器学习算法无法直接处理文本，所以需要对文本进行向量化处理。为完成这一操作，需要将文本分词，然后再将单词映射为向量，本文采用独热编码方法，为下一步词嵌入做准备。将所有文献中出现过的单词汇总为一个词袋（word bag），每个单词获得一个编号。然后将全文表示为一列由单词编号组成的数组，并保留单词与编号的索引。鉴于实验数据中部分题名包含主题国别信息，如“country reference November 1997: New Zealand”，这类报告全文所含国别信息较少，因而需要进一步处理，从而使主题国别特征获得更好的表示。为此，本文对标题和全文分开处理，使用定长独热编码表示标题；全文部分则通过自然语言处理和命名实体识别，抽取包含地名的句子重组，经过编码后与标题合并作为模型输入。

经过数据规范、清洗后，剩余有效数据20 331条。数据国别及权重分布情况见图4。

数据集分为训练集、验证集和测试集。首先将数据打乱顺序，然后随机选取其中的13 000条数据作为训练数据，2 300条数据作为验证数据，5 000条数据作为测试数据。训练数据用于训练模型，使模型在不断迭代中慢慢学习分类所需的参数权重，最后达到结果最优（精度最高，同时损失最小）。验证数据不直接参与训练，只用于每一轮模型训练完成后的验证，通过评估当前模型在验证数据上的损失率和精确度，为模型训练的效果及超参数调整提供一个比较客观的参考。测试数据完全不参与训练，在模型训练过程结束后，利用模型对测试数据进行分类，然后将分类结果与人工标注的结果进行比对得出精确度和损失率，以此评估模型的可用性。与验证数据相比，测试数据避免了信息泄露，对模型的可用性评估更加客观、准确。对模型来说，测试数据都是“新面孔”，测试结果反映的是模型的泛化能力，泛化能力越好，模型在前所未见的新数据上的表现越好。

实验环境及平台情况如下。

系统及硬件：win10 x64；Intel Core i7-7700HQ @2.8GHz CPU；DDR4 32GB；SSD（256G）+HDD（2TB）；NVIDIAQuadroM1200 GDDR5 4GB。

软件平台：python 3.7.1；keras 2.2.4；TensorFlow-GPU 1.14；CUDA 10.0.130-411.31；cuDNN 10.0-windows10-

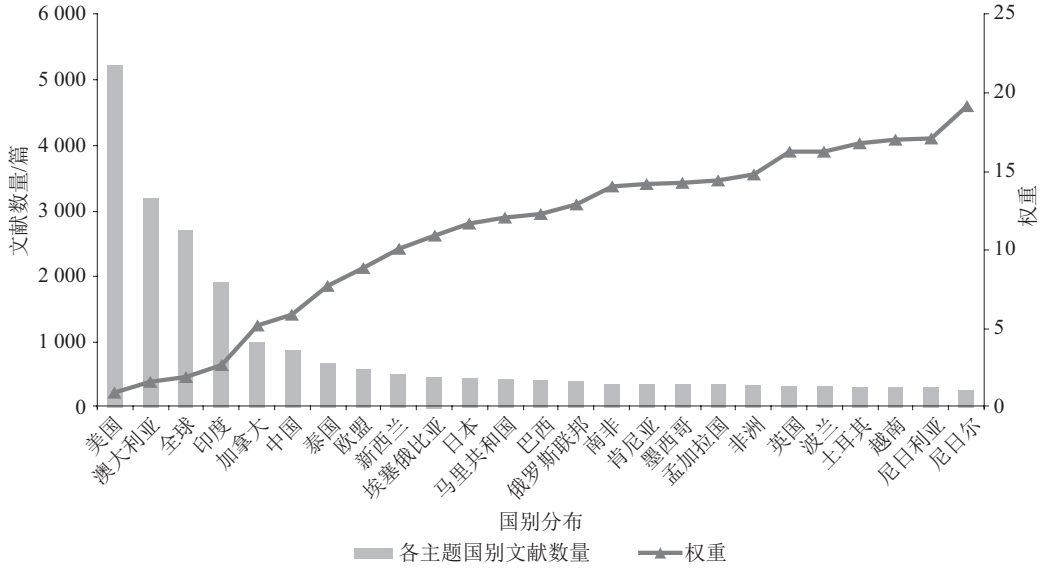


图4　预处理后数据及权重分布情况

x64-v7.6.1.34。

自编python软件：mysql数据库读取程序；数据预处理程序；DCNN深度学习网络。

3.3　词嵌入

经过文本向量化处理，所有全文都被映射为向量空间的向量。向量是一维结构（1阶张量），无法考虑到平面信息，但平面信息对文本特征处理至关重要。如一个关键词汇在文中多次出现，要获得其位置及相邻词汇关系，就需要2阶张量表示；如果要获取句子、段落语义及更抽象层次关系的表示，就需要更高维度的张量才能做到。此外，独热编码生成的文本向量是高维、稀疏和硬编码（0和1）的，没有携带语义信息，而且不利于计算。因此，本文使用预训练的词嵌入，预训练的词嵌入是在大规模文本上通过学习训练后获得的相对低维、稠密的向量表示，而且携带了单词的语义信息。经过词嵌入处理，每一篇全文由向量转换为富含语义信息的2阶张量。预训练的词嵌入数据库有很多，本文选用斯坦福大学于2014年开发的Glove（词表示全局向量）[23]。

为便于计算，本文选用100维的词嵌入数据。设输入文献样本数量为10 000，选择所有文本全部单词中出现频率最高的前5 000个单词作为词袋，经过100维词嵌入编码后，得到输入数据为形状（10 000，5 000，100）的3阶张量。词嵌入过程见图5。

3.4　基于DCNN的全文特征学习与多标签分类网络

DCNN从结构上主要包括卷积层、池化层和全连接层。卷积层与池化层交替使用组成深度网络，主要用于文本深层特征的自动学习，全连接层根据标签学习分类的权重。DCNN的整体架构见图6。

卷积层（convolution）定义了一组卷积核，它可以像窗口一样在句子串上滑动，卷积核的宽度与词向量的维度一致，这样就能以单词为单位进行卷积，即以单词为最小粒度学习文本的局部特征。以本文实验为例，由10 000篇训练集文献生成的输入数据是大小为（10 000，5 000，100）形状的3阶张量，其中每一篇文献样本被表示为（5 000，100）的2阶张量，卷积核的宽度为100，假设设定卷积核的大小为（9，100），就能一次扫描前后相邻的9个单词，每次扫描后输出为一个纯数字的标量。一篇文献的2阶张量切片扫描完成后的特征图（feature map）是形状（4998，）的向量，该向量是这篇文章中所有前后相邻的9个单词之间语义关系的特征表示。通过卷积核的大小控制局部感受野的窗口大小，从而实现不同粒度信息特征的学习。一般来说，越大的卷积核感受野越大，就能学习更全局、语义层次更高的特征，但是相应待学习参数的数量也是倍乘增长的关系。

池化层（pooling）的主要作用有两方面：一方面是通过下采样计算，保留特征集合中对分类作用显著的特征，丢弃不显著的特征，从而降低输出结果的维度；另

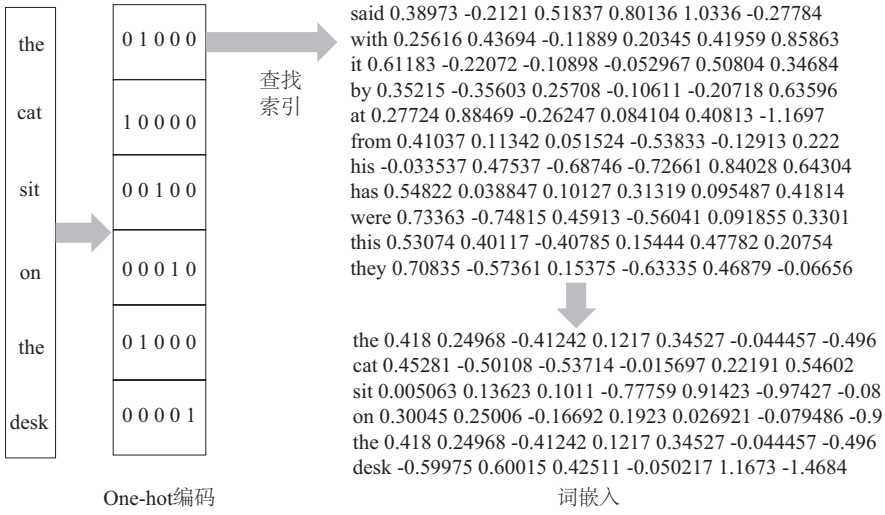


图5 词嵌入过程示意

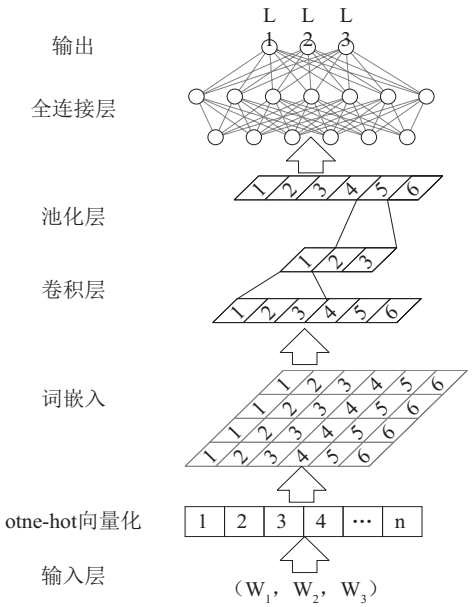


图6 DCNN模型架构

一方面是通过池化层对上一卷积层所得特征图的“蒸馏”，使下一层卷积层的窗口“扩大”。例如，假设在一个由30个单词组成的句子中，经过（9，100）卷积核的卷积计算后，可以获得一个形状为（22，）的向量，经过形状为（2，）的池化计算后，结果为（11，）的向量，该向量就是句子中最重要的11种连续9字的语义关系表示。在此基础上再次卷积，卷积核只需要不小于3，扫描的范围就可以比较容易地扩大到句子，从而学习到句子乃至更高级别的语义表示。池化层有多种算法，本文采用最大池化（max-pooling）算法，即抽取每个特征向量的最大值表示该特征，一般来说，最大的值表示的是最重要的特征。

全连接层（dense）接受一个一维向量作为输入，然后通过学习各个节点的权重，通过激活函数输出结果。由于本次实验任务是多标签分类，所以选择sigmoid函数作为激活函数，针对每一篇文献输出一个在全部国别标签上的概率分布。相应的，损失函数为二元交叉熵（binary-cross entropy）。此外，本文采用的优化算法为RMSprop。

4 结果分析

4.1 深度学习框架搭建

本文初始模型使用3个卷积层，4个全连接层，所有层的通道数量都设为64。DCNN拓扑结构信息见图7。

4.2 参数优化

卷积神经网络的参数有模型深度、卷积核的大小、学习批次（batch size）和学习率η。初始化模型的参数需要在实验中不断调整，获取最优参数，最终得到最佳模型。本文就以上参数对模型进行了测试，结果如下。

4.2.1 模型深度

为测试模型深度对本次学习任务的影响，本文分别对3种配置进行了比较，见表1。其他参数固定不变，分别为卷积核大小为9，学习批次大小为64，学习率为0.001。

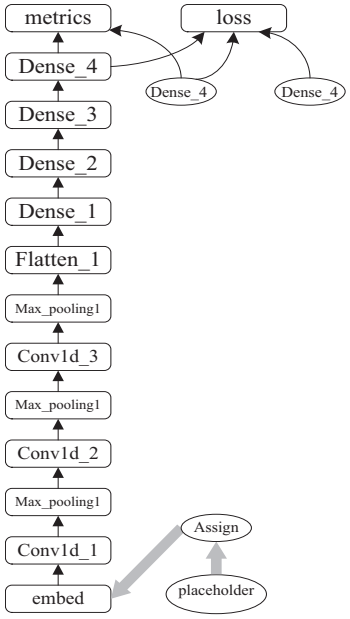
从表1数据记录可以发现，与其他模型深度相比，

采用3层卷积的网络结构可以达到模型收敛速度最快、验证精度高且验证损失低。

4.2.2 卷积核

为测试卷积核的大小对模型的影响，将卷积核分别设置为9、5、3进行测试；其他参数分别为3层卷积，学习批次大小64，学习率0.001，结果见表2。

通过表2数据可以发现，卷积核为9的时候模型可以取得最佳效果。



Layer (type)	Output Shape	Param #
embed (Embedding)	(None, 5000, 300)	600000
conv1d_1 (Conv1D)	(None, 4992, 64)	172864
max_pooling1d_1 (MaxPooling1	(None, 2496, 64)	0
conv1d_2 (Conv1D)	(None, 2488, 64)	36928
max_pooling1d_2 (MaxPooling1	(None, 1244, 64)	0
conv1d_3 (Conv1D)	(None, 1236, 64)	36928
max_pooling1d_3 (MaxPooling1	(None, 618, 64)	0
flatten_1 (Flatten)	(None, 39552)	0
dense_1 (Dense)	(None, 64)	2531392
dense_2 (Dense)	(None, 64)	4160
dense_3 (Dense)	(None, 64)	4160
dense_4 (Dense)	(None, 26)	1690

Total params: 3,388,122
Trainable params: 2,788,122
Non-trainable params: 600,000

图7 DCNN拓扑结构

表1 使用不同模型深度参数的网络性能对比

模型深度	参数数量	收敛到最佳的迭代次数	训练精度	训练损失	验证精度	验证损失
3层卷积	2 788 122	23	0.966 3	0.099 55	0.966 3	0.117 2
4层卷积	1 543 002	26	0.961 6	0.118 70	0.953 1	0.112 6
2层卷积	5 315 290	27	0.968 0	0.095 87	0.948 2	0.111 5

表2 卷积核大小对模型训练及性能的影响

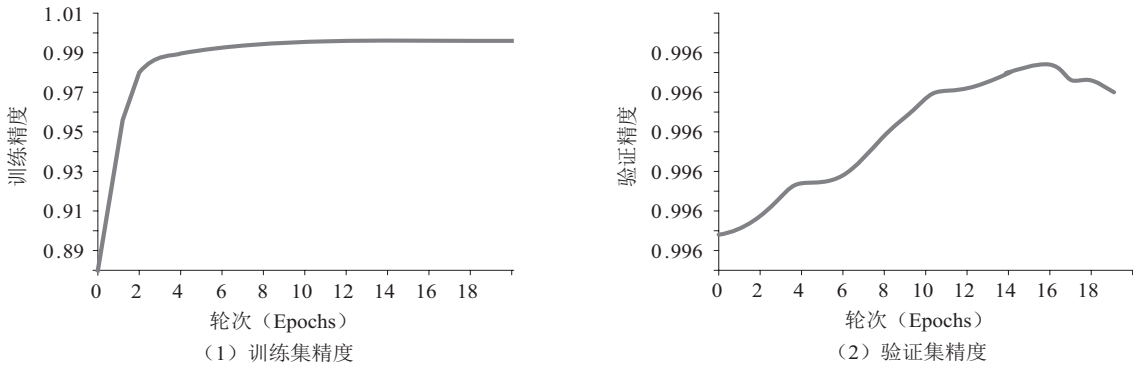
卷积核	参数数量	收敛到最佳的迭代次数	训练精度	训练损失	验证精度	验证损失
9	2 788 122	13	0.961 8	0.114 1	0.957 3	0.108 4
5	2 690 842	28	0.957 7	0.130 1	0.956 3	0.116 7
3	2 644 250	49	0.976 6	0.111 8	0.951 2	0.152 2

4.2.3 学习率和批次大小

学习率和批次大小都影响模型的收敛速度，一般来说，学习率越小，模型收敛需要的时间越久，但精度会不断提升，而学习率过大会直接导致模型不收敛或学习曲线振幅很大且收敛极慢。经过反复实验，确定最佳学习率为0.001。批次大小与输入数据的量有关，本实验适用的批次大小为64。

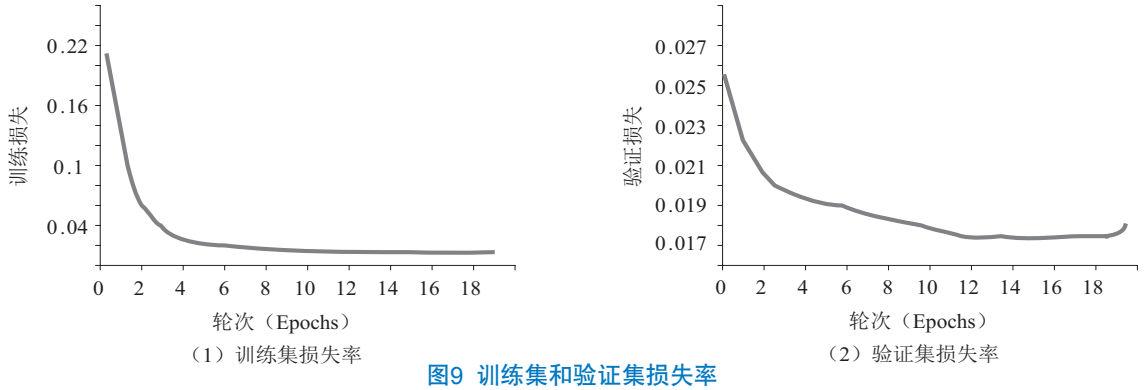
最终，模型的参数配置为3层卷积，卷积核为9，学习率为0.001，批次大小为64。训练结果见图8和图9。

从图8和图9结果可以看出，训练集精度不断提升，模型在训练到第10轮之后，训练精度趋于稳定，训练损失还在逐步降低；模型在验证集上精度在第16轮达到最高值，同时验证损失率在第16轮后不变，第18轮开始反弹，证明模型开始出现过拟合。因此可以将epoch的值定为16，获得分类效果最优、泛化能力最好的模型。



（1）训练集精度　　（2）验证集精度

图8 训练集和验证集精确度



（1）训练集损失率　　（2）验证集损失率

图9 训练集和验证集损失率

4.3 结果验证

最后，将未参与训练的人工标注数据作为测试集，对训练完成的模型实际分类效果进行测试，精确度达到了99.6%，证明模型在新数据上具有良好的泛化能力，具体结果见图10。

```
4896/5000 [============================>.] - ETA: 0s
4960/5000 [============================>.] - ETA: 0s
5000/5000 [==============================] - 5s 1ms/step
测试集loss: 0.01843527851700783
测试集accuracy: 0.9962499816894531
```

图10 测试集精确度和损失率

5 总结

本文将主题国别标引问题转换为适于机器学习的多标签分类问题，利用自然语言处理和深度卷积神经网络技术，以多来源的2.8万条文献全文数据为实验对象，探讨了问题转换的可行性，基于深度卷积神经网络设计构建了主题国别识别模型，并通过数据实验表明，深度卷积神经网络在主题国别标引任务上具有高精度的识别能力和良好的泛化能力，验证该方法可行、有效。由此可见，在“互联网+大数据”时代背景下，深度学习作为一种技术手段，在知识组织研究领域有广阔的应用前景，对主题标引等相关研究具有重要的参考价值。当然，本文研究还存在不足之处，如并未进行其他方法在该任务上的效果对比研究，对现有成果改进和完善也是下一步的研究方向。

参考文献

[1] LANDHUIS E. Scientific literature: Information overload [J]. Nature, 2016, 535 (7612): 457-458.

[2] RICHARD V N. Scientists may be reaching a peak in reading habits [EB/OL]. [2019-06-01]. https://www.nature.com/news/

scientists-may-be-reaching-a-peak-in-reading-habits-1.14658.

[3] EVANS J A. Electronic publication and the narrowing of science and scholarship [J]. Science, 2008, 321(5887): 395-399.

[4] CARTHY J. Lexical chains versus keywords for topic tracking [J]. Computational Linguistics and Intelligent Text Processing, 2004, 2945: 507-510.

[5] MICHAEL-SCHULTZ J, LIBERMAN M. Topic Detection and Tracking using idf-Weighted Cosine Coefficient [C] // Proceedings of the Darpa Broadcast News Workshop. San Francisco: Morgan Kaufmann, 1999: 189-192.

[6] 王曰芬，宋爽，卢宁，等. 共现分析在文本知识挖掘中的应用研究 [J]. 中国图书馆学报，2007(2)：59-64.

[7] HU C P, HU J M, LIU Y. A co-word analysis of library and information science in China [J]. Scientometrics, 2013, 97(2): 369-382.

[8] 叶春蕾，冷伏海. 基于共词分析的学科主题演化方法改进研究 [J]. 情报理论与实践，2012，35(3)：79-82.

[9] 丁晟春，王鹏鹏，龚思兰. 基于社区发现和关键词共现的网络舆情潜在主题发现研究——以新浪微博魏则西事件为例 [J]. 情报科学，2018，36(7)：78-84.

[10] WANG X, MCCALLUM A. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends [C] //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 424-433.

[11] GLYNN C, TOKDAR S T, HOWARD B, et al. Bayesian analysis of dynamic linear topic models [J]. Bayesian Analysis, 2019, 14(1): 53-80.

[12] JAGARLAMUDI J, DAUMÉ H. Extracting Multilingual Topics from Unaligned Comparable Corpora [C] //European Conference on Information Retrieval. Heidelberg: Springer, 2010: 444-456.

[13] 王曰芬，傅柱，陈必坤. 基于LDA主题模型的科学文献主题识别：全局和学科两个视角的对比分析 [J]. 情报理论与实践，2016，39(7)：121-126.

[14] 王曰芬，傅柱，陈必坤. 采用LDA主题模型的国内知识流研究结构探讨：以学科分类主题抽取为视角 [J]. 现代图书情报技术，2016(4)：8-19.

[15] 关鹏，王曰芬. 科技情报分析中LDA主题模型最优主题数确定方法研究 [J]. 现代图书情报技术，2016(9)：42-50.

[16] 关鹏，王曰芬. 基于LDA主题模型和生命周期理论的科学文献主题挖掘 [J]. 情报学报，2015，34(3)：286-299.

[17] 关鹏，王曰芬，傅柱. 不同语料下基于LDA主题模型的科学文献主题抽取效果分析 [J]. 图书情报工作，2016，60(2)：112-121.

[18] 夏火松，李保国，杨培. 基于改进K-means聚类的在线新闻评论主题抽取 [J]. 情报学报，2016，35(1)：55-65.

[19] ZHANG Y, LU J, LIU F, et al. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding [J]. Journal of Informetrics, 2018, 12(4): 1099-1117.

[20] 祝清松，冷伏海. 基于引文内容分析的高被引论文主题识别研究 [J]. 中国图书馆学报，2014，40(1)：39-49.

[21] 孟令恩，李颖，何彦青，等. 基于语义角色标注的专利主题提取研究 [J]. 图书情报工作，2014，58(19)：19-24.

[22] BIRD S, KLEIN E, LOPER E. Natural language processing with Python [M]. Cambridge: O'Reilly, 2009: 479.

[23] PENNINGTON J, SOCHER R, MANNING C. Glove: Global Vectors for Word Representation [C] //Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

作者简介

王新，1986年生，男，博士，馆员，研究方向：数字资源管理、信息组织，E-mail：wangxin@caas.cn。

A Country Topic Indexing Method Research Based on Neural Network

WANG Xin
(Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: In order to solve the problem of country topic indexing of massive literature, and to explore the use of deep learning in the field of knowledge organization under the background of "Internet + Big Data", this paper proposes a country topic indexing method based on deep convolutional neural network. On the basis of exploring the feasibility of converting the country topic indexing task into a multi-label classification task, this method use the natural language processing method to vectorize the full text of the document as the first step, and then use pre-trained word embedding to transform the document vector into a tensor rich in semantic relationships between words. Thirdly, using deep convolution neural networks to automatically extract text features from vocabulary, sentences, paragraphs, and chapters layer by layer, generates a volume rich in full-text semantics. Finally, the probability of the country label being output by the full connection layer. The experimental results show that the method achieves the desired effect, has a high accurate classification performance and good generalization ability, and provides a valuable reference for the application of deep learning algorithm in the field of knowledge organization.

Keywords: Knowledge Organization; Subject Indexing; Deep Learning; Deep Convolution Neural Networks

（收稿日期：2019-05-21）