研究论文
DOI：10.13998/j.cnki.issn1002-1248.23-0059

结构信息增强的文献分类方法研究

安　波[1,2]
（1. 中国社会科学院 民族学与人类学研究所，北京　100081；2. 中国科学院软件研究所，北京　100190）

摘　要：［目的 / 意义］针对传统文献分类方法未能充分利用文献结构信息的问题，本文提出使用关键词–文献图网络构建文献之间的结构信息，并用于增强传统基于文献内容的分类方法。［方法 / 过程］本文借助图卷积神经网络建模关键词 – 文献图数据，学习文献在图网络中的节点表示。同时使用 Bert+BiLSTM 学习文献的内容表示。然后，我们将文献的节点表示与内容表示进行拼接，得到融合文献结构信息和文本语义信息的表示，并基于该表示开展文献分类。［结果 / 结论］实验结果表明，文献的结构信息能够提升文献分类的性能，但单一的结构信息并不能很好地实现文献分类。通过错误分析，我们发现模型在处理包含新兴交叉科学和新概念的文献时容易出现分类错误，表明模型在处理这类数据时还有一定的局限性，是未来需要继续优化的方向。

关键词：文献分类；图卷积神经网络；关键词–文献图；语义关联；知识组织；自然语言处理

中图分类号：TP393　　**文献标识码：**A　　**文章编号：**1002-1248（2023）03-0015-10

引用本文：安波. 结构信息增强的文献分类方法研究[J]. 农业图书情报学报, 2023, 35(3)： 15-24.

1 引　言

文献分类是图书情报领域一项基础且重要的任务，对信息资源管理、文献检索与获取等工作具有重要价值，也是文献情报库建设的核心内容之一[1]。目前，中国文献分类主要依据《中国图书馆分类法》[2]，该分类系统包含 22 个一级分类，500 多个二级分类，1 700 多个三级分类和数千个四级分类。基于机器学习的文献分类方法是当前的主流，其中基于深度学习的文献分类方法取得了当前最好的效果。

传统的深度学习方法主要利用文献的文本信息将文献的标题、关键字、摘要等信息进行拼接，然后利用如卷积神经网络（CNN）[3]、长短时记忆神经网络（LSTM）学习文本的表示并进行分类[4]。近年来，随着大规模预训练语言模型（PLM）[5]的快速发展，基于预训练语言模型的文献分类，例如 Bert 能够显著提升文献分类的性能[6]。

与其他类型的文本不同，文献通常包含关键词。关键词反映了文献的核心内容，如主题、观点、方法等，相同类别的论文往往具有相似的关键词，因此相同类别的文献可以通过关键词相互关联。例如，当不同文献直接或间接共享多个关键词时，文献有更大的概率属于相同的类别。因此，这种相互关联所蕴含的

收稿日期：2023-02-08

基金项目：国家自然科学基金项目“知识增强的中文复述识别关键技术研究”（62076233）；国家社会科学基金项目“藏汉双语藏文古籍知识图谱构建研究”（22BTQ010）

作者简介：安波（1986– ），男，博士，副研究员，研究方向为自然语言处理、知识图谱。E-mail：anbo@cass.org.cn

DOI：10.13998/j.cnki.issn1002-1248.23-0059

文献之间的结构信息对文献分类具有潜在价值，由于相同类别的文献由于关键词重叠度较高则更有可能形成紧密关联的子网络。目前基于深度学习的文献分类方法通常使用文献内容进行分类，未能利用文献之间的结构信息。

图神经网络（Graph Neural Network，GNN）[7]用于处理由节点和边构成图结构的数据，能够学习图的网络拓扑结构和节点内容信息，将图中的节点转化为低维向量。由于图神经网络能够较好地处理网络拓扑信息，被广泛用于网络数据表示及应用[8]。例如研究者通过词汇的共现构建词汇网络，并用于文本分类，取得了不错的效果。

基于上述分析，本文提出一种结构增强的文献分类方法（Structural Information Enhanced Literature Classification，SIELC），通过关键词、文献信息构建图数据，建模文献之间的结构信息。然后利用图神经网络学习文献节点的表示，将文献之间的结构信息建模到向量表示中，并用于增强传统文献分类方法。实验结果表明，文献的结构信息能够有效提升文献表示与分类的性能。

2 相关工作

本文探索使用关键词增强文献分类中的效果，主要的相关工作包括文献分类和图神经网络，本节将分别对主要的相关工作进行介绍。

2.1 文献分类

文献分类指的是将学术文献分到指定的类别中，由于文献规模大、类目多等原因，完全基于人工的文献分类方法需要大量的专业人员[9]，消耗大量的人力成本和时间成本。因此，自动文献分类一直是图书情报领域的研究热点[10]。文献分类是一种典型的文本分类[11]，大致经历了基于规则的分类方法[12]、基于统计学习的分类方法[13]和基于深度学习的分类方法[14]3 个阶段。文献[12]在分类索引知识的基础上使用规则实现文献分类，基于规则的文献分类方法实现简单、可解释性强，但需要文献分类专家整理大量的规则，且无法直接迁移到其他分类，不能很好地应对新增分类等问题。

统计机器学习兴起后，基于统计学习文本分类方法迅速成为主流，文献[15]基于 N-Gram 语言模型实现中文文献自动分类，文献[12]进一步利用关联规则实现文献分类。随着 SVM 等分类模型的发展，文献[13]提出基于 SVM 实现文献分类。基于统计学习的文献分类方法具有分类速度快，准确率高等特点，但是需要专家设计特征及大量的训练数据。

传统的文本表示以独热表示（One-Hot）为主，Word2vec[4]将词汇表示发展为连续的向量表示，可以在向量空间中实现文本的表示和处理。同时，由于深度学习拥有更好的建模能力，基于深度学习的文本分类成为主流，如基于卷积神经网络（CNN）的文献分类方法[3]、基于迭代神经网络（RNN）的文本分类和基于 TextCNN 的文献分类方法[14]等。近期，以 Bert[16]为代表的大规模预训练语言模型在文本分类等应用中取得非常好的效果，基于预训练语言模型的文献分类方法也受到学者的关注。

在分类方法的基础上，学者们针对文献分类的特点提出针对性的解决方案，如文献[17]针对文献分类层次性的特点，提出使用多层分类器的方法实现文献分类。文献[18]利用文献中词、句子等信息构建关系图，利用图神经网络实现文本分类。由于文献分类的基础性和必要性，该任务一直是图书情报领域的热点任务。

从上述研究可知，当前的文献分类方法主要依赖于文献的内容信息，将关键词作为普通文本进行处理，忽略了文献之间的相关性。本文探索通过关键词建立文献之间的相互关联，借助于不同文献之间的结构信息增强文献分类的性能。

2.2 图神经网络

图神经网络是针对图结构数据（节点、边）进行学习的模型，将深度学习应用到图网络数据中。图神经网络最早由 GORI 等[19]在 2005 年提出。BRUNA 等[20]于 2013 年基于卷积神经网络实现了图神经网络，将图神经网络从理论推向了应用。图卷积神经网络目前被

广泛应用于文本分类、节点分类、推荐系统等应用。如杨旭华等[21]基于依存句法构建文本的图结构，并基于该图结构实现文本分类。王婷等[22]通过单词、文档、外部实体的方式构建图结构，然后利用图神经网络进行节点的学习，通过这种方式将外部知识融合到文本分类任务中。胡春华等[23]将图神经网络应用于社交电商中的信任和声誉信息的学习，能够提升对社交电商的识别效果。邵云飞等[24]通过“度”优化图神经网络信息在节点之间的将图神经网络用于社交网络表示与传播。

图神经网络在进行表示学习时通常需要在整个图网络中进行遍历和学习，计算复杂度较高。针对该问题，研究者提出了许多的图神经网络优化算法，如基于影响力剪枝的优化方法[25]、使用多 GPU 进行并行计算的方法[26]。

由于图神经网络的强大表示能力，其在文献分类与推荐任务中得到应用。如丁恒等[27]将文献表示建模为一个无监督学习任务，学习文献的表示并用于下游的文献分类与文献推荐。张晓丹[18]通过马尔科夫链采样算法对特征降维，以提升图神经网络在文献分类中的效率。黄学坚等[28]通过自注意力机制学习不同节点的权重，根据不同的权重进行神经网络的学习，以提升图神经网络学习的效率。

综上所述，目前基于图神经网络进行文献分类的工作还较少，主要集中在如何优化图神经网络的学习速度，未能对文献之间的信息进行挖掘。本文尝试将利用图神经网络充分挖掘文献之间的结构信息，并用于增强文献分类任务。

3 模型设计

本节主要介绍模型的整体架构和实现细节，包括如何学习文献的结构表示，以及结构信息如何增强文献分类（图 1）。

3.1 文献图建模

文献通常包含题目、关键词、摘要、正文等部分，其中题目、关键词和摘要经常被用于文献分类。有多种方式可以构建文献之间的关联，如引用网络、作者网络、关键词网络等。然而文献之间的引用数据分散在不同的论文数据库中较难完整获取，且有大量的论文引用率较低容易产生离散节点。而通过作者网络构建的文献图也存在网络稀疏问题，不利于学习高质量的节点表示。关键词由于可以被大量的文献共享，容易形成高质量的文献图数据。此外，关键词是作者对文献核心内容的凝练，对于文献分类具有重要作用。因此，本文利用文献中的关键词构建文献图数据。然而由于关键词通常有数量限制，除文献列出的关键词外，其他部分关键词可能蕴含在标题和摘要中。为此，本文针对待分类文献集合首先构建关键词词典，抽取所有文献的关键词。并从文献的标题和摘要抽取匹配的关键词，形成文献对应的关键词集合。

具体地，本文实现双向最大匹配算法，利用关键

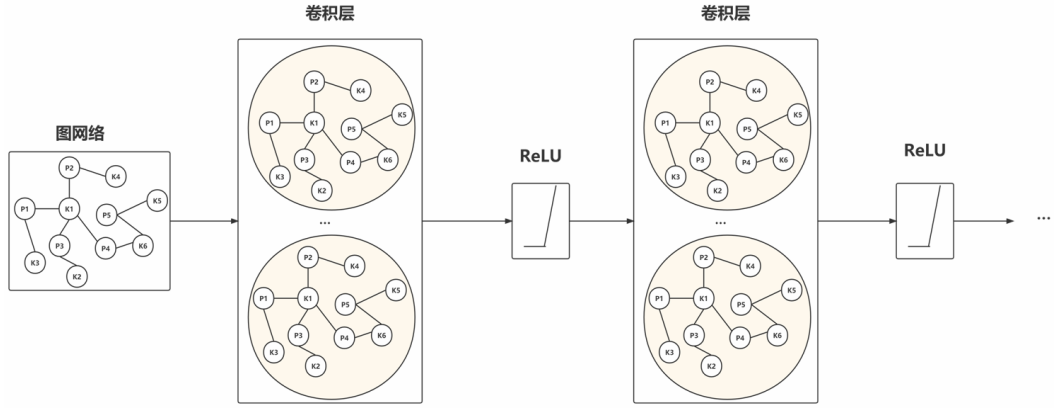

图 1 文献图卷积神经网络结构图

Fig.1 Illustration of convolution neural network of literature graph

DOI：10.13998/j.cnki.issn1002-1248.23-0059

词词典从文献的标题和摘要中提取匹配的关键词集合。然后，我们将文献和关键词作为网络节点，用边连接文献与关键词之间的包含关系，形成关键词 - 文献网络，文献之间则通过关键词建立间接关系，进而形成文献图数据。

3.2 文献结构信息的表示学习方法

本文旨在通过文献间的结构信息增强文献分类，为了利用相同类别文献关键词之间的相似性，本文构建了关键词 - 文献图数据，并尝试利用图卷积神经网络（GCN）学习图数据中的文献表示。

图卷积神经网络（GCN）[22]能够在文献图上进行卷积操作，根据节点及其关联关系学习节点的向量表示，其网络结构如图 1 所示。GCN 对输入的节点根据相邻情况进行卷积，并通过激活函数（如 ReLU）[29]更新向量表示，通过迭代上述两个步骤将节点转化为向量表示。

形式化地，给定一个文献图 $G=(V, E)$，其中 V 代表文献和关键词节点集合，E 表示连接文献和关键词边的集合。文献图节点的初始化表示 $H(0)$：关键词节点通过 Bert[6]学习其初始表示，文献节点通过 Bert 学习其标题的表示作为其初始表示，然后通过 GCN 学习节点的表示。具体地，GCN 从第 l 层到第 $l+1$ 层的推导如公式（1）所示，其中 $H^{(l)} \in R^{N\times D}$ X 为节点的初始向量表示（D 维向量），N 为节点的数量。$\widetilde{A}=A+I_N$ 为添加了自连接的邻接矩阵，D 为度矩阵，其计算方式为 $\widetilde{D}_{ii}=\sum_j \widetilde{A}_{ij}$，$W^{(l)} \in R^{N\times D}$ 为可训练参数。σ 为激活函数（如 ReLU）。

$$H^{(l+1)}=\sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad \text{公式（1）}$$

GCN 的整体计算流程为：①计算得到文献 / 关键词节点的初始化表示；②根据节点的向量表示计算邻接矩阵，$\widetilde{A}=\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ 基于邻接矩阵计算得到下一层文献 / 关键词节点的表示 $H^{(l+1)}$；重复过程②得到文献节点的最终表示。

3.3 文献文本信息的表示学习方法

对于文献的文本表示，我们使用目前文本表示学习的主流方法 Bert+BiLSTM。具体地，我们将文献的标题、关键词和摘要文本进行拼接，使用 Bert 学习其词汇表示，然后使用 BiLSTM 对词汇表示进行组合，得到文献表示 T。

3.4 文献分类

基于文献图学习的文献表示能够反映文献之间的结构信息。基于文本的文献表示方法能够学习文献自身蕴含的语义信息。我们通过将文献的结构表示和文本表示进行融合，得到结构增强的文献表示，并基于此表示开展文献分类工作。具体地，我们将文献的结构表示向量和文本表示向量进行拼接得到结构增强的文献表示。然后，通过一个分类器（Softmax）[30]对文献进行分类。本文提出方法的整体框架如图 2 所示，

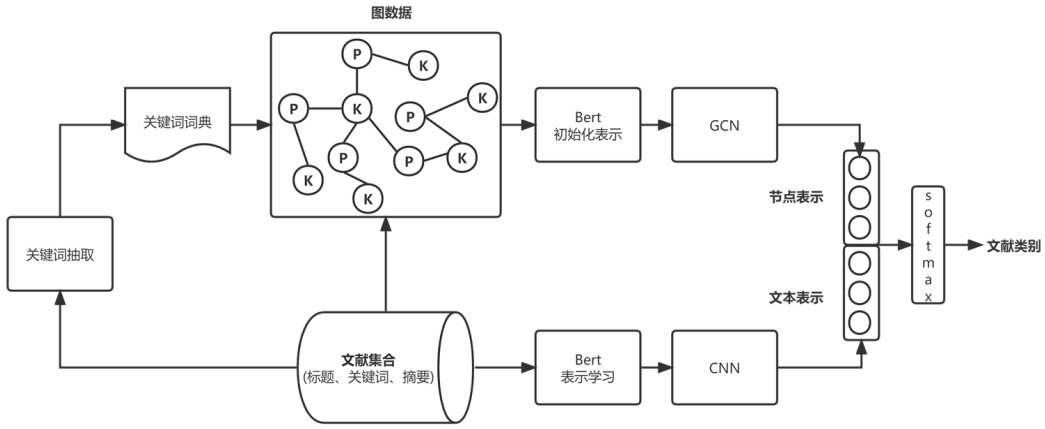

图 2 SIELC 整体架构图

Fig.2 Overall structure of SIELC model

文献通过两个通路分别学习其结构表示和文本表示 N。模型的具体操作流程如图 3 所示。

Algorithm 1: KEGText

Input: 文献分类数据集
Output: τ 分类结果

```
关键词典构建;
抽取关键词并构建文献图网络;
基于 Bert 学习节点初始化表示;
基于 GCN 学习节点表示 N_i;
for 文献 i of 文献数据集 do
    基于 Bert 初始化文献标题、关键词、摘要的文本表示;
    基于 CNN 学习文本表示 T_i;
    拼接 N_i 和 T_i 得到文献的最终表示 P_i;
    基于 P_i 进行文献分类;
```

图 3 SIELC 伪代码

Fig.3 The code of SIELC

通过上述过程，我们利用 GCN 学习文献间的结构表示，并用该表示增强传统的文献分类方法。

4 实验设计

本节旨在验证结构信息对文献分类的作用。本节将从实验数据、基线模型、评价方法、实验结果和错误分析等方面进行详细阐述。

4.1 实验数据

本文通过数据抓取与合作等方式获取约 93 万篇图书情报领域的中文文献数据，每篇文献包含标题、关键词、摘要、文献分类号 4 个字段。文献分类号采用《中国图书馆分类法》。通过对数据进行预处理，保留文献数在 150 篇以上的分类，最终数据包含 423 个文献类别，共包含 749 996 条数据，本文按照 8 : 1 : 1 分配训练数据、验证数据和测试数据，详细的数据情况如表 1 所示。

表 1 文献分类实验数据

Table 1 Experimental data of literature classification

类别数/个	训练集/条	验证集/条	测试集/条
423	599 996	75 000	75 000

4.2 基线模型及评价方法

为了验证本文提出方法的有效性，我们使用多种主流深度学习方法作为基线模型。具体的模型包括：TextCNN 是基于卷积神经网络实现文本分类的方法，该方法具有速度快，准确率高等特点，是目前被广泛使用的文本分类方法。TextRNN 是将循环神经网络引入到文本分类任务中，对文本序列进行建模进而实现文本分类的方法，使用 BiLSTM 作为序列建模模型。TextRCNN 利用循环神经网络及最大池化方法对文本序列进行建模，是结合了 RNN 和 CNN 的方法。BertMLP 是利用大规模预训练语言模型学习文本的表示，然后利用全连接层实现文本分类的方法。BertCNN 是将 TextCNN 的词向量替换为 Bert 学习的词汇向量表示，然后使用 CNN 实现文本分类。ERINE 是百度提出的中文预训练语言模型，本文基于该模型加 Softmax 实现文本分类。BertBiLSTM 是在大规模预训练语言模型的基础上使用 BiLSTM 进行文本序列建模以实现文本分类的方法，是本文提出方法的直接对比模型。KeyGCN 是在关键词 - 文献图数据的基础上直接使用图卷积神经网络进行建模和分类的方法。SIELC 是本文提出使用关键词图数据增强文献分类的方法。

本文使用准确率 P（Precision）、召回率 R（Recall）和 $F1$ 值（F1-value）作为评价指标进行模型的评价，其计算方法如公式（2）所示。由于文献分类是多分类问题，因此本文使用权重平均计算所有类别的准确率、召回率和 $F1$ 值，其计算方法如公式（3）所示，其中 n 为类别数，C_i 为一个类别的文献占测试集文献的比例。本文后续所述准确率、召回率和 $F1$ 值，未做特殊说明的情况下均为所有类别的权重平均。权重平均的能够更好地处理不同类别下数量不同导致的指标变化的问题。

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

TP：将正类预测为正类数

TN：将负类预测为负类数

DOI：10.13998/j.cnki.issn1002-1248.23-0059

FP：将正类预测为正类数

FN：将正类预测为负类数 （2）

$$weighted\_P = \frac{1}{n}\sum_{i=1}^{n} Pi * Ci / n$$

$$weighted\_R = \frac{1}{n}\sum_{i=1}^{n} Ri * Ci / n$$

$$weighted\_F1 = \frac{1}{n}\sum_{i=1}^{n} F1i * Ci / n \quad (3)$$

4.3 主实验

本节旨在通过实验验证结构信息对文献分类任务的作用，在 4.1 节数据的基础上开展文献分类实验。实验设置如下：其中 Bert 采用的是谷歌发布的 Bert-Base-Chinese 版本，ERINE 采用百度发布的官方模型，EPOCH 的次数设置为 20 次，batch_size 的大小设置为 16，dropout 设置为 0.5，pad size 设置为 32，学习率设置为 0.001，TextCNN 的 filter size 为(2,3,4)，filter 数量设置为 256；BiLSTM 的隐藏层设置为 128 维，层数设置为 2；Bert 模型的隐藏层维数为 768；BertCNN 的设置与 CNN 的设置相同，ERNIE 采用默认设置，其隐藏层的维数也是 768。所有实验中文献的文本均为多个字段的拼接（标题、关键词、摘要），输出的目标数据为文献分类对应的分类。SIELC 的卷积层嵌入大小为 768，窗口大小设置为 15，学习率设置为 0.001，dropout 设置为 0.5，epoch 设置为 100。所有模型均使用加权准确率、召回率和 *F1* 值作为评价标准，在验证集上取得的最好参数，在测试集上进行性能评价。整体的实验结果如表 2 所示。

实验结果表明：①本文提出的方法在准确率、召回率和 *F1* 值上均取得了最好的效果，表明基于关键词-文献构建的图数据能够增强文献分类的性能。验证了文献之间的结构信息对文献分类具有增强作用；②通过对比是否使用预训练语言模型（Bert、ERINE）可见预训练语言模型对文献分类的提升作用明显；③实验结果还表明基于卷积神经网络（CNN）的文献分类方法（TextCNN 和 BertCNN）相比其他同类模型能够取得更好的效果，说明 CNN 在文本分类任务上有显著优势。

表 2 主试验结果

Table 2 Main experiments results

模型	准确率	召回率	*F1* 值
TextCNN	0.834 8	0.804 8	0.793 7
TextRNN	0.587 3	0.501 2	0.540 8
TextRCNN	0.577 7	0.476 7	0.390 9
BertMLP	0.897 5	0.901 2	0.899 3
BertCNN	0.901 4	0.914 2	0.907 8
ERNIE	0.894 2	0.899 5	0.896 8
BertBiLSTM	0.887 5	0.897 5	0.892 5
SIELC	0.912 4	0.918 7	0.915 5

4.4 剥离实验

本节希望通过剥离实验来分析本文提出模型不同模块对文献分类的作用，对比模型包括：BertCNN 和 KeyGCN，实验结果如表 3 所示。

表 3 剥离实验结果

Table 3 The results of ablation experiments

模型	准确率	召回率	*F1* 值
BertCNN	0.901 4	0.914 2	0.907 8
KeyGCN	0.756 8	0.785 4	0.770 8
SIELC	0.912 4	0.918 7	0.915 5

实验结果表明：①KeyGCN 实验结果表明基于关键词 - 文献图网络的文献分类方法能够在一定程度上实现文献分类，但是其效果弱于基于预训练语言模型的文献分类方法，说明文本信息对于文献分类的作用更显著；②基于卷积神经网络的 TextCNN 和 BertCNN 分别在单纯使用词向量和预训练语言模型的方法上取得了最好的效果，表明文本卷积神经网络在文本分类任务上具有较高的性能。

4.5 错误分析

为了分析模型的局限性，我们对 SIELC 模型错误分类数据进行分析，表 4 列出了出错类别较多的类别对应的例子。我们发现文献分类的错误主要集中在交叉学科研究的文献，如“信息加工”技术可以应用于多个领域，对于出现较少的文献类别容易分到研究对象本身所属的类别，如表 4 中的“地方志”“社会生

表 4 错误实例

Table 4 Bad cases

分类	标题	预测类别
信息加工（检索机）	一种地方志资源的混合推荐模型	地方志
商品流通与市场	商情数据库——一种重要的竞争情报源	图书馆
信息资源及其管理	中国城乡青少年网民网络信息行为比较研究	社会生活与社会问题

活”。此外，一些新的概念也会导致模型预测错误，如“商情数据库”。

综上所述，本文提出的结构增强的文献分类方法对于具有交叉学科且领域特征显著的文献容易出现分类错误，对于新概念的表示能力还具有一定局限性。

5 结论与展望

本文提出了一种结构增强的文献分类方法，通过关键词 - 文献网络建模文献在数据中的结构信息，利用图卷积神经网络学习图中文献节点的表示。通过融合结构表示和文本表示增强文献分类性能。实验结果表明，结构信息能够有效提升文献分类的性能。模型对于交叉科学文献和一些新概念的分类还有待提高，未来我们希望通过小样本分类方法来解决这部分问题。

参考文献：

[1] 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.
ZHANG Z X, ZHAO Y, LIU H. Construction of a practical application–oriented automatic classification engine for scientific literature [J]. Journal of library science in China, 2022, 48(4): 104–115.

[2] 李清, 侯荣理, 张馨.《中国图书馆分类法》类目注释问题探讨[J]. 数字图书馆论坛, 2022(1): 47–51.
LI Q, HOU R L, ZHANG X. Discussion on some problems of class annotation in Chinese library classification[J]. Digital library forum, 2022(1): 47–51.

[3] 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.
LEI B, LIU X, ZHONG Z. Research on fine–grain classification method of academic literature based on bibliographies [J]. Library and information service, 2021, 65(14): 128–137.

[4] 谢红玲, 奉国和, 何伟林. 基于深度学习的科技文献语义分类研究[J]. 情报理论与实践, 2018, 41(11): 149–154.
XIE H L, FENG G H, HE W L. Research on semantic classification of scientific and technical literature based on deep learning [J]. Information studies: Theory & application, 2018, 41(11): 149–154.

[5] 陈德光, 马金林, 马自萍, 等. 自然语言处理预训练技术综述[J]. 计算机科学与探索, 2021, 15(8): 1359–1389.
CHEN D G, MA J L, MA Z P, et al. Review of pre–training techniques for natural language processing [J]. Journal of frontiers of computer science and technology, 2021, 15(8): 1359–1389.

[6] 沈立力, 姜鹏, 王静. 基于 BERT 模型的中文期刊文献自动分类实践研究[J]. 图书馆杂志, 2022, 41(5): 109–118, 135.
SHEN L L, JIANG P, WANG J. A study on the automatic classification of Chinese literature in periodicals based on BERT model[J]. Library journal, 2022, 41(5): 109–118, 135.

[7] 马帅, 刘建伟, 左信. 图神经网络综述[J]. 计算机研究与发展, 2022, 59(1): 47–80.
MA S, LIU J W, ZUO X. Survey on graph neural network[J]. Journal of computer research and development, 2022, 59(1): 47–80.

[8] 宁懿昕, 谢辉, 姜火文. 图神经网络社区发现研究综述[J]. 计算机科学, 2021, 48(s2): 11–16.
NING Y X, XIE H, JIANG H W. Survey of graph neural network in community detection[J]. Computer science, 2021, 48(s2): 11–16.

[9] 侯汉清, 黄刚. 电子计算机与文献分类[J]. 计算机与图书馆, 1982 (1): 5–14.
HOU H Q, HUANG G. Computer and document classification [J]. Data analysis and knowledge discovery, 1982(1): 5–14.

[10] 叶新明, 徐进鸿. 中文文献自动分类研究[J]. 情报科学, 1992(5): 31–34.
YE X M, XU J H. Research on automatic classification of Chinese

DOI：10.13998/j.cnki.issn1002-1248.23-0059

documents[J]. Information science, 1992(5): 31–34.

[11] 庞观松, 蒋盛益. 文本自动分类技术研究综述[J]. 情报理论与实践, 2012, 35(2): 123–128.
PANG S G, JIANGS Y. A survey of automatic text classification technology[J]. Information studies: Theory & application, 2012, 35 (2): 123–128.

[12] 周丽红, 刘勘. 基于关联规则的科技文献分类研究 [J]. 图书情报工作, 2012, 56(4): 12–16, 119.
ZHOU L H, LIU K. Research on classification of scientific and technological documents based on association rules[J]. Library and information service, 2012, 56(4): 12–16, 119.

[13] 王方, 阮梅花, 朱海刚, 等. 基于向量空间模型的科技文献自动分类研究[J]. 情报探索, 2013(12): 1–3, 8.
WANGF, RUAN M H, ZHU H G, et al. Research on vector space model–based automatic classification of sci–tech document[J]. Information research, 2013(12): 1–3, 8.

[14] 李彦轩. 基于摘要的论文分类与推荐模型的研究与实现[D]. 北京: 北京邮电大学, 2019.
LI Y X. Research and implementation of abstract –based paper classification and recommendation model [D]. Beijing: Beijing university of posts and telecommunications, 2019.

[15] 何浩, 杨海棠. 一种基于 N–Gram 技术的中文文献自动分类方法[J]. 情报学报, 2002(4): 421–427.
HE H, YANG H T. Approach of chinese document automatic classification based on the frequency of N–Gram[J]. Journal of the China society for scientific and technical information, 2002 (4): 421–427.

[16] 王颖. 科技文献内容语义描述模型研究[J].农业图书情报学报, 2020, 32(8): 12–24.
WANG Y. Semantic models for the content of scientific literature[J]. Journal of library and information science in agriculture, 2020, 32 (8): 12–24.

[17] 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.
ZHAO Y, ZHANG Z X, LIU H. Research on chinese medical literature classification based on hierarchical classification[J]. Research on library science, 2021(21): 49–55, 61.

[18] 张晓丹. 改进的图神经网络文本分类模型应用研究——以 NSTL 科技期刊文献分类为例[J]. 情报杂志, 2021, 40(1): 184–188.
ZHANG X D. The application of improved graph convolutional neural network in big data classification of scientific and technological documents[J]. Journal of intelligence, 2021, 40(1): 184–188.

[19] GORI M, MONFARDINI G, SCARSELLI F. A new model for learning in graph domains [C]. Proceedings of the IEEE international joint conference on neural networks, IEEE, 2005: 729–734.

[20] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected net–works on graphs[J/OL]. arXiv Preprint, arXiv: 1312.6203.

[21] 杨旭华, 金鑫, 陶进, 等. 基于图神经网络和依存句法分析的文本分类[J]. 计算机科学, 2022, 49(12): 293–300.
ZHANG X H, XIN J, TAO J, et al. Text classification based on graph neural networks and dependency parsing[J]. Computer science, 2022, 49(12): 293–300.

[22] 王婷, 朱小飞, 唐顾. 基于知识增强的图卷积神经网络的文本分类[J]. 浙江大学学报(工学版), 2022, 56(2): 322–328.
WANG T, ZHU X F, TANG G. Knowledge–enhanced graph convolutional neural networks for text classification[J]. Journal of Zhejiang university(engineering science), 2022, 56(2): 322–328.

[23] 胡春华, 邓奥, 童小芹, 等. 社交电商中融合信任和声誉的图神经网络推荐研究[J]. 中国管理科学, 2021, 29(10): 202–212.
HU C H, DENG A, TONG X Q, et al. A graph neural network recommendation study combing trust and reputation in social e–commerce[J]. Chinese journal of management science, 2021, 29(10): 202–212.

[24] 邵云飞, 宋友, 王宝会. 基于社交网络图节点度的神经网络个性化传播算法研究[J/OL]. 计算机科学: 1–10[2023–02–08]. http://kns.cnki.net/kcms/detail/50.1075.TP.20221228.1215.008.html.
SHAO Y F, SONG Y, WANG B H. Study on personalized propagation algorithm of neural network based on graph node degree of social network[J]. Computer science: 1–10[2023–02–08]. http://kns.cnki.net/kcms/detail/50.1075.TP.20221228.1215.008.html.

[25] 顾希之, 邵鋆侠. 基于影响力剪枝的图神经网络快速计算图精简[J]. 计算机科学, 2023, 50(1): 52–58.
GU X Z, SHAO Y X. Fast computation graph simplification via influence–based pruning for graph neural network[J]. Computer science,

2023, 50(1): 52-58.

[26] 苗旭鹏, 王驭捷, 沈佳, 等. 面向多 GPU 的图神经网络训练加速[J/OL]. 软件学报: 1-14[2023-02-08]. DOI:10.13328/j.cnki.jos.006647.
MIAO X P, WANG N J, SHEN J, et al. Graph neural network training acceleration for Multi-GPUs[J]. Journal of software: 1-14 [2023-02-08]. DOI:10.13328/j.cnki.jos.006647.

[27] 丁恒, 任卫强, 曹高辉. 基于无监督图神经网络的学术文献表示学习研究[J]. 情报学报, 2022, 41(1): 62-72.
DING H, REN W Q, CAO G H. Using unsupervised graphs of neural networks for constructing learning representations of academic papers[J]. Journal of the China society for scientific and technical information, 2022, 41(1): 62-72.

[28] 黄学坚, 刘雨飏, 马廷淮. 基于改进型图神经网络的学术论文分类模型[J]. 数据分析与知识发现, 2022, 6(10): 93-102.
HUANG X J, LIU Y Y, MA T H. Classification model for scholarly articles based on improved graph neural network[J]. Data analysis and knowledge discovery, 2022, 6(10): 93-102.

[29] 蒋昂波, 王维维. ReLU 激活函数优化研究[J]. 传感器与微系统, 2018, 37(2): 50-52.
JIANG A B, WANG W W. Research on optimization of ReLU activation function[J]. Transducer and microsystem technologies, 2018, 37 (2): 50-52.

[30] 黄光红, 林广栋, 吴尔杰, 等. 深度神经网络 Softmax 函数定点算法设计[J]. 中国集成电路, 2022, 31(7): 60-64.
HUANG H L, LIN G D, WU E J, et al. Design of fixed-point algorithm for softmax of DNN[J]. China integrated circuit, 2022, 31(7): 60-64.

Literature Classification Methods based on Structural Information Enhancement

AN Bo[1,2]

(1. Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing 100081;
2. Institute of Software, Chinese Academy of Sciences, Beijing 100190)

Abstract: [Purpose/Significance] Literature classification is a fundamental task in library and information service, which is of great value for information resource management, and literature retrieval and acquisition. Deep learning-based literature classification methods are the current mainstream methods in text classification, which employ neural networks to model and use the textual content for literature classification. This approach only utilizes the information of the literature itself, but ignores the knowledge of the association between the literature. By observing the data, we found that literature in the same category tends to share more keyword information. The literature can build association networks through keywords to form structural relationships between literature. We attempt to utilize this structural in-formation to improve the performance of literature classification. [Methods/Process] This paper proposes a method that can model the structural representation of the literature and employ this representation to enhance traditional literature classification methods. Specifi-cally, we first constructed a large-scale keyword dictionary based on the collected data from about 930,000 documents. Second, we extracted the keyword set from the titles and abstracts of papers by a two-way maximum matching algorithm and constructed the keyword-literature graph data with the literature and keywords as nodes and the inclusion relationship between the documents and keywords as edges. The literature was connected with each other by keywords. Furthermore, we employed graph convolutional neural

DOI：10.13998/j.cnki.issn1002-1248.23-0059

network to model the literature graph and learn the representation of literature and keywords in the keyword-literature graph. The literature representation generated by graph neural network contained the structural relationships between the literature. In addition, we employed Bert+BiLSTM to model the textual content representation of literature. Finally, the structural and textual representations of the literature were concatenated, and the classification of the literature was performed based on this representation. [Results/Conclusions] We constructed a literature classification dataset containing 423 classes and divided the training set, validation set and test set according to the ratio of 8:1:1. We conducted literature classification experiments on this dataset. The experimental results show that the structural information of literature can effectively enhance the performance of traditional literature classification methods. The results of the stripping experiments also show that the structural information alone is insufficient for the literature classification task. Through detailed analysis of the error data, we found that the model still has problems in handling some less frequent keywords and concepts. In the future, we plan to use small-sample learning methods to solve the classification problem for literature categories with less data.

Keywords: literature classification; graph convolution network; keyword-literature graph; semantic association; knowledge organization; natural language processing