计算机科学
COMPUTER SCIENCE

基于文本及历史数据的多标签专利分类算法研究

徐雪洁，王宝会

引用本文

徐雪洁，王宝会. 基于文本及历史数据的多标签专利分类算法研究[J]. 计算机科学，2024，51(5)：172-178.

XU Xuejie, WANG Baohui. Multi-label Patent Classification Based on Text and Historical Data[J]. Computer Science, 2024, 51(5): 172-178.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200199

基于文本及历史数据的多标签专利分类算法研究

徐雪洁　王宝会
北京航空航天大学软件学院　北京 100191
(x_xuejie@buaa.edu.cn)

摘　要　专利分类是专利数据挖掘领域一项非常重要的任务,该任务的目标是为给定专利文献分配若干个国际专利分类(IPC)号,近几年针对该任务的很多研究都集中在通过挖掘专利文本表示对IPC分类体系中部级或大类级分类号的多分类预测。而实际场景中,一篇专利往往有多个分类号,是一种多标签分类任务,且除了专利的文本内容外,每个专利都有对应的专利权组织,专利权组织的历史专利申请行为会有一定的业务倾向,这种申请行为的偏好表示能有效提高专利分类准确度。然而,目前专利分类的相关研究中并没有充分利用到专利的历史数据,针对IPC体系小类的多标签分类问题,提出了一个综合考虑专利内容的专利自动分类模型。首先用BERT预训练语言模型初始化专利文本表示,再利用Text-CNN捕捉局部特征获得将其输出作为专利文本的最终表示;其次,通过Bi-LSTM对历史专利文本及专利标签进行双通道聚合,学习该组织的历史专利申请行为表示;最后,将专利的文本表示与历史专利申请行为表示进行融合后做预测。在真实专利数据集上,将所提模型与基于专利文本挖掘的不同基线进行了对比实验,结果表明基于专利文本和历史数据建模的深度学习分类算法在精确度上有很大的提升。
关键词:深度学习;多标签专利自动分类;IPC分类号;专利
中图分类号　TP312

Multi-label Patent Classification Based on Text and Historical Data

XU Xuejie and WANG Baohui
College of Software,Beihang University,Beijing 100191,China

Abstract　Patent classification,which is used to assign multiple international patent classification(IPC) codes to a given paten,is a very important task int the field of patent data mining. In recent years,many studies on this task focus on mining patent text to predict the first or second level codes for IPC. In real scenarios,a patent often has multiple IPC codes which is a multi-label classification task. Apart from the texts,each patent has a corresponding assignee and the assignee's historical patent application behavior may have a certain business tendency. The preference representation of this behavior can effectively improve the precision of patent classification. However,previous methods fail to make full use of patent historical data. A classification model is proposed for patent automatic classification. Main processing of this model is as follows:firstly,initialize the patent text representation with BERT pretraining language model,then use Text-CNN model to capture local features and take the output as the final patent text representation;secondly,Bi-LSTM is used to learn the preference representation by aggregating historical patent texts and labels through dual channels;finally,we fuse the texts and assignee's sequential preferences for prediction. Experiments on real data set and comparisons with different baselines show that the proposed patent classification algorithm based on patent text and historical data has a great improvement in precision.
Keywords　Deep learning,Automatic classification of multi-label patent,IPC codes,Patent

1　引言

专利分类是根据国际专利分类(IPC)体系,为给定的专利分配多个专利号,它是专利管理的重要内容。传统专利分类任务主要依赖人工结合专业领域知识来进行。

但是,某些原因使得专利分类变得越来越困难。首先,随着IPC分类体系的不断细化,专利分类之间的概念差异变得更不易区分;其次,为了保护专利权,专利文献不仅在内容上专业性强,而且会使用众多复杂且罕见的声明描述[1],需要花费更多的成本去学习专业的领域知识;最后,专利数量不断激增,因此需要投入更多的人力以提升专利分类工作的效率。

除此之外,对专利分类准确度的要求也越来越高。精准的专利分类对于专利申请方而言,可以快速检索同类别专利

到稿日期:2023-02-26　返修日期:2023-06-26
通信作者:王宝会(wangbh@buaa.edu.cn)

信息、参考相关技术、调整声明措辞、评估专利批准机会；对于管理者和审查人而言，能够更好地管理专利、依据相似分类下的专利声明更快速地评估专利是否符合专利申请条件，在提高工作效率的同时减轻工作量。可见，专利分类极具现实意义。

因此，与其他类型的文本分类任务不同，专利自动分类是一项既具有挑战性、又有现实意义且迫切需要提高分类精确度的任务。

近年来，学者们围绕专利自动分类任务进行的研究主要有 3 方面：1）试图用特征工程对专利文本进行分析，研究专利文本某个或多个内容组合对专利分类的影响。如 Li 等[2]通过分析专利引用网络进行分类；Derieux 等[3]将标题、摘要、描述和声明结合起来提取词和短语特征，并增加层级标签和文本特征的语义关联，输入 SVM 分类在 3 种语言的数据集上都有不错的效果；Verberne 等[4]通过提取摘要、元数据、描述以及描述的前 400 个词的特征，利用 Winnow 模型获取了比 KNN 以及 SVM 更好的效果；Bao 等[5]将多示例多标签 MIML 模型引入中文文本分类任务中，通过提取专利标题、摘要、专利说明和专利权力要求书中的特征对 MIML 中 5 种模型下的分类结果做对比，其效果优于传统 SVM 分类。2）通过传统机器学习方法或深度学习算法构建不同分类器，以探索分类效果。例如 Fall 等[6]在 WIPO-alpha 数据集上使用专利、摘要、权利要求书前 300 字作为特征，利用朴素贝叶斯、KNN 及 SVM 等机器学习模型作为分类器，其中 SVM 分类效果最好，在大类分类中准确率达到 67％；Dai 等[7]通过一种基于 XGBoost 的医疗专利大数据自主标引方法来自动提取特征，对比了 Bi-LSTM-Attention 和 TextCNN 两种模型，分类效果有明显提升。3）主要集中于采用深度学习技术捕获文本内容中更充分的表示。如 Roudsari 等[8]和 Jung 等[9]在多标签专利分类任务中基于微调预训练模型 BERT 以及 XLNet 等关注不同预训练模型在分类任务中的效果。

综上，很多学者针对专利分类任务的研究主要集中在利用专利文本挖掘方法实现对 IPC 体系部级和大类级的多分类任务，且主要是针对英文专利，对中文专利数据集、基于 IPC 比较细分的分类的多标签任务研究很少。

除了专利文本内容，每个专利都有所属的专利权组织，而该组织在一定时间内更倾向于发表与其已经发表的专利相似领域内的专利，因此学习专利权组织的历史专利申请行为模式可以提高专利分类的准确度，但是之前的研究中并没有利用到专利的历史信息。鉴于此，本文提出了一个综合考虑专利内容的专利自动分类模型，主要是在专利文本内容挖掘的基础上融合专利权组织历史专利申请行为表示进行预测，模型使用的方法与文本挖掘类研究的方法对比如图 1 所示。在真实专利数据集上，不同基线的实验结果表明，本文提出的专利分类算法在精确度上提升了 4％左右，也验证了融入专利权组织历史专利申请行为表示对提高分类算法准确度的有效性。

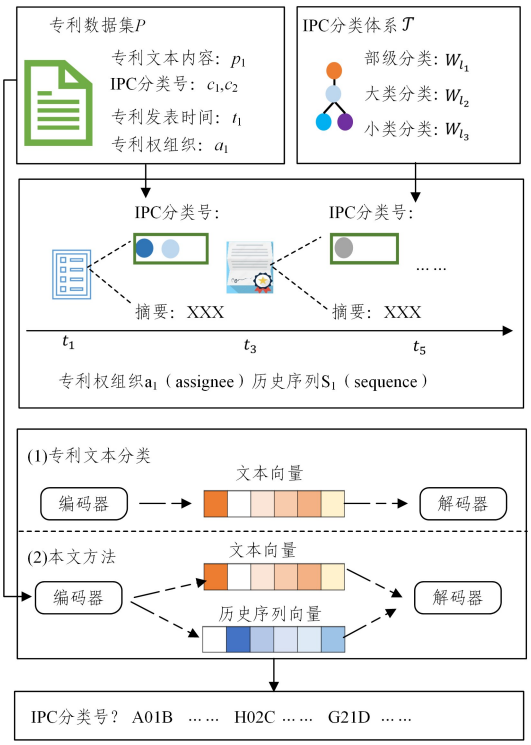

图 1 方法对比

Fig. 1 Methods comparison

2 研究问题定义与建模

本文主要研究的问题是基于深度学习算法学习专利数据充分的表示，实现对 IPC 体系第三级小类做多标签分类预测。

2.1 相关概念

2.1.1 专利权、专利文献

专利是专利权的简称，它指一项发明创造，即专利申请人向国家专利局提出专利申请，经过依法审查合格后，国家专利局向专利申请人授予的在规定的时间内对该发明创造享有的专有权，分为发明、实用创新型或外观设计 3 种。专利文献是在专利申请、审查、批准过程中所产生的各种有关文件的文件资料，由文献[10]可知，专利通常由以下主要部分组成：

1）标题：专利名称。

2）摘要：对发明的简短介绍。

3）描述：对发明更深入的解释说明。

4）著录项目：包含专利号、发布日期、专利权组织、发明人、申请人、国际专利分类码、区域分类码以及审查人员分配的参考文献等。

5）专利权项声明：明确专利保护范围。

6）发明书：技术背景、实施方案等介绍。

2.1.2 国际分类（IPC）体系

本文所用专利数据的分类方法采用国际上通用的专利分类体系——国际专利分类体系（International Patent Classification，IPC）。IPC 分类体系是由高至低依次排列的层级式结构，它将全部技术内容按部、类、大类、组和分组进行分类编码[11]，其层级概念由一般到具体分别是：

1)部:IPC 分类体系的第一级类,是对现有全部技术领域进行的总体分类。共分 8 个部,分类号用大写字母 A-H 表示。

2)大类:第二级类,部下面加以细分的分类,概括地指出专利所包含的技术范围。大类分类号是用部分类号再加上两位阿拉伯数字表示。

3)小类:第三级类,是大类下面加以细分的分类,较具体地规定了所包含的主题范围。小类分类号是用大类分类号加英文字母表示。

4)主组:第四级类,是小类下面加以细分的分类。主组分类号由小类号加 1～3 位阿拉伯数字,后面加一斜线("/"),斜线后面加两个零组成,如 C08F214/00。

5)分组:是主组下进一步细分的第五级类,更具体地规定了所适用的技术特征。分组的分类号是把主组号斜线后面的两个零改为 2～3 位阿拉伯数字表示,如 H04M3/08。

2021 年 1 月更新的 IPC 分类体系中,共包含 8 个部、131 个大类、646 个小类、7523 个主组和 68899 个分组。本文讨论的分类任务主要是针对 IPC 层级中的小类进行预测。

2.2 研究问题建模

2.2.1 研究对象定义

1)国际专利分类体系

本文使用 $T=\{C^{l_i}, l\in L\}$ 表示国际专利分类体系。C 表示 T 中所有可用的 IPC 分类号,L 代表 IPC 分类体系中层级总数,目前 L 为 5 层;l 表示 L 中的某一层级;C^{l_i} 表示分类体系 T 中第 l 层的第 i 个分类号。

2)专利

专利有唯一的专利号以及专利名称、描述、发布时间、专利权组织、IPC 分类号及一些其他信息。在本文中,专利号为 k 的一个专利用 $p_k=(W_k, a_k, t_k, Y_k)$ 表示。其中,W_k 表示 p_k 的文本信息,a_k 表示 p_k 所属的专利权组织,t_k 表示 p_k 的发表时间,Y_k 表示 p_k 的小类专利号集合。本文主要针对 IPC 分类体系中的第三层小类做预测,分类号集合表示为:$y_k=\{y_k^{l_1}, \cdots, y_k^{l_i}\}, y_{k,i}^{l}\in C^{l}, l=3\{0<i\leqslant N\}$。$l_i$ 表示 IPC 分类体系 T 的第 l 层的第 i 个分类,N 表示第 l 层的总共分类数量。当前专利 p_k 的历史专利序列表示为 $S_{u_k}^{t_k}=\{p_i \mid a_{k_i}=a_{k_k}, 0\leqslant t_i\leqslant t_k\}$,表示当前专利权组织 a_{k_k} 在 t_k 之前发表的历史专利序列。

2.2.2 研究问题定义

专利自动分类模型主要通过对已知专利集合 $P=\{W, a, t, Y\}$ 的学习,对新申请的专利 $p_k=(W_k, a_k, t_k)$ 待分配的分类号 Y_k 做预测。该任务被定义为 $\hat{Y}_k=f(W_k, S_{u_k}^{t_k})$。其中 $\hat{Y}_{k,j}$ 表示一个专利文本 p_k 将要被分配的 IPC 分类号,C^{l_i} 表示 IPC 第 l 层的第 j 个分类,$l=3, l\in[1, \cdots, L]$。

IPC 分类号在很大程度上取决于专利文本表示。除文本表示外,专利所属专利权组织的历史专利申请行为也对提升分类效果有帮助。本文融合对专利文本 W_k 及专利权组织的历史专利序列 $S_{u_k}^{t_k}$ 申请行为表示对 IPC 小类进行多标签预测。

3 模型建立

本文提出的专利自动分类模型主要流程如下。首先,采用 Word2Vec 预训练语言模型对训练集中全部专利的摘要进行训练并保存词向量信息,用于后续专利历史序列初始专利文本表示。根据输入的专利摘要文本,通过预训练语言模型 bert_base_chinese 生成专利摘要的文本表示,捕捉到文本序列上下文语义关系,再接入到 Text-CNN 模型,用其输出作为专利文本的最终表示。其次,利用当前专利发表日期之前的该组织的历史专利文本及分类号构建历史申请行为学习模块,以获得该组织的历史专利申请行为表示。最后,对两种表示进行加权融合做出预测。框架中每个过程的具体实现设计及计算将在下文中介绍。模型结构如图 2 所示。

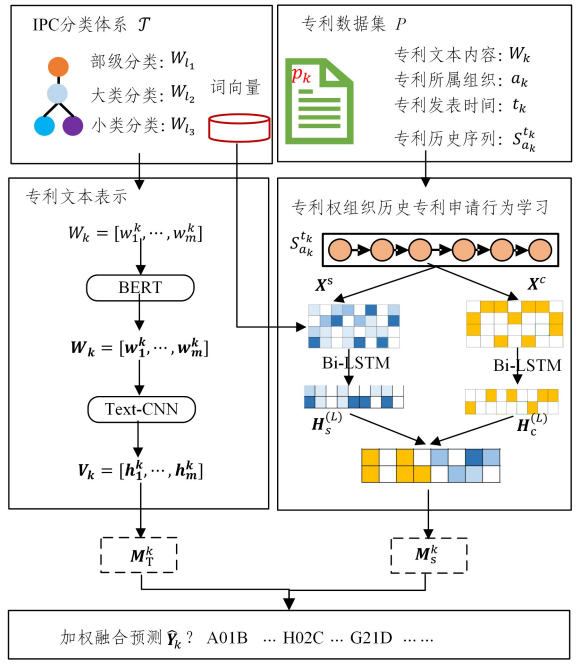

图 2 模型框架

Fig. 2 Model framework

3.1 专利文本表示学习

文本表示学习在自然语言处理中有着非常广泛的应用,如循环神经网络(RNN)系列方法[12-14]、Transformer[15]、Bert 预训练模型系列[16-21]等。已经有很多的方法被应用于解决多标签分类及专利分类问题[22-33]。

考虑到专利数据集数据量很大、文本内容多,同时也鉴于很多学者在预训练语言模型中取得了不错的分类效果的经验,本文采用预训练语言模型获取专利摘要的文本表示。首先,提取训练集中所有专利的摘要内容,将其分词后通过预训练语言模型生成词向量并保存。其次,对于一个专利 p_k,将其摘要部分进行分词后,选择前 M 个词作为输入的文本内容,记为 $W_k=[w_1^k, w_2^k, \cdots, w_M^k]$。根据预训练模型,每个词都将表示为一个 D 维的向量,因此 p_k 的文本向量可以表示为 $\boldsymbol{W}_k=[\boldsymbol{w}_1^k, \boldsymbol{w}_2^k, \cdots, \boldsymbol{w}_M^k]$,$\boldsymbol{W}_k\in R^{M\times D}$。最后,为捕捉词与词之间更多的语义、序列等信息,通过微调进一步获取文本语义表示 $\boldsymbol{V}_k=[\boldsymbol{v}_k^1, \boldsymbol{v}_k^2, \cdots, \boldsymbol{v}_k^M]$,将此语义表示作为专利文本的最终文本表示,记为 $\boldsymbol{M}_T^k$。

3.2 专利权组织历史申请行为学习

一般而言,任意一个专利权组织都有一定的业务范围,

也就意味着该组织在申请专利时会有一些业务相关的倾向，即有一定偏好的专利申请行为。例如，在一定时期内，专利权组织可能在若干个相关的领域内频繁地发表专利。同时，通过对现有数据集中的专利权组织的历史数据进行分析，可以发现专利权组织的历史专利存在一定的序列关系和行为模式。因此，本文设计了一个对专利历史序列进行学习的模块以学习专利权组织的申请行为偏好表示。

对于$p_k=(W_k,a_k,t_k)$，首先用 Word2Vec 词向量对a_k的历史序列$S_{a_k}^{t_k}$中每个专利的文本特征进行初始化，标签特征用 k-hot 生成标签向量，将两个向量分别通过 Bi-LSTM 捕获历史文本和标签文本的上下文信息。最后将两个特征进行融合输出向量，并将其作为历史行为学习表示，处理过程如图 3 所示。

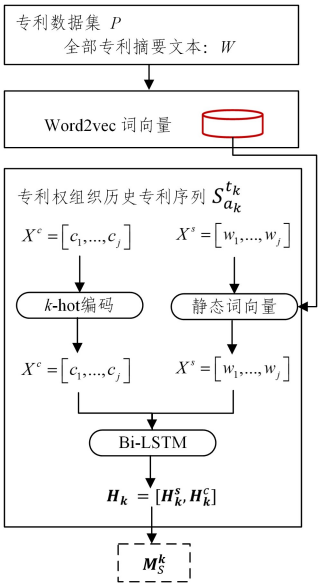

图 3　专利历史序列处理流程

Fig. 3　Process of historical patent data processing

该模块使用两种类型特征，一种是文本特征$\boldsymbol{X}^s\in R^{Q\times D}$，一种是标签特征$\boldsymbol{X}^c\in R^{Q\times N}$。前者是通过预训练语言模型编码的词向量，后者是由独热编码向量编码的向量。

对于专利历史序列$S_{a_k}^{t_k}$，针对每篇专利的文本特征和标签特征分别通过 Bi-LSTM 获取每一篇专利文献的表征，将两种特征进行融合获取到历史行为表示，如式(1)－式(3)所示。

$$\boldsymbol{H}_s^{(L)}=Bi\text{-}LSTM([\boldsymbol{X}_S^0,\cdots,\boldsymbol{X}_s^M])\in\mathbb{R}^{M\times D} \tag{1}$$

$$\boldsymbol{H}_c^{(L)}=Bi\text{-}LSTM([\boldsymbol{X}_c^0,\cdots,\boldsymbol{X}_c^N])\in\mathbb{R}^{M\times D} \tag{2}$$

$$\boldsymbol{M}_s=[\boldsymbol{H}_s^{(L)},\boldsymbol{H}_c^{(L)}] \tag{3}$$

其中，$\boldsymbol{M}_s\in R^{Q\times 2F}$是专利$p_k$基于标签特征和文本特征融合后的向量，在预测中以$\boldsymbol{M}_s^k$作为当前专利对应的专利权组织的申请行为偏好表示。

3.3　多标签专利分类预测

将专利文本表示和专利历史序列申请行为表示进行加权融合来做最终的预测。计算方式如式(4)所示：

$$\hat{Y}_k=\sigma(g_T(\boldsymbol{M}_T^k)+\lambda g_S(\boldsymbol{M}_S^k)) \tag{4}$$

其中，$g_T(\cdot)$和$g_B(\cdot)$分别是两个两层感知机网络实现的解码器；λ 表示历史专利申请行为权重。该专利分类任务属于多标签分类，也就是针对专利第三层的每个类别都将输出一个概率值$y_{k,i}^l$，所以 $\sigma(\cdot)$采用的是 sigmoid 函数。$\hat{Y}_k\in R^{C^l}$是当前专利在第 l 层所分配 IPC 分类号的概率。

4　实验验证

4.1　数据集

本文实验采用中文专利数据集，该数据集涵盖了中国技术公司从 1900－2020 年发表的 200 多万条专利数据。经过 4.2 节的数据预处理后的数据集统计信息如表 1 所列。

表 1　实验所用数据集

Table 1　Datasets used in experiment

数据集	CNPTD-200k	
专利权组织	1523	
专利数据	训练集	117726
	验证集	41357
	测试集	40966
国际专利分类	部	8
	大类	124
	小类	608

4.2　数据处理

本小节结合本文的算法设计以及数据集分布情况，对原始专利数据进行如下的数据清洗及预处理：

1)按照最新 IPC 体系编码规则，过滤掉不符合规则的数据，同时去除带有空值的数据。

2)删除数据中专利权组织一对多的数据，结合专利权组织历史数据的算法设计，去除发表专利数量少于 30 篇的专利权组织的专利数据。

3)根据年份，按“部”统计专利数量，如图 4 所示。为保证数据分布的均衡和准确，本文采用 2010－2019 年的专利数据，并将 2010－2017 年的专利数据作为训练集，2018 年的专利数据作为验证集，2019 年的专利数据作为测试集，按照这一年份划分，保留在 3 个时间范围内都发表过专利的专利权组织的专利信息。

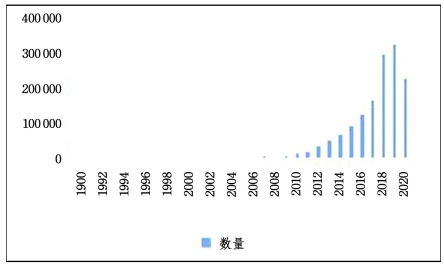

图 4　按“部”统计专利分类数量

Fig. 4　Patent number statistics by section

4)按照数据集划分的年份区间，对每个专利权组织发表的专利按照发表时间倒序排序，最多获取前 400 篇专利，其中最多 300 篇作为训练集，最多 50 篇作为验证集，最多 50 篇作为测试集。当数据集总数超过 20 万时停止并完成构建数据集。对于数据集分类号，用索引重定向将其处理为数值变量，最终获取专利数据集的总数为 200049。

5)利用训练集中的摘要数据，通过预训练语言模型 word2vec 保存专利词向量，用于初始化专利文本表示。

数据集经过上述预处理过程的数量变化如图5所示。

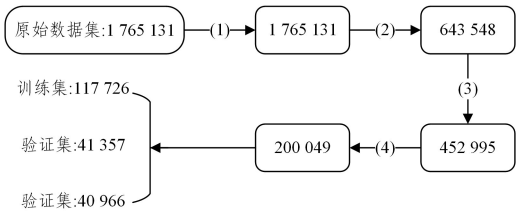

图5 专利数量变化

Fig. 5 Changes in the number of patents

4.3 实验环境

本文使用Pytorch10.2,在Python 3.8环境下实现代码,所有实验均在Tesla T4,CUDA10.2的Linux服务器上完成。

4.4 实验分析

4.4.1 实验设置

实验中epoch设置为100;LSTM隐藏层维度设置为256;基于word2vec的模型,输入前100词,学习率设置为0.001,batchsize设置为32;基于BERT的模型输入512字符,学习率设置为3×10^{-5},隐藏层768维,batchsize设置为8;将Adam作为优化器;为防止过拟合,dropout设置为0.3;并增加early stop机制,若训练效果经过5个epoch后没有提升,则提前停止训练。选择达到最佳性能的模型进行测试。

在训练过程中,将多标签文本分类任务进一步转化为多个二元分类问题,以最小交叉熵损失为优化目标,损失函数如式(5)所示:

$$L=-\sum_{p_i\in P}\sum_{p_c^j\in C^j}Y_{i,j}\log(\hat{Y}_{i,j})+(1-Y_{i,j})\log(1-\hat{Y}_{i,j}) \quad (5)$$

4.4.2 评估指标

本文主要使用4个评估指标对不同模型效果进行综合评估,分别是精确度、召回率、F1值和归一化折损累计增益(Normalized Discount Cumulative Gain,NDCG)。

1)精确率,衡量在预测的分类号中有多少是样本的真实分类号。将所有专利的平均精确度作为第一个评估指标,如式(6)所示:

$$P@K(p)=\frac{1}{\mathcal{N}}\sum_{i}^{\mathcal{N}}\frac{|\hat{C}_i\cap C_i|}{|\hat{C}_i|} \quad (6)$$

其中,$\hat{C}_i=\{c_i^1,\cdots,c_i^j\}$,$1<j<N$,表示针对专利$p_i$预测出的IPC分类体系中第三层级的分类号即预测标签集合,c_i^j表示第三层级分类中的第j个分类号,N为第三层中所有分类总数。C_i表示p_i的真实分类号即真实标签集合,$|C_i|$表示集合C_i的长度,N表示测试集中专利总数。

2)召回率,衡量在真实的分类号中有多少被准确预测出来。将所有专利的平均召回率作为第二个评估指标,计算式如式(7)所示:

$$R@K(p_i)==\frac{1}{\mathcal{N}}\sum_{i}^{\mathcal{N}}\frac{|\hat{C}_i\cap C_i|}{|C_i|} \quad (7)$$

其中,$\hat{C}_i$,$\mathcal{N}$和C_i与精确率中的意义相同。

3)F1值指标,其为精确率和召回率的调和平均数。F1值能够比较全面地反应分类器性能,将所有专利的平均F1值作为第三评估指标,如式(8)所示:

$$F1@K(p_i)=\frac{1}{\mathcal{N}}\sum_{i}^{\mathcal{N}}2*\frac{P@K(p_i)*R@K(p_i)}{P@K(p_i)+R@K(p_i)} \quad (8)$$

4)NDCG,通过所有标签的顺序来衡量排名质量。将所有专利的平均NDCG作为第四个评估指标,计算式如式(9)所示:

$$NDCG@K(p_i)=\frac{\sum_{k=1}^{K}\frac{\tau(\hat{C}_i^k,C_i)}{\log_2(k+1)}}{\sum_{k=1}^{K}\frac{\tau(C_i^k,C_i)}{\log_2(k+1)}} \quad (9)$$

以预测值的索引在真值中位置的值为增益。其中,$\hat{C}_i^k$表示专利p_i被预测的第k个标签;$\tau(\hat{C}_i^k,C_i)$表示当预测的分类号$\hat{C}_i^k$包含在真值集合C_i中时值为1,否则为0。

4.4.3 方法对比

本文中所有方法都基于Top-K性能进行对比,结果如表2所列。从对比结果中可以看到,所提模型与分类效果最好的基线模型相比仍有一定的提升。

表2 对比实验结果

Table 2 Comparison experiment results

Metrics	Word2VecBERT				BERT			Improvement/%
	Text-CNN	LSTM	Bi-LSTM	Ours	BERT	Text-CNN	Ours	
Precision@1	0.6402	0.6549	0.6752	0.6811	0.6783	**0.6853**	**0.6966**	**1.65**
Recall@1	0.5414	0.5543	0.5715	0.5761	0.5740	**0.5808**	**0.5908**	**1.72**
F1@1	0.5716	0.5850	0.6031	0.6081	0.6058	**0.6127**	**0.6231**	**1.70**
NDCG@1	0.6402	0.6549	0.6494	0.6811	0.6783	**0.6942**	**0.6966**	**0.35**
Precision@3	0.3192	0.3235	0.3401	0.3429	0.3374	**0.3395**	**0.3398**	**0.71**
Recall@3	0.7560	0.7657	0.8016	0.8078	0.7965	**0.8000**	**0.8020**	**0.69**
F1@3	0.4345	0.4401	0.4619	0.4656	0.4587	**0.4610**	**0.4618**	**0.68**
NDCG@3	0.7042	0.7160	0.7469	0.7522	0.7527	**0.7547**	**0.7563**	**0.27**
Precision@5	0.2119	0.2149	0.2237	0.2259	0.2216	**0.2228**	**0.2232**	**0.54**
Recall@5	0.8207	0.8303	0.8623	0.8701	0.8562	**0.8599**	**0.8603**	**0.43**
F1@5	0.3273	0.3316	0.3451	0.3483	0.3420	**0.3437**	**0.3442**	**0.50**
NDCG@5	0.7327	0.7446	0.7728	0.7797	0.7729	**0.7801**	**0.7783**	**0.23**

尤其在专利文本表示并不充分的情况下,提升更明显。通过对比分析得出了以下结论:

首先,对比基于word2vec的前四个模型,TextCNN通过卷积获取文本局部重要特征的分类效果比LSTM差,且LSTM的效果比Bi-LSTM差,表明通过学习专利文本上下文语序及语义关系的全局特征,更能充分得表示专利文本,获取更好的分类效果。其中本文模型是在Bi-LSTM的基础上融合专利历史序列申请行为表示,可见历史数据申请行为表示

的增加能够显著提升分类效果。

其次，对比基于 BERT 的前两个模型，其中单纯依靠 BERT 进行专利分类的效果已经接近本文模型，BERT 对比 word2vec 及 LSTM 等时序模型能够获取到更充分的文本表示。在 BERT 基础上再增加 Text-CNN 捕捉局部重要特征，分类效果有了一定提升。

最后，本文模型在基于 Bert＋TextCNN 获取到专利文本表示的基础上，融合以 Bi-LSTM 双通道聚合机制获取的专利历史序列申请行为表示，最终预测结果达到目前最佳，Top1 精确率达到 69.6%。这进一步表明了本文模型在多标签专利自动分类上的有效性。

4.4.4 消融实验和性能实验

本文的主要创新点在于，首先，专利文本表示除了通过 BERT 获取全局语义和语序特征外，还通过 TextCNN 以 100 个卷积核以及 2，3，4 这 3 种不同的卷积尺寸捕捉全局特征下的重点局部特征，经过融合后的专利文本表示，能够进一步提升专利分类效果；其次，本文在文本表示基础上对专利历史序列的申请行为表示做加权融合，进一步提高了专利分类准确率。消融实验对比结果如图 6 所示，可以看到，通过 TextCNN 捕捉局部特征后，在各个评估指标上都有一定提升，在 BERT＋TextCNN 基础上融合历史序列申请行为表示，分类效果又有了进一步提升。

通过对比基于 word2vec＋ours 与 BERT＋ours 方法的分类效果可以发现，专利文本表示越充分，历史序列申请行为表示的影响作用就越弱；同时，我们在实验中也对历史模块做了权重实验，以权值 0.1 融合历史序列申请行为表示，达到了最好的效果，不仅能弥补专利表示上的不充分，也尽可能降低了历史序列中领域跨度很大的专利带来的噪声影响。

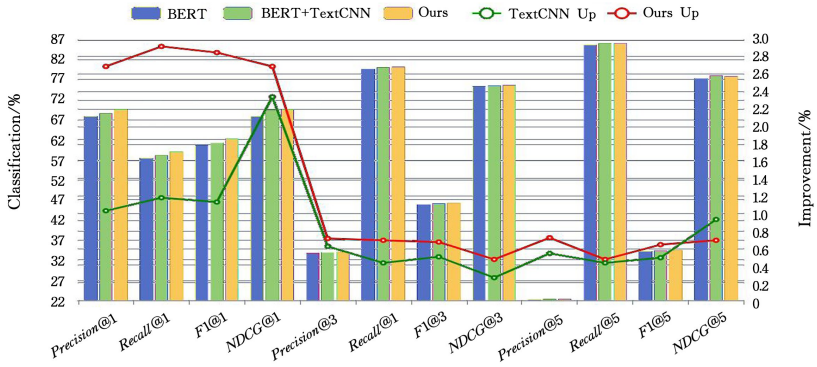

图 6 消融实验

Fig. 6 Ablation experiment

综上，本文的专利自动分类模型，不仅在文本上获得了更充分的表示，同时也验证了历史专利序列申请行为能够提升专利分类效果，在对 IPC 第三层小类多标签自动分类任务中取得了不错的效果。

本文为了防止过拟合，同时减少训练时间，设置 Early Stop 机制，基于参考文献常用参数值，设置如果连续 5 个 Epoch 中 NDCG 指标没有提升，则提前结束运行，该参数应用于本文模型和所有基线模型中。

按照常用 Early Stop 设置值，又针对 10，20 的不同值进行实验。整体上，对分类效果的影响差异很小，在值为 10 时分类效果 top1 指标相对有提高，值为 20 时分类效果反而有所下降，值为 5 时分类效果在 Top1 指标上稍有影响，但训练时间最短，收敛效果也不错。使用专利历史数据学习申请行为表示，数据量相对较大。综合考虑到性能、时间和效率等因素，最终选择 Early Stop 为 5 作为最终参数值。

结束语 专利文献复杂的结构、专业的内容以及不断增加和细化的专利分类体系使得专利自动分类更具挑战性。目前，很多学者的研究都集中在专利文本信息挖掘上，对 IPC 体系较高层级分类进行多分类预测，对于专利的其他内容如历史数据、IPC 层级小类级以下层级的、专利多标签分类等研究很少。本文充分考虑专利结构和内容，在一定数据分析基础上，提出了一个用于专利多标签分类任务的模型。首先，在专利文本表示方面，采用自然语言处理领域效果比较好的 BERT 预训练语言模型进行初始化，并通过文本卷积网络捕捉重要局部信息，得到更充分的文本表征。其次，除了考虑文本信息外，还考虑了专利权组织发表专利的历史序列，以学习专利权组织的历史专利申请偏好来增加预测效果。最终，实现了对 IPC 体系小类级别的多标签专利自动分类模型，通过精确率、召回率、F1 值和归一化折损累计增益等评估指标验证了本文模型在所有对比的模型中效果最优。

一些前沿的技术，如强化学习、迁移学习、动态图网络、chatGPT 等，推动了自然语言处理领域的飞速发展，目前，研究者们利用 chatGPT 做了基于问答的专利分类效果，其对于专利文本中心词的总结以及给定分类类别后对内容进行归类的效果都不错。OpenAI 提供了 chatGPT 调用接口，能够基于 chatGPT 通过输入自己的数据训练模型，未来可以基于 chatGPT 在构建专利自动分类模型方面做进一步的研究与实践。

参 考 文 献

[1] ABDELGAWAD L，KLUEGL P，GENC E，et al. Optimizing neural networks for patent classification[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham：Springer，2020：688-703.

[2] LI X，CHEN H，ZHANG Z，et al. Automatic patent classification using citation network information：an experimental study in nanotechnology[C]// Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. 2007：419-427.

[3] DERIEUX F，BOBEICA M，POIS D，et al. Combining semantics and statistics for patent classification[C]// DBLP. 2010.

[4] VERBERNE S, D'HONDT E. Patent classification experiments with the Linguistic Classification System LCS in CLEF-IP 2011 [C]// CLEF. 2011.
[5] BAO X, LIU G F, CUI J H. Application of Multi Instance Multi Label Learning in Chinese Patent Automatic Classification[J]. Library and Information Service, 2021, 65(8): 107-113.
[6] FALLC J, TÖRCSVÁRI A, BENZINEB K, et al. Automated categorization in the international patent classification[J]. ACM SIGIR Forum, 2003, 37(1): 10-25.
[7] DAI P J, HE C L, SHANYUE Y R. XGBoost-based Classification of Multi-label Texts of Pharmaceutical Patent[J]. Journal of Neijiang Normal University, 2021, 36(10): 55-60.
[8] HAGHIGHIAN ROUDSARI A, AFSHAR J, LEE W, et al. PatentNet: multi-label classification of patent documents using deep learning based language understanding[J]. Scientometrics, 2022, 127(1): 207-231.
[9] JUNG G, SHIN J, LEE S. Impact of preprocessing and word embedding on extreme multi-label patent classification tasks[J]. Applied Intelligence, 2023, 53(4): 4047-4062.
[10] GOMEZ J C, MOENS M F. A survey of automated hierarchical classification of patents[M]// Professional Search in the Modern World. Cham: Springer, 2014: 215-249.
[11] TIAN C, ZHAO Y J. A mapping model of patent and industry category based on similarity: A case study of International Patent Classification and Trade Classification of National Economy [J]. Library and Information Service, 2016, 60(20): 123.
[12] ELMAN J L. Finding structure in time[J]. Cognitive science, 1990, 14(2): 179-211.
[13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[14] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv: 1406. 1078, 2014.
[15] GRAVES A. Generating sequences with recurrent neural networks[J]. arXiv: 1308. 0850, 2013.
[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing SystemsDecember. 2017: 6000-6010.
[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301. 3781, 2013.
[18] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014: 1532-1543.
[19] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810. 04805, 2018.
[20] WOLF T, DEBUT L, SANH V, et al. Huggingface's transformers: State-of-the-art natural language processing[J]. arXiv: 1910. 03771, 2019.
[21] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv: 1907. 11692, 2019.
[22] GRAWE M F, MARTINS C A, BONFANTE A G. Automated patent classification using word embedding[C]// 2017 16th IEEE International Conference on Machine Learning and Applications(ICMLA). IEEE, 2017: 408-411.
[23] LI S, HU J, CUI Y, et al. DeepPatent: patent classification with convolutional neural networks and word embedding[J]. Scientometrics, 2018, 117(2): 721-744.
[24] SHALABY M, STUTZKI J, SCHUBERT M, et al. An lstm approach to patent classification based on fixed hierarchy vectors [C]// Proceedings of the 2018 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2018: 495-503.
[25] HUANG W, CHEN E, LIU Q, et al. Hierarchical multi-label text classification: An attention-based recurrent network approach[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1051-1060.
[26] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(1): 7370-7377.
[27] TANG P, JIANG M, XIA B N, et al. Multi-label patent categorization with non-local attention-based graph convolutional network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(5): 9024-9031.
[28] ROUDSARI A H, AFSHAR J, LEE C C, et al. Multi-label patent classification using attention-aware deep learning model [C]// 2020 IEEE International Conference on Big Data and Smart Computing(BigComp). IEEE, 2020: 558-559.
[29] GOMEZ J C. Analysis of the effect of data properties in automated patent classification[J]. Scientometrics, 2019, 121(3): 1239-1268.
[30] LYU L, HAN T. A comparative study of Chinese patent literature automatic classification based on deep learning[C]// 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2019: 345-346.
[31] FANG L, ZHANG L, WU H, et al. Patent2Vec: Multi-view representation learning on patent-graphs for patent classification [J]. World Wide Web, 2021, 24(5): 1791-1812.
[32] SHEN J, QIU W, MENG Y, et al. TaxoClass: Hierarchical multi-label text classification using only class names[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 4239-4249.
[33] ZHAO H Y, CAO J, CHEN Q K, et al. Methods for Hierarchical Multi-label Text Classification. Journal of Chinese Computer Systems. 2022, 43(4): 673-683.

XU Xuejie, born in 1988, postgraduate. Her main research interests include natural language processing and so on.

WANG Baohui, born in 1973, master, professor. His main research interests include big data, artificial intelligence and network information security.

(责任编辑:何杨)