浙江科技学院学报，第 33 卷第 6 期，2021 年 12 月
Journal of Zhejiang University of Science and Technology
Vol. 33 No. 6, Dec. 2021

doi：10.3969/j.issn.1671-8798.2021.06.003

基于 word2vec 和 logistic 回归的中文专利文本分类研究

程　盼，徐弼军
（浙江科技学院 理学院，杭州 310023）

摘　要：专利文本作为重要的信息载体，对其实现自动分类具有重要的研究意义。针对海量的专利文本，提出一种基于 word2vec 和 logistic 回归的中文专利文本分类模型的机器学习方法。本方法利用 word2vec 产生的词向量对专利文本进行表示，然后配合 logistic 回归模型，对专利说明和摘要合并的文本语料进行学习和训练，从而实现对专利文本的自动分类。试验结果表明，我们提出的机器学习方法能够得到较好的分类效果，其中个别类别的分类准确率达到了 84%；并且与 k 近邻算法相比，该模型在精确度、召回率及 F_1 值方面均有显著提高。本方法可为专利文本自动分类提供可靠的研究依据。

关键词：中文专利；文本分类；word2vec；logistic 回归；机器学习
中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1671-8798(2021)06-0454-07

Research on Chinese patent text classification based on word2vec and logistic regression model

CHENG Pan, XU Bijun
(School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: As the carrier of important information, patent text is of great research significance for its automatic classification. In response to the immense amount of patent text, a machine learning method was proposed on the basis of word2vec and logistic regression for Chinese patent text classification model. This method harnessed the word vector generated by word2vec to represent patent text, and then in combination with the logistic regression model, implemented machine learning and training on the text corpus integrating the patent description with the abstract, so as to realize automatic classification of patent text. The results show that the proposed machine learning method can achieve sound classification effect, among which the classification accuracy of individual categories reaches 84%. Moreover, compared with the k-

收稿日期：2020-11-25
基金项目：浙江省自然科学基金项目(Y20F050002)
通信作者：徐弼军(1979—　)，男，浙江省兰溪人，教授，博士，主要从事深度学习、知识产权统计分析研究。E-mail：xubijun@zust.edu.cn。

nearest neighbor algorithm, it has significant improvements in precision, recall and F_1 value. This method can provide reliable research basis for automatic classification of patent text.

Keywords: Chinese patent; text classification; word2vec; logistic regression; machine learning

专利作为重要的技术信息载体，包含着各种有价值的重要研究成果，其数量也在不断攀升[1]，于是面对海量的专利文本，对其进行合理的分类显得尤为重要[2]。目前，专利文本分类还是以人工为主，但随着专利数量的迅速增长，若仅靠专利审查员的专业素质与经验来进行分类，则无法满足高效和准确的实际需求[3]。

近年来，国内外对专利文本分类进行了大量的研究。Cassidy[4]提出了一种改进的朴素贝叶斯算法，并在来自世界专利信息(World Patent Information，WPI)测试集中的 7 309 项专利组成的语料库上进行了测试，结果表明在利用极少的数据进行训练时 F_1 值仍能够达到 34.26%。Li[5]等提出了一种基于卷积神经网络(convolutional neural networks，CNN)和单词嵌入向量的深度学习算法 DeepPatent，并将其在数据集 CLEF-IP 与新数据集 USPTO-2M 上分别进行了测试，其精确度分别达 83.98%与 73.88%。贾杉杉等[6]提出了一种多特征多分类器方法，对多种特征分别用多个分类器进行测试，准确率最高达 91.2%。胡云青[7]提出的改进的三体训练法半监督模式，能够动态改变分类器对相同未标记样本预测类别的概率阈值，并且在训练样本只有少数标记的情况下 F_1 值最高达 70.6%。

在海量的专利文本中，为了提高专利审查员的分类效率，提升对专利信息的组织管理水平，因此引进机器学习来对专利文本进行自动分类非常必要[8]。但是专利文本中大量专业术语的使用，特定于行业中的语言降低了词汇密度，并导致算法的搜索空间稀疏；此外，有意的非标准化语言虽然可以帮助申请人扩大专利范围或减少侵权的可能性，但这会给机器学习带来噪声，使其难以找到清晰的模式[9-10]。为了克服这些限制，我们提出了一种新的机器学习方法，利用 word2vec 模型的词向量进行文本表示，配合 logistic 回归模型来实现对专利文本的自动分类。

1　word2vec 模型与 logistic 回归模型的原理

1.1　word2vec 模型的原理

word2vec 是一种用来产生词向量的神经网络概率语言模型，由 Mikolov 等[11]在 2013 年提出。它可以根据给定的文本数据，在快速有效地优化训练模型后将一个词转换成向量形式。该算法依据连续词袋(continuous bag-of-words，CBOW)模型和 Skip-gram 模型来进行训练，两个模型的结构如图 1、图 2 所示，图中方框表示词汇的向量。CBOW 模型的输入层为当前词汇的上下文词汇的独热向量(one-hot 向量)，经过投影层对上下文词汇的词向量进行累加计算，最后输出层输出预测的当前词汇的词向量。Skip-gram 模型的输入层为当前词汇的 one-hot 向量，为了与 CBOW 模型对比，Skip-gram 模型也加入一个投影层，但此投影层只对输入的当前词汇的向量进行加权，最后经过输出层输出当前词汇的上下文词汇的词向量。

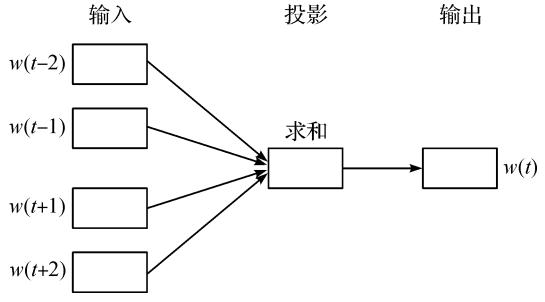

图 1　CBOW 模型的结构

Fig. 1　Structure of CBOW model

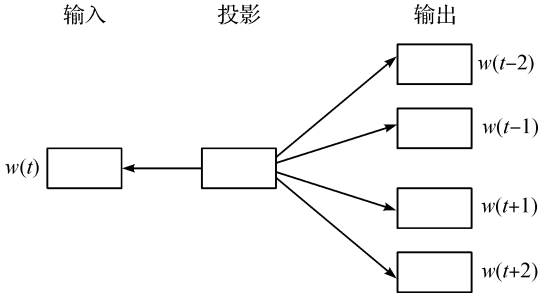

图 2　Skip-gram 模型的结构

Fig. 2　Structure of Skip-gram model

由图1可知,CBOW模型是在已知上下文词汇 $w(t-2)$、$w(t-1)$、$w(t+1)$、$w(t+2)$的前提下来预测当前词汇 $w(t)$,简言之,即通过上下文的内容来预测当前词汇。它的学习目标就是最大化对数似然函数,其数学表达式为

$$L_{\text{CBOW}} = \sum_{w \in C} \log_2 p(w \mid \text{Context}(w))。 \tag{1}$$

式(1)中:w 为当前语料库 C 中任意一个词;Context(w)为当前词语 w 的上下文。

而Skip-gram模型是在已知当前词汇 $w(t)$后,预测其上下文词汇 $w(t-2)$、$w(t-1)$、$w(t+1)$、$w(t+2)$。其目标函数的数学表达式为

$$L_{\text{Skip-gram}} = \sum_{w \in C} \log_2 p(\text{Context}(w) \mid w)。 \tag{2}$$

在本研究的实际操作过程中,Skip-gram模型的训练时间与CBOW模型相比更长,但其精度优于CBOW模型。因此,为了获得更好的分类效果,本研究选取Skip-gram模型[12]。

1.2 logistic回归模型的原理

logistic回归模型[13]是统计学中比较经典的分类算法。它虽然被称为回归,但实际上却是一种用于分类的模型。它的因变量有二分类、多分类,本研究利用它进行多分类。

设 X 是连续随机变量,X 服从logistic分布是指 X 具有下列分布函数和密度函数:

$$F(x) = P(X \leqslant x) = \frac{1}{1 + \mathrm{e}^{-\frac{x-\mu}{\gamma}}}; \tag{3}$$

$$f(x) = F'(x) = \frac{\mathrm{e}^{-\frac{x-\mu}{\gamma}}}{\gamma\left(1 + \mathrm{e}^{-\frac{x-\mu}{\gamma}}\right)^2}。 \tag{4}$$

式(3)~(4)中:γ 为形状参数;μ 为位置参数。

二项logistic回归模型由条件概率分布 $P(Y|X)$表示,形式为参数化的logistic分布。其中,随机变量 X 为实数,随机变量 Y 的取值范围为{0,1}。将 $\boldsymbol{x} \in \mathbb{R}^n$ 作为输入,$Y \in \{0,1\}$作为输出,则二项logistic回归模型可用以下条件概率分布来表示:

$$P(Y = 1 \mid \boldsymbol{x}) = \frac{\mathrm{e}^{\boldsymbol{w}\boldsymbol{x}+b}}{1 + \mathrm{e}^{\boldsymbol{w}\boldsymbol{x}+b}}; \tag{5}$$

$$P(Y = 0 \mid \boldsymbol{x}) = \frac{1}{1 + \mathrm{e}^{\boldsymbol{w}\boldsymbol{x}+b}}。 \tag{6}$$

式(5)~(6)中:$\boldsymbol{w}$、b 均为参数,$\boldsymbol{w} \in \mathbb{R}^n$,$b \in \mathbb{R}$;$\boldsymbol{w}$ 为权值向量;b 为偏置;$\boldsymbol{w}\boldsymbol{x}$ 为 $\boldsymbol{w}$ 和 $\boldsymbol{x}$ 的内积。

为了方便计算,有时会将权值向量和输入向量进行扩充,仍然记作 $\boldsymbol{w}$、$\boldsymbol{x}$,即 $\boldsymbol{w} = (w_1, w_2, \cdots, w_n, b)^{\mathrm{T}}$,$w_i$ 表示权值向量 $\boldsymbol{w}$ 的第 i 个分量,$\boldsymbol{x} = (x_{(1)}, x_{(2)}, \cdots, x_{(n)}, 1)^{\mathrm{T}}$,$x_{(i)}$ 表示输入向量 $\boldsymbol{x}$ 的第 i 个分量,$i = 1,2,\cdots,n$。这时,logistic回归模型如下:

$$P(Y = 1 \mid \boldsymbol{x}) = \frac{\mathrm{e}^{\boldsymbol{w}\boldsymbol{x}}}{1 + \mathrm{e}^{\boldsymbol{w}\boldsymbol{x}}}; \tag{7}$$

$$P(Y = 0 \mid \boldsymbol{x}) = \frac{1}{1 + \mathrm{e}^{\boldsymbol{w}\boldsymbol{x}}}。 \tag{8}$$

logistic回归模型在训练学习时,对于给定的训练数据集 $T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$ $(\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \{0,1\})$可以利用极大似然估计法来估计模型参数。

为了解决多类分类问题,将二项logistic回归模型进行推广。假设{1,2,…,K}为离散型的随机变量 Y 的取值集合,那么多项logistic回归模型为

$$P(Y = j \mid \boldsymbol{x}) = \frac{\mathrm{e}^{\boldsymbol{w}_j \boldsymbol{x}}}{1 + \sum_{j=1}^{K-1} \mathrm{e}^{\boldsymbol{w}_j \boldsymbol{x}}}, j = 1,2,\cdots,K-1; \tag{9}$$

$$P(Y = K \mid \boldsymbol{x}) = \frac{1}{1 + \sum_{j=1}^{K-1} \mathrm{e}^{\boldsymbol{w}_j \boldsymbol{x}}}。 \tag{10}$$

式(9)～(10)中：$\boldsymbol{x}\in\mathbb{R}^{n+1}$，$\boldsymbol{w}_j\in\mathbb{R}^{n+1}$。

2　试验设置与实现

2.1　试验环境

本研究所有的试验都是基于 Windows10 操作系统，CPU 型号为 Core i5，主频为 3.00 GHz，内存大小为 8 GB，编程语言使用Python3.7版本，用到了Sklearn、Jieba、Pandas等多方库。本试验数据为从万方数据知识服务平台下载的 2019 年中国已授权的专利文本数据，这些数据按照国际专利分类号(International Patent Classification，IPC)进行了标记。IPC 分类号采用“部—类—组”的层次分类方法，层次越低，文本的相似度就越高[14]。为了方便研究，本研究从数据库中获取的是“部”类别为 H 的电学领域的专利文本，分别为 H01、H02、H03、H04、H05，其中每个类别数量都为 6 000 篇，一共 30 000 条的专利数据作为语料库。各专利分类号的具体含义见表 1。

表 1　专利分类号含义

Table 1　Explication of patent classification number

分类号	含义
H01	基本电气元件
H02	发电、变电或配电
H03	基本电子电路
H04	电通信技术
H05	其他类目不包含的电技术

2.2　试验设计

由于专利说明是对该专利的进一步解释，因此，我们在试验中将专利说明和摘要的内容进行文本合并后，构成了最终参与试验的语料库[15]。为了方便试验，人工划分出了训练集与测试集，其中训练集包含每个类别的 5 000 条数据，测试集包含每个类别的 1 000 条数据。整个试验设计的流程如图 3 所示。

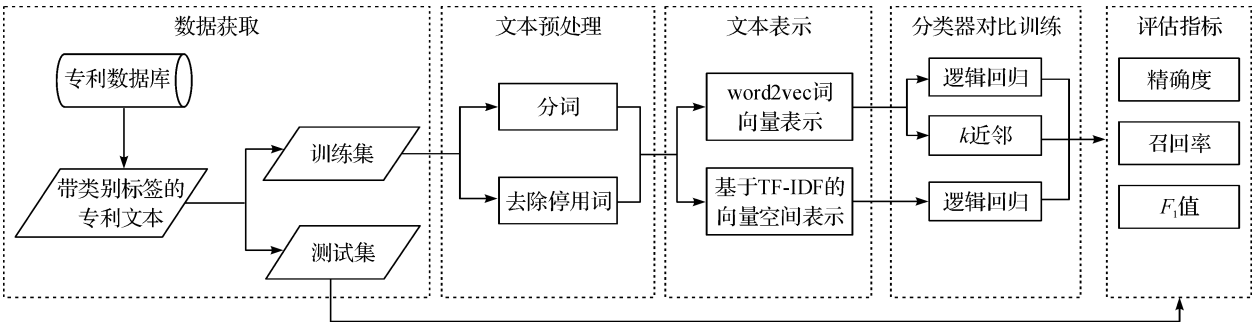

图 3　试验设计流程

Fig. 3　Flow chart of experiment design

对数据集进行人工标注后，开始进行文本预处理，其中主要包括分词和去除停用词[16]。由于文本处理中最基础的单元就是词汇，因此，需要将整个语料进行分词。我们使用 Python 的第三方库 Jieba 对专利文本进行有效的分词。然后再去除停用词，即助词、符号等对整个语义没有影响的词，本研究采用的是收录较为完整的《哈尔滨工业大学停用词表》。最后，将完成分词和去除停用词后的文档进行保存，该文档将全部由词汇组成。

将预处理之后的文本数据利用 Python 中的第三方库 Gensim 的 word2vec 模型进行训练，从而得到词汇的词向量表[17]。经过一系列的调试之后，我们发现将词向量维数设置为 200，迭代次数设置为 10，其他参数保持不变时，该模型的精度达到最优。

为了验证本研究提出的 word2vec＋logistic 模型的有效性和优越性，设计了两组试验进行对比。一组试验为在相同的分类器 logistic 回归模型下，采用不同的文本表示方法，即 word2vec 模型和向量空间模型进行比较。其中 word2vec 模型是将得到的词向量与分词之后的文档进行词汇映射，得到每个词汇具体的词向量值，然后求出文档的平均值，从而得到文档的词向量表示；而传统的向量空间模型主要采用的是 one-hot 编码，即将所有待分类的语句中不重复的词汇全部提取出来，形成词典，将词汇出现在词典中的位置记为 1，其他位置记为 0，然后使用词频-逆向文档频率(TF-IDF)

特征权重计算方法来加权表示[18]。另一组试验为在相同的文本表示方法 word2vec 模型下，采用不同的分类器 logistic 回归模型与 k 近邻(k-nearest neighbor，KNN)算法在相同数据集上进行对比试验。

3 试验结果评估与分析

3.1 评估指标

采用精确度 P、召回率 R 及 F_1 值来评价模型的分类效果。假设将关注的类别标签记作正类，其余的类别标签记作负类，则分类器的预测结果在测试集上有正确和不正确两种[19]。其中，精确度以预测结果为判断依据，召回率以实际样本为判断依据。

精确度可表示为

$$P=\frac{a}{a+b}。\tag{11}$$

式(11)中：a 为实际为正例的样本数；b 为实际为负例的样本数。

召回率可表示为

$$R=\frac{a}{a+c}。\tag{12}$$

式(12)中：c 为预测错误的样本数。

F_1 值是精确度和召回率的调和均值，其表达式为

$$F_1=\frac{2PR}{P+R}。\tag{13}$$

3.2 结果及分析

3.2.1 word2vec＋logistic 模型与 TF-IDF＋logistic 模型对比分析

为了验证词向量模型进行文本表示的优势，设计了在采用同种分类器下，利用词向量模型 word2vec 与向量空间模型 TF-IDF 分别进行文本表示的分类结果对比试验[20]。

在对利用 word2vec 模型进行文本表示的语料进行训练的过程中，采用了十折交叉验证，word2vec＋logistic 模型在训练集上的平均准确率达到了 71%。由图 4 可知，将模型保存之后应用在测试集上，各类别的准确率分别为 69%、64%、84%、76%、61%，平均准确率达到了 70%左右，这与在训练集上的结果相差不大。而用同样的数据在对利用 TF-IDF 进行文本表示的语料进行训练时，经过交叉验证之后，TF-IDF＋logistic 模型在训练集上的平均准确率仅为 42%，而在测试集上，平均准确率也就 40%左右。

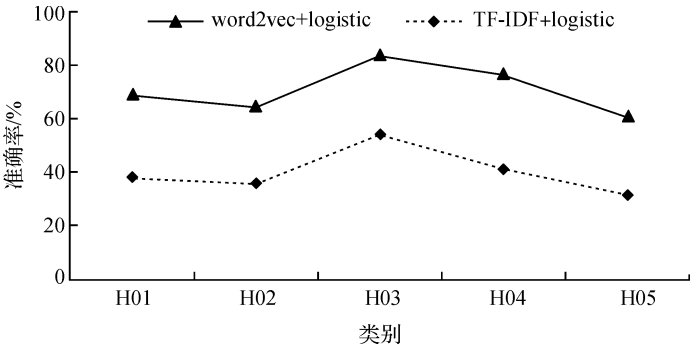

图 4 word2vec＋logistic 与 TF-IDF＋logistic 在测试集上各类别的准确率

Fig. 4 Accuracy of word2vec＋logistic and TF-IDF＋logistic classification on test set

表 2、表 3 所示的是 word2vec＋logistic 模型与 TF-IDF＋logistic 模型分类结果的精确度(P)、召回率(R)与 F_1 值，通过对比可以看出，使用向量空间模型的专利文本分类效果较差，除了 H03 类之外，其他类 F_1 值只达到了 35%左右。导致其分类效果低的原因可能是专利文本中不同领域拥有各种专有名词，而向量空间模型只会对词汇做比较简单的区分，并且向量的维度极高，样本之间的特征太过稀疏化。而基于 word2vec 的词向量文本表示，可以表达词汇之间的相似度，对近义词进行区分，其分类结果在各项指标上明显地要优于向量空间模型，相比之下，基本上所有类别的 F_1 值都提高了 30%左右。

表 2　基于 word2vec＋logistic 模型的文档分类结果

Table 2　Document classification results based on word2vec＋logistic model　%

类别	P	R	F_1 值
H01	66	67	66
H02	66	64	65
H03	78	84	81
H04	76	76	76
H05	66	61	63

表 3　基于 TF-IDF＋logistic 模型的文档分类结果

Table 3　Document classification results based on TF-IDF＋logistic model　%

类别	P	R	F_1 值
H01	38	39	38
H02	39	31	35
H03	42	67	52
H04	46	35	40
H05	34	28	31

3.2.2　word2vec＋logistic 模型与 word2vec＋KNN 模型对比分析

KNN 分类模型作为最简单的、经典的机器学习模型，在分类问题上被广泛使用，因此，选取 KNN 模型来与 logistic 回归模型进行比较。KNN 模型是在特征空间中通过计算待测样本与训练样本间的距离，得出与待测样本相邻最近的 k 个样本中的大多数属于哪一类别，则该待测样本也属于这个类别[21]。

图 5 所示的是 word2vec＋logistic 模型与 word2vec＋KNN 模型在测试集上各类别的准确率，从图中可以看出，在进行十折交叉验证之后，word2vec＋KNN 模型在整个测试集上的平均准确率为 63％左右，而 word2vec＋logistic 模型平均准确率为 70％左右，比 word2vec＋KNN 模型提高了 7％左右。

表 4 所示的是 word2vec＋KNN 模型在测试集上的分类结果，与表 2 所示的 word2vec＋logistic 模型的分类结果比较，可以发现 logistic 回归模型各个类别的 F_1 值最大提高了 10％。究其原因，在选取数据时，我们是基于部随机选取的专利样本，每个部下面还有很多类、组，其类组之间有的可能存在较大的差异，KNN 模型是靠邻近的 k 个点来判断，这就导致当出现样本不均衡问题时，其分类效果会变差；其次，KNN 模型中 k 值大小的选择没有理论上的最优值，在训练过程中发现，随着 k 值越来越大，模型精度的确有所提升，但是这只会让模型变得简单，这并不是一个较好提升模型精度的方式，而 logistic 回归模型不依赖于样本之间的距离。因此，在本试验中，logistic 回归模型充分表现出了它的优势。

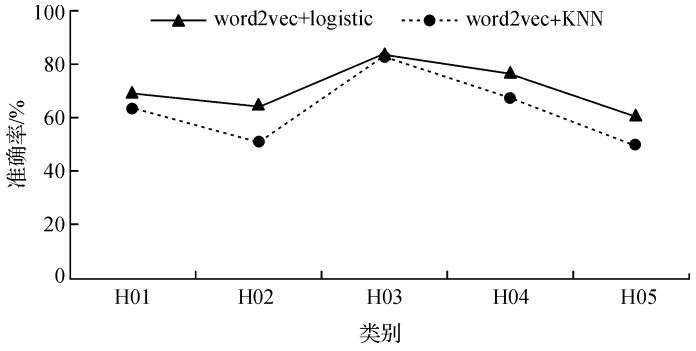

图 5　word2vec＋logistic 与 word2vec＋KNN 在测试集上各类别的准确率

Fig. 5　Accuracy of word2vec＋logistic and word2vec＋KNN classification on test set

表 4　基于 word2vec＋KNN 模型的文档分类结果

Table 4　Document classification results based on word2vec＋KNN model　%

类别	P	R	F_1 值
H01	60	64	62
H02	54	51	53
H03	74	83	78
H04	67	67	67
H05	56	50	53

4　结　语

针对中文专利文本的自动分类问题，本研究提出了一种新的机器学习方法，利用 word2vec 进行文本表示，用 logistic 回归作为分类器的专利文本分类模型，并与传统的向量空间模型进行文本表示及利用 KNN 模型作为分类器进行比较。经过理论分析和试验评估发现，与传统的向量空间模型采用 TF-IDF 进行文本表示相比，word2vec 模型在进行文本表示时，可以很好地区分专利文本中相似的特征，并且 logistic 回归模型与 KNN 模型相比，在分类效果上其精确度、召回率、F_1 值都有了显著的提高。此外，本

研究的模型还可以推广到其他专利类别的文本分类上，后续工作将是进一步研究优化，以获得更优的分类效果。

参考文献：

[1] 高颀. 基于“Effect-theme”共现网络的专利分类方法[J]. 信息技术与信息化，2020(4)：137.

[2] 胡杰，李少波，于丽娅，等. 基于卷积神经网络与随机森林算法的专利文本分类模型[J]. 科学技术与工程，2018，18(6)：268.

[3] 缪建明，贾广威，张运良. 基于摘要文本的专利快速自动分类方法[J]. 情报理论与实践，2016，39(8)：104.

[4] CASSIDY C. Parameter tuning Naïve Bayes for automatic patent classification[J]. World Patent Information，2020，61(1)：101968.

[5] LI S，HU J，CUI Y，et al. DeepPatent：patent classification with convolutional neural networks and word embedding[J]. Scientometrics，2018，117(2)：721.

[6] 贾杉杉，刘畅，孙连英，等. 基于多特征多分类器集成的专利自动分类研究[J]. 数据分析与知识发现，2017，1(8)：77.

[7] 胡云青，邱清盈，余秀，等. 基于改进三体训练法的半监督专利文本分类方法[J]. 浙江大学学报(工学版)，2020，54(2)：333.

[8] 王瑞杨. 基于深度学习的专利分类方法研究[D]. 天津：河北工业大学，2018.

[9] ZHU H，HE C，FANG Y，et al. Patent automatic classification based on symmetric hierarchical convolution neural network[J]. Symmetry，2020，12(2)：186.

[10] RISCH J，KRESTEL R. Domain-specific word embeddings for patent classification[J]. Data Technologies and Applications，2019，53(1)：108.

[11] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[C]// International Conference on Learning Representations. Scottsdale：ICLR，2013：4.

[12] 高明霞，李经纬. 基于 word2vec 词模型的中文短文本分类方法[J]. 山东大学学报(工学版)，2019，49(2)：36.

[13] 李航. 统计学习方法[M]. 2 版. 北京：清华大学出版社，2019：91-94.

[14] 薛金成，姜迪，吴建德. 基于 LSTM-A 深度学习的专利文本分类研究[J]. 通信技术，2019，52(12)：2890.

[15] 吕璐成，韩涛，周健，等. 基于深度学习的中文专利自动分类方法研究[J]. 图书情报工作，2020，64(10)：75.

[16] 包翔，刘桂锋，杨国立. 基于多示例学习框架的专利文本分类方法研究[J]. 情报理论与实践，2018，41(11)：144.

[17] KIM J，CHOI K. Patent document categorization based on semantic structural information[J]. Information Processing and Management，2007，43(5)：1200.

[18] 薛金成，姜迪，吴建德. 基于 word2vec 的专利文本自动分类研究[J]. 信息技术，2020，44(2)：75.

[19] 周志华. 机器学习[M]. 北京：清华大学出版社，2016：30-32.

[20] 谢剑芳，田英明，徐旭，等. 基于 FastText 的专利文本自动分类方法研究[J]. 仪器仪表标准化与计量，2020(4)：22.

[21] XIE J，HOU Y，WANG Y，et al. Chinese text classification based on attention mechanism and feature-enhanced fusion neural network[J]. Computing，2020，102(3)：683.