基于 CNN 的程序编译错误信息特征提取

何烨辛[1],谷　林[1],孙　晨[2]
(1. 西安工程大学 计算机科学学院,陕西 西安 710048;
2. 西安科技大学 管理学院,陕西 西安 710054)

摘　要:伴随着互联网行业的迅速发展,在自然语言处理领域中,有效地将输入表示为固定长度的特征向量是机器学习算法中的一个重要研究方向。海量的编译错误信息不仅可以用于程序错误相似度的研究,也可将编译错误信息进行聚类、分类之后给教师在计算机编程类课程的教育教学中给予针对性的指导。这些应用的根本在于高效地提取编译错误信息特征。该文提出了一种基于 word2vec 模型结合卷积神经网络(convolutional neural networks,CNN)对编译错误信息进行特征提取的方法,首先利用 word2vec 工具中的 skip-gram 模型以词向量的形式表示编译错误信息,然后利用 CNN 神经网络完整地表征编译错误信息特征向量。有效地从可变长度的编译错误信息中学习固定长度的特征表示。最后使用支持向量机(SVM)分类算法进行实验结果的验证。结果表明,该特征提取方法在编译错误信息中有显著的效果。

关键词:word2vec;编译错误信息;skip-gram 模型;CNN;支持向量机
中图分类号:TP391.1;TP18　　**文献标识码**:A　　**文章编号**:1673-629X(2021)05-0204-05
doi:10.3969/j.issn.1673-629X.2021.05.035

CNN-based Program Compilation Error Message Feature Extraction

HE Ye-xin[1],GU Lin[1],SUN Chen[2]
(1. School of Computer Science,Xi'an Polytechnic University,Xi'an 710048,China;
2. School of Management,Xi'an University of Science and Technology,Xi'an 710054,China)

Abstract:With the rapid development of the Internet industry,in the field of natural language processing,the effective representation of input as fixed length feature vectors is an important research direction in machine learning algorithms. Massive compilation error message can not only be used to study the similarity of program error,but also to cluster and classify the compilation error message to give teachers specific guidance in the education and teaching of computer programming courses. The essence of these applications lies in the efficient extraction of compiler error message characteristics. We propose a method of feature extraction of the compiled error message based on word2vec model and convolutional neural networks (CNN). At first,the compiled error message is represented by the skip-gram model in the word2vec tool in the form of word vector,and then the complete characteristic vector of the compiled error message is represented by the CNN. Effectively learn fixed-length feature representations from variable-length compile error message. Finally, SVM classification algorithm is used to verify the experimental results. It is showed that the feature extraction method is effective in compiling error message.

Key words:word2vec;compile error message;skip-gram model;CNN;SVM

0　引　言

在互联网飞速发展与数据大爆炸的时代,海量的数据如何有效地进行特征提取是自然语言处理领域一个重要的研究方向。在自然语言处理领域和人工智能领域中,高效地实现人与计算机之间用自然语言进行有效通信的各种理论和方法是现今学者们广泛研究的内容。在计算机学科相关教育教学中,编译错误信息是衡量学生代码正确性和代码质量的一个重要指标,而编译错误信息特征的提取不仅可以有效地进行特征的聚类与分类等研究,还可以为计算机学科的教育教学中学生编程问题提供针对性的指导,从而提高学生的编程效率,提升编程兴趣。在文本特征提取中,目前已有的特征提取方法包含词袋模型[1-3]、信息增益(information gain,IG)[4-5]、词频-逆向文件频率模型[6-9](term frequency-inverse document frequency,TF-IDF)以及 word2vec 文本特征提取[10-11]等相关方法。文献

收稿日期:2020-07-03　　**修回日期**:2020-11-05
基金项目:国家重点研发计划课题(2017YFF0210506)
作者简介:何烨辛(1994-),女,硕士研究生,研究方向为智能信息系统化管理;谷　林,副教授,研究方向为智能信息系统化管理。

[12]提出了一种方法用来解决从英语和意大利语文本中以无监督的方式提取关键词或短语的问题。这种方法的主要特征是由两种方法集成,单词嵌入模型(例如 word2vec 或 GloVe 能够捕获单词及其上下文的语义)和聚类算法(能够识别术语的本质)并选择较重要的一个或多个来表示文本的内容以更优的方式实现英语和意大利语关键词或短语提取。文献[13]使用 word2vec 将单词表示为矢量形式的模型计算英语单词之间的相似度。使用英语维基百科中的 320 000 条文章作为语料库,利用余弦相似度计算方法确定相似度值。然后,该模型通过测试集黄金标准 WordSim-353(多达 353 对单词)和 SimLex-999(多达 999 对单词)进行测试,并根据人类判断将它们标记为相似值。皮尔逊相关性用于找出相关性的准确性。文献[14]使用 word2vec 模型进行短文本的词向量生成,使用该方法的前提是忽略不同词性的词语对短文本的影响力,引入词性改进特征权重计算方法,将词性对文本分类的贡献度嵌入到传统的 TF-IDF 算法中计算短文本中词的权重,并结合 word2vec 词向量生成短文本向量,最后利用 SVM 实现短文本分类。文献[15]利用词间向量余弦的相似性构造了一种迭代算法来识别相似词。算法利用种子情感词在通用的汉语情感词汇(DSL、NSL 和 HSL)中,自动生成新的替代情感词。最后,情感词自动从备选词中选择相似情感词的词汇比较和统计分析方法,利用 word2vec 自动构建用于教育目的特定领域的汉语情感词典;文献[16]寻求通过将无监督的深度神经网络技术与词嵌入方法相集成来提高 ATS 的质量。首先,开发了基于单词嵌入的文本摘要,并且展示了 word2vec 表示比传统的 BOW 表示提供了更好的结果。其次,结合 word2vec 和无监督特征学习方法提出其他模型,以合并来自不同来源的信息。文献[11]是使用维基百科语料库,通过 word2vec 训练得到相应的词模型,然后加权分配模型中的词向量与对应词的 TF-IDF 值,建立了对应的短文本特征表达方法。

根据上述研究,目前基于 word2vec 文本特征提取方法的研究立足不同视角,涉足多个方面,国内外学者使用基于 word2vec 词向量模型的应用十分广泛,而对于编译错误信息文本数据的研究目前仍比较少。该文主要是针对 Java 程序的编译错误信息,希望能够提取该序列主要成分,即编译错误信息特征。相对于单一使用 word2vec 词向量表示模型,word2vec 和卷积神经网络相结合的模型可以增强数据的精度,增大数据的处理量,更好地表示编译错误信息特征向量。因此该文主要使用的方法是结合 word2vec 工具中的 skip-gram 模型与 CNN 神经网络,以无监督学习的方式从大量编译错误信息中学习语义信息,达到编译错误特征提取的目的。

1　基于 word2vec 的编译错误信息特征提取模型

该文构建了一种适用于编译错误信息的特征化表示方法,运用 word2vec 和 CNN 神经网络相结合的方法提取编译器错误信息的语义特征,最终形成编译错误信息空间向量模型的特征化表示。具体流程如图 1 所示。

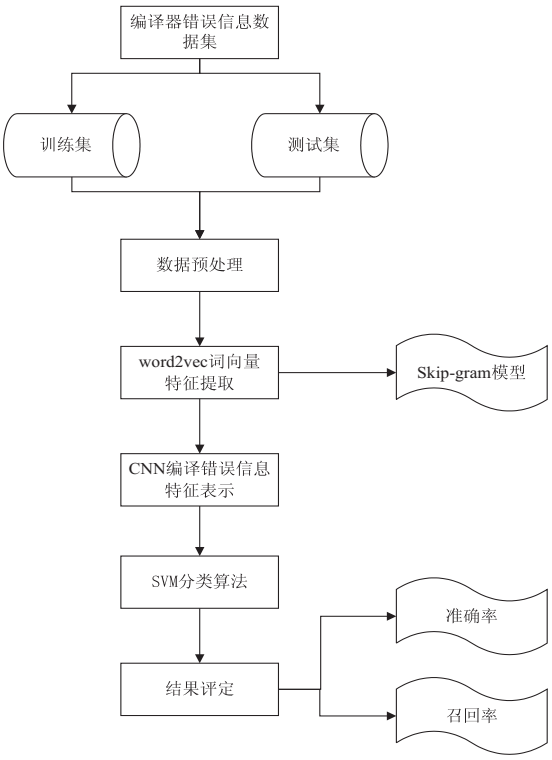

图 1　编译错误信息特征提取流程

1.1　word2vec

机器学习相关算法中实现计算机与人之间利用自然语言进行通信的首要任务是用数学数字表示文本特征,最常见的表示方式为使用词向量来表示某个词语。在现今的研究中,利用神经网络模型训练大量的数据是获得词向量表示的主要方法,神经网络模型表示词向量的基本思想不但可以包含词间的潜在语义关系,同时可以高效地避免维数灾难。word2vec 最先是由 Mikolov 在 2013 年提出的[17],是一种基于浅层神经的词向量生成模型。该模型的主要作用是求得低维度的词向量,并对相关数据词语与上下文语义之间关系进行建模。该方法得出的词向量的维度一般处于 100 ~ 300 之间,可以更好地解决传统空间向量模型高维稀疏的问题。近年来,word2vec 被广泛应用到文本特征提取工作中。word2vec 主要包含两个模型:跳字模型(skip-gram)和连续词袋模型(continuous bag of words,

CBOW)。skip-gram 是给定 input word 来预测上下文,而 CBOW 是给定上下文来预测 input word。

CBOW 模型包括输入层、投影层和输出层。假设以(context(w),w)为例,context(w)构成是由 w 前后各 x 个词。其中输入层中的词向量涵盖 context(w)中 $2x$ 个词 v(context(w)1),v(context(w)2),…,v(context(w)2x);投影层是求和累加输入层中的 $2x$ 个向量,如公式(1)。其技术原理如图 2(a)所示。

$$x_\omega = \sum_{i=1}^{2C} V(\text{Context}(\omega)_i) \in R^m \tag{1}$$

skip-gram 模型包括三层,分别为输入层、投影层和输出层。其中,输入层输入当前特征词,词的词向量 $W_t \in R^m$;输出为该特征词上下文窗口中词出现的概率;投影层的目的是使目标函数 L 值最大化。假定有一组词序列 $\omega_1,\omega_2,\cdots,\omega_N$,则:

$$L = \frac{1}{N}\sum_{j=1}^{N}\sum_{-c\leqslant i\leqslant c}\log p(\omega_{j+1} \mid \omega_j) \tag{2}$$

其中,N 为词序列的长度;c 为当前特征词的上下文长度;$p(\omega_{j+1} \mid \omega_j)$ 为在已知当前词 ω_j 出现的概率下,其上下文特征词 ω_{j+1} 出现的概率。通过 skip-gram 模型训练得到的全部词向量,组成词向量矩阵 $\boldsymbol{X} \in R^{mn}$。以 $x_i \in R^m$ 表示特征词 i 在 m 维空间中的词向量[18]。其技术原理如图 2(b)所示。

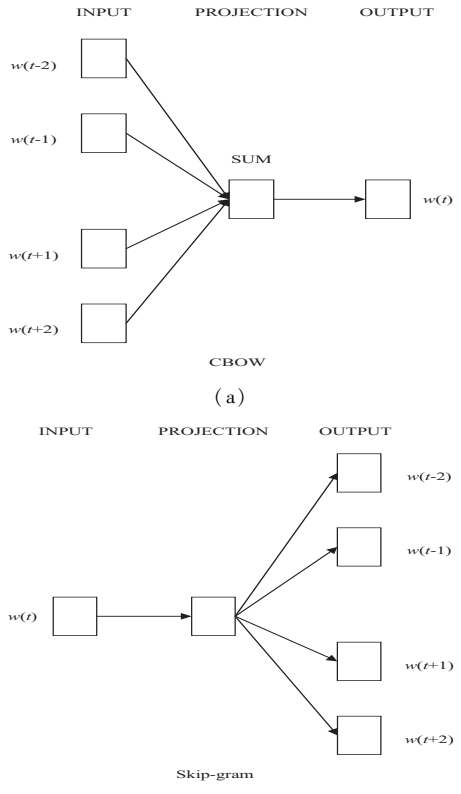

图 2 技术原理

该文将采用 word2vec 工具中的 skip-gram 模型进行编译错误信息文本词向量生成的相关训练。主要是利用 Hierarchical Softmax 构造的一棵 Huffman 树作为 word2vec 模型的输出层训练词向量。其具体流程如下所示:

步骤 1:取词完成。采用分词-滑动窗口的方式进行取词,如图 3 所示。

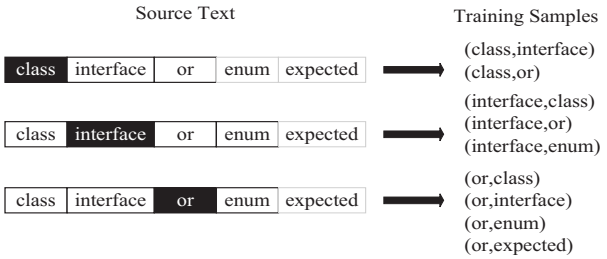

图 3 分词-滑动窗口取词

步骤 2:构造编译错误信息词典并统计词频。

步骤 3:构造树形结构。核心技术是 Hierarchical Softmax 构造的一棵 Huffman 树;生成节点并初始化各非叶节点的中间向量和叶节点的词向量,如图 4 所示。

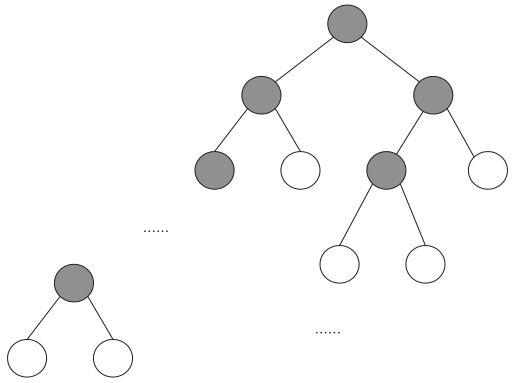

图 4 Huffman 树结构图

步骤 4:训练中间向量和词向量。

1.2 CNN 编译错误信息表示模型

卷积神经网络(convolutional neural network,CNN)是一种多层神经网络。其构建主要是模仿生物的视知觉,可用来进行监督学习和非监督学习,是深度学习的代表算法之一。卷积神经网络隐含层内的卷积核参数共享和层间连接的稀疏性使得卷积神经网络能够以较小的计算量对格点化(grid-like topology)特征。该神经网络模型包含 3 层结构,分别为输入层、卷积层和“池化+连接层”,如图 5 所示。

1.2.1 输入层

输入层又称数据输入层,主要是对原始数据进行预处理;其中包括去均值、归一化和 PCA 降维。

该文利用 word2vec 工具的 skip-gram 模型训练编译错误信息语料库中的数据,生成对应词向量。则输入层输入的数据为编译错误信息序列中各个词汇对应的词向量,是一个表示编译错误信息句子的矩阵,如公式(3)所示。其维度为 $m \times n$,其中每个词是由一个 n 维的词向量进行表示。

$$D_i = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{3}$$

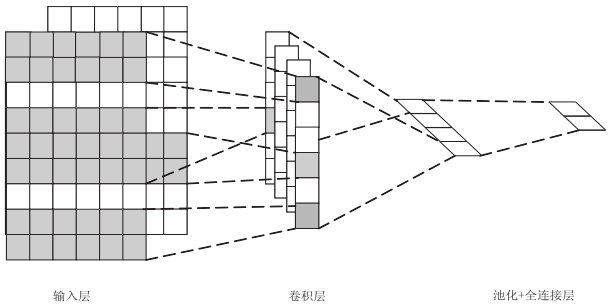

图 5　卷积神经网络结构

1.2.2　卷积层

卷积层是卷积神经网络最重要的一个层次,主要有两个关键的操作,分别是局部关联和窗口滑动。

在卷积层中输入的是公式(3)所示的维度为 $m \times n$ 的矩阵,用来表示编译错误信息句子矩阵。在卷积层进行卷积操作时,卷积核的宽度的选取与词向量的维度保持一致,这样的方式可以保证卷积核在每次滑动过的位置是一个完整的词向量。

卷积计算如公式(4)所示:

$$y_i = f(\sum \boldsymbol{w}_i \cdot x_{i:i+h-1} + b) \tag{4}$$

其中,$\boldsymbol{w}_i$ 表示为卷积核的权重矩阵,$\boldsymbol{x}_{i:i+h-1}$ 表示第 i 行到 $i+h-1$ 行的词向量矩阵,b 表示偏置,函数 f 表示激活函数。

当编译错误信息经过卷积相关操作后,最终可以得到一个 $n-h+1$ 维的向量 $\boldsymbol{y}$,如公式(5)所示。

$$\boldsymbol{y} = [y_1, y_2, \cdots, y_{n-h+1}] \tag{5}$$

1.2.3　池化+连接层

池化是存在于连续的卷积层中间的,主要作用是用于数据和参数量的压缩,充分减少过拟合现象,更方便优化。连接层是所有的神经元的权重连接,通常是处于卷积神经网络的尾部进行。

卷积层之后,通常需要在 CNN 之间添加池化。池化的主要作用是不断降低维数,减少 CNN 神经网络中的参数和计算次数。极大地缩短了训练时间并控制过度拟合。目前最常见的池类型是最大池化(max pooling),它在每个窗口中占用最大值。将卷积层输出的编译错误信息特征向量中提取最大值表示为最重要的编译错误信息特征,相同卷积核卷积池化后的标量组合得到这个窗口大小的特征向量,将所有窗口下的特征向量进行连接层的连接,最终组合成为完整的编译错误信息的特征向量。

1.3　SVM 分类方法

SVM(support vector machine),又称支持向量机,是二分类模型中的一种。它的主要原理是定义在特征空间上的间隔最大的线性分类器。是一种基于统计学习理论的机器学习方法,由贝尔实验室中的 Vapnik 首次提出。支持向量机学习的主要目的就是用于寻找类别间隔最大化的分类边界,最终将所求解问题转化为一个凸二次规划问题并进行求解。其步骤如下:

(1)将数据转为支持向量机包的格式;

(2)对数据进行归一化处理;

(3)优先选择径向基核函数;

(4)通过交叉验证寻找最佳函数;

(5)使用最佳参数训练编译错误信息数据。

1.4　结果评定

结果评定可以直观地证明方法使用的准确性。该文采用文本分类技术中经常使用的评价指标作为实验结果评定的依据。为了验证模型的编译错误信息特征提取效果,针对结果的测评,该文引入以下三个指标对分类器的结果进行评价,分别为准确率、召回率和 F1-Score 值,其中 F1-Score 值是准确率和召回率的综合指标,如式(6)、(7)所示。

(1)准确率(precision)。

$$\text{Precision}(c_j) = \frac{\text{TP}(c_j)}{\text{TP}(c_j) + \text{FP}(c_j)} \tag{6}$$

(2)召回率(recall)。

$$\text{Recall}(c_j) = \frac{\text{TP}(c_j)}{\text{TP}(c_j) + \text{FN}(c_j)} \tag{7}$$

(3)F1-Score 值。

$$\text{F1-Score}(c_j) = \frac{2\text{Precision}(c_j) \times \text{Recall}(c_j)}{\text{Precision}(c_j) + \text{Recall}(c_j)} \tag{8}$$

其中,$\text{TP}(c_j)$ 表示应为 c_j 的样本且被正确分成 c_j 类的样本数;$\text{FN}(c_j)$ 表示应为 c_j 类的样本没有被正确分成 c_j 类的样本数;$\text{FP}(c_j)$ 表示不为 c_j 类的样本但被分为 c_j 类的样本数。

2　实验过程与结果分析

2.1　数据来源与预处理

2.1.1　数据来源

该文采用某校 2018 级软件工程系四个班级共 151 名学生《面向对象程序设计课程》学习期间所提交的 Java 编程作业,主要选取的是在实现单一方法时学生编写 Java 代码的短片段所遇到的相关错误提示。将 Java 编译错误信息分为三类,分别为语法错误、语义错误和逻辑错误。语法错误是指程序中单词的拼写,标点和顺序等错误。语义错误是处理代码的含义,是由错误的编程语言解释方式导致的错误。这些类型的错误大多数是 Java 和类似语言所特有的,但是比语

法错误更抽象。逻辑错误一般是指编程过程中发生导致偏离程序本身意思的错误,程序可以通过编译并成功运行,但是运行结果与期望值不符。编译错误信息语料库中有错误信息 652 条,其中训练集文本数 520 条,测试集文本数 132 条。

2.1.2 数据预处理

在实验开始之前,首先需要对数据进行降噪处理。需要删除编译错误信息文本数据中的标点符号和一些无意义的常用词,在编译错误信息中利用正则表达式读取无意义常用词和数字并剔除它,同时将符号利用正则表达式替换为一个空格,用空格将数据进行分割。

2.2 实验结果分析

实验中利用机器学习的支持向量机(SVM)算法将编译错误信息特征进行分类实验,测评结果采用所有类别的平均值进行展示。实验结果如表 1 所示。使用 word2vec 结合 CNN 的方法可以大大减少直接对文本向量化而出现部分编译错误文本特征丢失,从而降低编译错误信息文本分类的准确性。相较单一使用 word2vec 进行编译错误信息特征提取的方法,word2vec 与 CNN 相结合的方法对编译错误信息特征表示的效果显著提升。

表 1 实验结果 %

类别	准确率	召回率	F1-Score 值
word2vec	91.3	91.2	91.2
word2vec+CNN	93.2	92.5	92.8

3 结束语

该文提出了一种基于 word2vec 与 CNN 神经网络相结合的编译错误信息特征提取的方法。为了更全面且准确地使用向量表示编译错误信息,采用 word2vec 工具中 skip-gram 模型对编译错误信息文本进行低维的词向量表示,然后结合卷积神经网络提取编译错误信息文本中最重要的特征表示,在池化+连接池层中组合成完整的编译错误信息特征表示。最后经过相关实验,通过 SVM 分类方法将 word2vec 模型和 word2vec+CNN 模型进行结果比对。结果表明 word2vec+CNN 方法能较好地适用于编译错误信息特征提取。下一步将针对该模型进一步进行优化,以提高编译错误信息特征提取的准确性,更好地得到编译错误信息的有效分类。

参考文献:

[1] POLAP D, WLODARCZYK-SIELICKA M. Classification of non-conventional ships using a neural bag-of-words mechanism[J]. Sensors, 2020, 20(6): 1608-1621.

[2] KADRIU A, ABAZI L, ABAZI H. Albanian text classification: bag of words model and word analogies[J]. Business Systems Research Journal, 2019, 10(1): 74-87.

[3] 黄春梅,王松磊. 基于词袋模型和 TF-IDF 的短文本分类研究[J]. 软件工程,2020,23(3):1-3.

[4] 邱宁佳,高 鹏,王 鹏,等. 基于改进信息增益的 ACO-WNB 分类算法研究[J]. 计算机仿真,2019,36(1):295-299.

[5] WEBER G F. Information gain in event space reflects chance and necessity components of an event[J]. Information, 2019, 10(11): 358-368.

[6] TU Shouzhong, HUANG Minlie. Mining microblog user interests based on TextRank with TF-IDF factor[J]. The Journal of China Universities of Posts and Telecommunications, 2016, 23(5): 40-46.

[7] 周 欣. 改进的 TF-IDF 特征选择和短文本分类算法研究[D]. 合肥:安徽大学,2020.

[8] 但唐朋,许天成,张姝涵. 基于改进 TF-IDF 特征的中文文本分类系统[J]. 计算机与数字工程,2020,48(3):556-560.

[9] ZHOU Zhuo, QIN Jiaohua, XIANG Xuyu, et al. News text topic clustering optimized method based on TF-IDF algorithm on Spark[J]. Computers, Materials and Continua, 2020, 62(1): 217-231.

[10] 毛郁欣,邱智学. 基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J]. 中国信息技术教育,2020(8):99-101.

[11] 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):34-41.

[12] GAGLIARDI I, ARTESE M T. Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods[J]. Multimodal Technologies and Interaction, 2020, 4(2): 30-50.

[13] JATNIKA D, BIJAKSANA M A, SURYANI A A. Word2Vec model analysis for semantic similarities in english words[J]. Procedia Computer Science, 2019, 157: 160-167.

[14] 汪 静,罗 浪,王德强. 基于 Word2Vec 的中文短文本分类问题研究[J]. 计算机系统应用,2018,27(5):209-215.

[15] FENG Xiang, QIU Longhui, ZHOU Chun, et al. Automatic construction of a Chinese sentiment lexicon in the field of education based on Word2vec[J]. Data Science and Industrial Internet, 2019, 2(1): 129-136.

[16] ALAMI N, MEKNASSI M, EN-NAHNAHI N. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning[J]. Expert Systems with Applications, 2019, 123: 195-211.

[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013, 11(2): 1103-1115.

[18] 张 弛,张贯虹. 基于词向量和多特征语义距离的文本聚类算法[J]. 重庆科技学院学报:自然科学版,2019,21(3):69-72.