情报学报　2020 年 6 月　第 39 卷　第 6 期
Journal of the China Society for Scientific and Technical Information, Jun. 2020, 39(6): 640-650

DOI: 10.3772/j.issn.1000-0135.2020.06.008

基于学术论文全文的研究方法句自动抽取研究

张颖怡，章成志

（南京理工大学经济管理学院信息管理系，南京　210094）

摘　要　研究方法是科技文献中的重要内容，是解决学科领域问题的方法、工具、手段或技术。研究方法的描述通常以句子为单位。将分散在科技文献中的研究方法句进行汇总，可以辅助科研工作者快速地搜寻合适的研究方法。根据方法使用主体，将研究方法句进一步分为论文使用方法句和论文引用方法句。论文使用方法句是指论文中使用的研究方法的描述句。论文引用方法句是指论文对前人使用过的研究方法的描述句。本文使用多种基于神经网络的句子分类模型从科技文献全文本中进行研究方法句抽取。在模型词向量表示层，论文使用BERT和word2vec两种词向量模型。在模型的特征选择层，本文选用三种不同的网络，分别为卷积神经网络、双向长短时记忆网络和注意力机制网络。另外，论文使用两种模型训练方式，分别为单层次结构和两层次结构。实验结果表明，基于BERT的单层次结构的双向长短时记忆网络模型取得了较优的性能。本文从《情报学报》已发表论文中进行研究方法句的抽取并分析研究方法句的分布情况。分析发现，《情报学报》逐渐重视情报学中理论的发展并关注建设情报学学科的理论体系。

关键词　研究方法句抽取；信息抽取；深度学习；BERT

Methodological and Automatic Sentence Extraction from Academic Article's Full-text

Zhang Yingyi and Zhang Chengzhi

(*Department of Information Management, School of Economics and Management, Nanjing University of Science & Technology, Nanjing*　210094)

Abstract: Research methods are essential in the scientific literature. These include methods, tools, or techniques for solving problems in the field. The research method's description is usually presented through sentences. Summarizing these scattered sentences in the scientific literature can help researchers to quickly explore appropriate research methods. According to the method's purpose in the research paper, the research method sentence is further divided into method used and method cited sentences. The method used sentence refers to the sentence that describes the research method used in the paper and the method cited sentence refers to that cited by the paper. In this study, a variety of neural network-based sentence classification models are used for extracting the method sentences from the scientific literature's full-text. At the word vector representation layer, the study uses two-word vector models: BERT and word2vec. In the feature selection layer, three different networks are utilized: convolutional neural network (CNN), bidirectional LSTM (BiLSTM), and attention mechanism network. In addition, the study uses two model training methods: a single-level structure and a two-level structure. The experimental results show that the BERT-based BiLSTM model with single-level structure achieves the best performance. This paper analyzes the distribution of research method sentences extracted from the *Journal of The China Society for Scientific and Technical Information*. The analysis indicates that this journal paid more attention to the theoretical developments of information science; in addition, the journal also focused on constructing theoretical systems for this discipline.

Key words: methodological sentence extraction; information extraction; deep learning; BERT

收稿日期：2019-10-13；**修回日期**：2020-02-19

基金项目：国家社会科学基金重大项目“情报学学科建设与情报工作未来发展路径研究”（17ZDA291）。

作者简介：张颖怡，女，1992年生，博士研究生，研究方向为自然语言处理与文本挖掘；章成志，男，1977年生，博士，教授，博士生导师，研究方向为信息组织、信息检索、文本挖掘及自然语言处理，E-mail：zhangcz@njust.edu.cn。

1 引　言

科研工作者的研究成果通常以学术论文、专利、报告或专著等文献形式作为交流与传播的载体。有研究指出，全世界产生的科技文献数量已达到百万级别，且每年以3%左右的速度持续增加[1]。海量的科技文献加剧了文献搜索的难度，增加了文献阅读的时间成本。研究方法是科技文献中的重要内容，是解决学科领域问题的方法、工具、手段或技术[2-4]，是作者提出的问题的解决方案[5]。研究方法句是包含研究方法的句子。将分散在科技文献中的研究方法句进行汇总，可以辅助科研工作者快速地搜寻合适的研究方法。因此，本文的目标是从学术论文中抽取研究方法句。

现有的研究方法句抽取工作主要在学术论文的摘要中进行。如Hirohata等[6]将学术文本的摘要分为四个部分，分别为目的、方法、结果和总结。其使用BI标注策略（如在标注目的句时，B为目的句的开始词，I为目的句的中间或结尾词），利用条件随机场进行分类。摘要是对学术文献全文的高度概括，关于研究方法更为详细的描述通常体现在全文内容中，特别是研究内容或研究方法论等章节。因此，一些学者考虑从学术文献全文本中进行研究方法句的抽取。如Liakata等[7]将学术文本分为假设、目标、背景、方法、结果和总结等多个部分，使用支持向量机和条件随机场进行内容分类。上述研究存在三方面的问题。首先，缺少对研究方法句的定义和分类。本文根据研究方法主体，将研究方法句分为论文使用方法句和论文引用方法句。论文使用方法句是对论文中使用的研究方法的描述。将分散在科技文献中的论文使用方法句进行汇总并形成摘要，可以减少学者的文献阅读时间，辅助科研工作者快速地搜寻合适的研究方法。论文引用方法句是对被引文献中的研究方法的归纳和总结。对论文引用方法句进行分析，可以发现学科领域中研究方法的演变和发展模式。其次，使用的分类模型性能不足，如Liakata等[7]工作中研究方法句抽取最优F_1值为30%。为提升研究方法句抽取性能，本文比较多种深度学习分类模型并采用性能最佳的模型进行研究方法句的抽取。最后，缺少对研究方法句抽取结果的分析。本文从《情报学报》已发表论文中进行研究方法句的抽取，并分析研究方法句在不同年份上的数量分布。分析发现，研究方法句的数量分布在一定程度上体现了情报学学科的发展趋势。

本文的创新点有三方面：①本文首次尝试将研究方法句分为论文使用方法句和论文引用方法句；②本文比较了多种深度学习模型在两类研究方法句抽取任务上的性能；③本文以《情报学报》已发表论文为例，考察研究方法的分布情况，以此梳理学科领域的发展趋势。

本文将在第2节中对相关文献进行述评。在第3节中介绍本文的研究思路、数据采集、文本预处理和研究方法句自动抽取模型的选择。在第4节中，本文使用4种评价指标比较了研究方法句自动抽取模型的性能，并使用性能最优的模型从《情报学报》近10年发表的学术论文中抽取研究方法句。根据研究方法句抽取的结果，本文对《情报学报》近10年研究方法句的分布情况进行了分析。本研究的结论将在第5节中给出。

2 相关工作概述

目前关于研究方法句自动抽取的研究较少。研究方法句抽取是一种句子分类任务，因此，本文对现有的句子分类方法进行梳理。句子分类任务是指将文本中的句子按要求进行类别划分。现有的句子分类方法分为以下三类：基于规则的方法、基于传统机器学习的方法和基于深度学习模型的方法。

基于规则的句子分类方法首先选择特征并确定特征所属的类别，接着利用特征匹配文中的句子，根据特征的类别将对应的句子划分到不同的类型。例如，Hayes等[8]通过专家制定规则进行句子分类。基于规则的方法能得到较高的准确率，但该方法召回率较低，且需要事先制定匹配规则，时间成本和人力成本均较高。

基于传统机器学习的方法包括序列标注方法和分类方法。序列标注方法和分类方法是以特征选择为前提的。常用的特征包括句子位置信息[9]、句子中是否包含关键词[10-11]、句子中是否包含引用[6]、句子长度[7]等。序列标注方法使用条件随机场和最大熵等序列标注模型。分类方法使用统计分类模型。Hirohata等[6]将学术文本的摘要分为四个部分：目的、方法、结果和总结，并利用条件随机场进行序列标注。Liakata等[7]将学术文本分为假设、动机、目标、主题、背景、方法、实验等多个部分，使用支持向量机和条件随机场模型，并比较了两种模型的结果。Nomponkrang等[10]使用关键词和词频等特征，使用决策树、朴素贝叶斯、K近邻和支持向量机模型对句子进行分类，最后发现支持向量机得到

最优的结果。基于统计模型的方法依赖于特征工程。特征选择需耗费大量的时间成本和人力成本，且特征选择的优劣会影响模型结果的好坏。

为解决特征选择的问题，深度学习方法在近年来被一部分学者使用并在句子分类任务中取得了优秀的结果。现有研究主要集中在卷积神经网络和循环神经网络两种模型的使用上。在模型使用中，一般具有两方面的区别：①不同的词向量表示层；②不同的特征选择层。在输入端的选择上，主要是改变输入的词向量。例如，Hsu 等[12]使用 2~5 个词的词组作为输入，而非将单个词输入；Limsopatham 等[13]使用两种不同的词向量，一种词向量从一般的文本中训练得到，一种词向量从特定领域的文本中训练得到，并将两种词向量进行合并。在特征选择模块的选择上，部分研究使用了循环神经网络中的 LSTM（long short-term memory）、GRU（gated recurrent unit）和 BiLSTM（bidirectional LSTM）模型进行分类。例如，Chung 等[14]使用 LSTM 和 GRU 进行模型训练，最终发现 LSTM 在句子分类中能够得到较好的效果；Limsopatham 等[13]使用 BiLSTM 模型进行特征选择。本文选择使用基于深度学习模型的句子分类方法，并尝试使用多种词向量表示层和特征选择模块。

3 研究方法

3.1 研究思路

本文设计研究方法识别模型框架如图 1 所示。该模型的第一阶段是模型选择，论文使用人工标注数据集对模型进行训练并测试，从而选择一种研究方法句抽取最优模型。第二阶段是句子分类，研究方法句抽取最优模型将未标注论文集中的句子分为三个类别，分别是论文使用方法句、论文引用方法句和非研究方法句。第三阶段是结果分析，本文分析论文使用方法句和论文引用方法句的数量在不同年份上的分布。

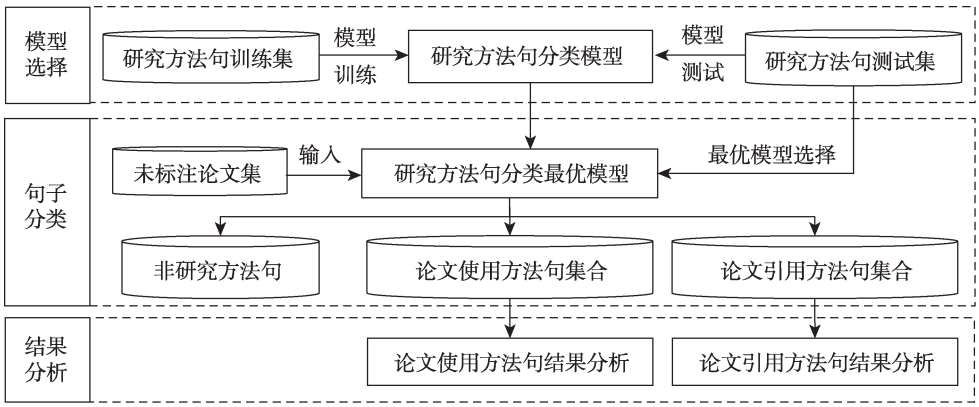


图 1 研究方法句抽取模型框架

3.2 数据集概述

3.2.1 研究方法句人工标注集构建

本文采集《情报学报》2009—2018 年发表的 1170 篇中文论文的全文内容，包括题名、摘要、作者信息、关键词与全文内容。然后从不同年份中随机选择相同数量的论文，总计 198 篇。本文首先将题名、作者信息与关键词信息剔除，保留摘要与全文内容。在全文内容中，删除表格标题与图标题。最后使用“。”、“；”和“？”等句间分隔符将论文分割成以句子为单位的形式。本文共招募 4 名情报学学科在读博士研究生和 4 名情报学学科在读硕士研究生。分为 4 组（每组 2 位标注人员）进行标注，其中 2 组分别标注 49 篇文献，另外 2 组分别标注 50 篇文献，共计文献 198 篇。标注分为两个阶段，第一阶段每组中每个标注人员独立进行标注；在第二阶段，筛选每组中两位标注人员标注不同的部分，然后两位标注人员进行讨论以确保标注统一性。

论文使用方法句判别标准分为两类：①句子中包含论文使用的研究方法（名词短语或简短描述）的句子。如，“本文以直接引用文献间的信息传递关系为基础，利用信息熵构建了反映领域知识演变转折的指标”，该句表明论文中使用了“信息熵”，因此属于论文使用方法句。②句子中包含论文使用的研究方法（名词短语或简短描述）和研究方法所需解决的问题的句子。例如，“利用社会网络的结构分析思想，对标签网络结构进行量化分析”，其中，“社会网络”是论文所使用的研究方法，“量化

分析”是该研究方法所需解决的问题。因此，该句子应标注为“论文使用方法句”。

论文引用方法句判别标准分为两类：①句子中包含论文引用的其他工作中所使用的研究方法（名词短语或简短描述）的句子；②句子中包含论文引用的其他工作中所使用的研究方法（名词短语或简短描述）和研究方法所需解决的问题的句子。例如，“基于标签的个性化推荐系统一般利用标签来表示用户兴趣模型”，其中，“标签”是其他工作所使用的研究方法，是一种工具；“表示用户兴趣模型”是该研究方法所需解决的问题。因此，该句子应标注为“论文引用方法句”。

3.2.2　数据集描述

本文将第 3.2.1 节构建的研究方法句人工标注数据集随机分为训练集、验证集和测试集，比例为 8∶1∶1。训练集、验证集和测试集中的非研究方法句、论文使用方法句和论文引用方法句的数量如表 1 所示。

表 1　研究方法句数据集统计信息

数据集	句子总数	非研究方法句数量	论文使用方法句数量	论文引用方法句数量
训练集	20936	18193	2110	633
验证集	2612	2275	260	77
测试集	2608	2306	230	72

3.2.3　文本预处理

文本预处理分为两个部分：①中文分词：使用 NLPIR[①]中文分词工具进行中文分词。为保证分词质量，论文使用 13720 篇中文图书情报学科论文[②]的关键词构建了用户分词词典。②非常用词替换：本文将文本中的引用标签替换为 CITE，URL 路径替换为 URLCOM。

3.3　关键技术描述

本文使用神经网络序列标注模型进行实验，并选择性能最优的模型进行研究方法句抽取。在神经网络序列标注模型中，本文选择使用两种不同的词向量表示层和三种不同的特征选择层。词向量表示层主要将输入的词语转化为模型能够理解的向量表示。如图 2 所示，假设需要词序列为 $\{p_{i,s,1},p_{i,s,2},\cdots,p_{i,s,w}\}$。该词序列输入到词向量层后转变为词向量表示 $\{v_{i,s,1},v_{i,s,2},\cdots,v_{i,s,w},\cdots,v_{i,s,w}\}$。经过特征选择层后经过 softmax 以得到最终的句子标签 $y_{i,s}$。在词向量表示层，本文使用两种预训练词向量模型，分别是 word2vec 领域词向量[15]和 BERT（bidirectional encoder representations from transformers）开源语言模型[16]。特征选择层主要从输入中自动学习有用的特征。如图 2 所示，特征选择层的输入为词序列的词向量表示 $\{v_{i,s,1},v_{i,s,2},\cdots,v_{i,s,w},\cdots,v_{i,s,w}\}$。特征选择层从词向量中学习到对分类任务有用的特征。在该层中，本文使用卷积神经网络、双向长短时记忆网络和结合注意力机制的双向长短时记忆网络。另外，在模型结构方面，本文使用两种结构的模型，分别是单层次模型和两层次模型。下面分别对不同的词向量表示层、特征选择层和模型结构进行介绍。

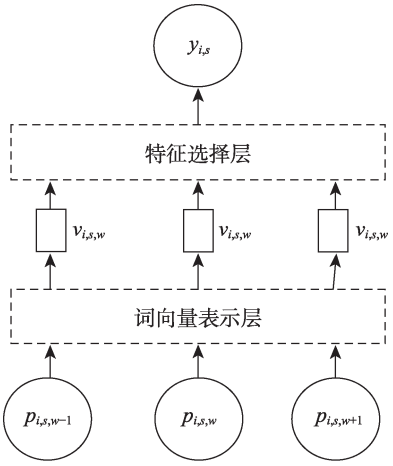


图 2　神经网络序列标注模型框架

3.3.1　词向量表示层选择

1）BERT 词向量训练模型

BERT 模型是一个语言编码器，把输入的句子和词语转化为特征向量。其输入是一个句子或一对句子，句子中的每个词有对应的词向量，该词向量包括词向量、位置向量和词标签向量。BERT 模型的训练算法采用双向 Transformer 模型[17]。论文使用 Google 开源的 BERT 中文预训练字向量语言模型。由于该语言模型以字为单位，因此当使用 BERT 来进行词向量表示时，在文本预处理阶段，本文将句子直接以中文字符为单位进行分割。

① https://github.com/NLPIR-team/NLPIR。

② 本文选择 CNKI 中的图书情报与档案领域的 CSSCI 来源期刊论文作为实验数据，采集有 HTML 全文的期刊论文全文。由于文献中掺杂着一些类似于“会议介绍”、“期刊投稿要求”等与研究无关的短文，本文对这些短文一一排除。最初共计有 14929 篇学术论文，通过筛选整理后剩余 13720 篇学术论文。

2）word2vec 词向量训练模型

word2vec 是 Google 推出的一个 NLP（natural language processing）工具，用于将所有的词向量化，挖掘词语间的联系。论文使用 word2vec[①] 工具训练词向量。使用 13720 篇中文图书情报学科论文[②] 作为词向量训练集。中文文本在词向量训练前需分词处理。本文应用 NLPIR 分词系统，将论文中作者标注的关键词作为用户词典加入分词系统中，共计 49108 个用户词。

3.3.2 特征选择模块选择

论文使用三种特征选择模块，分别是卷积神经网络、双向长短时记忆网络和基于注意力机制的双向长短时记忆网络。下面将分别介绍这三种模型。首先假设目标文献中的每个句子中的词表示为 $p_{i,s,w}$，词向量层将词 $p_{i,s,w}$ 转化为对应的词向量 $\boldsymbol{v}_{i,s,w}$。

1）卷积神经网络模型（convolutional neural networks，CNN）

卷积神经网络主要包括卷积层（convolution layer）和池化层（pooling layer）。卷积层的目的是自动选取输入的词向量中的特征；池化层的输入是卷积层的输出，其目的是对卷积层选取的特征进行压缩。其中，卷积层用于从数据中捕获对当前任务最重要的局部特征。论文使用一维卷积神经网络。对于一个句子 $p_{i,s}$，其中有 z 个词 $\{p_{i,s,1},p_{i,s,2},\cdots,p_{i,s,w},\cdots,p_{i,s,z}\}$。每个词 $p_{i,s,w}$ 有对应的词向量 $v_{i,s,w}$。使用大小为 k 的滑动窗口，对序列中的每个窗口使用同一个“滤波器”。其中，滤波器是该窗口向量与权重向量 u 的内积，其后使用一个非线性激活函数 σ。如式(1)和式(2)所示，该窗口中的内容使用 $x_{i,s,w}$ 表示，$h_{i,s,w}$ 为经过滤波器后的向量表示：

$$x_{i,s,w} = \text{Concatenate}(v_{i,s,w},v_{i,s,w+k-1}) \tag{1}$$

$$h_{i,s,w} = \sigma(x_{i,s,w}\cdot u) \tag{2}$$

式中，$h_{i,s,w} \in \mathbb{R}$，$x_{i,s,w} \in \mathbb{R}^{kd_{\text{emb}}}$，$u \in \mathbb{R}^{kd_{\text{emb}}}$。

2）双向长短时记忆网络（BiLSTM）

长短时记忆层是一种循环神经网络，其考虑序列顺序信息，能够将任意长度的序列表示为定长的向量。一般的循环神经网络在训练过程的反向传播阶段，会出现梯度爆炸或梯度消失的现象。长短时记忆网络为解决这一问题，引入了门机制和记忆单元。门机制决定有多少新的输入加入记忆单元，以及记忆单元中现有的多少记忆应该被忘记。记忆单元用来保存记忆和梯度信息。长短时记忆网络有三种门结构：i、f、o，分别控制输入、遗忘和输出。门的值由当前输入 $p_{i,s,w}$ 和前一个状态 $h_{i,s,w-1}$ 的线性组合通过一个 sigmoid 激活函数得到，

$$i = \sigma(p_{i,s,w}W^{pi} + h_{i,s,w-1}W^{hi}) \tag{3}$$

$$f = \sigma(p_{i,s,w}W^{pf} + h_{i,s,w-1}W^{hf}) \tag{4}$$

$$o = \sigma(p_{i,s,w}W^{po} + h_{i,s,w-1}W^{ho}) \tag{5}$$

式中，遗忘门 f 控制有多少先前的记忆被保留，输入门 i 控制有多少更新被保留。一个更新候选项 z 由当前输入 $p_{i,s,w}$ 和 $h_{i,s,w-1}$ 的线性组合通过一个 tanh 激活函数来得到。然后记忆 $c_{i,s,w}$ 被更新，

$$c_{i,s,w} = f\odot c_{i,s,w-1} + i\odot z \tag{6}$$

最后，$h_{i,s,w}$ 由记忆 $c_{i,s,w}$ 的内容通过一个 tanh 非线性激活函数并受输出门的控制来决定，

$$h_{i,s,w} = o\odot\tanh(c_{i,s,w}) \tag{7}$$

论文使用双向长短时记忆网络，双向是指从两个方向进行序列处理，分别是从前往后和从后往前。最后将两个方向的序列进行合并操作。

3）基于注意力机制的双向长短时记忆网络模型（attention-based BiLSTM，Att-BiLSTM）

注意力层（attention）决定输入序列中哪些部分应该得到更多的关注。在神经网络中，输入句子被编码为单一的向量，该结构强制所得到的向量包含生成时所需要的全部信息。然而，句子中的词语的重要性程度都是不同的。其中，注意力机制是一种能够决定网络中应该关注输入序列中哪些部分的模块。其计算公式为

$$\boldsymbol{v}_{i,s} = \sum_{w=1}^{|p_{i,s}|}\boldsymbol{\alpha}_{i,s,w}\boldsymbol{h}_{i,s,w} \tag{8}$$

式中，$v_{i,s}$ 为句子 $p_{i,s}$ 的向量表示，其通过句子中所有词 $p_{i,s,w}$ 的注意力权重系数 $\alpha_{i,s,w}$ 和当前状态 $h_{i,s,w}$ 相乘后连加得到。注意力权重系数 $\alpha_{i,s,w}$ 计算公式为

$$\boldsymbol{\alpha}_{i,s,w} = \text{softmax}(a(\boldsymbol{h}_{i,s,w})) \tag{9}$$

式中，a 表示当前状态 $h_{i,s,w}$ 经过一个非线性激活函数 tanh，表示为

$$a(\boldsymbol{h}_{i,s,w}) = \tanh(\boldsymbol{W}_a\boldsymbol{h}_{i,s,w}) \tag{10}$$

3.3.3 模型训练方式选择

论文使用两种训练方式，分别是单层次结构和

① https://code.google.com/archive/p/word2vec/，word2vec 是 Google 推出的一个 NLP 工具，用于将所有的词向量化，挖掘词语间的联系。

② 本文选择 CNKI 中的图书情报与档案领域的 CSSCI 来源期刊论文作为实验数据，采集有 HTML 全文的期刊论文全文。由于文献中掺杂着一些类似于“会议介绍”、“期刊投稿要求”等与研究无关的短文，本文对这些短文一一排除。原共计有 14929 篇学术论文，通过筛选整理后剩余 13720 篇学术论文。

两层次结构。两层次结构中的第一阶段将句子分为研究方法句和非研究方法句；在第二阶段，将识别得到的研究方法句分为论文使用方法句和论文引用方法句。两个层次使用相同的分类模型。单层次结构直接将句子分类为非研究方法句、论文使用方法句和论文引用方法句。

4　实验与结果分析

本文使用基于深度学习的神经网络模型进行实验并比较各模型性能。本节首先介绍参数设置情况，然后介绍评测标准并根据评测标准对各模型的性能进行分析，选择性能较优的模型来抽取研究方法句。

4.1　实验设置

4.1.1　参数设置

CNN 中的滑动窗口 K 分别设置为 3、4、5，滤波器大小设置为 128。长短时记忆网络的神经元设置为 150 个。Epoch 设置为 5，训练 Batch 设置为 32，学习率设置为 5e-5。在神经网络模型训练中，本文采用交叉熵损失函数和 RMSprop[18]优化函数。

4.1.2　基准模型

随机数方法（random）：该方法使用随机数生成器随机生成 0、1、2 三种数字。0 代表非研究方法句，1 代表论文使用方法句，2 代表论文引用方法句。为测试集赋予随机数。

朴素贝叶斯[19]（naive Bayesian，NB）：利用 NB 对句子进行分类。首先将训练语料进行分词和预处理（删除单字词），得到最终的词表，词表长度为 14116。使用 TF-IDF[20]进行特征选择并构建向量空间模型。将向量空间模型用于训练分类模型。最后，使用训练得到的模型进行测试并得到预测结果。

K 近邻[21]（k-nearest neighbor，KNN）：数据处理和特征选择同朴素贝叶斯模型。将训练语料转换为向量空间模型后，训练 KNN 分类模型。使用训练得到的模型进行测试并得到预测结果。

支持向量机[22]（support vector machine，SVM）：数据处理和特征选择同朴素贝叶斯模型。将训练语料转换为向量空间模型后，训练 SVM 分类模型。使用训练得到的模型进行测试并得到预测结果。

4.2　评测标准

本文的评测标准选择使用正确率、准确率 P 值、召回率 R 值、F_1 值、F_1 宏平均和 F_1 微平均。

正确率（accuracy，ACC）：在所有样本中，被正确分类到对应类别的样本数为 TA，所有样本数为 T，则

$$\text{accuracy} = \frac{TR}{\text{TA}} \tag{11}$$

$P/R/F_1$ 值：若存在类 C，本文定义属于类 C 的样本被正确分类到类 C，记这一类样本数为 TP；不属于类 C 的样本被错误分类到类 C，记这一类样本数为 FN；属于类别 C 的样本被错误分类到其他类，记这一类样本数为 TN。不属于类别 C 的样本被正确分类到了除类别 C 外的其他类，记这一类样本数为 FP，则

$$P = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{TN}} \tag{13}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{14}$$

F_1 宏平均（macro-averaging）：是先对每一个类统计指标值，然后再对所有类求算术平均值，即

$$\text{Macro\_}P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{15}$$

$$\text{Macro\_}R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{16}$$

$$\text{Macro\_}F_1 = \frac{2 \times \text{Macro\_}P \times \text{Macro\_}R}{\text{Macro\_}P + \text{Macro\_}R} \tag{17}$$

式中，P_i 和 R_i 分别表示第 i 类的 P 值和 R 值。

F_1 微平均（micro-averaging）：是对数据集中的每一个实例不分类别进行统计建立全局混淆矩阵，然后计算相应指标，即

$$\text{Micro\_}P = \frac{\sum_{i=1}^{n}\text{TP}_i}{\sum_{i=1}^{n}\text{TP}_i + \sum_{i=1}^{n}\text{FP}_i} \tag{18}$$

$$\text{Micro\_}R = \frac{\sum_{i=1}^{n}\text{TP}_i}{\sum_{i=1}^{n}\text{TP}_i + \sum_{i=1}^{n}\text{FN}_i} \tag{19}$$

$$\text{Micro\_}F_1 = \frac{2 \times \text{Micro\_}P \times \text{Micro\_}R}{\text{Micro\_}P + \text{Micro\_}R} \tag{20}$$

式中，TP_i、FP_i 和 FN_i 值分别代表第 i 类的 TP、FP 和 FN 值。

4.3　模型抽取结果的比较分析

研究方法句分类结果如表 2 所示。从表 2 可以得到如下结论。

单层次的双向长短时记忆模型（BiLSTM (Sin-

表 2 研究方法句分类模型性能

模型		ACC	F_1@0	F_1@1	F_1@2	Macro_F_1	Micro_F_1
基准模型	Random	35.20	50.11	16.64	5.61	24.12	35.20
	NB	86.60	92.86	25.77	7.27	44.68	86.60
	KNN	86.60	92.77	16.62	0.00	36.46	86.60
	SVM	87.17	93.12	6.52	0.00	35.21	87.17
word2vec	CNN（Binary）	91.73	95.70	61.86	40.71	66.09	91.73
	BiLSTM（Binary）	91.50	95.40	67.04	60.53	74.32	91.50
	Att-BiLSTM（Binary）	92.80	96.15	68.61	60.81	75.19	92.80
	CNN（Single）	91.62	95.44	61.54	20.78	63.49	91.62
	BiLSTM（Single）	92.11	95.66	67.47	59.74	75.78	92.11
	Att-BiLSTM（Single）	91.84	95.49	68.44	64.94	76.11	91.84
BERT	CNN（Binary）	93.30	96.27	72.90	72.85	80.67	93.30
	BiLSTM（Binary）	93.03	96.12	72.02	69.44	79.19	93.03
	Att-BiLSTM（Binary）	93.22	96.21	72.90	70.83	79.98	93.22
	CNN（Single）	93.19	96.12	74.06	73.08	81.09	93.19
	BiLSTM（Single）	**93.42**	**96.29**	73.26	**75.17**	**81.57**	**93.42**
	Att-BiLSTM（Single）	93.38	96.22	**74.95**	71.52	80.90	93.38

注：Single 代表单层次模型，Binary 代表两层次模型。ACC 代表句子分类正确率，F_1@0 代表类别为非研究方法句的 F_1 值，F_1@1 代表类别为论文使用研究方法句的 F_1 值，F_1@2 代表类别为论文引用研究方法句的 F_1 值，Macro_F_1 代表 F_1 宏平均，Micro_F_1 代表 F_1 微平均。

gle)）的句子分类的总体性能较高。如表 2 所示，BiLSTM (Single) 在准确率（ACC）、F_1 宏平均（Macro_F_1）和 F_1 微平均（Micro_F_1）等评测指标上的性能优于其他模型，分别为 93.42%、81.57% 和 93.42%。这 3 个评测指标用于评估模型在多个类别上整体的性能。因此，该模型在句子分类任务中的整体性能较优。

单层次的双向长短时记忆模型（BiLSTM (Single)）在识别非方法句和论文引用方法句上得到较高的性能。如表 2 所示，BiLSTM (Single)在非研究方法句识别上的 F_1 值（F_1@0），以及论文引用方法句识别上的 F_1 值（F_1@2）等评测指标上的性能优于其他模型，分别为 96.29% 和 75.17%。其中，F_1@2 值大幅度优于其他模型，例如，比 Att-BiLSTM (Single)和 CNN (Single)模型高出约 4% 和 2%。

单层次的基于注意力机制的模型（Att-BiLSTM (Single)）模型在识别论文使用方法句上性能较高。如表 2 所示，Att-BiLSTM (Single)模型在论文使用方法句识别上的 F_1 值（F_1@1）评测指标上的性能优于其他模型，为 74.95%。虽然其性能较优，但提升幅度较小，例如，比 BiLSTM (Single) 和 CNN (Single)模型高约 1%。

基于 BERT 预训练词向量的模型在各评测指标上得到较高的性能。如表 2 所示，使用 BERT 预训练词向量的分类模型在论文使用方法句抽取和论文引用方法句抽取任务上的性能均优于使用 word2vec 预训练词向量的模型的性能。

研究方法句的识别是研究方法实体识别任务的基础。较高的句子分类准确率能够保证研究方法实体识别的性能。另外，由于基于 BERT 的模型的性能较优，因此，本文进一步分析了基于 BERT 的各模型的准确率（P）值。如表 3 所示。

表 3 基于 BERT 的神经网络模型的准确率值

模型	性能指标		
	P@0	P@1	P@2
CNN（Binary）	96.10	73.62	74.32
BiLSTM（Binary）	95.77	73.02	74.63
Att-BiLSTM（Binary）	95.90	73.62	76.12
CNN（Single）	**96.42**	72.43	72.15
BiLSTM（Single）	96.10	73.83	**77.78**
Att-BiLSTM（Single）	96.22	**74.52**	72.97

注：P@0 代表类别为非研究方法句的准确率值，P@1 代表类别为论文使用方法句的准确率值，P@2 代表类别为论文引用方法句的准确率值。

在表 3 中，在非研究方法句识别中，单层次的卷积神经网络模型（CNN (Single)）能得到较高的准确率（P@0），为 96.42%。在论文使用方法句分类中，单层次的基于注意力机制的双向长短时记忆

网络模型（Att-BiLSTM (Single)）具有较高的准确率值（*P*@1），为 74.52%。在论文引用方法句分类中，单层次的双向长短时记忆网络模型（BiLSTM (Single)）得到较高的准确率值（*P*@2），为 77.78%。

虽然 CNN (Single)和 Att-BiLSTM (Single)在非研究方法句分类和论文使用方法句分类上的准确率值较高，但与其他模型相比并没有大幅度的提升。例如，在 *P*@1 中，Att-BiLSTM (Single)）的值仅比 BiLSTM (Single)高不足 1%。虽然 BiLSTM (Single)模型仅在论文引用方法句分类中的准确率较高，但与其他模型相比，提升幅度较大。例如，其 *P*@2 值超过 CNN (Single)和 Att-BiLSTM (Single)模型约 5%。另外，BiLSTM (Single)模型在非研究方法句识别和论文使用方法句识别中的性能也较优。因此，本文选择使用基于 BERT 的单层次结构的双向长短时记忆网络模型（BiLSTM (Single)）为最优模型并进行研究方法句分类。

4.4 《情报学报》近 10 年研究方法抽取结果分析

本文以《情报学报》近 10 年的学术论文为抽取对象，依据第 4.3 节中模型的性能分析结果，选择性能最优的基于 BERT 的单层次结构的双向长短时记忆网络模型，从中识别研究方法句。本节首先介绍用于研究方法句分类的语料，然后分析研究方法句识别的结果。最后，本文分析不同年份下论文使用方法句和论文引用方法句的数量分布情况。

4.4.1 数据集介绍

采集《情报学报》2009—2018 年共 10 年的学术论文，总计 1170 篇。本文将研究方法句人工标注数据集 198 篇标注文本作为训练集进行模型训练。该研究方法句人工标注数据集的统计情况如表 1 所示。接着，在剩余的 972 篇学术论文中，本文使用该模型进行研究方法句抽取。经统计，972 篇学术论文中共包含 163596 个句子。

4.4.2 研究方法句分类结果实例分析

本文使用基于 BERT 的单层次结构的双向长短时记忆网络模型进行模型训练。使用该模型，本文在 972 篇学术论文中共识别得到 15276 句论文使用方法句，5655 句论文引用方法句。本文以论文《基于 TAM/TTF 整合的网络信息资源利用效率模型与指标框架研究》为例，分析论文使用方法句和论文引用方法句的抽取情况，分别如表 4 和表 5 所示。

如表 4 所示，共抽取得到 8 句论文使用方法句。在论文使用方法句中，前 3 句论文使用方法句是关于 TAM/TTF 模型的描述，第 4 句和第 5 句是关于变量间相关关系的测度及任务-技术适配程度的测量方法，第 6 句表明该文中使用问卷调查进行实证研究，第 7 句表示采用李克特量表形式的问卷，第 8 句表示问卷的信度、效度检验方法以及问卷结果的分析方法。通过这 8 句论文使用方法句，可以对目标论文的内容有一个完整的认识。如表 5 所示，共抽取得到 5 句论文引用方法句。论文引用方法句总结了其他文献对 TAM 和 TTF 模型的使用情况。例如，第 5 句中的研究方法被作者所借鉴，第 3 句中描述了 TAM 模型优化的转折点。

根据以上分析，抽取得到的论文使用方法句是对论文的研究方法的概括。将论文使用方法句形成摘要，可以提升学者论文阅读的速度，帮助学者更快地搜寻合适的研究方法。另外，抽取得到的论文

表 4 《基于 TAM/TTF 整合的网络信息资源利用效率模型与指标框架研究》论文使用方法句实例

序号	论文使用方法句
1	在此基础上提出了一个基础性整合理论模型，用以描述影响网络信息资源利用效率各因素之间的总体逻辑结构，建立了研究假设，进行了测度指标体系的设计，初步探讨了模型后续实证研究的框架，为进一步深入研究创造了前提。
2	因此，我们拟以 TAM 与 TTF 两个较成熟的信息利用相关模型为理论基础，通过综合分析比较，尝试将其整合为一个能够更好地解释网络信息资源利用效率影响因素的基础理论模型，以便集成研究、深入探讨网络信息资源利用效率的“干扰变量”。
3	在此，我们引入一个基于 TAM/TTF 整合的用户网络信息资源利用效率研究基本理论模型，如图 3 所示。
4	为了测度并验证以上网络信息资源利用效率模型中各变量间存在的相关关系与显著性影响，我们参考国内外相关文献，设计了指标体系研究框架，其中，每个变量对应的测度指标分别反映用户对该变量感知的不同方面，见表 1。
5	本研究亦借鉴 Goodhue 于研究中发展出的相关指标来测量任务—技术适配程度。
6	实证研究方案可采用问卷调查方式，在研究模型与指标框架基础上进行指标量化与问卷设计，要注意指标设置中的准确性和可操作性，评价方法应易于被评价者和用户理解、掌握，并具有可移植性，即适用于新的资源评价和加入新的评价要求和因素。
7	可以李克特（Likert）量表形式，对各个测度指标进行 1~5 分或 7 分的评价。
8	问卷回收后进行甄别和变量指标测量，检验问卷的信度和效度，再采用各种统计方法和工具，如通过主成分分析法、结构方程模型、多元线性回归等方法来对模型所设计的假设进行验证，最终得出相关结论与建议。

表 5 《基于 TAM/TTF 整合的网络信息资源利用效率模型与指标框架研究》论文引用方法句实例

序号	论文引用方法句
1	Davis 为了有效解释与预测信息技术用户的使用行为，在理性行为理论的基础上，探讨了认知及情感因素与信息系统使用之间的关系，提出了易用认知（Perceived Ease of Use）和有用认知（Perceived Usefulness）两个概念，认为它们在解释用户对信息系统的采用和使用过程中是十分关键的影响因素，一些外部变量通过影响易用认知与有用认知而影响用户对使用信息系统的态度和行为意向，而行为意向又直接影响了用户对信息系统的实际使用，TAM 的模式架构如图1所示。
2	Goodhue 通过描述认知心理和认知行为来揭示信息技术如何作用于个人的任务绩效。
3	Davis 等通过实证研究发现，作为中介变量，“使用态度”并不能完全传递有用认知与易用认知对使用意向的影响，鉴于此，为了进一步简化原 TAM 模型，随后的许多研究者都舍弃了使用态度这一概念。
4	Mathieson、Klopping 和 Wu 等也通过实证研究证实了 TTF 对易用认知和有用认知的直接影响。
5	Goodhue、Dishaw 和 Wu 等还通过各自的实证研究证实了 TTF 对公司员工实际使用某些信息处理系统或工具的直接影响，这为我们认识 TTF 与用户网络信息系统实际使用的关系提供了借鉴。

引用方法句是对先前工作中的研究方法的总结。通过论文引用方法句，可以快速了解该领域的研究方法的研究现状和发展路径。

4.4.3 《情报学报》近 10 年研究方法句分布分析

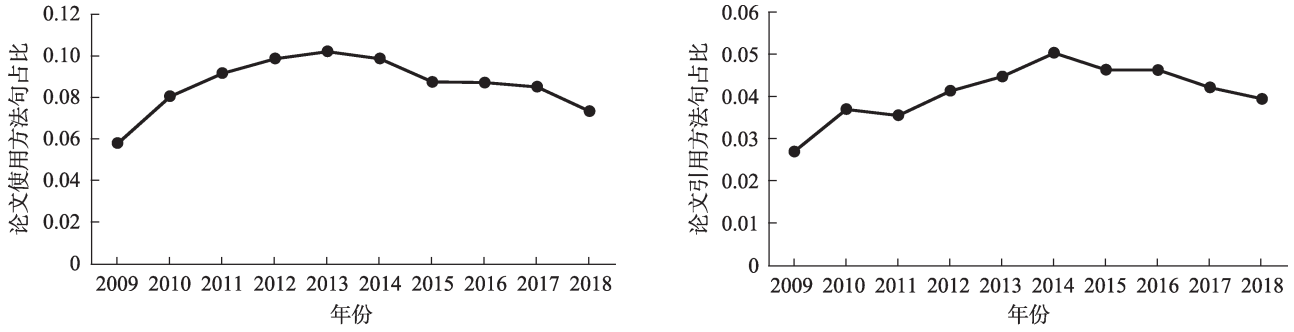
本节主要分析不同年份下研究方法句的分布情况。分析研究方法句在不同年份下的分布，需要统计不同年份下每篇论文的论文使用方法句数量、论文引用方法句数量和本文句子总数量。分析的基础建立在 1170 篇学术论文的基础上，其中包括研究方法句人工标注数据集的 2600 句论文使用方法句和 782 句论文引用方法句，以及通过自动抽取得到的 15276 句论文使用方法句和 5655 句论文引用方法句。

1）论文使用方法句年份分布分析

论文使用方法句占比是指论文使用方法句数量与论文句子总数量的比例。本文取每一年份下所有论文的论文使用方法句占比的平均数。具体分布如图 3a 所示。从图 3a 中可以发现，论文使用方法句的分布分为 4 个阶段。第一阶段，2009—2013 年，呈现整体上升趋势。在这个阶段，论文中论文使用方法句数量不断增加。由此表明，这一阶段，《情报学报》中的研究型和实验型论文的数量不断增加。第二阶段，2013—2015 年，呈现急速下降趋势。这个阶段，论文中论文使用方法句数量开始下降。由此表明，在这一阶段，《情报学报》开始增加理论型论文的录用和发表。第三阶段，2015—2017 年，呈现平缓的趋势。在这个阶段，论文中论文使用方法句的比例较平缓。该现象表明，在这一阶段《情报学报》中研究型论文的比例趋向于固定。第四阶段，2017—2018 年，呈现急速下降趋势。在这个阶段，论文中论文使用方法句的占比呈现下降趋势。由此表明，《情报学报》逐渐重视情报学学科中理论的发展。情报学基础理论是推动情报学范式形成和学科体系健康发展的决定性因素，加强对情报学基础理论的研究对于完善情报学学科体系具有重要意义[23]。现今，学科理论体系的建设受到情报学学科学者们的关注。因此，以上趋势在一定程度上体现了情报学学科的发展趋势。

2）论文引用方法句年份分布分析

论文引用方法句是论文中对先前工作的引用。引用包括多种情况，如与已有方法进行比较或使用已有研究工作中的方法的情况都需要引用对应的文献。论文引用方法句占比是指论文引用方法句和本文句子总数量的比例。本文取每一年份下所有论文的论文引用方法句占比的平均数。论文引用方法句



a. 论文使用方法句年份分布情况　　b. 论文引用方法句年份分布情况

图 3 研究方法句年份分布情况

占比越高，则表示论文中较倾向于使用已有方法或与已有方法进行比较研究。具体分布如图3b所示。从图3b中可以发现，论文引用方法句的分布分为3个阶段。第一阶段，2009—2010年，呈现急速上升的趋势。在这个阶段，论文引用方法句比例上升。这一现象表明，在这一阶段《情报学报》中的论文较倾向于使用已有研究工作中的方法或是提出新方法与已有方法进行比较。第二阶段，2010—2014年，呈现波动上升的趋势。在这个阶段，论文引用方法句比例缓慢上升。这一现象表明，在这一阶段，《情报学报》中的论文中仍习惯于使用已有研究工作中的方法或与已有方法相比较。第三阶段，2014—2018年，呈现波动下降的趋势。在这个阶段，论文引用方法句比例波动下降。这一现象与图3a中这一阶段论文使用方法句比例下降的趋势相同。因此，可以看出在这一时期，《情报学报》中有关研究型和实验型的论文占比下降，理论研究型论文数量上升，从而使论文引用方法句的数量也随之下降。

5　结论与展望

本文使用深度学习模型从学术文献全文本中抽取论文使用方法句和论文引用方法句。在输入端，本文对比使用了BERT和word2vec预训练词向量模型；在特征选择层，论文使用卷积神经网络、双向长短时记忆网络和基于注意力机制的双向长短时记忆网络这三种网络并比较了各网络的性能；另外，本文引入了两种结构的神经网络模型，分别是单层次结构和两层次结构。实验结果表明，基于BERT的单层次结构的双向长短时记忆网络模型在论文使用方法句和论文引用方法句抽取任务中得到了较优的性能。使用该最优模型，本文从《情报学报》已发表的1170篇论文中抽取研究方法句并进行统计分析。分析结果表明，《情报学报》逐渐重视情报学中理论的发展，关注于建设情报学学科的理论体系。

学术论文研究方法句抽取任务面临的严峻问题之一是标注数据（训练数据）的不足。人工标注方式的时间成本和人力成本高。为更好地应对该问题，在后续工作中，我们将引入半监督学习的方式，以减少对标注数据的依赖。另外，本文还将收集除《情报学报》之外的情报学期刊论文以进一步扩大研究方法句抽取范围。

参　考　文　献

[1] Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references[J]. Journal of the Association for Information Science and Technology, 2015, 66(11): 2215-2222.

[2] Gupta S, Manning C. Analyzing the dynamics of research by extracting key aspects of scientific papers[C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, 2011: 1-9.

[3] Kovačević A, Konjović Z, Milosavljević B, et al. Mining methodologies from NLP publications: A case study in automatic terminology recognition[J]. Computer Speech & Language, 2012, 26 (2): 105-126.

[4] Singh M, Dan S, Agarwal S, et al. AppTechMiner: Mining applications and techniques from scientific articles[C]// Proceedings of the 6th International Workshop on Mining Scientific Publications. New York: ACM Press, 2017: 1-8.

[5] 蒋婷. 学科领域本体学习及学术资源语义标注研究[D]. 南京: 南京大学, 2017.

[6] Hirohata K, Okazaki N, Ananiadou S, et al. Identifying sections in scientific abstracts using conditional random fields[C]// Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 2008: 381-388.

[7] Liakata M, Saha S, Dobnik S, et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications[J]. Bioinformatics, 2012, 28(7): 991-1000.

[8] Hayes P J, Andersen P M, Nirenburg I B, et al. TCS: A shell for content-based text categorization[C]// Proceedings of the Sixth Conference on Artificial Intelligence for Applications. IEEE Press, 1990: 320-326.

[9] 华秀丽, 徐凡, 王中卿, 等. 细粒度科技论文摘要句子分类方法[J]. 计算机工程, 2012, 38(14): 138-140.

[10] Nomponkrang T, Sanrach C. The comparison of algorithms for Thai-sentence classification[J]. International Journal of Information and Education Technology, 2016, 6(10): 801-808.

[11] Yamamoto Y, Takagi T. A sentence classification system for multi biomedical literature summarization[C]// Proceedings of the 21st International Conference on Data Engineering Workshops. IEEE Press, 2005: 1163-1163.

[12] Hsu S T, Moon C, Jones P, et al. A hybrid CNN-RNN alignment model for phrase-aware sentence classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017, 2: 443-449.

[13] Limsopatham N, Collier N. Modelling the combination of generic and target domain embeddings in a convolutional neural network for sentence classification[C]// Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 136-140.

[14] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[OL]. https://

arxiv.org/pdf/1412.3555v1.pdf.

[15] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019, 1: 4171-4186.

[16] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2013, 2: 3111-3119.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2017: 6000-6010.

[18] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5-6): 602-610.

[19] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification[C]// Proceedings of the AAAI/ICML -98 Workshop on Learning for Text Categorization. Palo Alto: AAAI Press, 1998: 41-48.

[20] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24 (5): 513-523.

[21] Keller J M, Gray M R, Givens J A. A fuzzy K-nearest neighbor algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1985, 15(4): 580-585.

[22] Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features[C]// Proceedings of the European Conference on Machine Learning. Heidelberg: Springer, 1998: 137-142.

[23] 王芳, 陈锋, 祝娜, 等. 我国情报学理论的来源、应用及学科专属度研究[J]. 情报学报, 2016, 35(11): 1148-1164.

（责任编辑 魏瑞斌）