中文文本分类方法综述

于游，付钰，吴晓平

（海军工程大学信息安全系，湖北 武汉 430033）

摘　要：如何高效地文本分类是当前研究的一个热点。首先对文本分类概念及流程中的分词、特征提取和文本分类方法等相关技术及研究现状进行了介绍和阐述，然后分析了现有文本分类相关技术面临的挑战，最后对文本分类的发展趋势进行了总结。

关键词：文本分类；分词；特征选择

中图分类号：TP391

文献标识码：A

doi: 10.11959/j.issn.2096−109x.2019045

Summary of text classification methods

YU You, FU Yu, WU Xiaoping

Department of Information Security, Naval University of Engineering, Wuhan 430033, China

Abstract: How to effectively classify text has become a hot topic. Firstly, the concept of text classification, word segmentation, feature extraction and text classification methods were introduced, and the research actuality was summarized. And then the challenges of text classification related technologies were analyzed. Finally, the development trend of text classification was summarized.

Key words: text classification, word segmentation, feature selection

1　引言

随着大数据、云计算等现代信息技术的发展，传统的纸质文档快速向电子化、数字化转变。面对大量的数据和信息，人们越来越倾向于利用计算机对数据和信息进行处理，不但可以提高相关操作的效率，还可以在一定程度上提高相关操作的准确度。信息挖掘和检索、自然语言处理是目前数据管理的关键技术，而文本分类则是这些技术进行操作的重要基础，是目前研究的一个热点，也是一个难点。传统的文本分类主要依靠人工完成，费时费力，为提高文本分类的效率、降低成本，文本自动分类技术已成为当前研究的一个热点。

收稿日期：2019−05−25；**修回日期：**2019−08−09

通信作者：于游，874354471@qq.com

基金项目：国家自然科学基金资助项目（No.61672531）

Foundation Item: The National Natural Science Foundation of China (No.61672531)

论文引用格式：于游, 付钰, 吴晓平. 中文文本分类方法综述[J]. 网络与信息安全学报, 2019, 5(5): 1-8.

YU Y, FU Y, WU X P. Summary of text classification methods[J]. Chinese Journal of Network and Information Security, 2019, 5(5): 1-8.

2　文本分类的概念和过程

2.1　文本分类的概念

文本分类是指按照一定的分类体系或规则对文本实现自动划归类别的过程，在信息索引、数字图书管理、情报过滤等领域有广泛的应用[1]。文本分类一般包括文本预处理、分词、模型构建和分类几个过程。随着互联网技术的快速发展，文本和词汇呈现出多元化、更新快的特点，这给文本分类带来了巨大的挑战。为更加清晰地了解文本分类算法的发展，本文针对文本分类过程中的相关技术和分类方法进行详细的梳理和分析。

2.2　文本分类流程

文本分类的一般流程可分为 5 步，如图 1 所示。

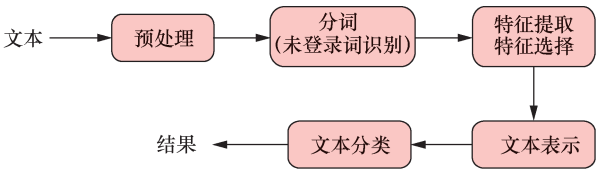


图 1　文本分类的一般流程

Step1　对文本进行预处理，去掉文本中多余的部分，如标点、介词等。

Step2　对文本进行分词操作，对预处理后的文本进行词切分操作，并识别其中的未登录词。

Step3　特征提取和特征选择，得到文本分词结果后，选择文本特征提取方法，并对特征进行选择，约简特征，尽量降低维度，减少后续计算量。

Step4　文本表示，选择合适的方法表示选择的特征，作为分类的依据。

Step5　文本分类，选择合理的分类方法对文本进行分类，得到文本类别。

其中，分词方法、特征选择以及分类算法的选择是关键。结合当前文本分类研究现状，本文主要对分词方法、特征提取与特征选择、文本分类方法进行综述。

3　分词

分词是中文文本处理的第一步，指通过一定的规则和方法将文本中的语句分割成词。相比于英文，中文词与词之间没有严格的分界符，增加了中文分词的难度。

3.1　分词的一般方法

目前，中文分词方法主要分为：基于字符串匹配的、基于理解的和基于统计的分词方法。

（1）基于字符串匹配的分词方法

基于字符串匹配[2]的分词方法是指在已有字典的基础上，按照指定的规则进行匹配，直到完成规则中的“最大”匹配，则识别出一个词。按照匹配的方向不同，基于字符串匹配方式的不同又可以分为：正向最大匹配、逆向最大匹配、双向最大匹配。

（2）基于理解的分词方法

基于理解的分词方法是指利用计算机模拟人对文本的理解，结合语义、句法等因素处理文本，从而实现分词。基于理解的分词方法需要大量的语言知识，由于中文文本自身的复杂性，该方法目前还难以实施。

（3）基于统计的分词方法

基于统计的分词方法[3]是指计算机通过计算字符串在语料库的出现频率对其是否构成词进行判断。随着大量语料库的出现及机器学习的不断发展，基于统计的分词方法是目前使用最广泛的一种分词方法。

3.2　分词研究现状

中文文本不同于英文文本，词与词之间没有明显的区分，增加了中文分词的困难。在文本进行分词处理时，常用的手段主要是：利用分词工具直接对文本进行分词操作、利用现有词典进行分词操作和通过算法建立分词模型进而进行分词

操作。常用的分词工具有：张华平等开发的ICTCIAS 分词系统[4]，其主要思想是通过隐马尔可夫模型进行分词，实现已有词识别、简单未登录词识别、词性标注等功能，提高了分词的准确性和效率，NLPIR 分词工具也是以此为基础开发的，缺点是标准版本需要付费，提供的接口难以适用于 JAVA；Jieba 分词工具是基于 Trie 树结构采用动态规划查找最大概率路径的方法得到分词结果，并采用基于隐马尔可夫模型和 Viterbi 算法进行未登录词识别，是国内使用最多的中文分词工具，缺点是在未登录词识别上存在缺陷，大部分需要用户手动加入词典；THULAC 分词工具[5]是由清华大学自然语言处理与社会人文计算实验室研发，具备分词和词性标注等功能，计算能力强、速度快、准确率高，缺点是只支持 UTF8 编码的中文文本。目前，常用的分词词典主要有《同义词词林（扩展版）》[6]《现代汉语语义词典》《现代汉语语义词典》《知网》[7]和《人民日报语料库》等。

许多学者针对分词工具存在的不足展开探索和研究，不断完善中文文本分词方法。针对未登录词识别问题，文献[8]提出网络舆情中的新词识别方法，利用网络舆情中未被词典收录的主题词的局部高频这一特性，通过计算异常分词与周围分词之间的粘结度，识别出未被词典收录的主题词，但该方法仅仅通过单个字分词对异常分词进行判断和召回。文献[9]针对短文本的特点，通过对条件随机场中的标记选择和特征做出了优化，提出一种基于条件随机场的中文文本分词方法，该算法可有效解决传统 CRF 算法标记冗余的问题，并有良好的未登录词识别效果，但由于标记选择的原因，其在不同长度词的识别上有一定的局限。文献[10]对未登录词识别方法做了进一步改进，利用互信息改进算法，提出一种非监督的词识别方法，结合规则，可以在大规模语料中识别出指定长度的新词。文献[11]和文献[12]都通过 LSTM 记忆单元和神经网络模型，对分词方法进行了改进，改进后的方法有效利用序列长距离信息和上下文信息，但算法复杂且神经网络具有黑箱特性，不易于理解。

为提高中文文本分词的速度，文献[13]为提高中文分词的速度，结合分组散列和正则表达式进行字符截取技术，提出基于分组 Hash 与变长匹配的中文分词技术，大幅度降低了算法的时间复杂度。文献[14]利用对抗训练的思想，通过多目标集成学习的方法来学习多个异构标准的分词语料集，利用不同标准的语料来提升分词的效果，突破传统侧重于改进使用单个标准的语料下的分词性能的方法，并通过实验证明，相比单标准学习方法，模型在每个语料集上的性能都获得了显著改进。随着深度学习技术不断发展，越来越多的学习方法被应用到文本分类中。文献[15]提出了一种 CNN 双向 GRU-CRF 神经网络模型，突破了传统方法窗口的限制，有效地利用上下文信息，并通过偏差变量权重贪婪策略解决了在神经网络学习中偏量的影响，缩短了中文分词中模型的训练时间，但易出现特征稀疏导致的过拟合现象。

近年来，许多学者在基于传统分词的基础上，通过对算法改进和整合，提出多种分词新方法。但目前实际应用中，广泛使用的分词手段还是已有的分词工具进行初步分词，再结合未登录词识别算法进一步进行操作。

随着互联网技术的迅速发展，网络用语日新月异，海量新词汇出现，给未登录词识别技术带来了巨大的挑战，且未登录词识别没有一套标准的规范，增加了未登录词的识别难度。消除歧义词往往需要利用上下文、语义等信息，而传统的分词方法往往忽略了文中的关联信息，给歧义词消除带来了困难。分析可得出，中文分词面临的

困难主要有 3 点：未登录词识别、歧义消除、效能提升。处理中文文本时，分词往往是处理的第一步，如何快速实现对文本的精准分词、提高效率，也是当前研究的一个热点。

4 特征提取与特征选择

特征提取和特征选择作为特征工程的两部分，是文本分类算法中的重要一步。特征提取主要是通过属性间的关系，改变原特征空间，如组合不同属性得到新的属性；特征选择则是对原特征空间中的特征进行筛选，没有改变其原属性。但两者的核心目的都是为降低特征向量维度，目前常用的特征提取方法有 PCA、LDA、SVD；常用的特征选择方法主要有 Filter、Wrapper、Embedded；本文对特征选择方法做详细介绍。

4.1 特征选择的一般方法

(1) Filter 法

Filter 法[16]的主要思想是通过对每个特征赋予权重，根据其重要程度对特征进行选择。目前常用的 Filter 法主要有：基于文档频率的方法、x^2 统计量法、互信息方法和信息增益方法。

(2) Wrapper 法

Wrapper 法[17]的实质是将特征选择问题作为寻优的问题，通过对不同组合进行评价和比较，选择出最优的特征集合。目前常用的 Wrapper 方法主要有遗传算法（GA）、粒子群优化（PSO）、优化蚁群算法（ACO）。

(3) Embedded 法

Embedded 法[18]是通过在建立模型的过程中，筛选出对提高模型准确度最有用的特征。

4.2 特征选择研究现状

针对传统特征提取方法存在的不足，众多学者展开了相关研究。针对传统 Filter 算法存在的不足，文献[19]提出一种基于特征重要度的文本特征加权方法，结合实数粗糙集理论定义特征重要度，在特征权重中引入特征对分类的决策信息。文献[20]和文献[21]分别采用基于类别分布信息和改进期望交差熵的方法，对特征提取算法进行了改进，可以提取出有较强区分能力的特征，有效地提高系统的分类效能。文献[22]针对网页内容，提出一种基于正则表达式的特征选取方法，能够有效提取其中的强特征。

针对传统 Wrapper 方法在特征选择时存在的不足，对传统算法进行了改进，取得了显著效果。文献[23]为解决高维数据处理困难，结合机器学习方法，提出一种多重遗传算法的特征选择方法，可以从大量冗余数据中提取出有用的特征信息，但文中的实验数据是通过模拟产生的，在实际数据中使用的实验效果有待进一步验证。文献[24]和文献[25]对粒子群算法进行了优化，分别通过社团划分和开方检验的方法对文本特征进行初步筛选，进而通过粒子群方法得到有效特征，提高了算法的准确率，对特征进行降维处理。基于传统蚁群算法存在的不足，文献[26]和文献[27]对蚁群算法做了改进，分别结合 SVM 评价方法和线性递减的方法动态调整观察半径，制定蚁群算法策略，提高了蚁群算法的效能，避免了传统算法依赖最大迭代次数而消耗时间和运算空间的问题。

随着机器学习和深度学习的不断发展，越来越多学者倾向于采用学习的方法对文本特征进行筛选。文献[28]提出卷积神经网络增强特征选择模型，将传统特征评价方法对特征重要性的理解结合到神经网络的学习过程中，能有效地对特征进行选择。文献[29]针对如何结合上下文信息挖掘信息重点，提出了一个序列匹配网。该网络通过二维卷积神经网和循环神经网的耦合可以很好地对上下文建模并且抓住上下文中的关键点，可以过滤很多冗余信息。文献[30] 将 N-gram 信息引入多种主流的词向量模型中，不仅可以学习到

更好的词向量，同时还能得到高质量的 N-gram 向量，通过构建共现矩阵的方法降低 N-gram 的训练复杂性，这些预训练的向量对于后续 NLP 任务都是非常有用的资源。为解决文本特征稀疏的问题，学者开始将文本主题引入特征中。文献[31]在 LDA 主题模型中引入了词与词的关系，提出了一种基于 Topical N-Gram Model 的特征提取方法，可通过对分词粒度的调整，更加精确地对文本特征进行选择，大大提升了短文本分类的效果，但该算法仍是以 LDA 算法为基础进行词和主题向量的嵌入，无法避免 LDA 算法的缺陷，不能有效解决原始文本特征稀疏的问题。文献[32]提出基于 Biterm Topic Model 的文本主题表示方法，使用结构化的事件来表示主题，有效地解决了事件稀疏性问题。

面对大量文本样本时，对其进行处理后得到的特征往往多而杂，这些特征中的大部分是一些无关特征，如何将其中的有效特征筛选出来，在高维特征中合理地选出高效特征，实现特征降维，从而提高后续操作的能效，是文本特征选择面临的一大挑战。并且，随着机器学习和深度学习的不断发展，越来越多的方法和手段被应用于特征选择算法中，但其融合的算法往往结构复杂、计算量大，如何有效降低算法的复杂度成了特征降维中的又一问题。

5　文本分类

文本分类是指利用计算机按照一定的分类标准或体系自动将文本分门别类[33]，它不仅是自然语言处理问题，也是一个模式识别问题。所以，研究文本分类问题不仅可以推动自然语言研究的发展，对人工智能技术的研究也有重大意义。

5.1　文本分类一般方法

文本分类一般分为两种：基于知识工程（KE，knowledge engineering）的分类方法和基于机器学习（ML，machine learning）的分类方法。基于知识工程的分类方法是指通过专家经验，依靠人工提取规则进行的分类；基于机器学习的分类方法是指通过计算机自主学习、提取规则进行的分类。应用最早的机器学习方法是朴素贝叶斯[34]，随后，几乎所有重要的机器学习算法在文本分类领域得到了应用，如支持向量机(SVM)[35]、神经网络[36]和决策树[37]等。

5.2　文本分类研究现状

针对传统文本分类方法存在的不足，众多学者对文本分类方法展开研究，对其进行修正和改进。基于神经网络算法在自然语言领域处理的优越性，文献[38]分别使用神经网络算法、KNN 算法及 SVM 算法对 Web 文本进行分类，结果显示神经网络算法的准确度优于其他算法。相比于传统的分类主要采用有监督的方法，依赖于现有的自然语言处理工具容易导致处理过程中的误差累积问题，文献[39]提出了基于卷积深层神经网络的文本语义特征学习方法，利用卷积深层神经网络，自动学习表征实体语义关系的词汇特征、上下文特征以及实体所在的句子文本特征等，该方法不需要利用 NLP 处理工具抽取特征，极大地改善了特征抽取过程中多个处理环节所带来的误差累积问题，提高了文本分类的准确性。文献[40]提出了一种基于表观语义和 ASLA 的中文文本分类方法。利用百度百科对中文文本的表观语义进行提取，进而采用 pLSA 挖掘潜在语义，并计算根据表观语义和潜在语义与文档对类别的相关程度，该方法能够很好地处理中文网络短文本等不规则文本的分类。为直接表达文本，文献[41]提出了一种基于密集网的短文本分类模型，采用 one-hot 编码，通过合并和随机选择的方法扩大文本特征选择，解决了特征稀疏、维文本数据和特征表示等方面问题。文献[42]和文献[43]分别采用改进 TF-IDF 修改词向量权重和人工建立词典的

表 1　各分类算法的优缺点

算法	优点	缺点
朴素贝叶斯算法	算法简单，分类效果稳定；适用于小规模数据的训练；所需估算的参数少，对缺失数据不敏感	算法前提假设属性之间相互独立，而实际中往往难以成立；属性多或者属性之间相关性较大时，分类效果不好；分类效果依赖于先验概率；对输入数据的表达形式很敏感
支持向量机算法	可用于小样本数据学习；具有较高的泛化能力；可用于高维数据的计算；可以解决非线性问题；可以避免神经网络结构选择和局部极小点问题	对缺失数据敏感；对非线性问题没有通用解决方案
神经网络算法	并行处理能力强；学习能力强、分类准确度高；对数据噪声有较强的顽健性和容错能力；能解决复杂的非线性关系，具有记忆的功能	神经网络训练过程中有大量的参数需要确定；不能观察网络之间的学习过程，输出结果难以解释；学习时间长，且效果不可保证
决策树算法	易于理解，逻辑表达式生成较简单；数据预处理要求低；能够处理不相关的特征；可通过静态测试对模型进行评测；能够短的时间内对大规模数据进行处理；能同时处理数据型和常规型属性，可构造多属性决策树	易倾向于具有更多数值的特征；处理缺失数据存在困难；易出现过拟合；易忽略数据集属性的相关性
模糊决策算法	顽健性强；可解决非线性、强耦合等问题；容错能力强；通过模糊选择，结果更加精确	不适用于简单信息的处理；无法精确定义目标

方法，对文本分类算法进行优化，最终利用卷积神经网络构造分类器，提高了文本分类的精度，但其对高阶特征未进行合理的处置，导致学习的时间复杂度远高于传统的机器学习方法，还有待进一步改善。文献[44]提出了一种基于深度学习的特征融合模型的文本分类方法，使用卷积神经网络和双向门控循环单元提取文本的上下文信息和本级信息，有效地提取文本间的语义特征信息，降低文本表示对分类结果的影响。

各分类算法的优缺点如表 1 所示。

现有的常用分类方法虽然在某些方面性能上能达到对文本分类的目标要求，但依旧存在算法效率不高、领域针对性差、学习过程易出现过拟合等问题，如何降低学习的时间、提高分类的效率、将不同分类方法的优点进行有机结合、实现高效准确的文本分类已是自然语言处理领域研究的热点问题。

6　结束语

本文介绍了文本分类概念、流程、关键技术和分类方法，综述了现有的研究和解决方法，结合目前文本分类过程中面临的挑战，总结了文本分类相关技术发展趋势，如下。

1) 对文本特征的表示从离散、高维到连续、低维发展。传统文本分类方法对文本进行描述时，一般通过词的方式对其进行表示，随着自然语言处理方法的不断发展，文本表示越来越倾向于以短语、句子为中心的主题表示方法，该类方法可有效解决词表示过程中的稀疏性问题。

2) 对文本的学习由浅层向深层发展。随着机器学习和深度学习的不断发展，文本处理方法开始由传统的步骤式向整体学习转变，对文本的理解由浅层分析到深度理解发展，大量机器学习和深度学习的方法被应用到文本分类过程中，如模糊神经网络、卷积神经网络[45]、循环神经网络[46]等在文本分类中的应用越来越广泛。

3) 文本分类方法由单一向集成发展。随着文本分类技术的日益成熟，各种文本分类方法的优点和不足显露出来，通过合理地融合不同分类方法，如 boosting 改进算法等，可以进一步优化文本分类方法。

参考文献：

[1] 郑捷.NLP 汉语自然语言处理原理与实践[M]. 北京：电子工业出版社,2017.

ZHENG J. NLP chinese natural language processing principle and practice [M]. Beijing: Publishing House of Electronics Industry, 2017.

[2] 常建秋, 沈炜. 基于字符串匹配的中文分词算法的研究[J]. 工业控制计算机, 2016, 29(2): 115-116+119.
CHANG J Q, SHEN W. Research on chinese word segmentation algorithm based on string matching[J]. Industrial control computer, 2016, 29(2): 115-116+119.

[3] 邹佳伦, 文汉云, 王同喜. 基于统计的中文分词算法研究[J]. 电脑知识与技术, 2019, 15(4): 149-150+153.
ZOU J L, WEN H Y, WANG T X. Research on chinese word segmentation algorithm based on statistics[J]. Computer Knowledge and Technology, 2019, 15(4): 149-150+153.

[4] 刘群, 张华平, 俞鸿魁, 等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004(8): 1421-1429.
LIU Q, ZHANG H P, YU H K, et al. Chinese lexical analysis model using cascading hidden horse model[J]. Journal of Computer Research and Development, 2004(8): 1421-1429.

[5] 孙茂松, 陈新雄, 张开旭, 等. THULAC: 一个高效的中文词法分析工具包[R]. 2016.
SUN M S, CHEN X X, ZHANG K X et al. THULAC: an efficient Chinese lexical analysis toolkit[R]. 2016.

[6] 熊回香, 叶佳鑫. 基于同义词词林的社会化标签等级结构构建研究[J]. 情报杂志, 2018, 37(1): 126-131.
XIONG H X, YE J X. Research on *t* Structure construction of social tags TongYiCi CiLin[J].Journal of Information,2018, 37(1) : 126-131.

[7] XIE R, YUAN X, LIU Z, et al. Lexical sememe prediction via word embeddings and Matrix factorization[C]//The 26th International Joint Conference on Artificial Intelligence. 2017: 4200-4206.

[8] 唐籍涛, 李飞, 郭昌松. 网络舆情监控中新词识别问题的研究[J]. 计算机技术与发展, 2012, 22(1): 119-121+125.
TANG J T, LI F, GUO C S. Research on new word pattern recognition in network monitoring public opinion[J].Computer Technology and Development, 2012, 22(01): 119-121+125.

[9] 刘泽文, 丁冬, 李春文. 基于条件随机场的中文短文本分词方法[J]. 清华大学学报(自然科学版), 2015, 55(08): 906-910+915.
LIU Z W, DING D, LI C W. A chinese short text word segmentation method based on conditional random fields[J]. Journal of Tsinghua University(Science and Technology), 2015, 55(8): 906-910+915.

[10] 杜丽萍, 李晓戈, 于根, 等. 基于互信息改进算法的新词发现对中文分词系统改进[J]. 北京大学学报(自然科学版), 2016, 52(01): 35-40.
DU L P, LI X G, YU G, et al. New word detection based on an improved PMI algorithm for enhancing segmentation system[J]. Journal of Peking University(Natural Science), 2016, 52(01): 35-40.

[11] 胡婕, 张俊驰. 双向循环网络中文分词模型[J]. 小型微型计算机系统, 2017, 38(3): 522-526.
HU J, ZHANG J C. Bidirectional recurrent network for chinese word segmentation[J]. Journal of Chinese Computer Systems, 2017, 38(3): 522-526.

[12] ZHANG Y N, XU J N, MIAO G Y, et al. Improving neural chinese word segmentation using unlabeled data[J]. IOP Conference Series: Materials Science and Engineering, 2018, 435(1).

[13] 杨光豹, 杨丰赫, 毛贵军. 基于分组 hash 与变长匹配的中文分词技术[J]. 计算机时代, 2019(4): 52-55.
YANG G B, YANG F H, MAO G J. Chinese word segmentation technology based on group hash and variable length matching[J]. Computer Age, 2019(4): 52-55.

[14] CHEN X C, SHI Z, QIU X P, et al. Adversarial multi-criteria learning for chinese word segmentation[C]//ACL2017:Computation and Language. 2017.

[15] YU C H, WANG S P, GUO J J. Learning chinese word segmentation based on bidirectional GRU-CRF and CNN network model[J]. International Journal of Technology and Human Interaction (IJTHI), 2019, 15(3).

[16] DARSHAN S L, JAIDHAR C D. Performance evaluation of filter-based feature selection techniques in classifying portable executable files[J]. Procedia Computer Science, 2018, 125: 346-356.

[17] HANCER E. Differential evolution for feature selection: a fuzzy wrapper–filter approach[J]. Soft Computing,2019,23(13): 5233-5248.

[18] MALDONADO S, LÓPEZ J. Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification[J]. Applied Soft Computing, 2018(67): 228-246.

[19] 刘赫, 刘大有, 裴志利, 等.一种基于特征重要度的文本分类特征加权方法[J]. 计算机研究与发展, 2009, 46(10): 1693-1703.
LIU H, LIU D Y, YAN Z L, et al. A Feature weighting scheme for text categorization based on feature importance[J]. Journal of Computer Research and Development, 2009, 46(10): 1693-1703.

[20] 靖红芳, 王斌, 杨雅辉, 等. 基于类别分布的特征选择框架[J]. 计算机研究与发展, 2009, 46(9): 1586-1593.
JING H F, WANG B, YANG Y H, et al. Category distribution-based feature selection framework [J]. Journal of Computer Research and Development, 2009, 46(9): 1586-1593.

[21] 单丽莉, 刘秉权, 孙承杰. 文本分类中特征选择方法的比较与改进[J]. 哈尔滨工业大学学报, 2011, 43(S1): 319-324.
SHAN L L, LIU B Q, SUN C J. Comparison and improvement of feature selection methods in text categorization[J].Journal of Harbin Institute of Technology,2011,43(S1):319-324.

[22] 王正琦, 冯晓兵, 张驰. 基于两层分类器的恶意网页快速检测系统研究[J]. 网络与信息安全学报, 2017, 3(8): 48-64.
WANG Z Q, FENG X B, ZHANG C. Study of high-speed malicious web pages detection system based on two-type classifier[J].Journal of Network and Information Security,2017,3(8):48-64.

[23] 蒋胜利. 高维数据的特征选择与特征提取研究[D]. 西安: 西安电子科技大学,2011.
JIANG S L. Research on feature selection and feature extraction of high dimensional data[D]. Xi'an: Xidian University, 2011.

[24] 李炜, 巢秀琴. 改进的粒子群算法优化的特征选择方法[J]. 计算机科学与探索, 2019, (6): 990-1004.
LI W, CAI X Q. Improved particle swarm optimization method for feature[J]. Journal of Frontiers of Computer Science and Technology, 2019, (6): 990-1004.

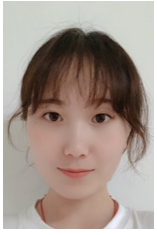
[25] 路永和, 曹利朝. 基于粒子群优化的文本特征选择方法[J]. 现代图书情报技术, 2011, (Z1): 76-81.
LU Y H, CAO L C. Text feature selection method based on particle swarm optimization[J]. New Technology of Library and Information Service, 2011, (Z1): 76-81.

[26] 张杰慧, 何中市, 王健, 等. 基于自适应蚁群算法的组合式特征选择算法[J]. 系统仿真学报, 2009, 21(6): 1605-1608+1614.
ZHANG J H, HE Z S, WANG J, et al. Hybrid feature selection algorithm based on adaptive ant colony algorithm[J]. Journal of System Simulation, 2009, 21(6): 1605-1608+1614.

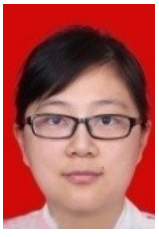
[27] 张海涛. 基于文本降维和蚁群算法的文本聚类研究[D]. 合肥:

安徽大学, 2016.
ZHANG H T. Research on text clustering based on text dimension reduction and ant colony algorithm [D]. Hefei: Anhui University, 2016.
[28] 卢泓宇, 张敏, 刘奕群, 等. 卷积神经网络特征重要性分析及增强特征选择模型[J]. 软件学报, 2017, 28(11): 2879-2890.
LU Y Y, ZHANG M, LIU Y Q, et al. Convolutional neural networks importance analysis and feature selection enhanced model [J]. Journal of Software, 2017, 28(11): 2879-2890.
[29] WU Y, WU W, XING C, et al. Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots[C]//ACL 2017: Computation and Language. 2017.
[30] ZHAO Z，LIU T, LI S, et al. Ngram2vec: learning improved word representations from ngram co-occurrence statistics[C]//Conference on Empirical Methods in Natural Language Processing. 2017: 244-253.
[31] 赵芃. 基于主题模型与深度学习的短文本特征扩展与分类研究[D]. 天津: 天津工业大学, 2018.
ZHAO P. Research on short text feature expansion and classification based on topic model and deep learning [D]. Tianjin: Tianjin Polytechnic University, 2018.
[32] 孙锐, 郭晟, 姬东鸿. 融入事件知识的主题表示方法[J].计算机学报,2017,40(4):791-804.
SUN R, GUO W, JI D H. Topic representation integrated method of integrating with event knowledge[J].Chinese Journal of Computers, 2017, 40(4): 791-804.
[33] 李森, 马军, 赵嫣, 等. 对数字化科技论文的自动分类研究[J]. 山东大学学报(理学版), 2006(3): 81-84.
LI S, MA J, ZHAO Y, et al. Automatic classification of digital science and technology papers[J]. Journal of Shandong University(Science Edition), 2006(3): 81-84.
[34] CUI W. A chinese text classification system based on naive bayes algorithm[C]//MATEC Web of Conferences. 2016: 1015.
[35] ZHANG M Y, AI X B, HU Y Z. Chinese text classification system on regulatory information based on SVM[C]//IOP Conference Series: Earth and Environmental Science. 2019: 252.
[36] SAHA D. Web text classification using a neural network[C]//2011 Second International Conference on,2011.
[37] 雷飞. 基于神经网络和决策树的文本分类及其应用研究[D]. 成都: 电子科技大学, 2018.
LEI F. Text research on text classification based on neural network and decision tree and its application [D]. Chengdu: University of Electronic Science and Technology of China, 2018.
[38] 周朴雄.基于神经网络集成的 Web 文档分类研究[J]. 图书情报工作, 2008, (7): 110-112.
ZHOU P X. Study on Web document classification based on neural network integration[J]. Library and Information Service, 2008, (7): 110-112.
[39] ZENG D J, LIU K, LAI S W, et.cl. relation classification via convolutional deep neural network[C]//The 25th international Computational Linguistics: 2014: 2335-2344.
[40] CHEN Y W, WANG J L, CAI Y Q , et al. A method for chinese text classification based on apparent semantics and latent aspects[J]. Journal of Ambient Intelligence and Humanized Computing, 2015, 6(4): 473-480.
[41] LI H M, HUANG H N, CAO X, et al. Falcon: a novel chinese short text classification method[J]. Journal of Computer and Communications, 2018, 6: 216-226.
[42] 王根生, 黄学坚. 基于 Word2vec 和改进型 TF-IDF 的卷积神经网络文本分类模型[J]. 小型微型计算机系统, 2019, 40(5): 1120-1126.
WANG G S, HUANG X J. Convolutional neural network text classification model based on Word2vec and improved TF-IDF[J]. Journal of Chinese Computer Systems, 2019, 40(5): 1120-1126.
[43] 王磊. 基于混合神经网络的中文短文本分类方法研究[D]. 杭州: 浙江理工大学, 2019.
WANG L. Research on chinese short text classification based on hybrid neural network[D]. Hangzhou: Zhejiang University of Science and Technology, 2019.
[44] JIN W Z, ZHU H, YANG G C. An efficient character-level and word-level feature fusion method for chinese text classification[C]// Journal of Physics: Conference Series. 2019: 12057.
[45] WANG P, XU B, XU J M, et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification[J]. Neurocomputing, 2016, 174: 806-814.
[46] 龚千健. 基于循环神经网络模型的文本分类[D]. 武汉: 华中科技大学, 2016.
GONG Q J. Text Classification based on recurrent neural network model[D]. Wuhan: Huazhong University of Science and Technology, 2016.

[作者简介]



于游（1995− ），女，山东威海人，海军工程大学硕士生，主要研究方向为信息安全。



付钰（1982− ），女，湖北武汉人，博士，海军工程大学副教授，主要研究方向为信息安全风险评估。

吴晓平（1961− ），男，山西新绛人，博士，海军工程大学教授、博士生导师，主要研究方向为系统分析与决策。