作者识别研究综述

张 洋 [1] 江铭虎 [1]

摘 要 作者识别是根据已知文本推断未知文本作者的交叉学科. 其传统研究通常基于文学或语言学的经验知识, 而现代研究则主要依靠数学方法量化作者的写作风格. 近些年, 随着认知科学、系统科学和信息技术的发展, 作者识别受到越来越多研究者的关注. 本文主要站在计算语言学的角度综述作者识别领域现代研究中的方法和思路. 首先, 简要介绍了作者识别的发展历程. 然后, 详述了文体风格特征、作者识别方法以及该领域中多层面的研究. 接着介绍了与作者识别相关的一些评测、数据集及评价指标. 最后, 指出该领域存在的一些问题, 结合这些问题分析并展望了作者识别的发展趋势.

关键词 作者识别, 文体学, 写作风格, 评价指标

引用格式 张洋, 江铭虎. 作者识别研究综述. 自动化学报, 2021, **47**(11): 2501−2520

DOI 10.16383/j.aas.c200654

A Review on Authorship Identification Research

ZHANG Yang[1] JIANG Ming-Hu[1]

Abstract Authorship identification is an interdisciplinary subject of inferring the author of unknown texts based on the known texts. The traditional research of authorship identification is generally based on the empirical knowledge of literature or linguistics, while the modern research mostly relies on mathematical methods to quantify the author′s writing style. In recent years, with the development of cognitive science, system science and information technology, more and more researchers pay attention to authorship identification. This paper mainly reviews the methods and ideas in modern research in the field of authorship identification from the perspective of computational linguistics. First, the development history of authorship identification is introduced briefly. Then, the stylometry, authorship identification methods and multi-faceted research in this realm are expounded. Next, some evaluations, data sets and evaluation metrics related to authorship identification are explicated. Finally, some problems in this domain are pointed out, while the development trend of authorship identification is analyzed and forecasted combined with these problems.

Key words Authorship identification, stylometry, writing style, evaluation metrics

Citation Zhang Yang, Jiang Ming-Hu. A review on authorship identification research. *Acta Automatica Sinica*, 2021, **47**(11): 2501−2520

大数据时代, 各种信息服务给人们的生活带来很多便捷, 人们足不出户就能知晓天下事. 然而, 人们在获取信息的同时也饱受信息泛滥的困扰. 垃圾短信、虚假信息、网络诈骗等严重影响人们的日常生活. 因此, 准确而及时地识别垃圾信息、阻止虚假信息或低俗作品的传播, 对于维护互联网生态健康以及保障人们的正常生产生活具有非常重要的意义. 作者身份识别 (Authorship identification) 又简称作者识别, 是通过分析未知文本的文体学特征或写作风格, 推断作者归属的一类研究. 有些研究者也称其为作者身份归属 (Authorship attribution), 其主要思路是将文本中隐含的作者无意识的写作习惯通过某些可以量化的特征表现出来, 进而凸显作品的文体学特征或写作风格, 以此确定匿名文本的作者[1].

19 世纪以前, 科学研究的社会化程度较差, 数学等相关工具的应用不普及, 作者识别尚处于传统研究的历史阶段. 在这一时期, 西方研究者通过韵律节奏的变换推断未知十四行诗的作者归属. 其判别标准多基于研究者的主观经验, 而缺乏客观参数衡量. 进入 19 世纪, 随着生产力的不断发展, 科学研究的社会化程度不断增强, 数学等相关工具也逐渐被应用到作者识别研究中. 最早尝试用数学工具去量化作者写作风格的是 Mendenhall[2], 他利用词谱和特征曲线对莎士比亚的戏剧等不同作品进行作

收稿日期 2020-08-14 录用日期 2021-02-09
Manuscript received August 14, 2020; accepted February 9, 2021
国家自然科学基金 (62036001) 资助
Supported by National Natural Science Foundation of China (62036001)
本文责任编委 张家俊
Recommended by Associate Editor ZHANG Jia-Jun
1. 清华大学人文学院计算语言学实验室 北京 100084
1. Lab of Computational Linguistics, School of Humanities, Tsinghua University, Beijing 100084

者归属研究, 标志着作者识别现代研究的开端. Yule[3] 在 Mendenhall 基础上进行了改进, 他利用文本句子长度作为识别散文等文学作品作者的有效特征. 最有影响力的早期研究是 Mosteller 和 Wallace[4] 合作完成的, 他们首次提出利用少数特殊词出现的频率以及分布作为特征, 识别联邦主义者论文的作者. Damerau[5] 在分析前人方法的基础上, 首次提出基于功能词 (Function words) 频率的作者识别方法, 有效地拓展了词汇特征. Efron 和 Thisted[6] 通过作品的词汇量推断未知文本是否为莎士比亚所作. 从此, 词汇成为作者识别以及作者风格分析一个重要的研究方向. 随后研究者不断尝试新的文本特征, 字符、句法、语义等特征均被研究者用于作者识别研究中, 取得了一些进展. 20 世纪 90 年代随着计算机技术和网络应用的发展, 大量电子文本产生, 于是便有了处理这些信息的需求. 这使得作者识别在情报分析和计算机取证等领域的地位越来越重要. 作者识别的意义主要体现在相关技术的应用上. 在法医学中, 作者识别技术可以对关键文字证据进行识别, 从而确定当事人的身份, 这对进一步侦破案件有着重要的作用[7]. 在文学研究中, 作者识别技术可以用来识别匿名作品的作者, 或者推断争议文本的作者, 给作者身份的确定带来新思路[8]. 在互联网领域, 作者识别技术可以追溯垃圾邮件、谣言以及计算机病毒等非法程序源代码的作者, 对于打击网络违法行为和维护网络安全具有重要的意义[9].

作者识别是一个涉及众多学科的交叉学科, 为了简化问题和便于研究, 研究者常常提出一些假设. 首先, 第一个假设就是, 作者的写作风格会受到作者自身特征的影响, 比如作者的身份地位、性别、性格、年龄和受教育程度等[10]. 这个是作者识别研究的首要假设. 第二个假设就是作者的这些特征能够从他的写作风格中看出来[11]. 这个假设是作者识别研究中最重要的一个假设. 在现代研究中, 研究者常常需要量化作者写作风格. 这个假设为量化作者写作风格提供了理论依据. 然而, 在一般情况下, 作者的写作风格并非一成不变, 它会受到很多外部条件的影响, 比如社会背景、时间、文本主题、传播媒介、受众等因素. 但研究者也一致认为作者写作风格的某些特征元素始终存在, 无论这些因素是什么, 它们都能够被研究者通过特定手段进行量化. 研究者所要做的是尽可能多地保持潜在的相互作用因素恒定, 而并非将它们剥离出来, 因为这会损失更多的信息[12].

作者识别领域有两个大的研究方向, 大多数作者识别研究都是从这两个方向进行的, 其中一个方向是数字人文 (Digital humanities) 方向, 而另一个方向则是计算语言学 (Computational linguistics) 方向[13]. 这两个方向的研究内容并无太多差异, 但在目的和侧重点上则有显著不同. 在数字人文研究中, 重点主要放在实际有争议的作者身份或文学风格分析的案例上; 而在计算语言学研究中, 研究者则更多地关注已知作者身份的数据集的表现以及确定最可靠的技术[13]. 计算语言学中更系统的方法允许严格控制与作者身份相互作用的因素, 比如主题和流派, 这种设置通常在有争议的作者身份的情况下无法实现. 一些模拟大规模作者身份归属的研究, 比如增加作者集合大小或减少训练数据大小, 允许系统地评估在各种情况下的技术水平. 以数字人文为导向的研究的主要优点之一是注重结果的解释以及对作者写作风格的分析. 这种类型的分析目前缺乏以计算语言学为导向的研究. 用一句话来概括数字人文和计算语言学这两个大方向的不同点: 数字人文学科更注重可解释性, 研究者常常希望通过模型解释作者识别结果或者分析作者风格特点; 而计算语言学更关注算法本身的正确率、鲁棒性、运行效率等性能, 而并非可解释性.

如果进一步细分, 作者身份识别任务通常有如下 3 种不同的形式: 闭集归属 (Closed-set attribution)、开集归属 (Open-set attribution) 和作者身份验证 (Authorship verification)[14]. 也有研究者给出了不同的分类标准, 他们把作者身份识别任务分为闭集归属、开集归属以及作者身份概述 (Authorship profiling), 而把作者身份验证视为开集归属的一种[12]. 一般而言, 闭集归属指的是未知文档的作者包含在候选作者集中的一类问题. 这是相对比较简单的一种情况, 也是学者们研究得最多的一类问题. 而开集归属则是未知文档的作者不一定包含在候选作者集中的一类问题. 这是比闭集归属更加困难的情况, 在很多与互联网相关的作者归属研究中, 研究者常常会面临庞大的候选作者集合以及未知文本不包含在候选作者集合中的情况. 作者身份验证是确定给定的文本是否由某位作者撰写的任务. 它与开集归属的主要区别在于, 作者身份验证样本数量少、候选作者集合单一. 所有作者身份归属问题都可以转换为一组单独的作者身份验证问题[14]. 因此, 作者身份验证问题是作者识别中的基本问题, 研究有效处理此类问题的方法对于作者识别研究至关重要.

本文后续章节的具体内容如下: 第 1 节介绍了作者识别中的文体风格特征, 主要包括字符特征、词汇特征、句法特征和语义特征在内的多元文体特征; 第 2 节阐述了常见的作者识别方法, 主要分为

无监督的方法和有监督的方法; 第 3 节总结了作者识别中的一些多层面的研究, 主要包括数据规模、跨域研究和特殊方法; 第 4 节介绍了与作者识别相关的一些评测; 第 5 节综述了作者识别领域的一些公开数据集以及各种评价指标; 第 6 节指出作者识别领域存在的一些问题; 第 7 节针对作者识别领域存在的问题, 分析并展望了该领域未来可能的发展趋势.

1 文体风格特征

文体风格是指作者在创作过程中表现出的一切行文方式的总和. 作者的写作风格来源于作者思想表达的方式. 在表达过程中, 作者会无意识地将其个性及社会背景融入进去. 虽然作者的写作风格会随着时间的推移而慢慢改变, 但研究者通常都假定衡量作者写作风格的特征元素始终存在, 并且可以通过某种技术手段进行量化. 文体风格特征主要分为一元文体特征和多元文体特征. 一元文体特征主要包括单词长度、句子长度、段落数、总词汇量等, 具有简单、便于统计等特点, 因此早期的作者识别采用的都是一元文体特征. 然而, 一元文体特征过于简单, 无法进行更深入的分析, 因此研究者又提出多元文体特征. 多元文体特征往往是一些简单特征的进一步组合, 研究表明多层面的文本特征能够有效提高作者识别的准确率[15]. 根据文体风格特征对语言学计算的需求和复杂度, 可以将多元文体特征分成字符特征、词汇特征、句法特征和语义特征等[1]. 有关一元文体特征的研究在上一部分已经简单叙述, 本部分主要针对几类典型的多元文体特征进行阐述.

1.1 字符特征

字符是指文本中使用的字母、数字、字和符号. 根据字符的种类可以定义各种字符级别的度量: 字母字符数、数字字符数、大写和小写字符数、字母频率、标点符号数等. 这种类型的度量很容易用于任何自然语言和语料库, 并且已被证明对量化写作风格非常有效[13]. 更高阶的字符特征是基于字符组合的特征, 研究者称其为字符 n-gram. 字符 n-gram 即为 n 个连续字符的组合, 这种高阶字符特征具有很多优秀的性质. 它可以捕捉到作者风格的细微差别, 包括由词汇、上下文、标点符号以及大小写变动所带来的差别[16]. 而且, 字符 n-gram 比单一字符抗干扰能力强, 特别适合短文以及风格多变的网络文章、电子邮件等.

很多研究者尝试使用字符 n-gram 来进行作者识别研究. Keselj 等[17] 提出一种通过计算和比较字符 n-gram 频率识别作者的方法. 该方法由 1976 年的开创性方法衍生, 首先选择少量频繁出现的字符 n-gram 构建文档轮廓, 然后选择包含在轮廓中的最佳 n 元组计算文档的相似度. 在对英语、希腊语和中文数据进行的实验中证明了该方法的有效性和语言独立性. Houvardas 和 Stamatatos[18] 在 Keselj 研究的基础上做了改进, 他提出了一种可变长度的 n-gram 方法, 用于选择可变长度的单词序列. 研究结果表明该方法至少与选择最重要的 n-gram 的信息增益一样有效.

Keselj 等的研究是作者识别领域中 n-gram 特征与作者轮廓相结合的早期研究, 最初的作者轮廓只包含单一类型的特征, 比如只包含字符 n-gram 或者词汇频率, 后面也逐渐发展出包含不同类型特征的作者轮廓. Stamatatos[19] 提出一种基于特征集子空间的作者识别方法, 把每个文本表示为字符 n-gram 的频率向量, 产生了具有高准确率的分类模型. 这是基于字符 n-gram 的集合模型, 给后续研究提供了新的思路.

除了由字符组成的 n-gram 之外, 一些研究者也会探究由单词、词性 (Part of speech, POS) 标签、标点符号、词缀等元素组成的 n-gram 在作者识别中的应用. Sapkota 等[20]研究了与不同语言特征相对应的字符 n-gram 子组, 结果表明关于词缀和标点符号的 n-gram 几乎占据了字符 n-gram 的所有功能, 为将来的作者识别工作和其他分类任务使用 n-gram 提供了新的见解. Sari 等[21] 使用连续的字符和单词 n-gram 表示研究作者身份归属, 与使用离散特征表示的工作相比, 模型可以通过神经网络与分类层一起学习 n-gram 特征的连续表示, 进而产生较优的的结果. Gomez-Adorno 等[22] 利用字符、单词和 POS 标签的 n-gram去学习文档段落向量, 获得了优于基于单词嵌入和基于字符 n-gram 线性模型的结果.

1.2 词汇特征

词汇是一种语言里所有词语和固定短语的总和. 最初的作者识别研究就是对词汇进行简单地统计分析, 这种方法简单易行, 适用于任何语言和任何语料库. 然而, 对于某些自然语言, 还需要一些其他辅助手段. 比如, 汉语需要首先进行分词, 然后才能进行词汇的统计分析. 某些大量使用缩写或首字母缩写的文本, 应当加入相应的识别规则, 尽管在这一过程中可能引入相当大的噪声.

词汇的丰富程度被认为是衡量作者写作风格的

一个重要因素, 有些研究者提出了各种各样函数来衡量词汇丰富度. 后续研究者更多的是通过实验验证不同组合的有效性. Burrows[23] 概述了使用常用词的相对频率来比较书面文本和测试其可能的作者身份的方法, 其中涉及的程序为区分超过 1 500 字长的文本作者提供了一个简单但相对准确的补充. Hoover[24] 探讨了使用词汇丰富度进行作者识别的效果, 并测试了词汇丰富度的适当度量可以捕捉作者写作风格或身份的假设. 实验表明, 词汇丰富度在文体和作者研究中具有边际价值, 而它对于大型文本群体是无效的, 因为不同文本之间存在极大的可变性. Garcia 和 Martin[25] 根据前人在词汇领域里提出的诸多参数, 研究它们实际表征的文本特定特征, 以寻求一种可靠的表达方式来衡量作者的词汇丰富度. 实验证实, 不同参数可以互相补充, 富文本往往通过其低功能来表征密度, 反之亦然.

然而, 词汇丰富度往往与很多因素密切相关, 比如文本的主题、内容、类别等因素. 因此研究者需要进一步考虑根据何种词汇特征来衡量特定作者的写作风格. 功能词被认为是区分作者的有效特征之一. 由于功能词不携带任何语义信息, 与文本主题无关, 作者很大程度上是在无意识的情况下使用它们, 因此功能词能够捕捉不同作者的写作风格. Zhao 和 Zobel[26] 研究功能词在新闻专线文章作者归属中的性能, 并通过增大数据量进一步观察其表现. 实验证实基于功能词特征的方法具有较好的可扩展性, 随着问题规模的增加, 其性能只有适度的下降. Coyotl-Morales 等[27] 通过组合功能词和内容词的一组词序列来表征文档, 并用诗歌进行分类实验, 得到了优于大多数方法的结果.

还有一种与功能词类似的思路, 就是为特定作者定义词汇特征集. 一种简单且常见的方法是在语料库中提取常用词, 然后再决定用作特征的频繁单词的数量. 不同研究者所定义的词汇特征集大小不同, 除了他们个人对衡量作者文本风格的因素的把握之外, 所使用的分类算法也会在很大程度上限制特征集的规模. 因为当问题的维度增加时, 许多分类器会出现过拟合. 并且, 特征集维度增加时, 一些特定于内容的单词也会包括在该特征集中.

Stamatatos[28] 提取 1 000 个最常用的单词构建特征集, 研究基于特征集子空间的分类器集合. 结果表明, 使用穷举的不相交子空间构造的集合在两个基准语料库上得到了较优的结果. Koppel 等[29] 使用 250 个最常用的单词构建特征集, 利用基于学习的方法表征两个示例集之间的 “差异深度”, 并证明了该方法以非常高的准确率解决了作者身份验证问题. Savoy[30] 提出一种计算标准化 Z 分数的技术, 该分数能够定义未知文本中的特定词汇. 与其他方法相比较, 该方法优于基于最常用词的 Delta 方法、基于词汇和标点符号的卡方距离以及基于预定义的方法. Akimushkin 等[31] 引入一种通用的相似性度量来比较文本, 通过考虑对应于节点的单词来增强复杂网络中文本的表示. 在 3 个书集上的实验表明, 该方法获得了超过 90 % 的准确率, 比基于词频–逆文本频率指数 (Term frequency-inverse document frequency, TF-IDF) 的传统方法要高得多, 也比不考虑节点标签的其他网络方法要高.

1.3 句法特征

句法是句子各个组成部分排列规则的统称. 研究者一般认为作者的写作风格在很大程度上由其遣词造句的模式决定. 因此, 句法特征在很长一段时间内都受到研究者的重视. 句法特征分为浅层句法特征和深层句法特征. 浅层句法特征是指不需要经过句法解析就能提取的特征, 比如词汇 n-gram; 而深层句法特征则是必须要经过句法解析才能提取的特征, 比如依存句法. 浅层句法特征多是一些词汇特征组合, 在之前的章节中已有叙述, 本节着重叙述深层句法特征.

深层句法特征能够表达隐含的文本结构, 并且在更高维度上刻画作者写作风格. 因此, 与词汇特征和浅层句法特征相比, 深层句法特征被认为是更可靠的作者指纹. Raghavan 等[32] 为每个作者构建概率上下文无关文法, 并使用该文法作为分类的语言模型进行作者归属. 该方法在几个数据集上的性能优于基线模型, 并且还具有一定的扩展性.

句法树是描述句子中各种不同成分之间相互关系的树状结构, 在句法特征的研究中有着重要的应用. 常见的两种句法树是短语结构树 (Constituent tree) 和依存句法树 (Dependency tree), 二者的主要区别在于短语结构树用来描述句子的句法结构, 叶子结点与输入句子中的词语相关联, 中间结点都是标记短语成分; 而依存树用来表达句子中词与词的依存关系, 其每个结点都是一个词语, 词语之间通过有向依存弧连接, 依存弧上标有相应的依存关系.

有些研究者尝试使用短语结构树研究作者识别. Tschuggnall 和 Specht[33] 提出一种通过分析作者的句法来增强作者识别的方法. 该方法先计算文本中每个句子的短语结构树, 再使用 pq-gram 将其分成长度无关的模式, 然后使用最常用的 pq-gram 来组成作者的样本, 再利用各种距离度量和相似性得分进行作者识别. 使用三个不同且独立的数据集

进行的评估得到了有希望的结果. Patchala 和 Bhatnagar 等[34] 提出了一种有效的基于模板的方法, 用于组合文档的各种句法特征以进行作者分析. 基于短语结构树的特征独立于文档主题, 能够反映作者固有的写作风格. 结果表明, 使用包括解析树子树的模板以及其他句法特征可以提高作者识别正确率. Zhang 等[35] 提出一种将句子的短语结构树编码为可学习的分布式表示形式的方法. 该方法为句子中的每个单词构造一个嵌入向量, 在对应于该单词的句法树中对路径进行编码. 此方法在五个数据集上获得了更高的准确率.

Sidorov 带领的研究团队提出句法 n-gram 的概念. 传统的 n-gram 是文本中若干同类元素的顺序组合, 这些元素可以是字符、单词、POS 标签等. 而句法 n-gram 则是句法树中若干同类元素在句法路径上的顺序组合. 换句话说, 句法 n-gram 是根据句法树中的路径构造的 n-gram, 而不是在文本的表面表示中获取的. 从本质上来说, 传统的 n-gram 是对文本局部信息的描述; 而句法 n-gram 则是对句法树或句法结构局部信息的描述. 因此, 与传统 n-gram 相比, 句法 n-gram 将句法知识引入机器学习方法中.

Sidorov 等[36] 利用基于句法关系 (Syntactic relation, SR) 标签的句法 n-gram 特征搭配支持向量机 (Support vector machine, SVM)、朴素贝叶斯 (Naive Bayes, NB) 和树分类器 J48 进行作者识别. 实验结果表明, 与多种传统的 n-gram 相比, 基于 SR 标签的句法 n-gram 获得了更好的结果. 并且在绝大多数情况下, SVM 要优于 NB 和 J48. 句法 n-gram 把特征组合的思想从链式结构拓展到树形结构上, 扩展了 n-gram 特征的维度. 同时为研究者提供了一种衡量句法树相似程度的思路, 研究者可以通过衡量句法树之间的距离间接判断不同文本的相似程度.

受此思想的影响, 学者们进一步探究了不同的句法 n-gram 特征在作者识别中的应用. Posadas-Duran 等[37] 提出了一种基于完整的句法 n-gram 作为风格标记的作者身份归属方法. 该方法利用 SR 标签、POS 标签以及词根的句法 n-gram 等特征刻画作者的写作风格, 并利用 SVM 进行分类. 实验结果表明, 完整的句法 n-gram 是比字符 n-gram 更有效的识别作者的特征, 使用该方法可以在较小的样本集中获得更准确的结果. 在另外两篇文章中, Posadas-Duran 等又把多种基于句法的 n-gram 特征用于 PAN 2015 作者身份验证任务[38] 和作者身份概述任务[39] 上. 结果表明, 在作者身份验证任务中, 荷兰语获得了较低的分数, 而英语和西班牙语获得了适中的分数; 而在作者身份概述任务中, 在预测个人特征时, 将句法 n-gram 与其他特定的推文特征结合使用可以获得良好的结果; 但在预测年龄和性别特征时, 它们的使用则并不成功.

1.4 语义特征

语义特征是根据文本语言所蕴含的意义而提取的特征. 由于语义特征与文本的内容和主题相关性强, 并不容易借助它捕捉作者自然流露出的写作风格, 因此语义特征在作者识别领域内的应用较少. 应用语义特征进行作者识别的研究者往往也会把语义特征和字符、词汇、句法等特征结合起来使用, 以提高作者识别的准确率.

Gamon[40] 提出了一些特征集和分类方法, 并使用了一种能够生成语义依赖图的工具, 实验结果表明深度语言分析特征可以在更常用的浅层特征上实现显著的误差减少. 武晓春等[41] 依据文体学理论, 利用 HowNet 知识库, 提出一种基于词汇语义分析的相似度评估方法, 利用功能词以外的其他词汇, 达到了较好的作者识别效果. Argamon 等[42] 基于确定的词或短语的各种语义功能, 提出一种词汇特征用于文体分类. 实验证明, 这些特征对于确定作者身份和国籍的分类任务具有重要作用. Hedegaard 和 Simonsen[43] 使用基于框架语义的分类器研究作者身份归属, 并测试它们对翻译文本的适用性. 结果表明, 对于翻译文本而言, 框架是有用的, 并且频繁词和框架的组合方法可以胜过仅基于传统标记的方法. 而对于未翻译文本, 频繁词和 n-gram 则是首选.

1.5 对比分析

本节从特征细分、获取难易度、应用广泛度等其他方面来比较不同的文本特征. 表 1 给出了这些方面的比较. 作者识别与文本分类、情感分析、关系抽取等自然语言处理任务均属于文本理解范畴, 而它们所关注的文本知识类型不同. Daelemans 区分了可以从文本中提取的三种知识类型: 客观知识、主观知识和元知识. 客观知识主要是回答谁、什么、什么地方、什么时候等问题的知识; 主观知识是回答谁对什么有何看法等问题的知识; 而元知识是除了内容本身以外, 能从文本中提取到的关于作者个人信息或者个人写作风格等方面的知识[44]. 按照这个分类标准, 文本分类和关系抽取提取的是客观知识, 情感分析提取的是主观知识, 而作者识别提取的是元知识. 因此, 研究者倾向于选择与文本内容无关的特征来进行作者识别, 而其他自然语言处理

表 1 文体风格特征对比表

Table 1 Comparative table of stylometry

文体特征	特征细分	获取难易度	应用广泛度	其他
字符特征	字符数量, 字符 n-gram, 字符错误	非常容易, 可直接提取	很高	主题独立, 可捕捉书写错误, 特征维度容易过大, 导致数据稀疏
词汇特征	词长, 词频, 词汇丰富度, 单词 n-gram, 词拼写错误	容易, 直接提取或分词后提取	很高	主题相关, 可捕捉书写错误
句法特征	短语或句子结构, 词性 n-gram, 句法 n-gram, 重写规则频率	较难, 深层句法特征需借助句法解析器	低	主题独立, 通常不具有连续性, 解析器容易引入噪声
语义特征	同义词, 语义依赖	困难, 需借助语义分析工具	很低	主题相关, 通常作为其他特征的补充, 很少独立使用

任务通常与文本内容相关. 具体来说, 文本分类需要根据文本内容将文本分配给一个或多个类, 因此文本分类的特征通常是文档中的单词[45]; 情感分析需要识别文本中带有意见和情感的句子, 因此情感分析的特征常常是评论性短语或单词[46]; 关系抽取是从文本中识别实体并对这些实体进行关系分类的任务, 它的特征通常是单词、字符串以及各种关系短语[47]

2 作者识别方法

一般情况下, 作者识别的过程可以分为两个步骤, 第一个步骤就是提取能够衡量特定作者写作风格的文本特征集, 第二个步骤就是建立由特征集预测作者归属的模型. 研究者通常称第一个步骤为作者风格分析 (Authorship style analysis), 第二个步骤为作者身份建模 (Authorship modeling). 有些时候, 作者身份建模也指由文本建立预测作者归属模型的过程. 图 1 展示了一般的作者识别流程: 将已知作者的文本经过特征提取器生成特征向量, 这些特征向量结合特定的作者分类算法经过训练得到作者识别模型, 该模型可以识别未知作者的文本. 将未知作者的文本也通过一个特征提取器得到特征向量, 再利用之前生成的作者识别模型分类这些向量, 即可得到作者识别结果. 在这里, 已知文本经过的特征提取器与未知文本经过的特征提取器对应同一个特征集. 该流程几乎涵盖了绝大多数作者识别研究, 可以说通过建立特征集来识别作者的研究都可以用该流程来描述. 后面会叙述一些不通过构建特征集实现作者识别的特殊方法, 这些方法不能用该流程表述.

在传统的作者识别研究中, 作者身份建模主要依靠相关专家的经验. 随着计算机技术的不断发展与进步, 研究者提出了很多建模方法. 从大的层面来分, 作者身份建模主要分为基于轮廓的建模 (Profile-based modeling) 和基于实例的建模 (Instance-based modeling). 二者都是基于训练文本构建作者

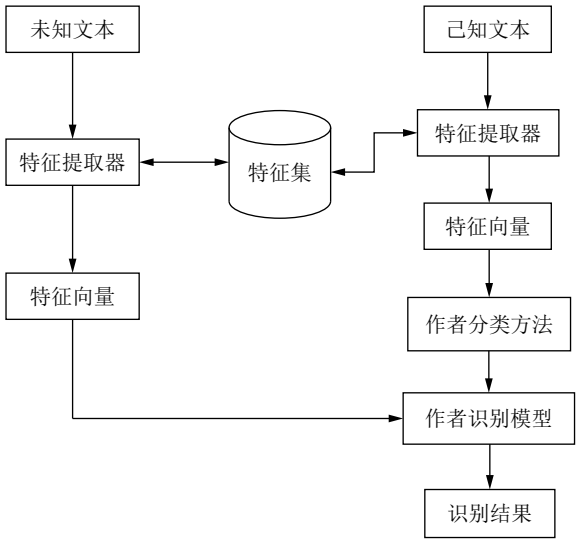

图 1 作者识别流程图

Fig. 1 Flow diagram of authorship identification

归属模型的过程, 不同的是在基于轮廓的建模中, 每位作者的所有文本会被累计处理, 即它们会在一个大文档中连接, 然后提取单个表示作为作者的轮廓; 而在基于实例的建模中, 每位作者的所有文本将单独处理, 每个文本样本都有自己的表示[14]. 通常情况下, 基于实例的建模要与机器学习算法相结合, 每个类常常需要多个实例. 因此, 当每个作者有多个文本可用或者可以将长文本拆分成多个样本时, 基于实例的建模会更有效. 另一方面, 当每个作者仅有较短或有限的文本样本时, 基于轮廓的建模会更有效[48].

除此之外, 还可以根据使用的数据是否有标注而把作者识别方法分为无监督的方法 (Unsupervised method) 和有监督的方法 (Supervised method). 其中, 无监督的方法使用的是无标注的数据, 比如聚类、主题建模等; 而有监督的方法使用的则是有标注的数据, 比如朴素贝叶斯、支持向量机、决策树、k 近邻方法、神经网络等. 有监督的方法还可以进一步细分为生成方法和判别方法. 首先学习联

合概率分布, 进而求得条件概率分布的方法是生成方法, 对应的模型是生成模型; 直接学习条件概率分布或决策函数的方法为判别方法, 对应的模型是判别模型[49]. 朴素贝叶斯属于生成方法, 而支持向量机、决策树、k 近邻方法、神经网络等属于判别方法. 本节采用这种分类方式论述作者识别方法.

2.1 无监督的方法

无监督的方法是从无标注的数据中学习统计规律或内在结构的方法, 它的基本思想是对给定数据进行某种 "压缩", 从而找到数据的潜在结构, 假定损失最小的压缩得到的结果就是最本质的结构[49]. 无监督的方法不借助先验的类别知识, 机器自己寻找数据中的规律. 与有监督的方法相比, 无监督的方法通常需要更多的训练数据才能发现其规律. 在作者识别领域, 研究者大多基于标记的文本进行研究, 因此多选用有监督的方法, 无监督的方法很少, 本部分主要介绍聚类和主题建模等方法.

2.1.1 聚类方法

聚类是根据样本的相似程度将其划分为若干子集的数据分析方法. 这些子集被称为 "类"或 "簇", 它们通常是不相交的. 与分类算法不同的是, 聚类算法不借助事先定义的类别, 而让机器自己进行划分, 使得每一类或簇中的样本相似, 而不同类或簇中的样本相异. 聚类主要包括 k 均值聚类、层次聚类、高斯混合聚类等方法. 有些研究者利用聚类来研究作者识别.

Jin 和 Jiang[50] 使用基于标点符号特征的文本聚类方法研究现代作家的身份识别问题. 该方法融合了句子节奏特征的信息, 同时具有低维的特性. 实验结果表明, Kullback-Leibler 散度优于欧氏距离和余弦距离, Ward 层次聚类优于 k 均值聚类. 基于 Kullback-Leibler 散度的 Ward 层次聚类可以达到 96 % 的准确率. Hacohen-Kerner 和 Margaliot[51] 使用最频繁词 (含功能词)、最频繁过滤词 (不含功能词) 和方差值最高的词以及 k 均值聚类和期望最大化算法研究犹太文本的作者识别. 实验结果表明, 最频繁词 (含功能词) 是效果最好的单词列表, 期望最大化算法优于 k 均值聚类, 最佳结果获得 98 % 的精度, 并且改善率超过 40 %. Fifield 等[52] 提出一种利用多个聚类组合识别文本作者的方法, 并展示了其在具有多种风格的文本上的应用. 该方法重复进行 m 次聚类, 每次都使用相对于上次偏移的片段, 在群集内重新分配标签, 以使群集尽可能一致, 把 m 个重新标记的聚类的平均值作为结果. 所提出的方法在少量作者的情况下表现出较低的一致性, 有待后续改进.

Mansoorizadeh 等[53] 选择单词 n-gram、词性标签 n-gram、句长、标点符号 n-gram 等作为特征, 组合不同的特征构成特征空间, 并将其用于文档聚类. 实验结果表明, 所提出的方法精度较低, 问题可能出在群集编号选择或特征空间上. 因此, 未来的工作可以使用更复杂的聚类方法以及更优的群集参数选择方法. Bagnall[54] 使用多头循环神经网络实现作者身份聚类, 该方法使用由多个语言模型共享的循环状态, 以相对熵的形式生成分数, 将神经网络的输出转换为聚类决策. 实验结果表明, 所提出的方法时间成本较高, 在一些困难问题上似乎表现良好, 但很难与其他方法进行比较. Agarwal 等[55] 将文档表示为对应于每个单词的嵌入向量的 TF-IDF 加权总和, 并使用层次聚类进行作者归属. 结果表明, 所提出的方法在作者聚类和作者身份链接排名任务上具有良好的性能, 超过 PAN 2017 作者聚类任务的最佳结果.

2.1.2 主题建模方法

主题建模 (Topic modeling) 是通过对语料进行分析, 学习、识别和提取文档主题的过程. 在文本信息处理领域, 传统方法是以单词向量表示文本内容, 以单词向量空间中的度量衡量文本之间的相似度; 而主题建模的基本思想是以主题向量表示文本内容, 以主题向量空间中的度量更准确地衡量文本之间的相似度[49]. 本部分主要介绍潜在语义分析 (Latent semantic analysis, LSA) 和潜在狄利克雷分配 (Latent Dirichlet allocation, LDA) 等主题建模方法以及它们在作者识别中的应用.

1) LSA

LSA 将文本集合表示为单词–文本矩阵, 通过对其进行奇异值分解, 把单词和文本映射到一个低维的语义空间, 从而实现对单词和文本更本质的表达. 有的研究者把 LSA 用于作者识别研究. Nakov[56] 使用 LSA 来研究德国文学作品, 并验证该方法能否区分作者以及自动发现散文和诗歌. 结果表明, 在一般情况下, 使用 LSA 可以区分所选的德国作者, 但对于某些作者来说似乎很难. 同时, 实验结果为自动发现散文和诗歌的假设提供了有力的支持. Satyam 等[57] 在基于字符 n-gram 的统计模型上应用 LSA, 以获得文档对之间的相似性, 并使用文档相似性的统计分析来确定阈值. 该方法运行时间很短, 整体性能与大多数其他方法相当, 在英文小说文本中达到了最好的效果, 而在西班牙文和希腊文中效果欠佳.

2) LDA

LDA 是基于贝叶斯理论的主题模型, 它假设每个文档都可以表示为潜在主题的概率分布, 并且所有文档的主题分布都具有相同的狄利克雷优先级; 同时每个潜在主题可以表示为单词的概率分布, 并且主题的单词分布也具有相同的狄利克雷优先级[58]. 有的研究者使用 LDA 研究作者识别. Seroussi 等[59] 利用 LDA 对文本和作者进行建模, 并使用基于 LDA 表示形式的文本距离对测试文本进行分类. 实验结果表明, 当训练文本足够多且存在有效作者时, 该方法的准确率超过基准方法, 而运行时间大大降低. Savoy[60] 利用 LDA 把每个文档建模为主题分布的混合, 每个主题指定单词的分布, 根据争议文本距离确定可能的作者归属. 实验结果表明, 基于 LDA 的分类方案优于基于 Delta 规则的分类方案, 同时, 基于 LDA 的方案在考虑更多术语时可以提供更好的有效性. Anwar 等[61] 使用 LDA 与 n-gram 结合的方法生成乌尔都语语料库的降维主题表示, 并使用该主题表示与改进的平方根余弦距离度量对测试文档进行分类. 结果表明, 所提出的方法具有很高的精度, 在由 6 000 个文档组成的数据集上达到了 92 % 的 $F1$ 测量值.

2.2 有监督的方法

有监督的方法是从标注的数据中学习模型预测的方法, 其中标注数据表示输入和输出的对应关系, 预测模型对给定的输入产生相应的输出, 因此从本质上来说, 有监督的方法学习的是输入到输出映射的统计规律[49]. 与无监督的方法相比, 有监督的方法可以利用先验的类别知识, 因此准确率通常较高, 这使其成为作者识别研究中的主流方法. 有监督的方法可以按照模型类型进一步细分, 比如可以分为概率模型与非概率模型、线性模型与非线性模型、参数化模型与非参数化模型、生成模型与判别模型等. 本小节把有监督的方法分为生成方法和判别方法, 并着重介绍一些作者识别中常用的方法.

2.2.1 生成方法

生成方法是先学习联合概率分布, 进而求得条件概率分布的方法, 在监督学习中, 概率模型是生成模型[49]. 本部分主要介绍朴素贝叶斯方法.

朴素贝叶斯是基于贝叶斯定理与特征条件独立假设的分类方法[49]. 具体来说, 它是在类条件概率密度和先验概率已知的情况下, 通过贝叶斯公式比较样本属于两类的后验概率, 将类别归为后验概率较大的一类, 这样可以使总体错误率最小[62]. 有些学者利用朴素贝叶斯研究作者身份识别. Zhao 和 Zobel[63] 选取 55 位作者的 634 篇文章, 采用功能词和 POS 标签作为特征, 使用朴素贝叶斯方法进行作者识别. 结果表明, 以功能词为特征的分类效果高于 POS 标签以及二者混合的结果. 同时也证实, 作者具有可识别的写作风格, 并且简单的标记就足以识别特定的作者. Boutwell[64] 使用朴素贝叶斯分类器, 利用基于字符 n-gram 的特征构建作者集统计模型识别短信的作者归属. 研究表明, 把推文或者短信息聚在一起容易提取文本特征, 更有利于作者识别. 在最差的情况下, 连接多个文本到一个文档比起单独检测准确率提高了 50 %. Altheneyan 和 Menai[65] 使用简单朴素贝叶斯、多项式朴素贝叶斯、多变量伯努利朴素贝叶斯和多变量泊松朴素贝叶斯等 4 种方法研究阿拉伯文本的作者识别. 实验结果表明, 多变量伯努利朴素贝叶斯达到了最高的准确率 97.43 %, 它与多项式朴素贝叶斯适合用来研究作者身份归属. Howedi 和 Mohd[66] 选择字符 n-gram 和单词 n-gram 作为文本特征, 使用朴素贝叶斯分类器进行阿拉伯文本的作者识别, 并与支持向量机进行对比. 实验结果表明, 朴素贝叶斯整体优于支持向量机, 基于单词 1-gram 的朴素贝叶斯达到了最高的准确率 96.67 %.

2.2.2 判别方法

判别方法是直接学习条件概率分布或决策函数的方法, 在监督学习中, 非概率模型是判别模型[49]. 本部分主要介绍支持向量机、决策树、k 近邻方法、神经网络等判别方法.

1) 支持向量机

支持向量机的基本原理是找到一个最优的分类面, 使得两类中距离这个分类面最近的点和分类面之间的距离最大[67]. Diederich 等[68] 利用支持向量机对德国报纸文本进行作者归属, 实验结果表明, 支持向量机在识别任务中始终具有良好的性能, 不需要特征选择, 并且可以处理文本所有单词的频率向量. Schwartz 等[69] 利用支持向量机研究微小信息在推特语料上的作者识别. 结果表明, 微小信息能够取得好的识别效果, 单个推文的作者可以在一系列短文本作者识别任务中被准确识别. Mikros 和 Perifanos[70] 提取多级 n-gram 轮廓, 利用多类支持向量分类算法进行分类, 并使用 10 次交叉验证和 500 条实际推文的外部数据集评估分类性能. 结果表明, 与单个 n-gram 特征组相比, 该方法获得了更好的准确性. Li 等[71] 采用支持向量机方法研究了 Facebook 的短社交网络帖子的作者身份验证问题. 测试结果显示, 样本大小、特征和用户书写风格对作者身份验证有较大影响, 带有线性内核的支持向

量机方法可以达到 79.6 % 的准确率, 超过 k 近邻方法. Martin-del-Campo-Rodriguez 等[72] 结合传统字符 n-gram 引入标点符号 n-gram 作为文档特征表示, 从不同文本特征开始对多个 SVM 进行训练, 并用所有 SVM 结果的平均值作为基准确定作者归属. 在针对开集跨域作者识别的 PAN 2019 竞赛中, 此方法获得了 0.642 的 $F1$ 分数. Soler-Company 和 Wanner[73] 使用面向表面的、句法依赖的以及包含话语结构特征的 188 个特征构建特征集, 并利用带内核的支持向量机进行作者识别. 结果表明, 句法依赖和话语特征的使用可以使总特征数量减少到小于 200 个, 而识别结果仍然能达到较高水平.

2) 决策树

决策树 (Decision tree, DT) 是机器学习中广泛研究的方法之一, 它是一种以实例为基础的逼近离散函数的归纳学习方法[74]. 决策树本质上是从训练数据集中归纳出一组分类规则, 它的模型是非参数的、无分布的, 并且对于异常值和不相关属性具有鲁棒性[75]. 有些研究者利用决策树分类方法研究文本作者识别. Frery 等[76] 采用基于文本的几种表示和优化决策树的机器学习方法进行 PAN 2014 作者身份识别任务. 该方法综合性能排名第二, 实验表明, 构建有效的属性会大大提高算法在某些语料库上的准确性. Digamberrao 和 Prasad[77] 使用序贯最小优化与基于规则的决策树相结合, 在五位作者撰写的马拉地语文章中进行作者识别, 并基于不同标准评估了该方法的性能. 结果表明, 虽然在训练集减小时精度会降低, 但该方法可以适用于英语、马拉地语、孟加拉语等多种语言. 也有的研究者利用随机森林 (Random forest, RF) 研究作者身份归属问题. 随机森林是包含多棵决策树的分类器, 它通过集成学习把若干棵决策树的输出集合起来, 综合评定产生最终输出. 因此, 随机森林在处理缺少变量的不均匀数据集时表现会非常好, 它往往比决策树具有更低的分类误差和更好的 F 分数. Maitra 等[78] 利用随机森林分类器根据基于单词和风格的特征对未知文档进行分类, 得到了较优的结果.

3) k 近邻方法

k 近邻 (k-nearest neighbor, KNN) 算法的目标是将对象分类为由机器学习创建的样本组的预定义类之一, 具体来说, 算法基于某种距离度量找出训练样本中与测试样本最接近的 k 个样本, 然后再基于这 k 个训练样本进行预测. 通常而言, 会根据 k 个样本中的大多数样本的类别来预测结果. 该算法不需要使用训练数据来执行分类, 可以在测试阶段使用训练数据[79]. 有些研究者采用 k 近邻方法进行作者识别研究. Halvani 等[80] 利用基于 k 近邻的方法研究 PAN 2013 作者识别任务. 该方法利用 k 近邻分类器计算真实作者的训练文档与未知文档之间的风格偏差分数, 根据分数以及给定的阈值确定作者归属. 该方法具有语言独立、运行时间短、易于扩展和修改等优点, 在 PAN 2013 作者识别任务上得到了 80 % 的总体准确率, 在个人数据集上的准确率是 77.50 %. Anwar 等[81] 利用 LDA 模型在文本 n-gram 上生成文档的主题表示, 然后使用余弦相似度和 KNN 分类器进行分类. 在不使用任何标签的情况下, 即可在英语和乌尔都语新闻语料中获得令人满意的结果. Sarwar 等[82] 基于词汇、句法和结构等特征, 使用概率 k 近邻分类器研究泰语文档的作者识别. 实验结果表明, 将所有特征类别组合在一起可以提高作者识别过程的准确率.

4) 神经网络

神经网络 (Neural networks, NN) 是简单处理元件、单元或节点的互连系统, 其网络的处理能力体现在通过适应或学习一组训练模式的过程中获得的单元间连接强度或权重上[83]. 针对一些实际情况复杂、背景知识不清楚、规则不明确的问题, 神经网络算法具有很强的处理能力. 有些学者利用神经网络方法研究文本作者识别. Bagnall[84] 使用循环神经网络同时对几个作者的语言进行建模, 每个作者的文本由依赖于共享循环状态的单独输出表示. 实验结果表明, 循环神经网络可以成为作者身份识别中的有用工具. 该方法更多地基于信息理论而不是传统的聚类, 并且能够避免特征选择和过拟合的泥潭. Ruder 等[85] 利用卷积神经网络进行大规模作者身份归属, 以处理特征级别信号并进行快速预测. 该方法结合了字符和单词通道, 利用了文本风格和主题信息, 获得了较优的结果. Qian 等[86] 使用门控循环单元、长短期记忆网络和孪生网络等三种深度学习模型识别作者身份, 并使用孪生网络验证作者身份. 结果表明, 文档级别的门控循环单元在作者身份识别方面表现最好, 孪生网络在作者身份验证上达到很高的准确率. Shrestha 等[87] 使用基于字符 n-gram 的卷积神经网络对推文进行作者识别, 并通过估计输入文本片段在预测分类中的重要性来提高模型的可解释性. 实验结果表明, 卷积神经网络在推文的作者识别方面具有很好的性能, 使用字符 n-gram 而不仅仅是字符序列也可以提高作者识别的性能. Jafariakinabad 等[88] 引入句法循环神经网络来编码层次结构中文档的句法模式. 该模型首先从词性标签序列中学习句子的句法表示. 随后, 使用循环神经网络将句子的句法表示聚合成文档表示.

实验结果表明, 句法循环神经网络在精度方面优于具有相同架构的词汇模型.

2.3 对比分析

识别方法在自然语言处理任务中具有通用性. 换句话说, 本节所述的方法可以应用到文本分类、情感分析、关系抽取等其他自然语言处理任务中. 本小节对比分析无监督的方法和有监督的方法. 其中, 表 2 给出了无监督方法之间的对比, 表 3 给出了有监督方法之间的对比[49].

3 多层面研究

从 19 世纪后期研究者开始使用统计等数学工具研究作者识别以来, 作者识别已经经过一百多年的发展. 在漫长的发展过程中, 作者识别研究呈现出两个明显的趋势: 文本特征丰富化以及方法和思想多元化. 文体特征从最初的一元单一特征逐渐发展为多元混合特征, 分类方法也从简单数学公式的应用而逐渐发展出复杂的神经网络. 研究者越来越倾向于多特征组合的研究方式, 比如 Khomytska 和 Teslyuk[89] 使用不同音素特征搭配统计模型进行作者识别, Grabchak 等[90] 提出基于广义 Simpson 索引的轮廓来判断两个样本是否由同一作者所撰写. Srinivasan 和 Nalini[91] 选择句法、结构和 n-gram 作为特征, 使用 4 种不同的机器学习分类器研究亚马逊评论的作者识别. 针对目前作者识别领域的发展状况, 下面主要从数据规模、跨域研究、特殊方法等三个方面阐述作者识别的多层面研究.

3.1 数据规模

作者识别的研究结果常常受到数据集规模或作者数量的影响. 研究者提出的方法在小数据集以及少数几个作者的情况下表现良好, 而一旦扩大数据集规模或者增加作者数量, 其执行结果往往不确定. 换句话说, 利用小数据集以及少数作者验证方法的研究者可能高估了其方法的准确性, 甚至高估了他们所选择的文本特征的重要程度[92]. 鉴于此, 一些研究者专门研究数据集大小以及作者数量对作者识别实验结果的影响.

Luyckx 和 Daelemans[93] 在一个有 145 位作者的语料库上针对特征选择进行研究, 实验结果表明, 当增加作者数量时, 系统性能显著下降. 功能词和句法特征的组合可以使系统性能显著提高, 部分作者高估了他们方法的准确率以及所选特征的重要性. Eder[94] 使用基于 k 近邻的 Delta 方法研究文本尺寸对作者归属的影响, 以希望找到可以用于作者归属的文本样本的最小尺寸. 实验结果表明, 对于现代英语, 最小稳定样本为 5000 个单词, 使用 2500 词的样本几乎不能提供可靠的作者识别结果. Koppel 等[95] 使用训练文本的各种子集进行实验, 以研究大数据集和大作者集上的作者归属问题, 同时确定样本尺寸对候选作者数量、每个候选作者的已知文本量以及未知文本长度的影响. 结果表明, 基于相似性的方法以及多个随机特征集可以在大数据集和大作者集上实现较高的精度. Luyckx 和 Daelemans[96] 系统地研究了作者集规模和数据集规模对作者识别性能和特征选择的影响. 实验结果表明,

表 2 无监督方法对比表

Table 2 Comparative table of unsupervised method

方法	模型	策略	算法
k 均值聚类	k 中心聚类	样本与类中心距离最小	迭代算法
层次聚类	聚类树	类内样本距离最小	启发式算法
高斯混合聚类	高斯混合模型	似然函数最大	期望最大化算法
LSA	矩阵分解模型	平方损失最小	奇异值分解
LDA	LDA 模型	后验概率估计	吉布斯抽样, 变分推理

表 3 有监督方法对比表

Table 3 Comparative table of supervised method

方法	模型类型	模型特点	学习策略	稳定性	准确率
NB	生成模型	特征与类别的联合概率分布, 条件独立假设	极大似然估计, 最大后验概率估计	高	低
SVM	判别模型	分离超平面, 核技巧	极小化正则化合页损失, 软间隔最大化	中	高
DT	判别模型	分类树、回归树	正则化的极大似然估计	中	中
KNN	判别模型	特征空间, 样本点	无	低	中
NN	判别模型	神经元拓扑结构	目标函数最小化	中	偏高

在小数据集上实现 95 % 准确率的方法无法在大数据集上达到相同或者类似的性能, 并且随着作者数量的增加, 方法的准确率降低到不具有实际意义的程度. 在大多数情况下, 字符 n-gram 的识别结果要优于其他文本特征.

3.2 跨域研究

作者识别研究常常关注特定作者在无意识的情况下表现出的写作风格, 这种风格往往与文章的内容无关. 然而, 一个不可否认的事实是, 文章的类型、主题甚至所用的语言会在更高维度上影响作者的表达方式. 换句话说, 同一位作者在不同类型或者不同主题的文本中可能表现出不同的行文风格. 因此, 一些研究者在跨主题作者识别方面进行研究, 希望发现更一般的规律. Stamatatos[97] 研究字符 n-gram 在跨类型和跨主题条件下的作者识别, 并与基于单词的方法进行比较. 结果表明, 当训练和测试语料库之间存在显著差异时, 字符 n-gram 能够更好地捕获文本的风格属性. Markov 等[98] 提出一个改进的跨主题作者归属算法, 以研究字符 n-gram 在跨主题作者归属中的性能. 结果表明, 通过执行简单的预处理步骤和适当调整特征数量, 可以显著提高字符 n-gram 在跨主题条件下的性能. 高频阈值能够有效排除与主题特定信息相关联的最不频繁的 n-gram, 进而提高准确率. Rahgouy 等[99] 基于文档不同表示形式的模型组合研究跨领域的作者识别. 该方法使用文档的 TF-IDF、Word2Vec 和 n-gram 表示来训练三种类型的模型并使用整体进行预测. 文中还使用临时网格搜索对模型和集合参数进行调整, 以达到最优效果. 实验结果表明, 该方法非常有能力区分不同作者.

以上这几篇文章是 n-gram 特征与文本主题相结合的研究. 可以看出, 在跨主题的研究中, n-gram 特征仍然能够充分捕捉文本特征, 进而获得较好的识别结果. 也有研究者选择词汇或者多种混合特征研究跨主题的作者识别. Mikros 和 Argiri[100]创建了由两位作者在两个不同主题中撰写的 200 篇现代希腊新闻专线文章组成的特殊语料库, 研究了作者身份归属中一些广泛使用的风格变量的主题中性特征, 以探讨文本主题对作者归属的影响. 实验结果表明, 大多数变量与文本主题具有很大的相关性, 在作者分析中应该谨慎使用. Sari 等[101] 对 4 个数据集进行分析, 以探讨不同类型的特征如何通过影响主题或风格影响作者归属的准确性. 随后他们将分析得出的结论应用到作者识别方法上, 在 4 个数据集中的两个数据集上, 得到了更好的结果. 有些研究者会借助主题模型进行研究. Seroussi 等[102] 对比分析了 SVM、LDA、作者感知主题模型以及不连贯的作者文档主题模型等 4 种作者识别模型, 发现作者感知主题模型胜过 LDA, 而该文提出的不连贯的作者文档主题模型胜过以上 3 种方法. Seroussi 等又在另一篇文章[103] 中进一步发展并完善了该方法. Yang 等[104] 提出了主题漂移模型, 用来描述个人作者的兴趣和写作风格的变化. 与之前的作者归属方法不同的是, 该模型对时间信息和单词顺序敏感, 因而能从文本中获取更多的信息. 实验结果表明, 与其他模型相比, 该方法获得了更高的准确率.

一些学者研究跨语言下的作者识别. Halvani 等[105] 提出一种作者验证方法, 该方法为每种语言提供一个通用阈值, 用于接受或拒绝所谓的文档作者身份. 在荷兰语、英语、希腊语、西班牙语和德语等 5 种语言 16 种类型和混合主题上的 28 个语料库上的实验获得了接近 75 % 的中位数准确率. 由于该方法不涉及自然语言处理技术以及机器学习库, 它可以灵活地扩展到新语言或者新类型上. Bacciu 等[106] 利用基于字符、单词、词干和失真文本的 n-gram 作为文本特征, 并使用组合的单分类器对不同语种的文档进行识别. 实验结果表明, 所提出的方法在几乎所有问题中都优于基线模型. 使用此模型, 在 PAN 2019 作者识别竞赛中获得了 0.68 的 $F1$ 分数.

也有研究者利用文本失真掩盖主题相关信息的方法进行作者识别. Stamatatos[107] 提出一种基于文本失真来压缩主题相关信息的方法. 该方法将输入文本转换为适当的形式, 并保持与作者个人风格相关的文本结构, 同时掩盖与主题信息相对应的最不频繁的单词的出现. 实验结果表明, 与其他作者身份归属方法相结合时, 该方法可以显著提高其在闭集归属和作者身份验证中跨主题条件下的效果. 而在另一篇文章中, Stamatatos[108] 再次利用基于文本失真的方法掩盖与主题相关的信息, 通过将输入文本转换为更加主题中立的形式, 尽量多地保持与作者个人风格相关联的文本结构. 使用包含细粒度主题和类型的受控语料库将文本失真方法用于跨领域的作者识别任务上. 实验结果表明, 在跨主题的作者身份归属中, 该方法显著提高了作者识别的性能; 而在跨类型的作者身份归属中, 该方法仅增强了一类方法的性能. 在以上二者结合的作者身份归属中, 结果与跨类型实验相近, 这表明类型是作者身份归属中比主题更重要的因素.

3.3 特殊方法

作者识别的交叉特点使得其他领域的思想和方法能够应用到该领域的研究中来, 从而产生一些比较特殊的研究方法. 一般的作者识别研究会选择某个或者某些文本特征来刻画作者的行文风格, 本部

分介绍两种特殊的方法——压缩方法和频率混沌游戏表示方法, 它们不借助文本特征就能实现作者识别.

1) 压缩方法

压缩方法是作者识别领域中的一种比较特殊的方法. 一般情况下, 作者识别研究需要根据文本特征确定作者归属. 而压缩方法避免了定义特征, 甚至在有些情况下仅依赖压缩算法、相异度度量和阈值就能完成整个识别过程. 利用压缩方法进行作者识别的一般步骤为: 使用压缩算法构建处理文档的模型或字典, 经过多次压缩产生较高的压缩率, 利用压缩率衡量训练文本和新文档之间的交叉熵, 新文档则被分配给训练文本中使交叉熵最小的类[109]. 压缩方法的思想可以简述为, 如果在一个文档中可以显著地压缩另一个文档, 那么这两个文档被认为是接近的. 换句话说, 如果两个文档很相似, 则可以用一个文档来更简洁地描述另一个文档. 压缩方法不使用关于数据的任何特征或背景知识, 因而其无参数、简单易用, 可以避免由于人为选择特征而引入的噪声以及信息丢失等问题[110].

一些学者利用压缩方法来研究文本的作者识别. Cerra 等[111] 使用快速压缩距离 (Fast compression distance, FCD) 研究基于压缩的相似性度量在文本作者分析方面的表现. FCD 能够捕获字典中单词的重复组合, 描述文本规则, 以及比较任何两个文档之间的共享信息. 实验结果表明, 该方法具有普适性, 可以在英语、意大利语、希腊语、西班牙语和德语文档中直接使用. 相对于传统的基于压缩的方法, FCD 计算复杂性低, 而准确率更高. Halvani 等[112] 提出一种基于压缩模型的简单且高效的作者身份验证方法. 实验结果表明, 部分匹配预测胜过所有其他测试压缩机, 基于压缩的余弦测量产生了最高的结果, 在针对所有训练语料库测试的 5 个压缩机中表现稳定. 基于压缩模型的方法超过基于支持向量机或神经网络的许多方法, 并且可以很容易地应用到其他语言上.

2) 频率混沌游戏表示方法

混沌游戏表示是一种从核苷酸序列创建图像的方法, 它被用来从大量文本文档中制作图像. Lichtblau 等用经过特殊处理过的图像特征代替传统的文本风格特征进行作者识别. 从结果上来看, 这种特征可以成为作者识别领域新的分类特征, 为后续作者识别研究提供了一种新思路. 具体来说, Lichtblau 和 Stoean[113] 使用频率混沌游戏表示从文本产生灰度图像, 然后用图像训练机器学习分类器, 利用所学的模型识别这些灰度图像, 以区分不同文本的作者. 实验结果表明, 该方法在英语和葡萄牙语语料库上获得了令人信服的结果. 联邦主义文档和葡萄牙语数据集上的验证结果与文献中的最佳结果相当. 此外, 该方法对少于 1 000 字符的文本也有较好的识别结果, 因此可以用于识别匿名电子邮件或博客文本的作者. 而在另一篇文章中, Lichtblau 和 Stoean[114] 再次利用混沌游戏表示将文本转换为图像, 再将图像压成向量, 通过奇异值分解进一步减小尺寸. 再用神经网络学习与每个作者相关的特征, 并建立模型对样本进行分类. 实验结果表明, 在 3 个基准数据集上, 所提出的方法明显优于频率混沌游戏表示的线性回归方法. 与其他成熟的作者识别方法相比, 该方法可以获得更好或者相似的结果.

4 相关评测介绍

随着互联网的发展, 网络文本大量增加, 作者识别的研究重点逐渐从传统文学作品转向人们接触更多的网络文本. 评测是采用统一数据集和评价标准进行测试和评价的活动. 由于数据集和评价标准都是统一的, 其结果对衡量算法的真实性能具有很强的说服力. 最近几年, 越来越多的研究者开始参与到网络评测中, 进而产生了一些持续多年的、受到学者们广泛关注的评测. 这些评测因其高质量的数据、评价和算法总结而在相关领域的影响较大. 本节主要介绍作者身份验证、作者身份概述、作者身份混淆等与作者识别相关的评测, 以期望为作者识别的研究带来新的方法和思想.

4.1 作者身份验证评测

作者身份验证又简称作者验证, 是数字文本取证研究的一个分支, 旨在确定两个文档是否由同一作者撰写. 评估论坛实验室大会 (Conference and Labs of the Evaluation Forum, CLEF) 在 2013~2015、2020 年组织过作者身份验证评测, 本小节主要关注 2020 年的评测. 在 2020 年的评测中, 有 10 个团队提交了 13 个系统, 下面介绍其中性能最优的几个.

Boenninghoff 等[115] 提出一种将神经特征提取与统计建模相结合的方法, 该方法采用具有孪生网络结构的深度学习框架生成特征, 然后在概率线性判别分析层执行贝叶斯因子评分, 以衡量两个文档之间的相似性. 评测结果表明, 所提出的方法在小型数据集和大型数据集上均取得了优异的总体性能评分. Halvani 等[116] 选择标点符号、功能词、缩写词、过渡短语等与主题无关的项作为文本特征, 并使用基于曼哈顿度量的距离函数以及基于相等错误率的

阈值处理程序作为分类器. 结果表明, 该方法具有出色的性能, 在所有提交的方法中排名第三. Kipnis[117] 提出一种无监督的分类方法, 该方法利用两个文档之间的单词二项式分配模型逐个计算单词的 p 值, 并使用较高的批评度将它们组合为一个分数统计. 通过评估与文档对相关的高级批评的经验分布, 将产生的分数转换为相似性得分. 该方法比较简单, 在跨域作者身份验证中取得了有竞争力的结果.

4.2 作者身份概述评测

作者身份概述又简称作者概述, 是通过对文本的分析来找出其作者尽可能多的个人信息的任务, 包括但不限于年龄、性别、母语、教育水平、社会地位等, 它在取证、市场营销和网络安全方面有着广泛的应用[118]. 与作者识别类似, 作者概述也需要对作者风格进行分析. 因此, 作者识别上的风格分析方法可以应用到作者概述上. 最近几年, 作者概述发展非常迅速, 这得益于作者分析方法的大量提出以及各种网络评测的开展, 特别是 CLEF 和信息检索评估论坛 (Forum for Information Retrieval Evaluation, FIRE) 组织的评测. CLEF 在 2013 ~ 2020 年连续 8 年组织了作者概述评测, 而 FIRE 则在 2018 ~ 2019 年组织了相关的评测. 由于相关研究众多, 无法一一列举, 本小节只介绍最近 4 年评测中排名相对靠前的作者概述方法.

CLEF 在 2017 年组织了确定推特作者的性别和语言种类的评测, 有 3 种方法获得了整体最佳结果, 它们之间没有显著的差异[119], 分别是 Basile 等使用字符和 TF-IDF n-gram 组合训练支持向量机, Martinc 等[120] 使用字符、单词和词性 n-gram 组合训练逻辑回归分类器以及 Tellez 等[121] 使用表情符号、情感、字符流和每个变体的单词列表训练支持向量机来完成任务. CLEF 在 2018 年组织了根据推特的文本或图像确定作者性别的评测. 3 个最佳结果分别来自: Takahashi 等[122] 利用单词嵌入和循环神经网络识别文本, 同时利用基于 ImageNet 的卷积神经网络识别图像; Daneshvar 和 Inkpen[123] 基于单词和字符 n-gram 组合训练支持向量机; Tellez 等[124] 使用不同类型的 n-gram 训练支持向量机, 同时结合使用 DAISY 特征描述符的视觉词袋模型进行分类. 总体而言, 传统方法仍然保持竞争力, 而一些基于深度学习的新方法正在获得优势[125]. CLEF 在 2019 年组织了判断给定推特的作者是人还是机器 (如果是人的话, 确定其性别) 以及根据名人的推特, 确定其主人的年龄、名望、性别和职业的评测. 在第一个任务中, 性能最高的 4 个团队均使用了单词和字符 n-gram 与支持向量机的组合. 评测结果表明, 传统方法比深度学习方法获得了更高的准确率, 深度学习方法首次出现在排名中, 具体来说是卷积神经网络, 排在第 11 位[126]. 至于第二个任务, CLEF 在 2020 年也组织过. 这两年一共收到研究者提交的 10 种方法, 其中 2019 年的最佳方法和 2020 年的最佳方法分别是: Radivchev 等[127] 选择单词 2-gram 作为特征, 用 TF-IDF 进行向量化, 然后使用逻辑回归和支持向量机进行分类; Hodge 和 Price[128] 选择 POS 标签、停用词数、命名实体类型等特征并使用逻辑回归、随机森林和支持向量机进行分类.

FIRE 在 2018 年组织了识别乌尔都语和英语文本作者的性别和年龄的评测, 2019 年组织了确定阿拉伯语推特用户的年龄、性别和语言种类以及两种不同类型的阿拉伯语欺骗检测的评测. 这里主要关注阿拉伯语推特的作者概述. Siagian 和 Aritsugi[129] 选择单词 n-gram、字符 n-gram、二者的组合以及功能词作为特征, 并使用支持向量机进行分类. 该方法性能优越, 在性别、年龄和语言类别等三个方面的综合排名中位列第一. Nayel[130] 利用基于 n-gram 的词袋模型提取文本特征, 并使用线性分类器、支持向量机和多层感知器进行分类. 结果表明, 在绝大多数情况下, 线性分类器获得了最高的准确率. 这说明与作者身份相比, 其性别、年龄和语言类别等信息通常隐藏在更低维度的文本特征中. Sharmila 等[131] 分别使用单词和字符的 n-gram 以及单词嵌入, 经过 TF-IDF 加权, 再使用支持向量机和 fastText 进行分类. 该方法具有较好的性能, 在欺骗检测中排名第二, 而在作者概述中排名第三. 与传统模型相比, 加权嵌入获得的准确性较低, 其原因可能是给定数据集中的某些单词在预训练模型中不存在.

4.3 作者身份混淆评测

作者身份混淆 (Authorship obfuscation) 又简称作者混淆, 是作者识别的对抗性任务, 其目的是使基于作者写作风格的身份识别变得不可能或至少难以进行[132]. 由于作者识别和作者混淆互为对抗任务, 因此对于其中一项任务而言, 某种方法的成功与否取决于其对另一项任务中最有效方法的 “免疫力”[133]. 与作者识别相比, 作者混淆很少受到研究者的关注, 造成这种情况的原因很可能是作者混淆需要释义作为子任务, 从而给研究者进入该领域带来很大的障碍[133]. 从评价标准上来看, 作者混淆比作

者识别更复杂. 作者识别一般通过准确率等指标评估算法的优劣, 而作者混淆除了要评估安全性以外, 还需要对合理性和完整性进行评估, 甚至很多时候需要人工进行审核. 2016 ~ 2018 年, CLEF 连续组织了作者混淆评测任务, 产生了 7 种混淆方法, 促进了该领域的研究进展. 本部分主要介绍一些比较成功的方法, 以期望能为作者识别提供一些可以借鉴的思路.

Mihaylova 等[134] 对文本中可以表明作者身份的不同特征 (句长、标点、停用词、词性等) 进行评估, 然后使用多种基于规则和随机的文本操作, 将目标文本的这些特征的度量调整到平均水平, 同时保持文本的含义和完整性. 此外, 他们还尝试将随机噪声添加到文本中. 使用该方法的混淆器获得了当年的最佳性能, 在 2016 ~ 2018 提交的 7 种混淆器中排名第二, 与第一名评分很接近. 该方法的成功说明基于简单特征捕获作者写作风格的识别方法容易被混淆器击败, 要想对抗这种混淆方法, 必须考虑不容易改变的、更深层次的文本特征. Mansoorizadeh 等[135] 从 WordNet 上获得同义词来替换原始文本中的 200 个最常见单词. 他们通过词义和语义两个方面来衡量原始词和被替换词的相似度, 以便选择最佳同义词, 每个句子最多替换一个同义词. 从整体上来说, 该方法专注于改变文档的词频特征, 较少的改动能够保证很高的文本质量, 同时可以使基于词汇特征的作者识别方法失效. Keswani 等[136] 采用基于往返翻译的方法, 将英语译为中间语言, 再将中间语言译回英语, 以此来实现作者混淆. 在翻译的过程中, 由于翻译模型的差异以及翻译过程中的各种惩罚, 词汇、平均句长以及语言结构都会发生变化. 虽然该方法在评测中表现一般, 但是仍然具有相当的潜力, 结合成熟的商业引擎能够保证在较高文本质量的前提下达到混淆作者的目的.

Castro-Castro 等[137] 提出一种在无监督的情况下执行句子转换的方法, 同时使用字典和语义资源以及句法简化规则进行句法和语义更改. 具体操作是根据字典或更长的版本替换缩略语, 使用 FreeLing 替换同义词, 并通过省略括号中的部分、语篇标记以及并列成分来缩短句子. 该方法获得了很高的混淆性能, 在 CLEF 连续组织的评测中排名第一. Kocher 和 Savoy[138] 的方法基于 20 条规则, 这些规则将紧缩词与普通单词互换、替换了一些形容词和连词, 还通过重复拼写来引入错误. 总体来说, 该方法基于搜索和替换, 通过改变频率特征以欺骗识别器, 在保证原始文本质量的前提下, 可以达到一定的混淆度. Rahgouy 等[139] 从作者已知的文档中学习作者指纹, 然后利用相关统计信息有针对性地对原始文本进行定向转换和变形. 该方法主要改变句子长度、紧缩词和一般单词的使用, 并根据与原始单词的相似性、单词出现的可能性以及句子变体的单词移动距离对可能的替换术语进行评分. 对混淆文本进行的自动和手动评估显示了该方法的有效性.

5 数据集和评价指标

5.1 数据集

语料, 即语言材料, 是若干语言样本的统称. 在计算语言学中, 语料通常指研究者搜集的大规模语言实例. 语料经过进一步集合和加工可以形成语料库, 换句话说, 语料库是大量经过整理的、具有既定格式和标准的语料集合. 国内的作者识别起步较晚, 研究者数量少, 研究相对落后. 目前, 在作者识别领域, 尚无公开的汉语数据集. 国外的作者识别研究起步较早, 最近几年发展较快, 有一些公开的数据集. 下面简单介绍这些公开的数据集.

1) IMDb62 数据集[1]

包含互联网电影资料库中 62 位超级用户的 62 000 条电影评论和 17 550 个留言板帖子, 其中每个用户撰写了 1 000 条电影评论以及不同数量的留言板帖子.

2) 博客数据集[2]

包含 19 320 位作者的 681 288 篇博客文章, 总共超过 1.4 亿个单词, 平均每人 35 篇文章和 7 250 个单词.

3) 判决数据集[3]

包含 Dixon、McTiernan 和 Rich 等三名澳大利亚高等法院法官的判决, 其中有来自 Dixon 的 902 个文档, 来自 McTiernan 的 253 个文档和来自 Rich 的 187 个文档.

4) 电子邮件数据集[4]

包含大约 150 个用户的 50 万封电子邮件, 其中大多数用户都是 Enron 的高级管理人员.

5) CCAT10 数据集[5]

路透社语料库第 1 卷的子集, 包含 10 位作者的新闻专线报道, 其中每位作者有 100 篇文档, 总

[1] https://umlt.infotech.monash.edu/?page_id=266
[2] http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
[3] https://umlt.infotech.monash.edu/?page_id=152
[4] https://www.cs.cmu.edu/~./enron/
[5] https://drive.google.com/drive/folders/1hlIWVSt0dfy8fz8d4wRzZItlLCo5BH1?usp=sharing

共 1000 篇文档.

6) CCAT50 数据集[6]

路透社语料库第 1 卷的子集, 包含 50 位作者的新闻专线报道, 其中每位作者有 100 篇文档, 总共 5000 篇文档.

7) PAN 数据集[7]

包含各种有关数字文本取证和文体学评测任务的数据集.

5.2 评价指标

评价指标 (Evaluation index) 是衡量作者识别分类器或作者识别模型性能优劣的评价标准. 评价指标在自然语言处理任务中具有通用性. 换句话说, 用于作者识别的评价指标也可以用于文本分类、情感分析等其他自然语言处理任务. 评价指标分为很多种, 比如正确率 (Accuracy)、查全率 (Recall)、查准率 (Precision)、F 测量值 (F-measure)、宏平均 (Macro-average)、微平均 (Micro-average)、受试者工作特征 (Receiver operating characteristic, ROC) 曲线下的面积 (Area under ROC curve, AUC) 等. 下面逐一来介绍这些评价指标.

正确率, 也称为准确率或者精度, 是最常用的评价指标, 它被定义为测试集中正确预测的样本数量占整个测试集的百分比. 正确率 A 的公式表示为:

$$A = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

其中 TP、FP、TN 和 FN 分别代表真正类 (True positive)、假正类 (False positive)、真负类 (True negative) 和假负类 (False negative). 为了进一步细化分类器在某个特定类别上的分类性能, 查全率、查准率等评价指标被应用于作者识别研究中. 一般而言, 查全率 (又称为召回率) 被定义为某一特定类别中预测正确的样本数量占该类别样本数量的百分比, 而查准率则被定义为某一特定类别中预测正确的样本数量占实际预测为该类别样本数量的百分比. 查全率 R 和查准率 P 的公式表示为:

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

查全率和查准率是从不同的角度衡量分类器性能的, 为了综合二者的整体效果, 学者提出 $F1$ 测量值. $F1$ 测量值被定义为查全率和查准率的调和平均值, 用公式表示为:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

查全率、查准率和 $F1$ 测量值在正确率的基础上进一步细化了评价标准. 然而, 这 3 个评价指标只能针对特定的类别. 为了克服这一缺点, 学者们提出宏平均和微平均指标. 这两个指标可以给出平均意义下的查全率、查准率或者 $F1$ 测量值, 能够反映分类器对不同类别的整体分辨能力. 宏平均和微平均的差别在于, 宏平均先分别计算不同混淆矩阵的查全率和查准率, 然后通过取平均的方式得到宏查全率和宏查准率, 再根据宏查全率和宏查准率计算出宏 $F1$ 测量值; 而微平均先针对所有混淆矩阵求平均, 然后再计算微查全率、微查准率和微 $F1$ 测量值.

与查全率、查准率和 $F1$ 测量值相比, 宏平均和微平均在综合不同类别预测结果的基础上给出分类器的整体性能评价. 由于其计算比较复杂, 因此在作者识别研究中应用得并不多. 作者识别研究需要根据语料的特点选择相应的分类算法和评价指标. 通常情况下, 研究者所选择的语料, 其每个类别的样本数量相等, 或者即便不相等也差别不大. 此时, 选择宏平均或者微平均得到的结果差异并不大. 但是, 有些时候研究者需要利用一些不同类别样本数量差距较大的语料进行作者识别研究. 此时, 选择宏平均还是微平均得到的结果差异较大, 用它们就不容易反映分类器的整体性能了. 为了克服由于样本数量差异带来的影响, 研究者提出 AUC 指标. ROC 曲线是以假正类率为横坐标, 真正类率为纵坐标的曲线, 它反映了不同阈值对分类器泛化性能的影响[67]. ROC 曲线下的面积即为 AUC, 该指标同时考虑了分类器对正类和负类的分类能力, 因此在样本不平衡的情况下, 仍然能够对分类器的性能进行合理的评价.

6 存在的问题

计算机的出现和广泛应用使得作者识别在最近几十年中快速发展. 到目前为止, 作者识别已经发展成为一个涉及众多学科的交叉学科. 多学科交叉使得作者识别研究能够借鉴其他学科优秀的方法和思想. 与其他自然语言处理任务相比, 作者识别缺少一些应用场景, 从事作者识别的研究者数量相对较少, 相应的研究多基于理论探究. 目前的作者识别主要面临一个宽领域、缺乏应用、小众研究的局面. 在这个大背景下, 作者识别研究主要存在以下几个方面的问题.

1) 数据集的差异使得不同研究之间很难横向

[6] https://archive.ics.uci.edu/ml/datasets/Reuter_50_50

[7] https://pan.webis.de

比较. 在计算语言学领域的研究中, 基准数据集和评价指标是评估和分析算法性能的关键. 然而, 在作者识别领域缺乏基准数据集. 除了一些竞赛会采用统一的数据集外, 其他研究很少基于统一数据集, 多数研究者会选择自己感兴趣的数据集完成研究. 数据集的差异会导致很多问题, 常见的是其他的研究者无法重复论文的工作, 更无法在原有的基础上进行改进. 很多研究者都强调自己的方法更先进, 而由于无法排除数据集差异所带来的影响, 算法的实际改进效果无法确定.

2) 实验结果通常受很多因素的影响, 而多数文章未对这些因素进行详细叙述. 作者识别领域的实验会同时受到多种因素的影响, 比如语料的选择、预处理、特征提取、分类算法的选择及参数设置等. 目前该领域的一种常见情况是, 研究者对实验设计的描述不够清晰. 有的是对新提出的分类算法描述不清, 更多的则是缺乏分类算法之外的实验细节. 这样在不公布代码的情况下, 其他研究者很难了解具体的实验方案. 算法描述不清晰或者其他实验细节的缺乏会导致已有的工作很难被复现或评价.

3) 目前的大多数研究都侧重于对结果进行定量评估, 而缺乏对文本特征的进一步分析. 文体风格是一个很复杂的组合, 理论上可能有数千个特征组成. 研究特定的作者识别问题意味着只能选择有限数量的文本特征. 对文本特征的分析有利于研究者从庞大的特征组合中选出最有效的特征, 进而提高作者识别的正确率. 反之, 则不利于特征的筛选, 正确率的提升可能仅依赖算法的改进. 从另一个角度来讲, 文本特征直接和可解释性相关, 而可解释性又和法医学、文体学、心理学上的一些应用相关. 只进行结果评估而不详细讨论用于识别作者的文本特征, 既不利于作者识别研究的改进, 也不利于相关应用的发展.

7 未来发展趋势

作者识别研究经历了由 “文体学知识”到 “规则和统计”再到 “机器学习”的发展过程, 其主要的推动力来源于计算机技术的发展. 目前, 借助计算机强大的算力, 研究者可以处理大规模文本, 作者识别进入快速发展的时期. 从现有状况来看, 作者识别研究主要有以下几个可能的发展趋势.

1) 作者识别研究体系的建立和完善. 作者识别研究虽然已经取得了很多成果, 但从整体上来看, 该领域内的研究比较分散、缺乏对比、尚未形成体系. 主要表现在以下两个方面: 一是该领域缺乏基准数据集, 数据集的差异使得不同研究之间很难横向比较; 二是该领域的很多学者不断尝试提出新方法, 很少有人去检验或者规范旧方法, 而这个是建立完整学科体系所必不可少的工作. 因此, 未来首要的工作就是建立并推广使用基准数据集, 进一步完善评测标准, 使得同类型的研究能够放在一起进行比较. 然后再逐渐细化研究分支, 检验并规范已有方法, 通过公布成熟算法框架等方式使得该领域的研究进一步规范化和体系化.

2) 开发针对网络文本和大数据的作者识别模型. 随着互联网的不断发展和计算机的广泛应用, 数据量呈现爆炸式增长, 海量网络文本给作者识别研究带来一系列新的挑战. 与传统的文学作品相比, 网络文本通常具有创作周期短、文本短小、内容随意性强等特点. 这些特点意味着作者在创作文本时往往注重读写效率, 而忽略语句的准确性甚至语法规则. 因此, 网络文本的作者写作风格更难把握, 研究者们必须针对网络文本的具体特点寻找新的文本特征.

除此之外, 文本和潜在作者数量巨大也是需要解决的另一个难题. 由于互联网人数众多, 未知文本所面临的潜在作者集合巨大, 这给作者识别带来很大难度. 现有的作者识别方法大多适用于较小规模的数据集和少数几个候选作者的情况. 如果增大数据集规模或者潜在作者数量, 这些方法的准确率会大幅度下降. 因此, 研究者亟待开发新的作者识别技术, 以应对文本集合或作者集合过大等问题.

3) 对文本风格进行更深入的分析, 拓展跨学科应用. 现阶段研究者主要依靠机器学习提升模型的性能, 而忽视针对文本风格的进一步分析, 这一点在上一节也提到过. 计算机的发展加速了不同学科之间的交叉融合, 很多学科都尝试利用计算机技术改进本学科的研究模式. 在这一大背景下, 作者识别研究实际上承担着连接计算机科学与文体学、认知心理学等学科的桥梁作用. 因此, 对文本风格进行更深入的分析, 或者说对可解释性进一步探究, 有助于发展一些跨学科应用, 同时也会为相关领域提供很好的方法和思路. 由于可解释性问题一直都是作者识别中的一个难题, 因此该方向会在多学科交叉融合的基础上面临更多的挑战.

References

1 Qi Rui-Hua. *Text Authorship Identification*. Beijing: Tsinghua University Press, 2017. 1–2
(祁瑞华. 文本作者身份识别. 北京: 清华大学出版社, 2017. 1–2)

2 Mendenhall T C. The characteristic curves of composition. *Science*, 1887, **ns-9**(214S): 237–246

3 Yule G U. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 1939, **30**(3–4): 363–390

4 Mosteller F, Wallace D L. *Inference and Disputed Authorship: The Federalist*. Reading, Mass: Addison-Wesley Publishing Company, 1964.

5 Damerau F J. The use of function word frequencies as indicators of style. *Computers and the Humanities*, 1975, **9**(6): 271−280

6 Efron B, Thisted R A. Estimating the number of unseen species: How many words did Shakespeare know. *Biometrika*, 1976, **63**(3): 435−447

7 Chaski C E. Who′s at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 2005, **4**(1): 1−14

8 Hoover D L. Testing Burrows′s delta. *Literary and Linguistic Computing*, 2004, **19**(4): 453−475

9 Frantzeskou G, Stamatatos E, Gritzalis S, Katsikas S. Effective identification of source code authors using byte-level information. In: Proceedings of the 28th International Conference on Software Engineering. Shanghai, China: ACM, 2006. 893−896

10 Koppel M, Schler J, Argamon S, Winter Y. The "fundamental problem" of authorship attribution. *English Studies*, 2012, **93**(3): 284−291

11 Rudman J. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 1997, **31**(4): 351−365

12 Koppel M, Schler J, Argamon S. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 2009, **60**(1): 9−26

13 Luyckx K. *Scalability Issues in Authorship Attribution*. Antwerp: UPA University Press, 2010. 13−18

14 Potha N, Stamatatos E. A profile-based method for authorship verification. In: Proceedings of the 8th Hellenic Conference on Artificial Intelligence. Ioannina, Greece: Springer, 2014. 313−326

15 El Manar El Bouanani S, Kassou I. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 2014, **86**(12): 22−29

16 Johnson A, Wright D. Identifying idiolect in forensic authorship attribution: An N-gram textbite approach. *Language and Law*, 2014, **1**(1): 37−69

17 Keselj V, Peng F C, Cercone N, Thomas C. N-gram-based author profiles for authorship attribution. In: Proceedings of the Pacific Association for Computational Linguistics. Halifax, Canada: PACL, 2003. 255−264

18 Houvardas J, Stamatatos E. N-gram feature selection for authorship identification. In: Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications. Varna, Bulgaria: Springer, 2006. 77−86

19 Stamatatos E. Ensemble-based author identification using character N-grams. In: Proceedings of the 3rd International Workshop on Text-Based Information Retrieval. Seattle, WA, USA, 2006. 41−46

20 Sapkota U, Bethard S, Montes-y-Gomez M, Solorio T. Not all character N-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, USA: ACL, 2015. 93−102

21 Sari Y, Vlachos A, Stevenson M. Continuous n-gram representations for authorship attribution. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: ACL, 2017. 267−273

22 Gomez-Adorno H, Posadas-Duran J P, Sidorov G, Pinto D. Document embeddings learned on various types of N-grams for cross-topic authorship attribution. *Computing*, 2018, **100**(7): 741−756

23 Burrows J. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 2002, **17**(3): 267−287

24 Hoover D L. Another perspective on vocabulary richness. *Computers and the Humanities*, 2003, **37**(2): 151−178

25 Garcia A M, Martin J C. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 2007, **22**(1): 49−66

26 Zhao Y, Zobel J. Effective and scalable authorship attribution using function words. In: Proceedings of the 2nd Asia Information Retrieval Symposium. Jeju Island, Korea: Springer, 2005. 174−189

27 Coyotl-Morales R M, Villasenor-Pineda L, Montes-y-Gomez M, Rosso P. Authorship attribution using word sequences. In: Proceedings of the 11th Iberoamerican Congress in Pattern Recognition. Cancun, Mexico: Springer, 2006. 844−853

28 Stamatatos E. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 2006, **15**(5): 823−838

29 Koppel M, Schler J, Bonchek-Dokow E. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 2007, **8**: 1261−1276

30 Savoy J. Authorship attribution based on specific vocabulary. *ACM Transactions on Information Systems*, 2012, **30**(2): Article 12

31 Akimushkin C, Amancio D R, Oliveira O N. On the role of words in the network structure of texts: Application to authorship attribution. *Physica A: Statistical Mechanics and its Applications*, 2018, **495**: 49−58

32 Raghavan S, Kovashka A, Mooney R. Authorship attribution using probabilistic context-free grammars. In: Proceedings of the ACL 2010 Conference Short Papers. Uppsala, Sweden: ACL, 2010. 38−42

33 Tschuggnall M, Specht G. Enhancing authorship attribution by utilizing syntax tree profiles. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: ACL, 2014. 195−199

34 Patchala J, Bhatnagar R. Authorship attribution by consensus among multiple features. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: ACL, 2018. 2766−2777

35 Zhang R C, Hu Z Y, Guo H Y, Mao Y Y. Syntax encoding with application in authorship attribution. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018. 2742−2753

36 Sidorov G, Velasquez F, Stamatatos E, Gelbukh A, Chanona-Hernandez L. Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 2014, **41**(3): 853−860

37 Posadas-Duran J P, Sidorov G, Batyrshin I. Complete syntactic N-grams as style markers for authorship attribution. In: Proceedings of the 13th Mexican International Conference on Artificial Intelligence. Tuxtla Gutierrez, Mexico: Springer, 2014. 9−17

38 Posadas-Duran J P, Sidorov G, Batyrshin I, Mirasol-Melendez E. Author verification using syntactic N-grams. In: Working Notes of the Conference and Labs of the Evaluation Forum 2015. Toulouse, France, 2015.

39 Posadas-Duran J P, Markov I, Gomez-Adorno H, Sidorov G, Batyrshin I, Gelbukh A, et al. Syntactic N-grams as features for the author profiling task. In: Working Notes of the Conference and Labs of the Evaluation Forum 2015. Toulouse, France, 2015.

40 Gamon M. Linguistic Correlates of Style: Authorship classification with deep linguistic analysis features. In: Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: ACL, 2004. 611−617

41 Wu Xiao-Chun, Huang Xuan-Jing, Wu Li-De. Authorship iden-

tification based on semantic analysis. *Journal of Chinese Information Processing*, 2006, **20**(6): 61−68
(武晓春, 黄萱菁, 吴立德. 基于语义分析的作者身份识别方法研究. 中文信息学报, 2006, **20**(6): 61−68)

42 Argamon S, Whitelaw C, Chase P, Hota S R, Garg N, Levitan S. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 2007, **58**(6): 802−822

43 Hedegaard S, Simonsen J G. Lost in translation: Authorship attribution using frame semantics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: ACL, 2011. 65−70

44 Daelemans W. Explanation in computational stylometry. In: Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing. Samos, Greece: Springer, 2013. 451−462

45 Dasgupta A, Drineas P, Harb B, Josifovski V, Mahoney M W. Feature selection methods for text classification. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, CA, USA: ACM, 2007. 230−239

46 Lijo V P, Seetha H. Text-based sentiment analysis: Review. *International Journal of Knowledge and Learning*, 2017, **12**(1): 1−26

47 Cui M J, Li L, Wang Z H, You M Y. A survey on relation extraction. In: Proceedings of the 2nd China Conference on Knowledge Graph and Semantic Computing. Chengdu, China: Springer, 2017. 50−58

48 Ma J B, Xue B, Zhang M J. A profile-based authorship attribution approach to forensic identification in Chinese online messages. In: Proceedings of the 11th Pacific Asia Workshop on Intelligence and Security Informatics. Auckland, New Zealand: Springer, 2016. 33−52

49 Li Hang. *Statistical Learning Methods* (Second edition). Beijing: Tsinghua University Press, 2019. 6−12, 27−28, 59, 237, 245−253, 435−436
(李航. 统计学习方法. 第2版. 北京: 清华大学出版社, 2019. 6−12, 27−28, 59, 237, 245−253, 435−436)

50 Jin M Z, Jiang M H. Text clustering on authorship attribution based on the features of punctuations usage. In: Proceedings of the 11th International Conference on Signal Processing. Beijing, China: IEEE, 2012. 2175−2178

51 Hacohen-Kerner Y, Margaliot O. Authorship attribution of responsa using clustering. *Cybernetics and Systems*, 2014, **45**(6): 530−545

52 Fifield D, Follan T, Lunde E. Unsupervised authorship attribution. arXiv: 1503.07613, 2015

53 Mansoorizadeh M, Aminiyan M, Rahgooy T, Eskandari M. Multi feature space combination for authorship clustering. In: Working Notes of the Conference and Labs of the Evaluation Forum 2016. Evora, Portugal, 2016.

54 Bagnall D. Authorship clustering using multi-headed recurrent neural networks. In: Working Notes of the Conference and Labs of the Evaluation Forum 2016. Evora, Portugal, 2016.

55 Agarwal L, Thakral K, Bhatt G, Mittal A. Authorship clustering using TF-IDF weighted word-embeddings. In: Proceedings of the 11th Forum for Information Retrieval Evaluation. Kolkata, India: ACM, 2019. 24−29

56 Nakov P. Latent semantic analysis for German literature investigation. In: Proceedings of the International Conference on Computational Intelligence, Theory and Applications. Dortmund, Germany: Springer, 2001. 834−841

57 Satyam A, Dawn A K, Saha S K. A statistical analysis approach to author identification using latent semantic analysis. In: Working Notes of the Conference and Labs of the Evaluation Forum 2014. Sheffield, UK, 2014.

58 Jelodar H, Wang Y L, Yuan C, Feng X, Jiang X H, Li Y C, et al. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. arXiv: 1711.04305, 2018

59 Seroussi Y, Zukerman I, Bohnert F. Authorship attribution with latent Dirichlet allocation. In: Proceedings of the 15th Conference on Computational Natural Language Learning. Portland, Oregon, USA: ACL, 2011. 181−189

60 Savoy J. Authorship attribution based on a probabilistic topic model. *Information Processing & Management*, 2013, **49**(1): 341−354

61 Anwar W, Bajwa I S, Choudhary M A, Ramzan S. An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution. *IEEE Access*, 2019, **7**: 3224−3234

62 Zhang Xue-Gong. *Pattern Recognition* (Third edition). Beijing: Tsinghua University Press, 2010. 48−53
(张学工. 模式识别. 第 3 版. 北京: 清华大学出版社, 2010. 48−53)

63 Zhao Y, Zobel J. Searching with style: Authorship attribution in classic literature. In: Proceedings of the 13th Australasian Computer Science Conference. Ballarat, Victoria, Australia: ACS, 2007. 59−68

64 Boutwell S R. Authorship Attribution of Short Messages Using Multimodal Features [Master thesis], Naval Postgraduate School, USA, 2011

65 Altheneyan A S, Menai M E B. Naive Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University - Computer and Information Sciences*, 2014, **26**(4): 473−484

66 Howedi F, Mohd M. Text classification for authorship attribution using naive Bayes classifier with limited training data. *Computer Engineering and Intelligent Systems*, 2014, **5**(4): 48−56

67 Zhou Zhi-Hua. *Machine Learning*. Beijing: Tsinghua University Press, 2016. 33−35, 121−123
(周志华. 机器学习. 北京: 清华大学出版社, 2016. 33−35, 121−123)

68 Diederich J, Kindermann J, Leopold E, Paass G. Authorship attribution with support vector machines. *Applied Intelligence*, 2003, **19**(1): 109−123

69 Schwartz R, Tsur O, Rappoport A, Koppel M. Authorship attribution of micro-messages. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: ACL, 2013. 1880−1891

70 Mikros G K, Perifanos K A. Authorship attribution in Greek tweets using author′s multilevel N-gram profiles. In: Proceedings of the 2013 AAAI Spring Symposium Series. Palo Alto, USA: AAAI, 2013. 17−23

71 Li J S, Monaco J V, Chen L C, Tappert C C. Authorship authentication using short messages from social networking sites. In: Proceedings of the 11th International Conference on e-Business Engineering. Guangzhou, China: IEEE, 2014. 314−319

72 Martin-del-Campo-Rodriguez C, Alvarez D A P, Sifuentes C E M, Sidorov G, Batyrshin I, Gelbukh A. Authorship attribution through punctuation N-grams and averaged combination of SVM. In: Working Notes of the Conference and Labs of the Evaluation Forum 2019. Lugano, Switzerland, 2019.

73 Soler-Company J, Wanner L. On the relevance of syntactic and discourse features for author profiling and identification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: ACL, 2017. 681−687

74 Rokach L, Maimon O. *Data Mining with Decision Trees*: *Theory and Applications*. Singapore: World Scientific Publishing, 2008. 5−8

75 Apte C, Weiss S. Data mining with decision trees and decision rules. *Future Generation Computer Systems*, 1997, **13**(2−3): 197−210

76 Frery J, Largeron C, Juganaru-Mathieu M. UJM at CLEF in author verification based on optimized classification trees. In:

Working Notes of the Conference and Labs of the Evaluation Forum 2014. Sheffield, UK, 2014.

77 Digamberrao K S, Prasad R S. Author identification using sequential minimal optimization with rule-based decision tree on Indian literature in Marathi. *Procedia Computer Science*, 2018, **132**: 1086−1101

78 Maitra P, Ghosh S, Das D. Authorship verification — An approach based on random forest. In: Working Notes of the Conference and Labs of the Evaluation Forum 2015. Toulouse, France, 2015.

79 Trstenjak B, Mikac S, Donko D. KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, 2014, **69**: 1356−1364

80 Halvani O, Steinebach M, Zimmermann R. Authorship verification via *k*-nearest neighbor estimation. In: Working Notes of the Conference and Labs of the Evaluation Forum 2013. Valencia, Spain, 2013.

81 Anwar W, Bajwa I S, Ramzan S. Design and implementation of a machine learning-based authorship identification model. *Scientific Programming*, 2019, **2019**: 9431073

82 Sarwar R, Porthaveepong T, Rutherford A, Rakthanmanon T, Nutanong S. *StyloThai*: A scalable framework for stylometric authorship identification of Thai documents. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2020, **19**(3): Article No. 36

83 Gurney K. *An Introduction to Neural Networks*. London: CRC Press, 1997. 13−16

84 Bagnall D. Author identification using multi-headed recurrent neural networks. In: Working Notes of the Conference and Labs of the Evaluation Forum 2015. Toulouse, France, 2015.

85 Ruder S, Ghaffari P, Breslin J G. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. arXiv: 1609.06686, 2016

86 Qian C, He T C, Zhang R. Deep Learning based Authorship Identification, Department of Electrical Engineering, Stanford, CA, 2017.

87 Shrestha P, Sierra S, Gonzalez F A, Rosso P, Montes-y-Gomez M, Solorio T. Convolutional neural networks for authorship attribution of short texts. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: ACL, 2017. 669−674

88 Jafariakinabad F, Tarnpradab S, Hua K A. Syntactic recurrent neural network for authorship attribution. arXiv: 1902.09723, 2019

89 Khomytska I, Teslyuk V. Statistical models for authorship attribution. In: Proceedings of the 9th International Conference on Computer Science and Information Technologies. Lviv, Ukraine: Springer, 2019. 579−592

90 Grabchak M, Cao L J, Zhang Z Y. Authorship attribution using diversity profiles. *Journal of Quantitative Linguistics*, 2018, **25**(2): 142−155

91 Srinivasan L, Nalini C. An improved framework for authorship identification in online messages. *Cluster Computing*, 2019, **22**(5): 12101−12110

92 Qian T Y, Liu B, Chen L, Peng Z Y. Tri-training for authorship attribution with limited training data. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA: ACL, 2014. 345−351

93 Luyckx K, Daelemans W. Authorship attribution and verification with many authors and limited data. In: Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, UK: ACL, 2008. 513−520

94 Eder M. Does size matter? Authorship attribution, small samples, big problem. *Literary & linguistic computing*, 2015, **30**(2): 167−182

95 Koppel M, Schler J, Argamon S. Authorship attribution in the wild. *Language Resources and Evaluation*, 2011, **45**(1): 83−94

96 Luyckx K, Daelemans W. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 2011, **26**(1): 35−55

97 Stamatatos E. On the robustness of authorship attribution based on character N-gram features. *Journal of Law and Policy*, 2013, **21**(2): 421−439

98 Markov I, Stamatatos E, Sidorov G. Improving cross-topic authorship attribution: The role of pre-processing. In: Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing. Budapest, Hungary: Springer, 2017. 289−302

99 Rahgouy M, Giglou H B, Rahgooy T, Sheykhlan M K, Mohammadzadeh E. Cross-domain authorship attribution: Author identification using a multi-aspect ensemble approach. In: Working Notes of the Conference and Labs of the Evaluation Forum 2019. Lugano, Switzerland, 2019.

100 Mikros G K, Argiri E K. Investigating topic influence in authorship attribution. In: Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection. Amsterdam, Netherlands, 2007.

101 Sari Y, Stevenson M, Vlachos A. Topic or style? Exploring the most useful features for authorship attribution. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: ACL, 2018. 343−353

102 Seroussi Y, Bohnert F, Zukerman I. Authorship attribution with author-aware topic models. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Korea: ACL, 2012. 264−269

103 Seroussi Y, Zukerman I, Bohnert F. Authorship attribution with topic models. *Computational Linguistics*, 2014, **40**(2): 269−310

104 Yang M, Chen X J, Tu W T, Lu Z Y, Zhu J, Qu Q. A topic drift model for authorship attribution. *Neurocomputing*, 2018, **273**: 133−140

105 Halvani O, Winter C, Pflug A. Authorship verification for different languages, genres and topics. *Digital Investigation*, 2016, **16**: S33−S43

106 Bacciu A, La Morgia M, Mei A, Nemmi E N, Neri V, Stefa J. Cross-domain authorship attribution combining instance-based and profile-based features. In: Working Notes of the Conference and Labs of the Evaluation Forum 2019. Lugano, Switzerland, 2019.

107 Stamatatos E. Authorship attribution using text distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: ACL, 2017. 1138−1149

108 Stamatatos E. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 2018, **69**(3): 461−473

109 Ishikawa M, Kawakami H. Compression-based distance between string data and its application to literary work classification based on authorship. *Computational Statistics*, 2013, **28**(2): 851−873

110 Diamantini C, Panti M. An efficient and scalable data compression approach to classification. *ACM SIGKDD Explorations Newsletter*, 2000, **2**(2): 49−55

111 Cerra D, Datcu M, Reinartz P. Authorship analysis based on data compression. *Pattern Recognition Letters*, 2014, **42**: 79−84

112 Halvani O, Winter C, Graner L. On the usefulness of compression models for authorship verification. In: Proceedings of the 12th International Conference on Availability, Reliability and Security. Reggio Calabria, Italy: ACM, 2017. Article No. 54

113 Lichtblau D, Stoean C. Authorship attribution using the chaos game representation. arXiv: 1802.06007, 2018

114 Lichtblau D, Stoean C. Text documents encoding through images for authorship attribution. In: Proceedings of the 6th International Conference on Statistical Language and Speech Processing. Mons, Belgium: Springer, 2018. 178−189

115 Boenninghoff B, Rupp J, Nickel R M, Kolossa D. Deep Bayes factor scoring for authorship verification. In: Working Notes of the Conference and Labs of the Evaluation Forum 2020. Thessaloniki, Greece, 2020.

116 Halvani O, Graner L, Regev R. Cross-domain authorship verification based on topic agnostic features. In: Working Notes of the Conference and Labs of the Evaluation Forum 2020. Thessaloniki, Greece, 2020.

117 Kipnis A. Higher criticism as an unsupervised authorship discriminator. In: Working Notes of the Conference and Labs of the Evaluation Forum 2020. Thessaloniki, Greece, 2020.

118 Weren E R D, Kauer A U, Mizusaki L, Moreira V P, de Oliveira J P M, Wives L K. Examining multiple features for author profiling. *Journal of Information and Data Management*, 2014, **5**(3): 266−279

119 Rangel F, Rosso P, Potthast M, Stein B. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in twitter. In: Working Notes of the Conference and Labs of the Evaluation Forum 2017. Dublin, Ireland, 2017.

120 Martinc M, Skrjanec I, Zupan K, Pollak S. PAN 2017: Author profiling - gender and language variety prediction. In: Working Notes of the Conference and Labs of the Evaluation Forum 2017. Dublin, Ireland, 2017.

121 Tellez E S, Miranda-Jimenez S, Graff M, Moctezuma D. Gender and language-variety identification with MicroTC. In: Working Notes of the Conference and Labs of the Evaluation Forum 2017. Dublin, Ireland, 2017.

122 Takahashi T, Tahara T, Nagatani K, Miura Y, Taniguchi T, Ohkuma T. Text and image synergy with feature cross technique for gender identification. In: Working Notes of the Conference and Labs of the Evaluation Forum 2018. Avignon, France, 2018.

123 Daneshvar S, Inkpen D. Gender identification in twitter using N-grams and LSA. In: Working Notes of the Conference and Labs of the Evaluation Forum 2018. Avignon, France, 2018.

124 Tellez E S, Miranda-Jimenez S, Moctezuma D, Graff M, Salgado V, Ortiz-Bejar J. Gender identification through multimodal tweet analysis using MicroTC and bag of visual words. In: Working Notes of the Conference and Labs of the Evaluation Forum 2018. Avignon, France, 2018.

125 Rangel F, Rosso P, Montes-y-Gomez M, Potthast M, Stein B. Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in twitter. In: Working Notes of the Conference and Labs of the Evaluation Forum 2018. Avignon, France, 2018.

126 Rangel F, Rosso P. Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in twitter. In: Working Notes of the Conference and Labs of the Evaluation Forum 2019. Lugano, Switzerland, 2019.

127 Radivchev V, Nikolov A, Lambova A. Celebrity profiling using TF-IDF, logistic regression, and SVM. In: Working Notes of the Conference and Labs of the Evaluation Forum 2019. Lugano, Switzerland, 2019.

128 Hodge A, Price S. Celebrity profiling using twitter follower feeds. In: Working Notes of the Conference and Labs of the Evaluation Forum 2020. Thessaloniki, Greece, 2020.

129 Siagian A H A M, Aritsugi M. DBMS-KU approach for author profiling and deception detection in Arabic. In: Working Notes of the Forum for Information Retrieval Evaluation 2019. Kolkata, India, 2019.

130 Nayel H A. NAYEL@APDA: Machine learning approach for author profiling and deception detection in Arabic texts. In: Working Notes of the Forum for Information Retrieval Evaluation 2019. Kolkata, India, 2019.

131 Sharmila D V, Kannimuthu S, Ravikumar G, Anand K M. KCE_DALab-APDAFIRE2019: Author profiling and deception detection in Arabic using weighted embedding. In: Working Notes of the Forum for Information Retrieval Evaluation 2019. Kolkata, India, 2019.

132 Potthast M, Schremmer F, Hagen M, Stein B. Overview of the author obfuscation task at PAN 2018: A new approach to measuring safety. In: Working Notes of the Conference and Labs of the Evaluation Forum 2018. Avignon, France, 2018.

133 Potthast M, Hagen M, Stein B. Author obfuscation: Attacking the state of the art in authorship verification. In: Working Notes of the Conference and Labs of the Evaluation Forum 2016. Evora, Portugal, 2016.

134 Mihaylova T, Karadjov G, Kiprov Y, Georgiev G, Koychev I, Nakov P. SU@PAN′2016: Author obfuscation. In: Working Notes of the Conference and Labs of the Evaluation Forum 2016. Evora, Portugal, 2016.

135 Mansoorizadeh M, Rahgooy T, Aminiyan M, Eskandari M. Author obfuscation using WordNet and language models. In: Working Notes of the Conference and Labs of the Evaluation Forum 2016. Evora, Portugal, 2016.

136 Keswani Y, Trivedi H, Mehta P, Majumder P. Author masking through translation. In: Working Notes of the Conference and Labs of the Evaluation Forum 2016. Evora, Portugal, 2016.

137 Castro-Castro D, Bueno R O, Munoz R. Author masking by sentence transformation. In: Working Notes of the Conference and Labs of the Evaluation Forum 2017. Dublin, Ireland, 2017.

138 Kocher M, Savoy J. UniNE at CLEF 2018: Author masking. In: Working Notes of the Conference and Labs of the Evaluation Forum 2018. Avignon, France, 2018.

139 Rahgouy M, Giglou H B, Rahgooy T, Zeynali H, Rasouli S K M. Author masking directed by author′s style. In: Working Notes of the Conference and Labs of the Evaluation Forum 2018. Avignon, France, 2018.

张 洋 清华大学人文学院中文系博士研究生. 主要研究方向为作者识别, 文本分类, 情感分析.
E-mail: yumaoqiuq@163.com
(**ZHANG Yang** Ph. D. candidate in the Department of Chinese Language and Literature, School of Humanities, Tsinghua University. His research interest covers authorship identification, text categorization, sentiment analysis.)

江铭虎 清华大学人文学院中文系教授. 主要研究方向为自然语言处理, 脑与语言认知, 模式识别, 人工智能. 本文通信作者.
E-mail: jiang.mh@mail.tsinghua.edu.cn
(**JIANG Ming-Hu** Professor in the Department of Chinese Language and Literature, School of Humanities, Tsinghua University. His research interest covers natural language processing, brain and language cognition, pattern recognition, artificial intelligence. Corresponding author of this paper.)