文本作者身份识别

——基于机器学习与计算语言学

祁瑞华 著

华大学出版社

文本作者身份识别
——基于机器学习与计算语言学

祁瑞华 著

清华大学出版社
北京

内 容 简 介

文本作者身份识别广泛应用于文学作品、新闻稿、商品评论、垃圾邮件的作者身份鉴定以及法庭取证等领域。随着大数据时代网络文本的大量涌现，匿名文本的作者身份识别在网络取证、不良舆情监控等任务中的应用成为国内外学者关注的热点。

本书探讨了文本作者身份识别的关键问题、基本方法和最新研究进展，并应用于实践得以验证。全书共 7 章，分为 3 部分：第 1 部分包括第 1～2 章，介绍文本作者身份识别的基本概念、研究内容、建模基本方法和主要应用领域；第 2 部分包括第 3～4 章，介绍现有的作者身份文体特征、作者身份识别算法、性能评价指标、主要实验平台等；第 3 部分包括第 5～7 章，介绍本书对作者身份识别研究的贡献和在中英文博客、微博语料上的实验验证。

本书主要面向文本挖掘领域的研究生和相关专业的研究人员，既可以作为文本分析与处理研究的教科书，也可以作为政府相关部门产品研发人员的参考书。

图书在版编目(CIP)数据

文本作者身份识别：基于机器学习与计算语言学/祁瑞华著. —北京：清华大学出版社，2017
ISBN 978-7-302-45576-9

Ⅰ. ①文…　Ⅱ. ①祁…　Ⅲ. ①数据处理　Ⅳ. ①TP274

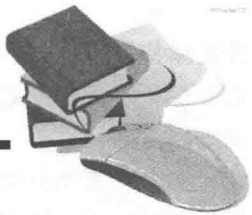
中国版本图书馆 CIP 数据核字(2016)第 283882 号

责任编辑：贾　斌　张爱华
封面设计：何凤霞
责任校对：胡伟民
责任印制：王静怡

出版发行：清华大学出版社
网　址：http://www.tup.com.cn，http://www.wqbook.com
地　址：北京清华大学学研大厦 A 座　**邮　编**：100084
社总机：010-62770175　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课　件　下　载：http://www.tup.com.cn，010-62795954
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　**印　张**：11.5　**字　数**：287 千字
版　　次：2017 年 2 月第 1 版　**印　次**：2017 年 2 月第1 次印刷
印　　数：1～2000
定　　价：49.00 元

产品编号：070384-01

前言



文本作者身份识别广泛应用于文学作品、新闻稿、商品评论、垃圾邮件的作者身份鉴定以及法庭取证等领域。随着大数据时代网络文本的大量涌现，匿名文本的作者身份识别在网络取证、不良舆情监控等任务中的应用成为国内外学者关注的热点。

文本作者身份识别研究主要通过文体风格特征建模表达作者无意识的写作习惯，从而自动映射匿名文本作者归属。相关研究经过百余年的发展，奠定了良好的理论和应用基础，已经广泛应用于文学作品或新闻报道等传统语料的作者身份识别。近年来网络文本作者身份识别成为研究热点，语料涉及电子邮件、网络评论、BBS 和博客等，出现了数据海量、特征维度巨大、每个用户可得训练文本少等新特点，这些都是文本作者身份识别研究面临的新挑战。

本书探讨了文本作者身份识别的关键问题、基本方法和最新研究进展，并应用于实践得以验证。

全书共 7 章，共分为 3 部分。

第 1 部分包括第 1～2 章，介绍文本作者身份识别的基础知识。其中，第 1 章介绍了作者身份识别的基本概念、研究内容、建模基本方法和面临的主要问题；第 2 章分类归纳了作者身份分析的主要应用领域。

第 2 部分包括第 3～4 章，介绍现有的作者身份文体特征和作者身份识别算法。其中，第 3 章介绍了作者身份文体特征类别和特征选择的一般方法；第 4 章介绍了作者身份识别的主要算法、性能评价指标和主要实验平台。

第 3 部分包括第 5～7 章，介绍本书对作者身份识别研究的贡献和实验验证。其中，第 5 章建立了英文博客作者身份文体特征模型，在公开博客语料上的实验证实了模型在短文本语料的有效性；第 6 章建立了中文微博作

者文体特征模型，在中文微博语料上证实了模型在短篇幅网络文本上的有效性；第7章在中文微博作者性别识别实验中进一步拓展了文体特征模型的应用范围。

本书主要面向文本挖掘领域的研究生和相关专业的研究人员，既可以作为文本分析与研究的教科书，也可以作为政府相关部门产品研发人员的参考书。

本书能够尽快完成出版，首先要感谢美国 Purdue 大学的 Marcus Rogers 教授、Julia Taylor 教授和我的同事霍跃红老师、刘彩虹老师、郭旭老师等，以及参与数据收集和整理的学生，本书的若干专题研究都与他们进行过深入的讨论。还要感谢清华大学出版社的编辑，是他们的鼓励和细致工作使得本书得以顺利出版。最后感谢在本书中所引用参考文献的作者和公开语料库的开发者，本书的写作从他们的研究成果中获取了很多营养，正是他们的勤奋以及分享的科研精神引领和启发我完成本书的写作。

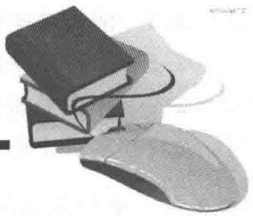
本书研究获得以下研究项目的资助：教育部第50批留学回国人员科研启动基金“典籍英译国外读者观点文本主题挖掘研究”（教外司[2015]1098）；国家社科基金一般项目“典籍英译国外读者网上评论观点挖掘研究”（15BYY028），在此表示感谢。

虽然我始终以认真严谨的态度对待本书的撰写工作，但很多研究尚属于探索阶段，书中难免有不足之处，恳请广大读者批评指正！

祁瑞华

2016年7月

目 录



第1章

绪　论

作者身份识别研究属于应用语言学和计算机科学的交叉领域，其主要思路是将文本中隐含的作者无意识写作习惯通过某些可以量化的特征表现出来，凸显作品的文体学特征和写作风格，以此确定匿名文本的作者。

最初作者身份识别研究应用于传统文学作品作者鉴定和法庭文本取证，目前已经广泛应用于包括 BBS、博客、电子邮件、聊天室、新闻群组等网络社交媒体平台上文本的作者身份分析。网络文本作为反映社会舆情的主要载体之一，在促进先进文化知识共享传播、及时有效反映公众意见中起到了重要作用。但网络文本中同时也存在着一些以匿名方式滥用互联网的问题，例如虚假商品信誉评价、垃圾邮件、恐吓信息、盗版软件、色情作品等，这些问题都具有潜在的社会危害。在很多利用互联网传播不良信息的案例中，网络文本作者都试图隐藏真实身份以逃避检测。

与传统文学作品的作者身份识别相比，网络文本的作者身份识别问题有着新的技术特征。网络文本以数字和符号化的形式流动，更具隐秘性，传统条件下的作者身份识别技术受到前所未有的挑战。如何使作者身份识别技术与在线信息中隐含的细微文体风格特征相匹配，从而为网上热点舆情和不良信息的监控与追踪提供分析依据，是当前理论研究和应用研究的前沿和热点。解决这一问题需要结合互联网应用环境的特点，从新的视野和角度延伸现有应用语言学理论和应用，结合计算机科学、统计学方法和应用语言学方法，以程序化、数字化和精密化的方式，通过从词汇、语法、结构和

语义多个层面定量分析和表示文本，使大规模在线信息和文学作品的作者身份识别成为可能。

1.1 基本概念

作者身份识别是作者身份分析的任务之一，作者身份分析广泛应用于文学作品、商品评论、垃圾电子邮件的作者身份鉴定以及网络舆情检测等领域，近年来成为国内外学者研究和关注的热点。作者身份分析主要有三类任务：作者身份识别(Auhor Identification)、作者身份描述(Author Profiling)和作者聚类分析(Author Clustering)。

1.1.1 作者身份识别

布封和斯皮彻等人认为，文体实际上是一种个人的行为方式，作家在写作行为中会自觉或不自觉地将其个性和个人社会背景融入或体现于作品中[1]。作者身份识别正是基于这一理论，以文本分类视角根据匿名文本的内容自动确定其作者归属的映射，属于模式分类和自然语言处理的交叉学科。

作者身份识别是作为最传统的作者分析任务，其研究以文体风格等特征为依据，进而自动确定文本作者归属的映射过程，可应用于法庭取证、文学分析等领域。本书的研究对象主要是作者身份识别任务。

作者身份识别研究可以追溯到 1887 年 Mendenhall[2] 对戏剧作品文体特征曲线的研究。经过国内外学者百余年的努力，作者身份识别的研究逐步深入，总结性的研究文献主要有：Holmes 从语言学和文学研究的视角，对传统文学作品作者身份分析研究进行总结[3]，之后 Stamatatos 侧重于计算需要和实验环境设置对 1999 至 2009 十年间的作者身份归属研究做了归纳[4]，是近年此领域颇具影响力的综述文献。近年来，随着大数据时代网络文本的大量涌现，作者身份识别领域出现的许多新特点导致作者身份识别难度大大增加。网络文本的作者身份识别任务框架如图 1.1 所示。

根据候选作者集是否开放，作者身份识别任务可以分为闭集任务和开集任务。在闭集任务中，候选作者集中的所有候选作者已经确定，作者身份

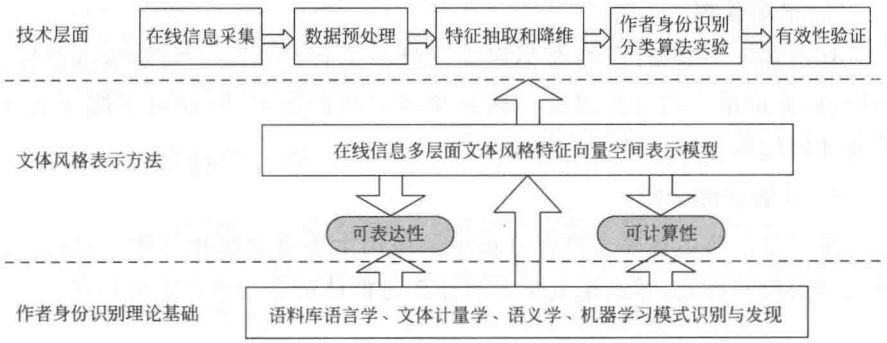


图 1.1 作者身份识别任务框架

识别的目标是从此候选作者集中选择最有可能的一个候选作者。一般来说，文学作品的作者身份识别任务属于闭集任务，例如 Gamon[5] 对勃朗特三姐妹英语作品的作者身份识别就是候选作者集仅包含三位作者的闭集实验。当候选作者集仅包含一位作者时，作者身份识别任务演化为作者身份查证(Author Verification)任务。从这个角度，所有的作者身份识别任务都可以分解为一系列的作者身份查证任务。

开集任务是作者身份识别任务在更广泛意义上的定义，开集任务中候选作者集是一个开放的集合，真正的作者可能不在已知候选作者集中，这类任务在法庭取证和网络文本作者身份识别中比较常见。与闭集任务相比，作者身份识别开集任务的难度和复杂度更高，是相关领域研究尚待解决的难点。

1.1.2 作者身份描述

作者身份描述的主要任务是抽取作者自然属性的统计信息，其基本属性包括性别、年龄、出生地、母语、教育背景、职业属性等。在用户市场分析等实践应用中，获取作者的自然属性往往是非常关键和有价值的。作者身份描述分析的主要应用领域如下所示。

1. 法庭取证

法庭取证的方法之一，就是对犯罪嫌疑人所写文本的特征进行分析，并与特定的年龄、性别群体或有犯罪倾向群体的文本语言学特征进行对比。作者身份分析技术的采用，能够帮助在法庭取证中确定犯罪嫌疑人。

2. 市场营销

从市场推广的角度，商业公司往往结合用户购买行为和网络商品评论等信息，分析用户的自然属性与购买偏好之间的联系，从而对不同年龄段、不同性别的客户群体进行个性化推荐。

3. 作者群体分析

作者身份描述在作者群体分析中的应用主要有文学作品作者身份与文学活动的关系研究，学术论文学术著作作者群体的分布和特点分析等。

1.1.3 作者聚类分析

作者聚类分析主要应用于剽窃检测和特定作者不同时期写作风格变化分析等。例如，通过聚类分析匿名文学作品或电子邮件的作者归属、风格差异、从众性、可读性、语法结构以及语言变化程度等[6]。

作者聚类依据同一作者所著文档相似度大的原则，将文本按照文体分析进行类别划分。作者聚类分析的基础是文本聚类算法，是在没有标注语料库的情况下进行无监督学习的有效途径。聚类分析属于无监督学习方法，不需要事先标注文档的类别，适合标注语料难以获取的情况，算法可分为划分聚类法、层次聚类法、密度聚类法、网格聚类法和模型聚类法等。

对于文学作品数量有限的长文本语料，应用聚类分析之前可以将其分割成若干文本片段，从而获得足够数量的训练样本。

1.1.4 机器学习

Tom Mitchell 提出[7]：机器学习这门学科所关注的问题是计算机程序如何随着经验积累自动提高性能。机器学习的形式定义为：对于某类任务 T 和性能度量 P，如果一个计算机程序在 T 上以 P 衡量的性能随着经验 E 而自我完善，那么称这个计算机程序在从经验 E 学习。

机器学习的研究，采用了统计学、计算机科学、信息论、人工智能、控制论、心理学、生物学等学科的理论与方法。从理论基础角度，机器学习方法一般可以分为线性模型、决策树、神经网络、支持向量机、贝叶斯分类、集成学习、聚类、强化学习等。从学习的反馈机制角度，机器学习可分为有监督学习、无监督学习和半监督学习。

目前，机器学习不仅已经广泛地应用于图形学、软件工程、网络通信、体系结构、芯片设计、计算机视觉、自然语言处理等计算机科学领域，还为许多交叉学科提供着重要的技术支持[8]。

1.1.5 计算语言学

“计算语言学”这个学科名称在1966年美国科学院的ALPAC报告中正式得到学术界的承认。我国语言学家冯志伟对计算语言学定义如下[9]：计算语言学是采用计算机技术来研究和处理自然语言的一门新兴学科，是一门介乎语言学、数学和计算机科学之间的边缘性的交叉学科，它同时涉及文科、理科和工科三大领域。

计算语言学对自然语言的研究和处理，一般应经过如下三方面的过程：

第一，把需要研究的问题在语言学上加以形式化，使之能以一定的数学形式，严密而规整地表示出来；

第二，把这种严密而规整的数学形式表示为算法，使之在计算上形式化；

第三，根据算法编写计算机程序，使之在计算机上加以实现。

计算语言学的研究内容包括以词汇、句子、话语或语篇及其词法、句法、语义和语用等相关信息为研究对象的系列处理技术[10]。

冯志伟将计算语言学最初的形式模型分为七种[11]：基于短语结构语法的形式模型、基于合一运算的形式模型、基于依存和配价的形式模型、基于格语法的形式模型、基于词汇主义的形式模型、语义自动处理的形式模型以及语用自动处理的形式模型。这些形式模型融合了不同学科的知识和技术，对语言学的建模大多使用数学或计算机科学领域的方法。

计算语言学与数学、计算机科学等学科的交叉融合，推动了相关领域理论和应用研究的进步。在当前大数据环境下，计算语言学领域的学者与统计机器学习领域的学者之间的交流和互动更加频繁[12]。这种交流和互动是双赢的：一方面，利用计算语言学的相关知识，可以从大规模语料中挖掘语法、语义和语用等不同层面挖掘语言知识，这些语言知识的合理运用，能够帮助机器学习领域提高数据处理的速度和准确率；另一方面，机器学习领域的技术方法正在帮助语言学家完成人工无法完成的海量语料中的语言知识的挖掘。计算语言学的研究，正在跨越狭义的定义，结合语言学、数学和计算机科学，形成并发挥着超学科研究的优势[12]。

1.2 作者身份识别研究

早在文本作品产生之前，口头传颂的诗歌作品的作者归属就已经备受关注。自从文本作品诞生之日起，各种关于文本作品创作者身份的争议随之而来。从西方的《圣经》、《联邦党人文集》的作者考证，中国古代的诗经、《孔雀东南飞》、《红楼梦》以及许多历史悠久民谣的作者考证，到现代的网络文本作者身份追踪，作者身份识别始终是社会各界关注的热点话题。

作者身份识别研究基于语言学研究领域中的文本分析，通过统计学方法分析文体风格，以此判断一段作品是否由某个特定作者创作。传统的作者身份识别技术主要应用于文学作品作者归属问题，近年来，国内外学者正尝试将其应用于网络文本，如电子邮件、博客、在线信息以及源代码的作者归属问题。作者身份识别研究有两项关键内容：文体风格特征和作者身份建模技术。

1.2.1 文体风格特征研究内容

文体风格特征是指能够有效识别作者身份的文档属性或写作风格标识，国内外相关研究中选择的文体风格特征主要有词汇特征、语法特征、结构特征和语义特征等。

词汇特征是基于字符和基于词的特征。主要包括平均词长、句子长度、词汇丰富度和高频词等[13]。词汇特征在传统文学作品作者身份识别任务中的效果较好。但当应用于网络文本作者身份识别任务时，由于在电子邮件、博客、微博等网络文本中大量使用缩略语和首字母缩写词，给基于词汇的特征数据造成相当大的噪声，词汇特征的效果受到影响。此外，由于文章中词汇的选择与主题高度相关，因此词汇特征在跨主题领域应用时的效果也会受到影响。

语法特征传统上包括功能词、词性标注 Ngram 和标点符号的使用方式等特征，近年来相关研究还包括：使用语法分析程序分析重写规则频率[5]、利用自然语言处理工具检测句子和短语边界[14]、统计 POS(Part of Speech)标签频率或 POS 标签 Ngram 频率[15]等。现有研究表明，功能词特征能够

有效反映作者在文体格式上的习惯，在作者身份识别中是很有效的。此外，也有文献证明检查上下文语法信息中的双连词频率对 200 词左右短文本的作者识别是有效的[16]。

结构特征包括与文本组织和布局相关的特征，如段落数目、段落长度、文本字体、字号、字的颜色、空格、缩格、签名档以及图像的使用等特征。由于在线信息篇幅较短小，作者倾向于使用更为丰富灵活的文本组织和布局方式，因此结构特征对于在线信息文体风格的表达尤为有效。

目前语义特征相关研究较少，主要有生成语义关系图[5]、基于 WordNet 抽取英文隐含语义分析词汇特征[17]、利用 HowNet 语义知识库筛选中文词汇作为作者写作风格特征[18]等，这些方法均对作品长度有一定要求。

为了提高作者身份识别的准确率，现有研究通常在短文本的文体风格特征中引入内容相关特征[19][20]，这类特征已经被证明是有效的，但缺乏在跨主题语料上的通用性。

国内外作者身份识别相关文献中使用的特征已经超过数千种，目前还没有公认的最有效的文体风格特征集。

1.2.2 作者身份建模技术研究内容

作者身份建模是指通过数理统计方法或机器学习等方法，建立文档与其作者归属的预测模型。此模型的输入是文本特征集，输出通常是文档属于某一作者的可能性预测估算值或者连续值。作者身份建模技术的理论介绍详见 1.3 节。

1.3 作者身份建模基本方法

传统的文本作者身份识别方法是笔迹鉴定，即通过对手写笔迹的比较，分析文本书写作者的真实身份，例如法庭取证中的遗书真实性鉴定等。笔迹鉴定的主要途径可以分为人工笔迹鉴定和计算机笔迹鉴定。

其中，人工笔迹鉴定依赖专业技术人员的经验判断，也称专家推测法。专家推测法主要依靠专业技术人员仔细阅读纸质或在线文本内容。对于纸质文本，专家通常通过分析笔迹图像学特征、纸张特征、墨水特征等推断某一纸质文本是否出于特定作者。对于电子文档，通常通过分析信息的写作

时间、信息发布 IP 地址所在区域、文本写作风格等特征，推断作者的真实身份。人工笔迹鉴定方法对专家的专业知识和经验要求高，实现成本很高。因此仅在特定领域，如法庭鉴定中采用。

计算机笔迹鉴定可以自动提取笔迹特征进行分析鉴定真伪[21]。随着互联网应用环境的普及，身份识别研究也从传统环境扩展到在线信息的作者身份识别，笔迹鉴定随之划分为离线电子文档和在线电子文档的笔迹鉴定。在线电子文档笔迹鉴定的典型应用案例是电子邮件的作者身份识别，其主要方法是分析电子邮件头信息、IP 地址等物理信息，以及分析邮件文本结构、语言风格特征等基于邮件内容的特征，结合这些特征来判断作者身份。

对于没有签名信息的文档，其作者身份识别的主要信息来源就是文档内容，即要从文档内容中抽取能够表达作者身份的字、词汇、语法结构或行文风格，然后通过一定的技术方法建立作者身份模型。

作者身份识别研究的基本思路是：给定一组候选作者及他们的训练文本，根据从训练文本中学习到的作者模型将匿名文本分配给某一位候选作者。现有的作者身份建模技术主要有：基于概率模型的朴素贝叶斯及其改进算法；基于向量空间模型的判别分析方法，支持向量机、决策树、神经网络和遗传算法；基于相似度度量的最近邻法等[22][23]。

根据训练文本的处理方式不同，作者模型的学习方法可分为基于侧面的方法和基于实例的方法。

1.3.1 基于侧面的作者身份建模

基于侧面的作者身份建模(Profile-based Authorship Modeling)方法中，首先将每位作者的训练文本累积成一个文档，然后从这个文档中抽取属性建立作者侧面模型。匿名文本通过一定的距离度量方式被分配给最相似的作者。这种方法中，不需要单独为某一训练文本建立表示模型，忽略每个作者的多个训练文本间的差别，因此，最终合成的文档可能与单个训练文本相去甚远。

基于侧面建模是早期作者身份识别的主要建模方法，其优点是训练过程相对简单、时间复杂度低，但对训练文本的长度要求较高，难以应用于多层面特征集。

近年来，随着短文本语料作者身份识别需求的大量出现，学者们尝试对