文本作者身份识别研究综述*

祁瑞华[1] 霍跃红[2] 胡润波[3]
[1]大连外国语大学计算机教研部 大连 116044 [2]大连外国语大学英语学院 大连 116044
[3]中国大连高级经理学院 大连 116024

摘要: [目的/意义] 鉴于传统的作者身份识别方法不适用于当前大量涌现的网络文本。综述近年文本作者身份识别的典型方法和关键问题,并进行客观分析和评价,以期为进一步研究提供新的思路。 [方法/过程] 分别从应用领域、文体特征选取、作者身份建模和性能评价指标等方面对国内外作者身份识别相关研究现状进行客观分析,梳理相关领域研究发展脉络和趋势。 [结果/结论] 作者身份识别需要适应短文本、不规范文本、海量、高维和多语种环境,需更具表现和刻画能力的多层面特征和相应的作者身份建模方法,并借助信息检索、机器学习和自然语言处理领域的最新研究成果提高效率和准确率。

关键词: 身份识别 文体特征 作者身份 建模性能 评价指标
分类号: TP393
DOI: 10.13266/j.issn.0252-3116.2015.16.021

1 引言

文本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 3 类问题: 作者身份识别、作者模型描述和作者聚类分析。作者身份识别是以文体风格特征为依据,自动确定文本作者归属的映射过程,可应用于法庭取证、文学分析等问题; 作者模型描述的主要任务是抽取作者的统计信息,如性别、年龄等,普遍应用于市场分析; 作者聚类是通过文体分析进行文本类别划分,主要应用于剽窃检测和特定作者不同时期写作风格的变化分析等。

作者身份研究可以追溯到 1887 年 T. C. Mendenhall[1] 对戏剧作品文体特征的研究,经过国内外学者 100 多年的努力,作者身份识别问题的研究逐步深入,D. I. 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综述[2],之后 E. Stamatatos 侧重于计算需要和实验环境设置对 20 世纪末至 21 世纪初的作者身份归属研究做了归纳[3],是此领域颇具影响力的综述。

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。为此,本文结合文献检索顺查法和追溯法,选择国际知名期刊和重点会议中引用率高或近期的文献,其中 2000 年以前代表性文献 4 篇、2000-2009 年文献 15 篇、2010 年后 15 篇、国别分布为美国 9 篇,中国 6 篇,希腊 4 篇,加拿大、澳大利亚和德国各 2 篇,英国、西班牙、荷兰、墨西哥、以色列、法国、韩国、伊朗和阿拉伯联合酋长国学者各 1 篇,在此基础上梳理文本作者身份识别发展的历史脉络,重点论述近年来该领域研究的典型方法和关键问题,分别从应用领域、文体风格特征、作者身份建模和性能评价指标等方面对国内外研究现状进行客观分析,对文本作者身份识别的未来发展趋势作出展望,以期为进一步研究提供新的思路。

2 应用现状

2.1 传统语料作者身份识别

文学作品是作者身份识别的传统语料,研究涉

* 本文系国家社会科学基金一般项目"典籍英译国外读者网上评论观点挖掘研究"(项目编号: 15BYY028)和教育部人文社会科学研究规划青年基金项目"基于多层面特征分析的在线信息作者身份识别研究"(项目编号: 11YJCZH131)研究成果之一。
作者简介: 祁瑞华(ORCID: 0000-0002-2583-3055),博士,副教授,E-mail: rhqi@dlufl.edu.cn; 霍跃红(ORCID: 0000-0001-8657-298X)博士,副教授; 胡润波(ORCID: 0000-0003-3210-6339),博士,讲师。
收稿日期: 2015-07-14 修回日期: 2015-08-06 本文起止页码: 143-148 本文责任编辑: 杜杏叶

及英语、中文、日语、希腊语、阿拉伯语、荷兰语和土耳其语等语种，研究成果较为丰富。其定量研究始于T. C. Mendenhall从单词长度规律的角度对英美文学作品写作风格的分析[1]，代表研究有G. U. Yule根据句长分析英文散文、传记和随笔等作品的写作风格[4]，H. Baayen基于重写规则频率语法对Nijmegen标注语料库中20世纪60年代戏剧、犯罪小说和文学评论的作者进行分析等[5]。这些早期研究主要基于一元文体特征，仅适用于特定语料。

为了增强通用性，随后学者们在作者身份识别中引入多元特征，如Zhao Ying等从句法角度以365个功能词为特征对美联社TREC语料库文章进行作者识别[6]，R. Goebel采用DepWords编码替代句法依存关系来识别侦探小说的作者[7]，F. H. Hassan等检测了文本单词首字符、中间字符、结束字符的Ngram，指出仅使用首字符Bi-gram和Tri-gram能有效识别作者[8]。

对多层面特征进行组合是进一步提高作者身份识别准确率的有效方法，相关研究有M. Gamon基于语法分析建立多层面组合特征集，应用于勃朗蒂三姐妹作品，验证了其有效性[9]。Zhang Chunxia等在21本英文作品和路透社语料上抽取多层面特征，证明了依存关系能够描述相对稳定的语法模式和谓词参数关系，有助于提高作者身份识别准确率[10]。

传统语料的作者身份识别研究经过100多年的发展，从最初的一元特征到多元特征，再到多层面组合特征集，作者身份识别准确度不断提高，为文本作者身份识别奠定了坚实的基础。但研究限于文学作品等长文本，候选作者通常为2－5人，当将传统方法应用于短文本语料或候选作者数量增加时，其准确率明显下降。

2.2 网络文本作者身份识别

最近10年，随着网络文本的大量涌现，以电子邮件、BBS、博客和网络评论等为语料的作者身份识别受到广泛关注，成为新的研究热点。最初的网络文本作者身份识别语料主要是电子邮件，如A. Abbasi等提出文本结构特征与传统特征相结合，并在滑动窗口中以Karhunen-Loeve变换发现文体风格变化构成笔迹特征，对25名作者的电子邮件和商品评论文本进行识别，获得了90%以上的准确率[11]；F. Iqbal等探索了词汇拼写错误和句法错误特征在电子邮件作者识别中的应用[12]。

随着网络应用的迅速发展，作者身份识别在网络评论、聊天记录、BBS和博客文本上都有应用研究，如N. Ali等基于字符Tri-gram提出新特征TF-ITF应用于聊天机器人语料，发现单独使用时其效果受语料规模影响显著[13]；Fan Mengdi等在对网上评论作者身份识别时指出，抽取词汇特征前不应该去停用词和还原词性，否则会丢失作者风格信息[14]；H. Zamani等提出以词汇和句法特征的极大似然估计分布模型作为特征集，并给出特征集间距离计算方法和特征选择方法，增强了多层面特征集的可解释性[15]。吕英杰等抽取词汇、句法、结构和内容特征构成多层面特征集，采用支持向量机在BBS论坛和博客文本上获得80%左右的作者识别准确率[16]。网络犯罪追踪也是作者身份识别的一个主要方向，代表研究如Zheng Rong等建立了中英文作者身份识别框架，对含非法销售盗版软件的新闻组和BBS语料进行作者身份识别[17]；A. Abbasi在英文和阿拉伯文网络文本的恐怖分子追踪中取得显著成果[11]。

上述研究针对网络文本篇幅短小、体裁结构变化丰富的特点，各自提出了新的文体风格特征作为有益的补充，但是仍存在以下问题有待解决：①短文本的特殊性使得文体风格特征集数据稀疏，影响作者身份识别的效率和准确率；②为提高作者身份识别准确率，现有研究通常在网络文本文体风格特征中引入内容相关特征[8,10]，这类特征虽然有效，但缺乏跨主题的通用性。

3 关键问题

本文主要聚集于作者身份识别研究的两个关键问题：文体风格特征选择和作者身份建模技术。

3.1 文体风格特征

文体风格特征是指能够有效识别作者身份的独特文档属性和写作风格标识等语言参数。理论文体学的作家决定论指出，作品风格产生于作者对其思想行为的合理安排[18]，作者在其作品中会自觉或不自觉地融入其个性和个人社会背景。作者身份识别研究的基础就是对文体风格特征的比较分析，关键问题是如何捕捉这些作者独特的文档属性和写作风格。

根据文体风格特征的语言学计算需求和复杂度，文学作品作者身份特征可以梳理分类为字符特征、词汇特征、句法特征、语义特征以及领域相关特征。

字符特征将文本看作字符序列，抽取诸如字母

大小写频率、数字频率、标点符号频率等文本特征。其中，字符 Ngram 能表现上下文信息、标点符号和字母大小写搭配习惯等，还能够捕捉到文本中的语法错误和拼写错误等细微特征，从而发现作者独特的痕迹，是传统文体风格研究中最有效特征之一[3]。这一类特征的优势在于对计算能力要求低、不需要特殊的分析工具并适用于多语种环境，其缺点在于统计字符 Ngram 特征的维度过高而导致包含冗余信息。

词汇特征包括单词的统计特征和频率，如词长、词汇丰富度、词频、单词 Ngram 以及特殊词汇等。值得注意的是，特殊词汇也可以体现作者特有的拼写错误习惯。词汇特征的抽取依赖于语料长度，因此通常不单独使用词汇特征。此外，词频等特征是主题相关的，不具备跨主题语料上的通用性。

句法特征包括功能词、词性标注 POS、重写规则和标点符号等，可分为浅层和深层句法特征。浅层句法特征是指不需要进行句法解析的，诸如句长、词性标注统计、类符形符比和功能词频率等特征。深层句法特征是经过完全或部分解析句子后获取的特征，如依存句法关系、词性标注 N-gram 等。句法特征能够表达隐含的文本结构，近年来成为文体特征研究的热点[19]。相关研究还表明句法特征单独使用的效果不如词汇特征，但与其他特征结合使用能够改进作者身份识别性能[3]。

对语义特征的研究主要包括生成语义关系图[3]、基于 WordNet 抽取英文隐含语义分析词汇特征[20]、利用 HowNet 语义知识库筛选中文词汇作为作者写作风格特征[21]等，这些方法均对作品长度有一定要求，通常与语料的主题相关，不具有通用性。

为提高准确率学者们还探讨了领域相关特征，主要思路是抽取主题相关的关键词加入到多层面特征集中[17]，但并未给出这些关键词的选择方法或选择的基本原则，不具有可解释性和可再现性。

3.2 作者身份建模

作者身份识别的基本思路是：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。从建立作者模型的方法可分为基于侧面的方法和基于实例的方法。

基于侧面的建模将每位作者的训练文本累积成一个文档，从这个文档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。这种方法不考虑为单一训练文本建立表示模型，每个作者的多个训练文本间的差别被忽略，最终合成的文档可能与单个训练文本相去甚远。其作者识别核心算法是匿名文本与作者侧面模型之间的距离计算函数，主要有概率模型如 Bayesian 模型[22]，压缩模型[23]如 CNG(Common N-Gram)模型[24]及变形 SPI(Simplified Profile Intersection)模型[25]等。基于侧面建模作为早期作者身份识别的主要建模方法，优点是训练过程相对简单、时间复杂度低，缺点是对训练文本的长度要求较高，累积成一个文档的建模方式难以应用于多层面特征集。

近年基于实例的作者建模成为主流，这种方法将每个训练文本看作一个实例单独处理，要求每个训练文本足够长以便能抽取充足的文体风格信息。如果文本长度的差距较大，需要对训练文本长度标准化，将长训练文本分解等长的片段使得特征获取的语料长度相对一致。基于实例建模方法可以分为向量空间模型、基于相似度模型和元学习模型。

基于向量空间模型的作者身份识别算法主要有：判别分析方法、支持向量机、决策树、神经网络和遗传算法等。从机器学习的角度看，作者身份识别是典型的多类别、单标签的文本分类任务，基于向量空间模型可以处理高维、稀疏、含有噪声的数据，尤其是支持向量机能够避免高维特征集与训练样本过拟合问题，被公认为最有效率的算法之一[3]。其他算法的研究也取得了多项成果，例如 J. Fréry 等以优化决策树算法在 2014 年剽窃、作者身份和社会软件滥用国际研讨会(International Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, PAN)作者身份识别评测任务中获综合性能第二名[26]。

基于相似度建模的典型算法是最近邻法，代表研究为 J. Burrows 采用的 Delta 方法，该方法应用于样本长度在 1 500 词以上文学作品的效果较好，但其性能随样本长度的缩短而迅速下降[27]。类似研究有 R. Cilibrasi 等以标准化压缩距离(normalized compression distance)度量文本相似度[28]，E. Castillo 等组合 LSA、Jaccard 相似度、Euclidean 距离、Chebyshev 距离和 Cosine 相似度判断文本相似性等[29]，后者在 2014 年 PAN 国际公开评测作者身份识别任务中综合性能排名第三。

元学习方法的思路是在全局范围内对局部学习获取的知识进行再次学习，即基于元知识的再次学

习。代表研究有 M. Koppel 等在 Unmasking 方法[30]中为每一个匿名文本建立将其与候选作者训练文本区分的 SVM 分类器，在随后的迭代过程中计算每次移除一些特征后的准确率下降值，下降最快的被认为是真实作者，此方法要求训练文本足够长以便能够切分成若干样本用于训练 SVM 分类器。特别值得注意的是 M. Khonji 等基于元学习思想改进的 General Impostors 方法[31]赢得了 2014 PAN 评测任务综合性能第一名，其不足之处在于时间复杂度较高。

3.3 述评

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。结合当前网络应用环境，文体风格特征需要适应两个变化：

一是大数据环境下的网络文本作者身份识别。候选作者数和训练样本数是作者身份识别的重要参数，网络文本的候选作者数量巨大，而每位作者可获得的训练样本非常有限，现有文体风格特征的适用性还需进一步探讨。针对这一需求，可以通过在传统多层面特征集上补充结构特征提高作者身份识别正确率。结构特征包括文本组织和布局相关的特征，涉及 HTML 标记分布、字体大小和颜色分布、段落数、段落长、平均句长等，尤其是与 Email、博客或微博等特定语料相关的一些结构特征，如表情符号及其出现位置，嵌入图像、超链接、引用、签名档、致敬语和告别语等，在短文本上尤为有效[11,17]。

二是文体风格特征需要适应跨语言环境。网络的国际化决定了在多语种环境下作者身份识别的必要性，而文体风格特征与语种高度相关，跨语言环境的文体风格特征需要细致研究。目前适应跨语言环境的有效途径主要是抽取底层的字符或词汇文体特征，例如字符 Ngram 和词汇 Ngram 等[24]。

总体上，基于实例的建模方法能适应网络文本，可以采用支持向量机等机器学习方法处理高维、噪声和稀疏的网络短文本语料，其优势在于对文本长度要求不高，但目前仍存在准确率依赖于训练样本数量、训练过程时间复杂度较高的问题。

4 性能评价指标

作为多类别分类问题，作者身份识别常用的性能评价指标包括正确率(Accuracy)、查全率(Recall)、查准率(Precision)、F 测量(F-measure)、AUC(Area Under Roc Curve)、C@1 指标和运行时间等。

正确率是分类模型做出正确预测的样本数占总样本数的比例，是最为通用的性能指标，在作者身份识别文献中被广泛采用作为性能比较的依据[10-11,13]。

查准率是指预测结果中某个特定作者正确预测的样本数占实际分为该作者样本数的百分比，可以考察各个作者的预测准确率。正确率与查准率的区别在于：正确率检测考察分类模型在所有类别上的预测结果偏离真实值的程度；查准率考察特定类别上分类模型的正确性。

查全率是指预测结果中正确判断属于某个作者的样本数占应该属于该作者样本数的百分比，可以考察对各个作者样本识别的完备性。查全率和查准率可以从各个类别衡量作者身份识别的完备性和准确率，适合考察算法的局部性能。

F 测量值可以平衡考察查全率和查准率来衡量分类算法的整体效果[11,32-33]。理想的分类结果是查全率和查准率越高越好。但查全率升高意味着完备性提高，分类模型要为所有的实例给出类别预测结果，即使是对某些实例并没有足够的依据给出明确的分类预测。因此在现实中随着查全率的升高，查准率通常呈下降趋势。F 测量值是二者的调和平均值，计算公式为：

$$\text{F 测量值} = \frac{2 \times \text{查全率} \times \text{查准率}}{\text{查全率} + \text{查准率}} \quad (1)$$

AUC 计算以 False Positive Rate 为横坐标，以 True Positive Rate 为纵坐标的 ROC 曲线下的面积，能够考察样本在不同类别上不平衡分布情况下算法的综合性能，由于现实世界中训练文本的作者分布通常是不平衡的，采用 AUC 作为性能指标是一个主要的趋势，因此近年的 PAN 评测和相关文献都普遍以 AUC 为主要的评测参数[31,33]，例如文献[32]同时采用了 F 测量值、查全率、查准率、AUC 作为性能指标。

在 PAN 等国际测评任务中还引入 C@1 指标来综合衡量未得到明确作者预测情况下的正确率，并将 C@1 与 AUC 的乘积作为评测名次的排序依据。C@1 最初是用于问答系统的性能指标[31,34]，其优点是能够将不能明确分类的样本计入评测性能，可以鼓励作者身份识别算法对样本做出明确的预测，减少预测的不确定性。计算 C@1 指标时首先将概率预测值为 0.5 以上的结果转换成“肯定”，反之为“否定”，预测值等于 0.5 的为“未预测样本”，设 n 为待预测样本数，n_c 为正确预测样本数，n_u 为未预测样本

数，则 C@1 的计算如公式(2)所示，其值越高，算法的综合性能越好。

$$c@1 = \frac{1}{n}\left(n_c + \frac{n_c}{n}n_u\right) \quad (2)$$

此外，运行时间也是一个重要的性能指标，尤其在处理大数据集时，算法的时效性决定了算法的实际应用可行性[31－32]。

5 总结与展望

现有研究为作品作者身份识别奠定了坚实的基础，同时也面临着互联网环境的一系列挑战：与文学作品相比，近年来网络文本大量涌现，社交媒体、电子邮件、博客、微博、在线论坛中的网络文本的长度有限，语言形式丰富，大量使用缩略语、标点符号和首字母缩写词等，导致传统的鉴别特征失效；网络环境使得匿名文本的潜在作者数量巨大，每位作者可获得的训练样本又是非常有限的，使得建立作者身份识别模型的难度大大增加；此外，文体风格特征的选择与语言环境是高度相关的，多语言环境成为网络文本身份识别的主要难点，目前尤其缺乏中文作者身份识别的研究。

综上所述，今后作者身份识别的研究趋势为：结合传统的多层面特征，针对网络文本特性抽取更具有表现和刻画能力的文体风格特征，借助信息检索、机器学习和自然语言处理领域的最新研究成果，提高处理海量、稀疏、高维和多语种文本作者身份识别的准确率和效率。

参考文献：

[1] Mendenhall T C. The characteristic curves of composition[J]. Science，1887(214S)：237－246.

[2] Holmes D I. The evolution of stylometry in humanities scholarship[J]. Literary and Linguistic Computing，1998，13(3)：111－117.

[3] Stamatatos E. A survey of modern authorship attribution methods[J]. Journal of the American Society for Information Science and Technology，2009，60(3)：538－556.

[4] Yule G U. On sentence-length as a statistical characteristic of style in prose：With application to two cases of disputed authorship[J]. Biometrika，1939：363－390.

[5] Baayen H，Van Halteren H，Tweedie F. Outside the cave of shadows：Using syntactic annotation to enhance authorship attribution[J]. Literary and Linguistic Computing，1996，11(3)：121－132.

[6] Zhao Ying，Zobel J. Effective and scalable authorship attribution using function words[C]//Information Retrieval Technology. Berlin Heidelberg：Springer，2005：174－189.

[7] Goebel R，Wahlster W. Using dependency-based annotations for authorship identification[C]//Text，Speech and Dialogue. Berlin Heidelberg：Springer，2012：314－319.

[8] Hassan F H，Chaurasia M A. Author assertion of furtive write print using character n-grams[C]//International Conference on Future Information Technology IPCSIT. Singapore：IACSIT PRESS，2011：212－216.

[9] Gamon M. Linguistic correlates of style：Authorship classification with deep linguistic analysis features[C]//Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg：Association for Computational Linguistics，2004：611－617.

[10] Zhang Chunxia，Wu Xindong，Niu Zhendong，et al. Authorship identification from unstructured texts[J]. Knowledge－Based Systems，2014：99－111.

[11] Abbasi A，Chen H. Applying authorship analysis to extremist-group web forum messages[J]. IEEE Intelligent Systems，2005，20(5)：67－75.

[12] Iqbal F，Binsalleeh H，Fung B C M，et al. Mining writeprints from anonymous e-mails for forensic investigation[J]. Digital Investigation，2010，7(1)：56－64.

[13] Ali N，Price M，Yampolskiy R. BLN-Gram-TF-ITF as a new feature for authorship identification[EB/OL]. [2015－07－02]. http://www.ase360.org/bitstream/handle/123456789/130/Poster60.pdf? sequence=1&isAllowed=y.

[14] Fan Mengdi，Qian Tieyun，Chen Li，et al. Authorship attribution with very few labeled data：A co-training approach[C]//Web-Age Information Management. Berlin：Springer International Publishing，2014：657－668.

[15] Zamani H，Esfahani H N，Babaie P，et al. Authorship identification using dynamic selection of features from probabilistic feature set[C]//Information Access Evaluation. Multilinguality，Multimodality，and Interaction. Berlin：Springer International Publishing，2014：128－140.

[16] 吕英杰，范静，刘景方. 基于文体学的中文 UGC 作者身份识别研究[J]. 现代图书情报技术，2013(9)：48－53.

[17] Zheng Rong，Li Jiexun，Chen Hsinchun，et al. A framework for authorship identification of online messages：Writing style features and classification techniques[J]. Journal of the American Society of Information Science and Technology，2006，57(3)：378－393.

[18] 胡壮麟. 理论文体学[M]. 北京：外语教学与研究出版社，2000：50－63.

[19] 祁瑞华，杨德礼，郭旭，等. 基于多层面文体特征的博客作者身份识别研究[J]. 情报学报，2015.6：628－634.

[20] McCarthy P M，Lewis G A，Dufty D F，et al. Analyzing writing styles with coh－metrix[C]//Proceedings of the Florida Artificial Intelligence Research Society International Conference. Menlo Park，California，USA：AAAI Press，2006：764－769.

[21] 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[J]. 中文信息学报，2006，20(6)：61－68.

[22] Peng Fuchun, Shuurmans D, Wang Shaojun. Augmenting naive Bayes classifiers with statistical language models [J]. Information Retrieval Journal, 2004. 7(1), 317 - 345.

[23] Marton Y, Wu Ning, Hellerstein L. On compression - based text classification [C] //Proceedings of the European Conference on Information Retrieval. Springer, Berlin German. 2005: 300 - 314.

[24] Keselj V, Peng Fuchun, Cercone N, et al. N - gram - based author profiles for authorship attribution [C] //Proceedings of the Pacific Association for Computational Linguistics. PACLING, Canberra, Australia. 2003: 255 - 264.

[25] Frantzeskou G, Stamatatos E, Gritzalis S, et al. Effective identification of source code authors using byte-level information [C] //Proceedings of the 28th International Conference on Software Engineering. New York: ACM, 2006: 893 - 896.

[26] Fréry J, Largeron C, Juganaru-Mathieu M. UJM at CLEF in author identification [J]. Notebook for PAN at CLEF, 2014, 1180: 1042 - 1048.

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

[28] Cilibrasi R, Vitanyi P. Clustering by compression [J]. Information Theory, IEEE Transactions on, 2005, 51(4): 1523 - 1545.

[29] Castillo E, Cervantes O, Vilariño D, et al. Unsupervised method for the authorship identification task [J]. Notebook for PAN at CLEF, 2014, 1180: 1035 - 1041.

[30] Koppel M, Schler J, Bonchek-Dokow E. Measuring Differentiability: Unmasking Pseudonymous Authors [J]. Journal of Machine Learning Research, 2007, 8(2): 1261 - 1276.

[31] Khonji M, Iraqi Y. A slightly-modified GI-based author-verifier with lots of features [J]. Notebook for PAN at CLEF, 2014, 1180: 977 - 983.

[32] Gollub T, otthast M P, ABeyer A, et al. Recent trends in digital text forensics and its evaluation [C] //Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Berlin Heidelberg: Springer, 2013: 282 - 302.

[33] Potha N, Stamatatos E. A profile-based method for authorship verification [C] //Artificial Intelligence: Methods and Applications. Berlin: Springer International Publishing, 2014: 313 - 326.

[34] Peñas A, Rodrigo A. A simple measure to assess non-response [C] //Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011(1): 1415 - 1424.

作者贡献说明:

祁瑞华: 提出研究思路,设计研究方案,论文起草和最终版本修订;

霍跃红: 设计研究方案、收集和分析文献;

胡润波: 收集和分析文献。

Review on Text Authorship Identification

Qi Ruihua[1] Huo Yuehong[2] Hu Runbo[3]

[1]Computer Education Department, Dalian University of Foreign Languages, Dalian 116044

[2]School of English Studies, Dalian University of Foreign Languages, Dalian 116044

[3]China Business Executives Academy, Dalian 116024

Abstract: **[Purpose/significance]** The traditional authorship identification methods are not applicable to web text. In this paper some typical methods and the key problems in recent years are reviewed in order to provide new ideas for further research. **[Method/process]** We objectively analyzed the authorship stylistic features selection, the authorship modeling and the performance evaluation indexes respectively, presenting the latest development of the related areas and trends. **[Result/conclusion]** Authorship identification should adapt to short, non - standard, mass, high - dimensional, sparse and multilingual text. More efficient multidimensional features models and corresponding authorship identification methods are required. The latest achievements in information retrieval, machine learning and natural language processing are the promising solutions to improve the efficiency and accuracy of authorship identification.

Keywords: authorship identification stylistic features authorship modeling performance evaluation indexes