文章编号：2095-2163(2021)01-0014-06 中图分类号：TP391 文献标志码：A

基于 fastText 的可视化作者归属模型

李 逍，顾长贵，杨雷鑫，陆祺灵
（上海理工大学 管理学院，上海 200093）

摘 要： 基于滑动窗口的方法，结合机器学习分类技术，可以判定文本的作者归属。但是此类方法需要精心挑选对应的文本特征，不同的文本特征选取可能会影响判定结果。针对以上问题，提出了一种基于快速文本分类（fastText）的文本作者归属判定模型。该模型融合滑动窗口的思想，引入词（字）向量、数据增强技术，从而充分利用文本信息、自动提取文本特征，并且以可视化的方式将结果呈现出来。使用该模型来检测《红楼梦》、《Roman de la Rose》的作者归属，实验结果表明《红楼梦》的前八十回与后四十回为不同作者所著、《Roman de la Rose》开篇 4 058 行（约 50 000 字）与后面 17 724 行（约 218 000 字）为不同作者所著。证明了 Rolling-fastText 模型判定文本作者归属的有效性。

关键词： 滑动窗口；作者归属；快速文本分类器；数据增强技术；可视化

A visual model of authorship attribution based on fastText

LI Xiao, GU Changgui, YANG Leixin, LU Qiling
(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 Some methods are based on sliding window and machine learning, which can determine the authorship attribution of text. However, these methods require careful selection of text features, and different text features may affect the outcome of the authorship attribution. In response to the above problems, this paper proposes a model based on fastText classification to determine authorship attribution. The model incorporates the idea of the sliding window, introduces word (character) vectors and data enhancement technology, so as to make full use of text information and extract text features automatically, and presents the results in a manner of visualization. Finally, this paper uses the model to detect the authorship attribution of《A Dream of Red Mansions》and《Roman de la Rose》. The experimental results show that the first 80 chapters and the last 40 chapters of《A Dream of Red Mansions》are written by different authors, the opening 4 058 lines (approximately 50 000 words) and the following 17 724 lines (approximately 218 000 words) of《Roman de la Rose》are written by different authors. It is proved that this model is effective to determine the authorship attribution.

【Key words】 sliding window; authorship attribution; fast text classifier; data enhancement technology; visualization

0 引 言

在文体学中，文本作者归属判定是指从有争议的文本中提取作者风格信息，然后在一组“候选者”中确定最佳匹配的过程；题材识别则是找到一组在风格上相似的作品。在研究这些问题时，许多经典方法的目标都是计算语料库中文本之间的相似度，以便发现隐藏的模式或规律[1]。

早在 20 世纪上半叶就有学者开始使用统计学方法对文本进行作者归属研究[2]，在研究《联邦党人文集》中 12 篇存在作者争议的文章时，Mosteller 等人[3]提出了一种以贝叶斯公式为核心的分类算法，基于文本中的文体特征，如句子长度、词长或高频虚词的分布，推断出 12 篇文章的作者。

机器学习建立在统计学的基础之上，近年来该领域发展迅猛，因而众多机器学习的方法也被用到了文本作者归属这一任务[4]。施建军[5]运用支持向量机技术（SVM）抽取《红楼梦》文本的虚字作为特征，认为该书的前八十回与后四十回分别为不同人所著。但是，现今《红楼梦》流传多种版本，每个版本的虚字字数也不尽相同，可能会影响实验结果。Burrows[6]提出 Delta 分类技术，用于分析文本内部差异，判定文本作者。Maciej[7]基于文本的大小分析 Delta 技术的可操作性，Rybicki 等人[8]将单词频率作为特征来分析 Delta 技术的有效性。Juola 等人[9]创建任务型软件 JGAAP3.0，来分析文本作者

基金项目： 国家自然科学基金（11875042）；上海理工大学大学生创新创业计划资助项目（SH2020072）。

作者简介： 李 逍（1997-），男，硕士研究生，主要研究方向：自然语言处理、机器学习；顾长贵（1982-），男，博士，副教授，博士生导师，主要研究方向：复杂网络、系统科学；杨雷鑫（1997-），女，本科生，主要研究方向：人工智能；陆祺灵（1998-），男，本科生，主要研究方向：自然语言处理。

通讯作者： 李 逍 Email：1253487438@qq.com

收稿日期： 2020-11-13

归属问题。以上所采用的模型虽然可以快速、准确地判断文本的作者归属，但是需要精心挑选文本的特征，只有通过大量的实验与参数调整，才能找到符合某文本的对应特征，进而得到相对准确的结果。

作为机器学习的子领域，深度学习也逐渐被越来越多的学者用来进行作者归属判定。Bagnall[10]使用多头递归神经网络（RNN）语言模型在 PAN2015 作者身份识别任务中估计每个作者的字符级概率，其结果优于其他模型。但是 RNN 模型面临着梯度消失与梯度爆炸问题。Julian 等人[11]运用卷积神经网络在一个科学出版物数据集上执行作者身份识别，能够预测同一学科的科学出版物的作者。Dainis 等人[12]探讨了卷积神经网络（CNNs）在多标签作者归属（AA）问题中的应用，并介绍了一种专门针对此类任务而设计的 CNN 模型。但是 CNN 模型无法提取文本序列的时序信息。深度学习技术在提取文本特征时更依赖于隐藏层，可由模型自动提取特征。但是深度学习对样本的数量要求较大。当样本数量较少时，很难训练出有效的模型。而且无论是递归神经网络、还是卷积神经网络，当样本长度较长、维数较大时，结果显示判定速度都较慢。上面所讨论的模型都是以相应的数据指标展示实验结果，然而将结果以图的形式可视化，可以展示文本的风格变化，具有很强的解释性。

如果将文本分割成样本，这些样本就可被视作基本的文学作品，因此语篇的局部部分也可能显示出某些文体特征的线性发展。基于这一点，带有滑动窗口序列分析方法逐渐应用到文本作者归属分析上。孙龙龙等人[13]把《红楼梦》文本转化为时间序列数据，运用滑动窗口的方法计算序列数据的赫斯特指数，从而得到《红楼梦》不同部分的作者归属。van 等人[14]在研究中世纪荷兰亚瑟王史诗《Roman van Walewein》时也成功应用了滑动窗口的概念，该研究是通过滑动窗口从《Roman van Walewein》中生成一系列子样本，以测试这些子样本是否在整个文本中的风格一致。

通过时间线很容易来表示文本内部的风格发展并且可视化。Maciej 等人[15-16]基于此点，提出了 Rolling Attribution 模型，该模型可用于解决某些争议作品的作者归属性问题。Rolling Attribution 模型将机器学习分类模型与序列分析思想相结合，通过对文本的线性切块，来查看文本内部的风格是否一致。该模型的目标不是对给定的整个文本进行作者归属判定，而是对每个通过滑动窗口产生的样本进行独立的归属判定，再将结果作为一系列有序的风格信号进行可视化比较。

但是，Rolling Attribution 模型需要精心提取文本的某些特征，如提取文本最频繁使用的字的频率、虚词出现的次数等等作为文本特征，继而采用机器学习算法（如支持向量机等）对每个片段进行作者归属判定。在提取文本特征时工作繁琐，不同的语言可能需要对应不同的特征提取，如果提取的某些特征不适合该文本或者统计错误，就可能得到错误的结果。基于以上问题，本文对 Rolling Attribution 模型进行修改，提出了 Rolling-fastText 模型，该模型具有以下特点：

（1）无需精心挑选文本特征，避免特征选取问题对测试结果造成影响。

（2）引入词向量，发挥词向量的优势，充分提取文本信息。

（3）可以解决小样本以及样本数据不平衡问题。

1　Rolling-fastText 模型

本节将分别对 Rolling-fastText 模型所涉及的相关技术进行介绍，对此拟做阐释分述如下。

1.1　滑动窗口

滑动窗口示意如图 1 所示。由图 1 可知，Rolling-fastText 模型所运用滑动窗口方法是指：以 S 个字（单词）为一窗口对文章进行切块，生成样本，允许样本之间可以重叠（重叠字（单词）数为 d），使文本中的某些字（单词）可以重复使用，以提高信息提取的完整性。

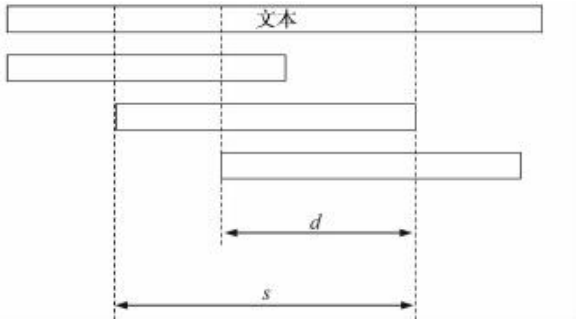

图 1　滑动窗口示意图

Fig. 1　Schematic diagram of sliding window

1.2　fastText 模型

Rolling-fastText 模型需要对样本进行类别判定，因此该模型基于 fastText 模型，对样本进行快速分类。fastText 模型是由 Armano 等人[17]提出的一种快速文本分类方法，该算法结构类似于 Mikolov

等人[18]的连续词袋模型(CBOW 模型)。fastText 模型把连续词袋模型中的中间单词由标签替换,在此基础上使用激活函数 f 来计算样本类别的概率分布。对于 N 个样本的集合,其损失函数为:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(f(\boldsymbol{BA}\boldsymbol{x}_i)), \quad (1)$$

其中,$\boldsymbol{x}_i$ 表示第 i 个样本向量化矩阵;$\boldsymbol{B}$、$\boldsymbol{A}$ 皆为权重矩阵;y_i 为样本标签。

同时,fastText 通过 n-gram 模型[19]提取样本的时序信息,丰富样本语义。本文中,是将原始样本与提取时序信息之后的样本拼接,并加以向量化,既能利用样本的时序信息,又能保留样本的原始信息。其损失函数为:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(f(\boldsymbol{BA}(\boldsymbol{x}_i, \boldsymbol{x}_{i,n}))). \quad (2)$$

其中,$\boldsymbol{x}_{i,n}$ 表示第 i 个样本提取时序信息后的向量化矩阵。

1.3 EDA 技术

由于文本年代久远、文本作者无其他作品等原因,可能会很难找到某些作者其他的大量的相关作品作为训练语料库,因此当研究中只有少量样本时,或者样本类型不平衡时,就很难得到较为理想的结果。为了解决上述问题,本文用到了一种数据增强技术:EDA[20](Easy Data Augmentation)。在随后的内容中,将会分析讨论 EDA 技术的主要细节。对于训练语料库中给定的句子,随机进行如下的 4 类操作之一:

(1)同义词替换(Synonyms Replace,SR):不考虑停用词,在句子中随机抽取 n 个词,再随机抽取该词同义词将其替换。

(2)随机插入(Randomly Insert,RI):不考虑停用词,随机抽选一个词,再随机抽取该词同义词,随机插入原句子中。

(3)随机交换(Randomly Swap,RS):随机交换句子中的 2 个词。

(4)随机删除(Randomly Delete,RD):以概率(可自行设定)随机删除句子中的每个词。

通过多次使用 EDA 技术,可以扩大样本数量,均衡样本类别。以上 4 种操作无操作先后顺序。

1.4 可视化

Rolling-fastText 模型检测的文本结果是通过可视化来展现的。可视化的目的是方便人们观察文本不同片段的作者归属,同时也展示了该片段的作者归属的概率,在视觉上强调该片段最有可能为哪位作者所著(下部条纹),并把不太可能的作者保留在阴影区用作参考(上部条纹)。例如,假设样本 x 可能属于作者"1"或者作者"0",当:

$$p(x=1) > p(x=0). \quad (3)$$

判断样本 x 属于作者"1",并且用下部条纹展示 $P(x=1)$、用上部条纹展示 $P(x=0)$(条纹宽度表示概率),反之亦然。在第 3 节实验结果分析时,只分析下部条纹。

2 Rolling-fastText 模型的训练与检测流程

2.1 模型训练流程

在训练 Rolling-fastText 模型时,首先使用滑动窗口把训练语料库分成 N 个子样本,通过 EDA 技术扩大样本数量、均衡样本类别,提取扩大后样本的时序信息形成新的样本,然后将 2 种样本拼接并且向量化,经过隐藏层之后计算损失函数,从而更新模型参数。模型训练流程见图 2。Rolling-fastText 模型的损失函数为:

$$Loss = -\frac{1}{\sum_{i=1}^{N} M_i}\sum_{i=1}^{N} y_i \sum_{j=1}^{M_i} \log(f(\boldsymbol{BA}(\boldsymbol{x}_j, \boldsymbol{x}_{j,n}))). \quad (4)$$

其中,第 i 个样本通过 EDA 技术扩展了 M_i 个样本(包括样本 i),$\boldsymbol{x}_j$ 表示其中的某一个样本的向量化矩阵。

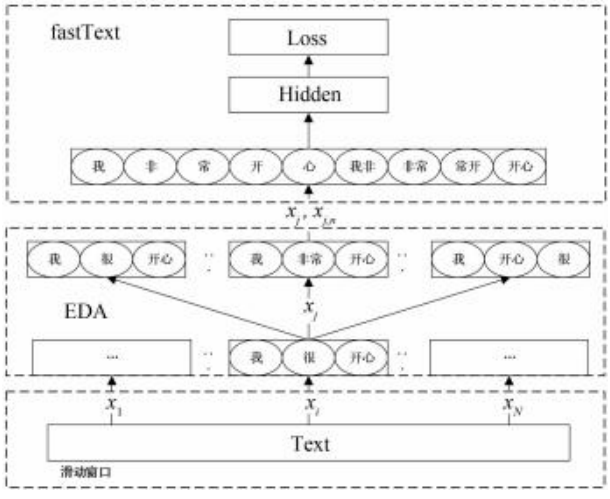

图 2 模型训练流程

Fig. 2 The process of model training

2.2 模型检测流程

模型检测流程如图 3 所示。由图 3 可见,当使用训练好的 Rolling-fastText 模型检测文本时,首先使用滑动窗口把待检测文本分成若干个子样本,这些样本无需经过 EDA 技术进行数据增强,直接提取

样本的时序信息生成新的样本，然后将其与原始样本拼接之后词（字）向量化，经过隐藏层，最终输出结果。该模型的输出结果为：

$$Output = f(BA(\boldsymbol{x}_i, \boldsymbol{x}_{i,n})) . \quad (5)$$

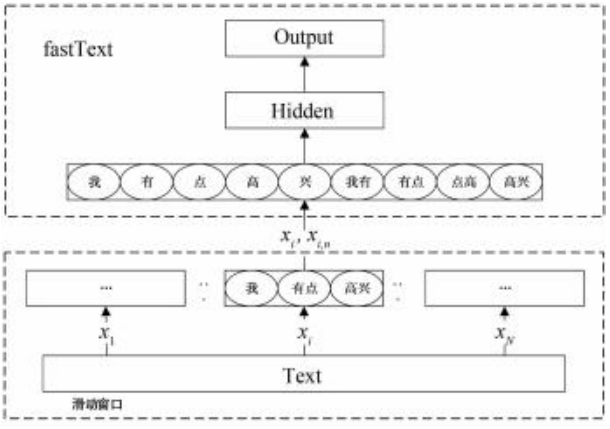

图 3　模型检测流程

Fig. 3　The process of model detection

3　Rolling-fastText 模型应用

本节将进行 Rolling-fastText 模型在《红楼梦》、《Roman de la Rose》两个案例上的应用研究。作为对比，本文同时用 Rolling Attribution 模型（SVM 作为分类器）分析这些案例。具体研究内容详述如下。

3.1　《红楼梦》的检测研究

《红楼梦》是中国古代章回体小说，中国古典四大名著之一，关于其作者一直存在争议。有不少学者通过各种方法证明《红楼梦》的前八十回与后四十回为不同人所著[5,21]。这些研究都为本文通过 Rolling-fastText 模型检测《红楼梦》提供了较好的范例。

本节使用的《红楼梦》的版本为程乙本，为了使模型能够快速、准确地判断样本的类别，同时也能充分利用样本的文本信息，经过测试，本节设定 S = 5 000，d = 500。因为很难找到《红楼梦》的作者的其他相关作品，所以本文与文献[5]相同，假设《红楼梦》的第一回到第十回（黑色）和第一百一十一回到第一百二十回（灰色）为不同作者所写，将其作为训练集来训练 Rolling-fastText 模型。而后用 Rolling-fastText 模型在整本《红楼梦》上进行测试，得到整本书的内部片段的作者归属。为了方便观察《红楼梦》的作者归属，本文在《红楼梦》的每二十回都会插入一条分界线。分界线之间的内容见表 1。实验结果如图 4 所示。本文运用同样的方法训练 Rolling Attribution 模型（提取样本中的 44 个虚词出现频率作为样本特征，如“之”等等[5]）将其对《红楼梦》进行测试，所得结果如图 5 所示。

表 1　《红楼梦》中分界线之间文本内容

Tab. 1　Text content between the dividing lines in 《A Dream of Red Mansions》

分界线	内容
-a	《红楼梦》第一回至第二十回
a-b	《红楼梦》第二十一回至第四十回
b-c	《红楼梦》第四十一回至第六十回
c-d	《红楼梦》第六十一回至第八十回
d-e	《红楼梦》第八十一回至第一百回
e-	《红楼梦》第一百零一回至第一百二十回

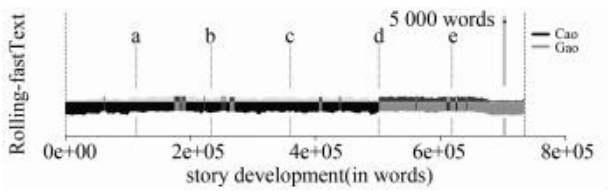

图 4　Rolling-fastText 模型检测《红楼梦》

Fig. 4　The detection of 《A Dream of Red Mansions》 by Rolling-fastText model

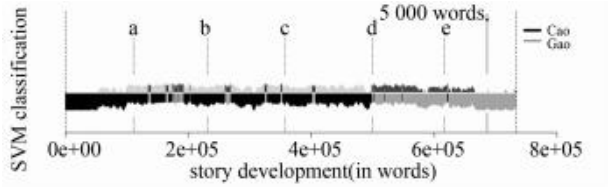

图 5　Rolling Attribution（SVM）模型检测《红楼梦》

Fig. 5　The detection of 《A Dream of Red Mansions》 by Rolling Attribution（SVM） model

由图 4、图 5 可见，2 种模型的结果大致相同，分界线 d 之前大部分为黑色，分界线 d 之后大部分为灰色，风格发生了改变，分界线 d 位于《红楼梦》第八十回与第八十一回之间。因此，利用这两个模型可以判断《红楼梦》的前八十回与后四十回分别为不同作者所著。另外，结果显示有噪声的出现，前八十回中某几个片段被判断为灰色、后四十回少许片段被判断为黑色。这可能是因为本文所用来测试的《红楼梦》版本为后世传抄本，可能传抄者在某些内容上做了少许的修改导致有噪声的出现。

3.2　《Roman de la Rose》的检测研究

《Roman de la Rose》是 13 世纪法国诗歌，学者普遍认为这本书的第一部分是由 Guillaume de Lorris 在 1230 年左右完成的，而第二部分是 Jean de Meun 在 1275 年左右完成的。并且该书中 2 部分的转折点很清楚，Guillaume de Lorris 是该书开篇4 058

行(约 50 000 字)的作者,Jean de Meun 是后面 17 724行(约 218 000 字)的作者[18]。本节选用《Roman de la Rose》的文本为 Marteau 的版本,可在古腾堡计划网站上公开获取(Marteau, 1878)。

Maciej Eder 已经运用 Rolling Attribution 模型对《Roman de la Rose》做了测试。过程中选取该书开头 10 000 个单词(大约 1 000 行)、该书中间 10 000 个单词(第 113 000 至 123 000 之间的单词)用作语料库训练模型,然后测试《Roman de la Rose》(实验参数: $S = 5\ 000$, $d = 500$),提取文本中前 100 个出现频率最高的单词的频数作为特征,可参阅文献[15]),得到结果如图 6 所示。本文同样使用上述语料库与实验参数训练 Rolling-fastText 模型,然后测试《Roman de la Rose》(实验参数: $S = 5\ 000$, $d = 500$),所得结果如图 7 所示。图 6、图 7 显示的分界线之间的内容见表 2。

表 2 《Roman de la Rose》中分界线之间文本内容

Tab. 2 Text content between the dividing lines in《Roman de la Rose》

分界线	内容
-a	《Roman de la Rose》开头 10 000 个单词
a-b	《Roman de la Rose》开头 5 000 行(不包括开头10 000 个单词)
b-c	《Roman de la Rose》第 5 000 行至该书结束(不包括第 113 000 至 123 000 之间的单词)
c-d	《Roman de la Rose》第 113 000 至 123 000 之间的单词

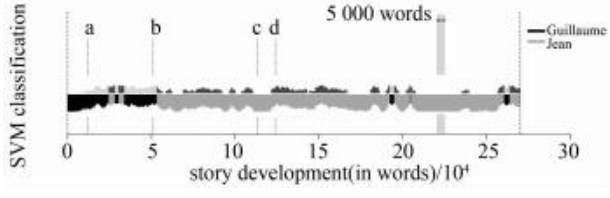

图 6 Rolling Attribution (SVM) 模型检测《Roman de la Rose》[15]

Fig. 6 The detection of《Roman de la Rose》by Rolling Attribution (SVM) model[15]

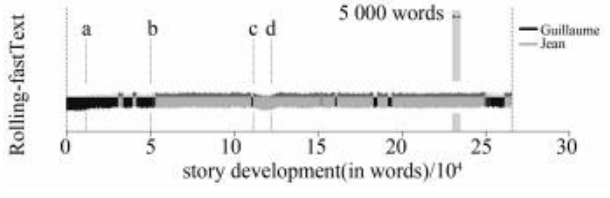

图 7 Rolling-fastText 模型检测《Roman de la Rose[15]》

Fig. 7 The detection of《Roman de la Rose》by Rolling-fastText model[15]

由图 6、图 7 可以看出,2 种模型都以 b 为分界线,较为准确地判断出了不同作者所写的部分。但是分界线 b 之前的内容仍有少许部分被判断为 Jean de Meun 所著、文章结尾也有少许内容被判断为 Guillaume de Lorris 所著。这种情况可能由以下原因造成:

(1) Jean de Meun 可能对 Guillaume de Lorris 所写内容作了某些修改。

(2) Jean de Meun 的写作风格可能受到了 Guillaume de Lorris 的影响。

(3) 由于成书年代距今较远,后人在整理出版时也可能会对内容有所修改。

总之,2 种模型都可以得到较为准确的结果,并且能够清晰地展示出《Roman de la Rose》两部分的转折点。

4 结束语

本文基于 fastText 方法,融合滑动窗口的思想,结合词(字)向量与数据增强技术,提出一种名为 Rolling-fastText 模型。该模型通过隐藏层提取样本特征,利用文本信息,同时扩大样本数量、平衡样本类别,能够快速、准确地对样本进行分类。运用 Rolling-fastText 模型在《红楼梦》与《Roman de la Rose》案例上做了实验,得出的结果与滚动归属法的结果大致相同,证明了 Rolling-fastText 模型的实用性与准确性。

《红楼梦》的前八十回与后四十回都在讲述同一个故事,具有相同的时代背景,后四十回更是继承了前八十回的故事情节,其中出场的人物大致相同,并且故事的场景大都是在"宁国府"、"荣国府"。因此,本文发现的《红楼梦》前八十回和后四十回的不同,可能是由于不同作者的写作风格导致的,而不是故事情节导致的。

在《Roman de la Rose》第一部分中,Guillaume de Lorris 叙述了"情人"追求心爱的女人,并从爱神那里学到了求爱的艺术,而且遇到了一系列寓言人物。每一种寓言人物都是其爱慕的对象的表现。此后就一起为浪漫爱情的心理学提供了迷人的评论。在 Jean de Meun 续写的《Roman de la Rose》第二部分中,寓言人物,如"自然"、"天才"和"理性",讨论了爱情的哲学,"情人"则达到了其目的,全书情节趋于完善。因此,本文发现的《Roman de la Rose》两部分的不同,可能是由于 Guillaume de Lorris 与 Jean de Meun 不同的写作风格导致的,而不是故事情节导致的。

本文的不足之处是 Rolling-fastText 模型依然依赖监督学习方法，要求有相应的训练集。未来可以将研究重点放在无监督学习上，进一步优化模型，提高模型精度。

参考文献

[1] 王春柳，杨永辉，邓霏，等. 文本相似度计算方法研究综述[J]. 情报科学，2019，37(3)：158-168.

[2] YULE G U. On sentence-length as a statistical characteristic of style in prose：With application to two cases of disputed authorship[J]. Biometrika，1938，30(3-4)：363-390.

[3] MOSTELLER F，WALLACE D L. Inference and disputed authorship：The federalist[J]. Revue de l Institut International de Statistique，1966，34(2)：277-279.

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

[5] 施建军. 基于支持向量机技术的《红楼梦》作者研究[J]. 红楼梦学刊，2011 (5)：35-52.

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

[7] MACIEJ E. Does size matter? Authorship attribution，small samples，big problem[J]. Digital Scholarship in the Humanities，2015，30(2)：167-182.

[8] RYBICKI J，EDER M. Deeper Delta across genres and languages：Do we really need the most frequent words? [J]. Literary and Linguistic Computing，2011，26(3)：315-321.

[9] JUOLA P，NOECKER J，RYAN M，et al. JGAAP3-Authorship attribution for the rest of us [C]//Proceedings of Digital Humanities 2008：Book of Abstracts. Oulu：University of Oulu，2008：250-251.

[10] BAGNALL D. Author identification using multi-headed recurrent neural networks[J]. arXiv preprint arXiv：1506.04891，2015.

[11] JULIAN H，BERG E V D，REHBEIN I. Authorship attribution with Convolutional Neural Networks and POS-Eliding[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. East Stroudsburg：ACL，2017：669-674.

[12] DAINIS B，ZHANG Yifan，MUKHERJEE A. Experiments with Convolutional Neural Networks for multi-label authorship attribution[C]//LREC 2018，11th International Conference on Language Resources and Evaluation. Miyazaki，Japan：LREC，2018：2576-2581.

[13] 孙龙龙，顾长贵，冯靖，等. 四大名著文本中的无标度规律[J]. 上海理工大学学报，2019，41(1)：77-83.

[14] VAN DALEN-OSKAM K，VAN ZUNDERT J. Delta for middle dutch：Author and copyist distinction in Walewein[J]. Literary and Linguistic Computing，2007，22(3)：345-362.

[15] MACIEJ E. Rolling stylometry [J]. Literary & linguistic computing：Journal of the Alliance of Digital Humanities Organizations，2016，31(3)：457-469.

[16] MACIEJ E，RYBICKI J，KESTEMONT M. Stylometry with R：A package for computational text analysis[J]. The R Journal，2016，8(1)：107-121.

[17] ARMANO J，GRAVE E，BOJANOWSKI P，et al. Bag of tricks for efficient text classification [C]//Proceedings of 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia，Spain：ACL，2017：427-431.

[18] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space [C]// Proceedings of the 1st International Conference on Learning Representations. Scottsdale，Arizona：ACL，2013：1-12.

[19] WANG S，MANNING C D. Baselines and bigrams：Simple，good sentiment and topic classification [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg，USA：ACL，2012：90-94.

[20] WEI J W，ZOU K. EDA：Easy data augmentation techniques for Boosting performance on text classification tasks [C]// Proceedings of Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (IJCNLP). Suzhou，China：AACL，2020：6382-6388.

[21] 李贤平.《红楼梦》成书新说[J]. 复旦学报：社会科学版，1987 (5)：3-16.

[22] PENNINGTON J，SOCHER R，MANNING C. Glove：Global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha，Qatar：ACL，2014：1532-1543.

(上接第 13 页)

[7] WANG H，HE H，YANG J，et al. Dual labeling：Answering graph reachability queries in constant time[C]//22nd International Conference on Data Engineering (ICDE'06). Atlanta，Georgia：IEEE，2006：75.

[8] WEI Hao，YU J X，LU C，et al. Reachability querying：An independent permutation labeling approach [J]. The VLDB Journal，2018，27：1-26.

[9] YANO Y，AKIBA T，IWATA Y，et al. Fast and scalable reachability queries on graphs by pruned labeling with landmarks and paths [C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco，CA，USA：ACM，2013：1601-1606.

[10] CHENG J，SHANG Zechao，CHENG Hong，et al. K-reach：Who is in your small world [J]. Proceedings of the VLDB Endowment，2012，5(11)：1292-1303.

[11] BUNDY A，WALLEN L. Breadth-first search [M]// Catalogue of Artificial Intelligence Tools. Symbolic Computation (Artificial Intelligence). Berlin /Heidelberg：Springer，1984：13.

[12] STEIER D M，ANDERSON A P. Depth-first search [M]// Algorithm synthesis：A Comparative Study. US：Springer，1989：47-62.

[13] TANG Xian，CHEN Ziyang，LI Kai，et al. Efficient computation of the transitive closure size[J]. Cluster Computing，2019，22：6517-6527.

[14] ZHOU Junfeng，ZHOU Shijie，YU J X，et al. DAG reduction：Fast answering reachability queries[C]//Proceedings of the 2017 ACM International Conference on Management of Data. Raleigh：ACM，2017：375-390.