情 报 学 报 ISSN 1000 - 0135
第 34 卷 第 6 期 628 - 634, 2015 年 6 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 34 No. 6, 628 - 634 June 2015

doi:10.3772/j.issn.1000-0135.2015.006.008

基于多层面文体特征的博客作者身份识别研究[1)]

祁瑞华[1,2] 杨德礼[1] 郭 旭[2] 刘彩虹[2]

(1. 大连理工大学管理与经济学部,大连 116024; 2. 大连外国语大学计算机教研部,大连 116044)

摘要 传统的文体风格特征模型不适用于当前大量涌现的网络文本。本文针对以博客为代表的网络文本篇幅短小、表达方式丰富灵活的特点,以内容无关为原则,分别抽取字符特征、词汇特征、句法特征和文本布局等特征,建立了由词汇特征、浅层句法特征、深层句法特征和结构特征组成的多层面文体风格特征模型,并选取朴素贝叶斯、决策树、序列最小优化支持向量机和大规模线性分类支持向量机算法在公开博客语料上进行对照实验。实验结果验证了各个层面特征在作者身份识别中的作用,表明了本文方法的准确性、通用性及其在短文本上的鲁棒性。

关键词 文体特征 博客 作者身份

Blogger Identification Based on Multidimensional Stylistic Features

Qi Ruihua[1,2], Yang Deli[1], Guo Xu[2] and Liu Caihong[2]

(1. *Faculty of Management and Economics*, *Dalian* 116024;
2. *Computer Education Department*, *Dalian University of Foreign Languages*, *Dalian* 116044)

Abstract Models for traditional stylistic features are not suitable for Web tents. Based on the principle of content-independent, we extracted character features, lexical features, syntactic features and text layout features, and established a multidimensional stylistic features model which consists of lexical features, shallow syntactic features, deep syntactic features and structure features. We tested the performance of this model with Naive Bayesian, Decision Tree, Sequential Minimal Optimization SVM and LIBLINEAR SVM on public blog corpus. The results verified the contribution of each feature-dimension. The experiments also proved the accuracy, versatility and robustness of the method proposed in this paper.

Keywords stylistic features, blogger, Identification

1 引 言

文本作者身份识别是以文体风格特征量化表示匿名文本作者的无意识写作习惯,并以此为依据自动确定其作者归属的映射过程,近年来已经广泛应用于文学作品、新闻稿、商品评价、垃圾邮件的作者身份鉴定以及计算机取证等领域。当前,随着网络文本的大量涌现,匿名文本的作者身份识别在不良舆情鉴别监控等任务中的重要性与日俱增。

作者身份识别有两个关键问题:文体风格特征表示和身份识别算法。作者身份识别算法效果较好的有支持向量机、朴素贝叶斯、决策树、最近邻法以及神经网络等,其中支持向量机、朴素贝叶斯、决策

收稿日期:2014 年 12 月 16 日。

作者简介:**祁瑞华**,女,1974 年生,博士,副教授,主要研究方向:自然语言处理和文本挖掘。E-mail:rhqi@dlufl.edu.cn。**杨德礼**,男,1939 年生,教授,博士生导师,主要研究方向:管理决策分析。

1) 本文系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);辽宁省高等学校优秀人才支持计划(项目编号:WJQ2013017);大连外国语大学科研项目"基于语言学特征的网络舆情信息挖掘"的研究成果之一。

树及这三种算法的改进算法应用尤为普遍;文体风格特征可以分为字符特征、词汇特征、句法特征、结构特征、语义特征和领域相关特征[1]。其中句法特征能够表达隐含文本结构信息,近年来成为学术界较为集中的探索方向,尤其是基于深层句法分析的文体风格特征成为新的研究热点[2]。

博客作为网络信息发布的主要途径之一,近年来在文本挖掘领域受到了广泛的关注。博客文本长度有限、特殊字符出现频率高、主题和体裁结构变化丰富,为传统的作者身份识别技术提出了新的挑战。为此,本文提出以多层面文体风格特征为基础、以深层句法分析为特点的一种主题无关的博客作者身份识别方法。

2 文体风格特征相关研究

所谓文体风格特征,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特征。理论文体学的作家决定论指出,作品风格产生于作者对自身思想的合理安排[3]。布封和斯皮彻等认为文体实际上是一种个人行为方式,作家在写作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。依据作家决定论可以假设每个作者都有其独特的语言使用特征,关键问题是如何寻找并表示这种特定个体的独有特征。

文体风格特征的定量研究可以追溯到 Mendenhall 依据单词长度对 Shakespeare 和 Bacon 小说风格的分析[4]。早期的文体风格研究主要基于一元特征,代表研究有 Yule 依据句长分析英文作品文体风格[5],李贤平根据虚词使用频率得出《红楼梦》前 80 回的语言风格明显不同于后 40 回的结论[6]等。一元文体风格特征通常仅适用于特定语料,为此学者们开展了多元文体风格特征的研究,如 Baayen 基于重写规则频率语法标注小说语料,比传统的基于词的分析方法获得更高的作者识别准确率[7],Zhao Ying 等采用 365 个功能词作为特征,分析美联社的 TREC 语料库文章的作者[8]等。多元特征增强了文体风格特征的通用性,但准确率和有效性有待提高。因此多层面多元特征成为当前文体风格研究的主要方法,代表性研究有 Gamon 基于语法分析建立多层面组合特征集,在勃朗蒂三姐妹英语小说作者识别中得到验证[9];Abbasi 证实了文本结构特征与传统特征结合能够提高作者身份识别的准确率[10]。

近年来学者们还尝试了基于深层句法挖掘隐含的文体风格,例如 Goebel 采用 DepWords 编码替代传统句法依存关系表示句子,并利用其统计特征识别侦探小说的作者身份[2];Zhang 在 21 本英文作品和路透社语料上抽取了结构化特征、常用词、代词、功能词、POS、短语类型以及依存关系特征,对照实验表明依存关系能够描述基本的、相对稳定的语法模式和谓词参数结构,有助于提高作者身份识别效率[11];此外依存关系还被应用于评论特征观点对抽取、评论极性分类以及文本分类等任务。

博客文本作者身份识别的难点在于短文本的文体风格特征表示,针对这一任务,Abbasi 在滑动窗口中通过 Karhunen-Loeve 变换扑捉文体风格变化构成笔迹特征,对 25 名作者的 Email 和商品评论文本进行识别,获得了 90% 以上的准确率[10];吕英杰抽取 BBS 论坛和博客文本用户生成内容的词汇、特征、结构和内容特征的组合特征集,采用支持向量机获得 80% 左右的作者识别准确率[12]。

综上所述,现有研究为文本作者身份识别奠定了坚实的基础,但是仍然存在一些局限:①以博客为代表的网络文本具有短文本的特殊性,使得传统的文体风格特征数据稀疏,很大程度上影响着作者身份识别的效率和准确率;②现有文体风格特征可分为内容相关和内容无关特征,内容相关特征主要是指从语料中抽取的、能有效表达主题的关键词,这类特征可能会揭示作者在某一主题领域的用词习惯,但当涉及跨主题语料时,需要重新定义和选择关键词以适应新的领域。内容无关特征则避免使用主题相关信息、通常不涉及具体词汇含义。为了提高作者身份识别的准确率,现有研究通常在短文本文体风格特征中包含内容相关特征[10,12],这类特征已经被证明是有效的,但缺乏在不同主题语料上的通用性。为此,本文尝试以内容无关为准则,选择主题无关的、基于字词或句法的统计特征和基于结构的文体风格特征。

3 多层面文体风格特征模型

设所有可能的作者集合为 $A = \{a_1, a_2, \cdots, a_k\}$,对于一组文本 $T = \{T_1, T_2, \cdots, T_r\}$,作者身份识别任务可以描述为给每个文本 T_i 指定一个最可能的作者 a_i 的过程,其中 $a_i \in A$。为完成这一任务,首先要将自然语言描述的非结构化的文本映射到特征向量表示空间。文体特征向量表模型应该具

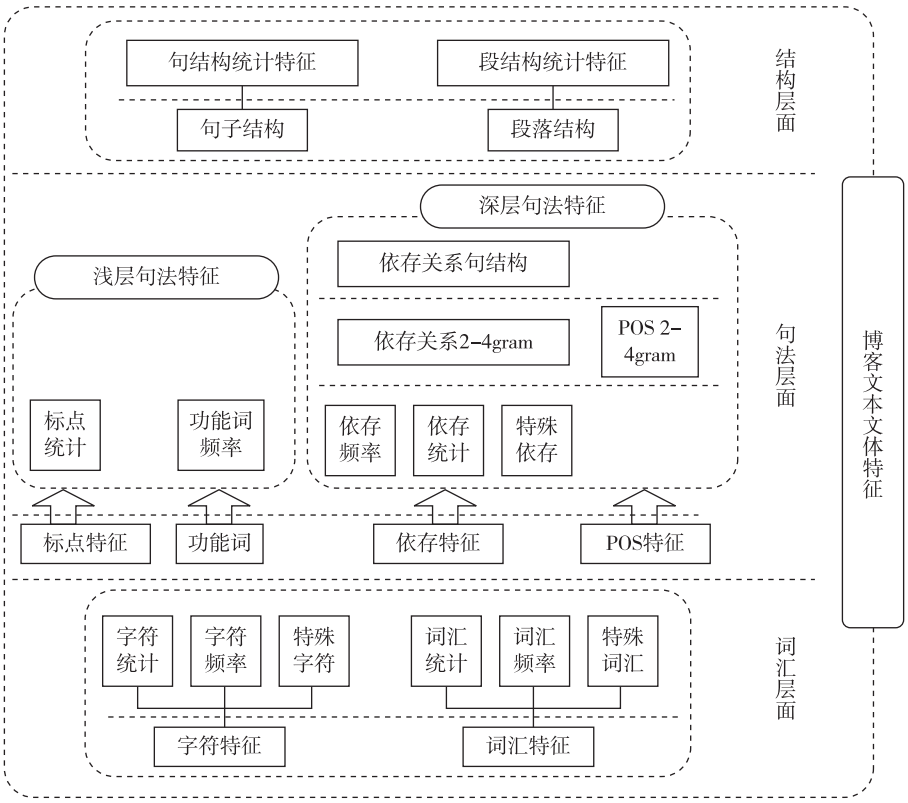

图 1　博客文本文体风格特征模型

有区分目标文本和其他文本的能力,特征值应该具有较好的可获取性。本文在博客文体风格特征模型中引入依存关系和词性标注特征,结合以往研究成果建立了以深层句法分析为特点的,包括词汇特征、浅层句法特征、深层句法特征和结构特征的多层面文体风格特征表示模型,如图 1 所示。

3.1　词汇层面特征

词汇层面特征指的是基于字符和词汇的特征,是目前普遍采用的文体风格特征,具体包括字符和词汇的统计特征和频率,以及特殊字符和词汇的统计特征。自然语言中词语长短和使用频率能够体现作者使用语言时是否力求经济简便的自然心理趋势。由于博客文本具有短文本的特殊性,单独使用词汇特征容易造成数据稀疏,导致不能有效描述文体风格。因此本节在选取词汇特征时,着重考虑博客文本篇幅短小、特殊字符和特殊词汇比例高的特点,兼顾文体风格特征的可获取性,选取基于字符的特征具体包括:字符总数、字母总数、数字字符频率、大小写字母频率、特殊字符如"~号"、"@号"、"#号"、"MYM 号"、"%号"、"^号"、"&号"、"*号"、"-号"、"_号"、"=号"、"+号"、">号"、"<号"、"[号"、"]号"、"{号"、"}号"、"/号"、"\号"、"|号"共 21 个特殊字符的频率。选取基于词汇的特征包括:单词总数、不同单词总数、字长大于 4 的单词个数、平均字长、出现一次单词频率和出现两次单词频率等。上述特征不涉及具体的词汇意义,是内容无关的特征。

3.2　句法层面特征

句法层面特征传统上包括功能词、词性标注(Part of Speech, POS)和标点符号的使用方式等。为了提高输入效率,博客作者往往频繁使用标点符号来表达特殊或强烈的感情色彩。此外,博客中的短句和语法不规则句的比例也很高。这些特点使得传统的句法特征不足以刻画博客的文体风格。

根据是否运用句法结构解析,句法层面文体风格特征可以划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进行句法解析就可以获得的、显而易见的特征,例如句长、词性标注统计、类符形符比和功能词频率等。深层句法特征是对句子进行完全或者部分句法解析后获取的特征,例如依存句法关系和词性标注 N-gram 等。深层句法特征的获取相对复杂,但能够表达潜在的句子结构信息。本节在句法层面分别抽取浅层句法特征和深层句法特征。

3.2.1 浅层句法特征

本文多层面文体风格特征模型中,浅层句法特征包括标点符号特征和功能词特征。

(1)标点符号特征:从语法分析和文体学角度,标点符号既能反映作者的文体习惯倾向性,也能反映作者显性或隐性衔接手段的运用习惯。本节选择的标点符号统计特征包括:标点符号的总数和不同标点符号总数,叹号、逗号、句号、冒号、分号、问号、双引号和单引号共 8 种标点符号的频率,以及空格的频率。

(2)功能词特征:功能词是指本身并没有独立完整词汇意义,而只表达语法意义或语法功能的词。由其定义可知,功能词具有与主题内容无关的特点。相对于实义词,功能词的数量较少、出现频率高,在构建句法逻辑关系中起到关键作用,已经被证实是有效的文体风格特征[8]。常用的功能词有代词(this, that, her, his, it, its, our, your, what 等)、冠词(a, an, the)、助动词(can, would, should, have, do 等)、介词(to, with, in, at, for 等)、连词(but, not, or, and 等),基本动词(be, have, do 等)、情态动词(must, can, may 等)、限定词(what, which, no, some, any, every, all 等),本文统计常用的 70 个功能词的频率。

3.2.2 深层句法特征

多层面文体风格特征模型中的深层句法特征包括基于依存关系的特征和词性标注 N-gram 特征。

(1) 基于依存关系的特征:依存句法是由法国语言学家 Tesniere 提出的利用词与词之间的从属和支配关系来描述句法结构的理论框架,近年广泛应用在多语言处理、文本挖掘、信息检索和语义标注等领域。其形式化描述基于四条公理:一个句子只有一个独立成分;句子中的其他成分都直接依存于某一成分;任何一个成分都不能依存于两个或两个以上成分;如果成分 X 直接依存于成分 Y,成分 Z 在句子中位于 X 和 Y 之间,则成分 Z 依存于成分 X 或者 Y,或者是依存于 X 和 Y 之间的某一成分。依存句法的形式化描述由句子成分的二元关系组成,即核心词和依存词之间的依存关系对。如果用 $S = \{w_0, w_1, \cdots, w_n\}$ 表示分词分句处理后的句子,其中 w_i 为句子中顺序为 i 的词。抽取句子的依存关系可将句子表示为 $S = \{r_1(w_{11}, w_{12}), r_2(w_{21}, w_{22}), \cdots, r_m(w_{m1}, w_{m2})\}$,其中由 (w_{i1}, w_{i2}) 组成的词对构成依存关系 r_i,$w_{ij} \in S$,$r_i \in \mathbf{R}$,R 为所有依存关系类型的集合。

采用依存关系作为文体风格特征具有三个优势:①可计算性好,存储结构简单,有利于适应网络文本大数据和跨语言环境;②依存句法分析直接突出句子成分之间的支配与被支配、修饰与被修饰的依存关系,不必局限于句子成分的顺序;③依存关系是对抽象句法结构信息的提取,与句子的具体字词含义无关,具有内容无关性。本节在深层句法层面引入依存关系的统计特征、频率特征和特殊依存特征,具体包括依存关系总数、不同依存关系总数、出现一次依存关系频率、出现两次依存关系频率、依存关系词频、依存关系 2 - 4 gram 频率、依存关系句中不同单词总数、平均依存关系句长、最长依存关系句长和最短依存关系句长。

(2) 基于词性标注 N-gram 的特征。词性标注是根据单词词形或句法上的行为和功能,对单词进行类型标注的分析方法,目前普遍应用于句法分析。词性标注关注单词的类型而非含义,具有内容无关的特性。在不同语言和语料上有多种词性标注体系,需要依据具体任务的需求来选择。例如,宾州树库(Penn Treebank)的英文标注集有 48 类词性标注,中文标注集将词性划分为 89 类。本文在博客文体特征模型的句法层面中引入宾州树库词性标注基础上的句法分析,即在获取单词的词性标注结果之后,对其进行 N-gram 分析,得到 POS 2 - 4gram 频率作为文体风格特征。

3.3 结构层面特征

结构层面特征包括文本组织和布局相关的特征,如段落数目、段落长度、整篇文章的平均句长等特征。博客文本篇幅短小,作者更倾向于使用丰富灵活的文本组织和布局方式,结构特征对文体风格的表达作用不容忽视。本文引入“平均句长(单词)”、“平均句长(字符)”、“最长句子”、“最短句子”、“句子数”、“换行数”作为结构层面特征。

4 博客作者身份识别实验

4.1 数据准备

本文选取 Koppel 的公开博客语料[13]作为作者身份识别实验数据来源,此语料收集了 Google 博客网站 Blogger. com 一个月期间的博客文本,包含超

过140 000 000词的语料,平均每位作者35篇约7250词的博客文本。数据预处理阶段的主要工作包括:去除语料中无关的XML标记和转载的博客文本。由于作者身份识别任务要求在训练集中每位作者有足够的训练样本,为此筛选并保留博客篇数600篇以上的作者样本。为避免因字符数过少而导致的文体风格信息不足,筛选并保留100字符以上的博客文本。最终得到15位作者10 895篇共1 642 228词的博客文本,平均每位作者726篇博客样本。

首先从上述博客样本集中随机抽取3位作者10次,分别组成10组博客样本数据,在每次对照实验中对每组博客样本数据执行十折交叉验证,即将每组样本随机分为10份,轮流将其中的9份作为训练样本,余下的1份作为测试样本,产生10对不同的训练集和测试集,对实验结果求平均值。我们最终记录10组博客样本数据实验准确率(Accuracy)、召回率(Recall)和F-Measure的平均值来评估作者身份识别性能。

4.2 实验方案

为考察本文提出的多层面文体风格特征模型,从两个维度进行对照实验:一是以词汇特征作为基准特征集,依次增加结构特征、浅层句法特征和深层句法特征,特征集的递增情况如表1所示。每当增加一个层面的特征,分别对10组博客数据集重新进行十折交叉验证,计算作者身份识别准确率、召回率和F-Measure,目的是验证依次加入的特征集的有效性;二是分别检测各个层面特征集在100字符、200字符和300字符博客文本上的作者身份识别结果,目的是测试各个层面特征在短文本上的鲁棒性。

表1 实验特征集

特征集	词汇特征	结构特征	浅层句法特征	深层句法特征
F1	√			
F1 + F2	√	√		
F1 + F2 + F3	√	√	√	
F1 + F2 + F3 + F4	√	√	√	√

在各组对照实验中分别应用四种分类算法:朴素贝叶斯(NBC)、决策树(C4.5)、序列最小优化支持向量机(SMO)[14]以及大规模线性分类支持向量机(LIBLINEAR)[15]。其中,SMO核函数选择Polynomial kernel,LIBLINEAR核函数选择Regularized L2-loss support vector classification(dual),实验环境为Weka3.7。

4.3 实验结果及分析

两组实验在100字符、200字符和300字符以上博客数据上的实验结果如表2所示,在各个特征集上的准确率、召回率和F-measure最高数值用加粗字体显示。

从分类算法的角度看:①四种分类算法中,基于支持向量机的LIBLINEAR和SMO的作者身份识别性能最好,尤其是LIBLINEAR,在四种特征组合、三种字符长度数据上的准确率、召回率和F-measure几乎都是最高值;②C4.5的性能居中;③朴素贝叶斯分类性能最差,分析原因是多层面文体风格特征不能满足朴素贝叶斯分类的独立性假设。

从各层面特征集对作者身份识别的影响来看:①随着新特征的加入,三种文本长度、四种分类算法实验中作者身份识别的准确率、召回率和F-measure都有增长,在所有层面都加入后,准确率、召回率和F-measure达到最高值。这验证了各个层面特征能够起到作者区分的作用。②从深层句法特征的作用来看,表3中所示的是移除深层句法特征后准确率的变化,三种文本长度、四种算法的实验数据中,移除深层句法特征后准确率都有明显下降。以LIBLINEAR为例,在移除深层句法特征后100字符、200字符和300字符数据上的准确率分别下降10.70%、9.27%和8.30%,进一步表明了深层句法特征在作者身份识别中的作用。③从各层面特征对准确率的提升程度来看,图2所示的是特征集逐步增加时LIBLINEAR算法准确率的变化趋势,可以看出在依次加入结构特征、浅层句法特征和深层句法特征的过程中,相对于其他特征,句法特征尤其是深层句法特征能够快速提高LIBLINEAR算法准确率,反映了引入深层句法特征更有利于抽象短文本隐含信息。

从文本长度对作者身份识别的影响来看:①从表2中的100字符博客文本全特征集实验数据可以看出,SMO和LIBLINEAR作者身份识别平均准确率分别为80.2%和85.2%,平均召回率分别为80.2%和85%,F-Measure分别为80.2%和85%,证明本文提出的多层面文体风格特征模型能较好适应短文本,具有较好的鲁棒性;②总体上,当博客文本

长度依次从 100 字符增加到 200 字符、300 字符时，作者身份识别的准确率依次有所提高，说明增加文本长度有助于提高多层面文体风格特征模型的表现力；③从作者身份识别效果最好的 LIBLINEAR 算法来看，表 4 所示的是字符增加时 LIBLINEAR 准确率的变化，表明了当字符数逐步增加，LIBLINEAR 算法准确率与文本长度是正相关的；④从各层面特征对文本长度的敏感度来看，图 2 展示了逐步加入各层面特征集时，100 字符、200 字符和 300 字符数据上 LIBLINEAR 算法准确率的变化，可以看出文本越短，则加入深层句法特征后准确率提升越明显，分析原因是词汇特征和结构特征在短文本上相对稀疏，使得深层句法特征对短文本作者身份识别的作用尤为明显。

表 2　四种分类算法在 100 字符、200 字符和 300 字符以上博客上的实验结果

		F1			F1 + F2			F1 + F2 + F3			F1 + F2 + F3 + F4		
		100ch	200ch	300ch	100ch	200ch	300ch	100ch	200ch	300ch	100ch	200ch	300ch
NBC	Accuracy	0.485	0.476	0.472	0.515	0.522	0.526	0.537	0.550	0.563	0.596	0.624	0.655
	Recall	0.483	0.476	0.472	0.515	0.522	0.526	0.537	0.550	0.563	0.596	0.596	0.655
	F-Measure	0.454	0.461	0.471	0.495	0.519	0.536	0.523	0.549	0.572	0.588	0.588	0.661
C4.5	Accuracy	0.656	0.666	0.688	0.689	0.706	0.724	0.704	0.726	0.748	0.769	0.769	0.816
	Recall	0.656	0.666	0.688	0.689	0.706	0.724	0.704	0.726	0.748	0.769	0.769	0.816
	F-Measure	0.656	0.665	0.686	0.689	0.706	0.724	0.704	0.726	0.748	0.769	0.769	0.816
SMO	Accuracy	0.550	0.582	0.605	0.632	0.665	0.685	0.728	0.771	0.800	0.802	0.802	0.849
	Recall	0.550	0.581	0.605	0.632	0.662	0.685	0.728	0.771	0.800	0.802	0.802	0.849
	F-Measure	0.506	0.533	0.551	0.616	0.638	0.650	0.726	0.769	0.794	0.802	0.802	0.849
LIB-LINEAR	Accuracy	0.670	0.699	0.730	0.692	0.725	0.754	0.745	0.780	0.807	0.852	0.852	0.890
	Recall	0.670	0.699	0.730	0.692	0.725	0.754	0.745	0.780	0.807	0.850	0.850	0.890
	F-Measure	0.664	0.691	0.720	0.688	0.719	0.745	0.742	0.776	0.801	0.850	0.850	0.890

表 3　移除深层句法特征准确率的变化

	NBC		C4.5		SMO		LIBLINEAR	
	全特征集	移除 F4	全特征集	移除 F4	全特征集	移除 F4	全特征集	移除 F4
100 字符	59.56%	-5.88%	76.88%	-6.51%	80.21%	-7.37%	85.22%	-10.70%
200 字符	62.44%	-7.43%	78.86%	-6.26%	83.25%	-6.18%	87.27%	-9.27%
300 字符	65.46%	-9.16%	81.59%	-6.77%	84.86%	-4.87%	89.01%	-8.30%

表 4　字符增加 LIBLINEAR 准确率的变化

特征集	100 字符	200 字符	300 字符
F1	67.00%	+2.88%	+3.15%
F1 + F2	69.20%	+3.32%	+2.90%
F1 + F2 + F3	74.52%	+3.48%	+2.71%
F1 + F2 + F3 + F4	85.22%	+2.05%	+1.74%

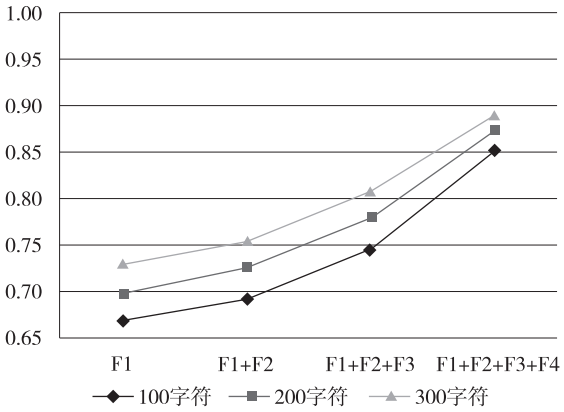

图2 字符增加和特征增加对LIBLINEAR准确率的贡献

5 结 论

以博客为代表的网络文本在促进先进文化知识传播中起到了重要作用,但同时也存在以匿名方式滥用互联网的问题,使传统条件下的作者身份识别方法受到前所未有的挑战。本文针对互联网应用环境的特点,结合文体计量学和机器学习方法,探索性地提出主题内容无关的多层面文体风格特征模型,尝试对现有文本挖掘理论进行延伸,为在线信息作者身份的自动识别提供了新的技术思路。实验结果表明此方法在博客作者身份识别任务上具有较好的性能。尽管本研究的尝试在一定范围内得到了验证,但在一些研究方向上仍需探索:

(1)大数据环境下的网络文本作者身份识别。国内外相关研究表明,潜在作者数目和每位作者已知的训练样本数量是作者身份识别的重要参数。网络文本数据海量,导致一条匿名文本的潜在作者数量巨大。而互联网用户经常以不同用户名在不同网络渠道登陆,使得每位作者可获得的训练样本非常有限,大大增加了作者身份识别的难度,大数据环境下的网络文本作者身份识别的可行性和适用性还需要进一步探讨。

(2)跨语言环境下的文体风格特征模型。Internet的国际化特质决定了在多语种上下文中研究作者身份识别的必要性,而文体风格特征与所处理的语言是高度相关的。因此,跨语言环境下的文体风格特征也是需要细致研究的方向。

参 考 文 献

[1] 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.

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

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

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

[5] 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.

[6] 李贤平.《红楼梦》成书新说[J].复旦学报(社会科学版),1987(5):2-4.

[7] 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.

[8] Zhao Y, Zobel J. Effective and Scalable Authorship Attribution using Function Words [M]//Information Retrieval Technology. Berlin:Springer, 2005: 174-189.

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

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

[11] Zhang C, Wu X, Niu Z, et al. Authorship identification from unstructured texts[J]. Knowledge-Based Systems, 2014:99-111.

[12] 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.

[13] Koppel M. The Blog Authorship Corpus [OL].[2014-05-28]. http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm.

[14] Platt J C. Fast training of support vector machines using sequential minimal optimization [C]. Advances in kernel methods. MIT press, 1999: 185-208.

[15] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research, 2008(9):1871-1874.

(责任编辑 赵 康)