基于风格特征融合的文档分割方法

刘 刚[1,2] 王 凯[1*] 刘汪洋[2] 曹 扬[2] 李 涛[1]

[1](哈尔滨工程大学计算机科学与技术学院 黑龙江 哈尔滨 150001)

[2](中电科大数据研究院有限公司 贵州 贵阳 550081)

摘 要 风格裂缝指的是多作者共同完成的文章中作者身份发生转变的位置。提出基于多特征风格的文档分割的剽窃检测方法。使用 7 种文本风格特征用于风格裂缝识别。通过特征提取的结果,利用无监督的机器学习算法,基于提取的特征进行分类。利用聚类算法对风格特征进行聚类,从而找到文章风格裂缝的位置。采用参数权重技术进行特征权重调节和多特征融合,对不同文章出现的特征冗余问题进行合理解决。分别使用滑动窗口与段落划分对不同类型的文本进行实验,得出基于段落的实验结果比基于滑动窗口的实验评估值高出 10% 左右的结论。

关键词 剽窃检测 文本风格 风格裂缝 多特征融合 机器学习

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.10.032

DOCUMENT SEGMENTATION METHOD BASED ON STYLE FEATURE FUSION

Liu Gang[1,2] Wang Kai[1*] Liu Wangyang[2] Cao Yang[2] Li Tao[1]

[1](*College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, Heilongjiang, China*)

[2](*CETC Big Data Research Institute Co., Ltd., Guiyang 550081, Guizhou, China*)

Abstract Style crack refers to the position where the author's identity has changed in the article completed by multiple authors. This paper proposes a plagiarism detection method based on multi-feature style document segmentation. Seven text style features were used for style crack recognition. Through the result of feature extraction, unsupervised machine learning algorithm was used to classify the features based on the extracted features. Clustering algorithm was used to cluster the style features to locate the position of style cracks. Parameter weighting technology was used to adjust the weights of features and fuse multiple features to solve the problem of feature redundancy occurred for different articles reasonably. Using the sliding window and paragraph partitioning, experiments are carried out on different types of text. The conclusion is that the experimental results based on paragraph are about 10% higher than the experimental evaluation values based on sliding window.

Keywords Plagiarism detection Text style Style cracks Multi-feature fusion Machine learning

0 引 言

随着数据时代的到来,人们获取相关资料,相互共享信息的途径也十分广泛。无论是文学作品还是学术论文,所产生的剽窃行为都屡禁不止,这种行为令人深恶痛绝。更有甚者,简单修改文章的内容以逃过现有的剽窃检测,为当今学术界造成恶劣影响。

基于此,人们对论文的剽窃检测研究不单单是停留在简单的字符串判断方面,对于作者的写作风格和写作习惯的研究也越来越受到人们的关注。写作风格不仅能反映一个作者的写作习惯,更能运用在剽窃检测系统和用户画像技术上,对作者的识别也有很好的帮助,给剽窃检测系统提供一个新的研究角度,对网上

收稿日期:2019-06-29。教育部人文社会科学研究规划基金项目(19YJAZH053);中央高校基本科研业务费专项资金项目(3072019CF0601)。**刘刚**,副教授,主研领域:人工智能,自然语言处理,数据挖掘,机器学习。**王凯**,硕士生。**刘汪洋**,博士。**曹扬**,硕士。**李涛**,硕士。

匿名文章作者的判断也提供了强有力的支持。不同人在进行写作时会形成自己独特的风格特点,主要体现在用词、句、段、修辞手法、情感等方面,这些是作者在不经意间养成的写作习惯,所以通过对文章的写作风格特征的提取来推断文章的所属是有效的。

1　相关理论与工作基础

国外语言学学家在文本风格研究方面早已起步,1985 年 Cary Taylor 通过写作风格发现一首 9 节诗歌是莎士比亚的作品,中国的语言学家也通过写作风格推测红楼梦的作者原创性。中文方面,对于四大名著之一的《红楼梦》是不是同一个人写的问题备受争议,华中师范大学博士生刘悦基于语料库对四大名著《红楼梦》的部分写作风格进行统计[1],验证了前 80 回和后 40 回的用词习惯有所差异,间接地证明了后 40 回可能出自其他作者。

1.1　风格特征提取

早期的风格研究主要是利用统计的方法,对词汇、句子、段落的规律进行统计,利用统计的规律来约定一个人的风格。风格特征提取最早是对单特征进行研究,随着单特征不能满足实验结果,多特征融合应运而生。近年来,把机器学习和神经网络的算法引入到风格提取和作者识别中,并且取得了好的结果。

由于中文的多变和困难,所以在对中文的风格提取上,比外文的风格提取明显更加困难,中文需要考虑到分词系统的准确性,句子结构也比较复杂。尽管中文的风格提取比外文更困难,但对于中文风格的研究仍然受到了广泛的关注。

1.2　文本分割技术

文本分割技术把一篇文章根据某些特征分成几个独立的片段,该技术在文本预处理、自然语言处理中占用很重要的比重。由于文本分割的目的不同,所以使用的方法和特征也有所不同。Tian 等[2]提出了一种多级 MSER 技术,该技术从一组不同颜色通道文本图像中提取的稳定区域中识别出最优质的文本候选。为了识别最优质的文本候选,定义了一个分割得分,利用四个度量来评估每个稳定区域的文本概率。该方法在 ICDAR2003 和 SVT 数据集上进行评估,实验表明它优于流行的文档图像二值化方法和最先进的场景文本分割方法[2]。

在中文方面,刘耀等[3]提出了一种基于领域本体对文本进行线性分割的方法。该方法利用初始概念自动获取结构化语义概念集合,并根据获取的概念、属性及属性词在文本中出现的频次、位置和关系等因素为段落赋予语义标签,挖掘文本的子主题信息,将拥有相同语义标注信息的段落划分为相同语义段落,实现了文本不同子主题之间的分割,分割效果能够满足实际应用需求,并优于现有的无须训练语料的文本分割方法。

2　多特征提取与融合

利用网络资源中的电子小说进行全文下载,选取风格差异比较明显的 20 个文档,作为实例进行实验效果的风格特征分析,其中 10 篇来自古龙,另外 10 篇来自琼瑶。

2.1　单维风格特征

2.1.1　词长度

在中文风格方面,可以使用分词之后的词汇长度,观察作者在用词方面对两字词语、三字词语、四字词语以及四字以上词语的使用习惯。有研究发现在双字词语的使用频率上,张爱玲的使用频率是 0.17,而鲁迅的使用频率高达 0.43。

在英文方面是统计单词字母个数,而在中文上统计的平均词长度的范围比较小,平均词长度基本范围在 2 ~4 之间,但是缩小比较范围同样可以看出实验效果。因此把词长度作为最后分类的一个参数。

2.1.2　平均句子长度

平均句子长度是统计作者对文本句子长度的使用习惯,对比长短句的使用频率。统计出每一个句子的长度,进行平均求和,平均句子长度以“。”、“!”和“?”为一组标记,统计句子中字数长度的平均值作为最后分类的参数,如图 1 所示。

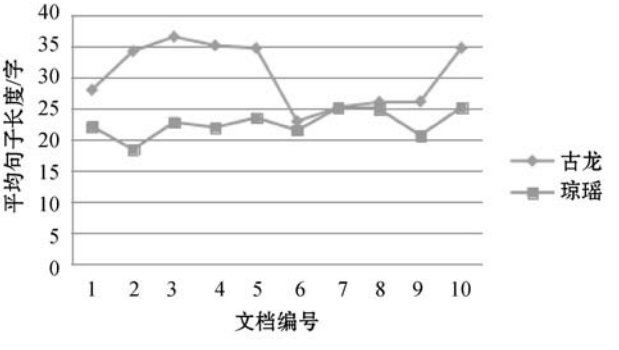


图 1　平均句子长度

可以看出,古龙文档的平均句子长度明显比琼瑶的平均句子长度长,可以通过这个特征来区分出不同写作风格。

2.1.3　情感偏向

情感分析一直都是对情感词的统计来分析文章情

感,Xia 等[4]在情感分析上使用神经网络等,并且对比了实验的结果,验证了其有效性。Sailunaz 等[5]使用机器学习基于各种基于用户和 Twitter 的参数来计算用户的影响分数,对 Twitter 进行情感分析。本文使用网络中训练的情感字典对文章进行情感分析。对使用情感词典来进行情感分析的算法的形式化描述如算法 1 所示。

算法 1 情感分析算法

输入:测试文本 D1。

输出: *Pos*,*Neg*,*AvgPos*,*AvgNeg*,*Res*。

1. BEGIN

2. 对中文语句进行分句,以句号为句子结束标志;

3. 查找分句中情感词,记录其是积极还是消极,及其位置;

4. 查找情感词前的程度词,匹配程度词表,找到即停止搜寻;

5. 为程度词设权值,乘以情感值;

6. 查找情感词前的否定词,匹配否定词表,直至找到全部否定词;

7. 若数量为奇数,乘以 -1;

8. 若数量为偶数,乘以 1;

9. 判断分句结束处是否存在感叹号;

10. 是,往前寻找情感词,且相应的情感值 +2;

11. 每个分句计算所得的情感值,存在数组(list);

12. 遍历所有分句,计算 *AvgPos*,*AvgNeg*;

13. END

14. 返回所有分句的 *Pos*,*Neg*,*AvgPos*,*AvgNeg*,*Res*;

其中:*Pos* 表示积极参数的结果;*Neg* 表示消极参数的结果;*AvgPos* 表示积极参数的平均值;*AvgNeg* 表示消极参数的平均值;*Res* 代表最后情感偏向结果取 *AvgPos* 和 *AvgNeg* 相加的结果。通过实验取 *Res* 作为最后的分类参数。文档风格提取情况如图 2 所示。

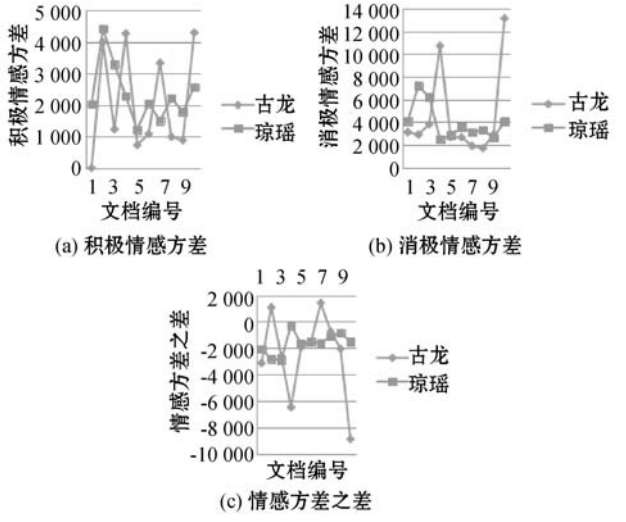


图 2 文档风格提取情况

其中情感方差之差为积极情感方差减去消极情感方差。通过分析图 2(c),发现古龙和琼瑶写作时的情感倾向,大多都是消极情感,而且琼瑶情感方差差值的波动范围,完全包含在古龙的差值波动范围之中。所以对于一个文档,即便情感倾向发生变化也无法判断是由于作者改变还是文档情节内容改变而导致的。因此,该特征中易出现特征冗余问题。

2.2 多维风格特征

2.2.1 词汇特征

用词方面可以体现一个人的文学功底,可以根据用词的丰富程度去评判作者的写作风格。词汇特征[6]可以定义为词的长度、词频、占比和密度等方向,词汇特征如表 1 所示。

表 1 词汇特征表

编号	词汇特征	编号	词汇特征
1	总词数	5	自造词/总词数
2	两个字的词个数	6	感叹词/总词数
3	三个字的词个数	7	不同词个数/总词数
4	四个字的词个数	8	词汇密度

这 8 个维度可以概括一个人在用词上的习惯,把这 8 个特征作为最后的分类参数的其中 8 个。

词长特征提取实例如图 3 所示,纵坐标表示文档中不同词长出现的频率。

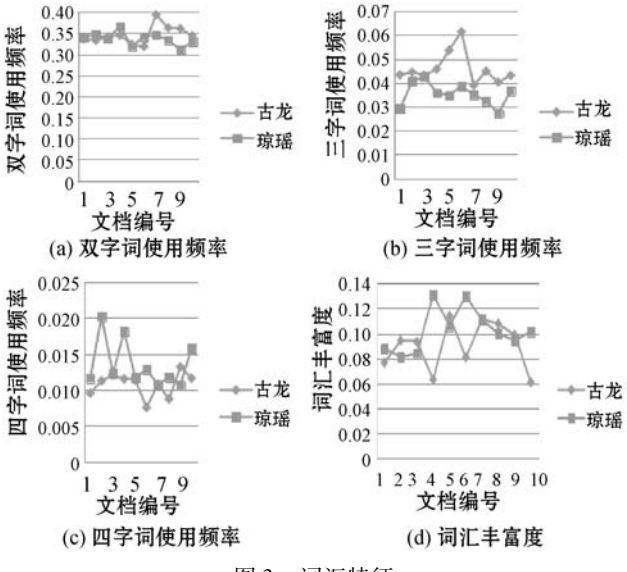


图 3 词汇特征

可以看出,双字词特征和词汇丰富度存在与情感偏向一样的特征冗余问题,两个作者在对这双字词的使用频率上重合范围很大,没有什么明显的个人特色。古龙的词汇丰富度大致在 0.06 ~ 0.12 之间,琼瑶的词汇丰富度大致在 0.08 ~ 0.14 之间,两位作者的词汇丰富度有很大范围的重合,差异度不大。但是从三字词

和四字词的使用频率上可以看出存在差别,琼瑶对三字词的使用明显没有古龙频繁,但是琼瑶对四字词使用要比古龙多得多,词长中的三字词和四字词可以体现出两位作者的写作风格差异。

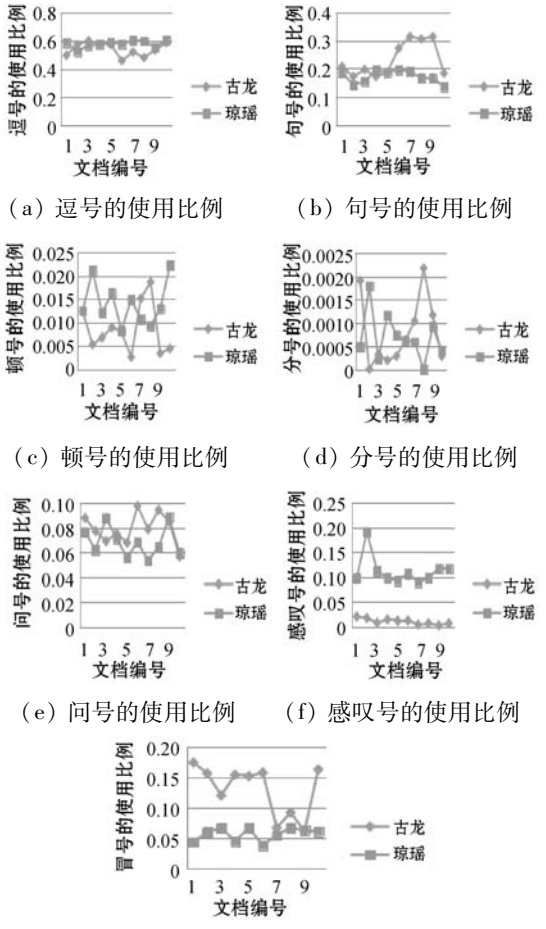
2.2.2　特殊标点符号

标点符号能反映作者写作时显性或隐性运用衔接内容的行文习惯。作者在写作过程中为了提高输入效率和精简篇幅,往往频繁地使用标点符号以表达特殊的情绪。在作者使用短句和非正式文法时,对标点的统计可以看出作者对句型的使用习惯,比如感叹句往往伴随着感叹符号一起使用,问句往往伴随着问号一起使用,但是常用的标点符号共性太强,不能统计出一个作者的使用情况,所以需要使用特殊标点符号。

特殊标点符号特征统计冒号、分号、千百分号、单位符号、左右引号、左右括号、叹号、省略号、破折号、问号和顿号,表示为 $P_1 - P_{11}$:

$$F = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}, P_{11}\} \quad (1)$$

特殊标点使用简单统计的方法,将特殊标点的统计作为最后分类算法的参数,需要删除标点符号频率为 0 的标点,所以该特征的维度最高是 11 维度,如图 4 所示。



(a) 逗号的使用比例　(b) 句号的使用比例

(c) 顿号的使用比例　(d) 分号的使用比例

(e) 问号的使用比例　(f) 感叹号的使用比例

(g) 冒号的使用比例

图 4　标点符号

可以看出,在感叹号和冒号的使用比例上,可以明显地体现出古龙与琼瑶写作风格的差异,其他标点符号的使用比例无法明显分辨出写作风格的差异,同样存在特征冗余问题。

2.2.3　同义词

同义词是中文文体中一个特有的分支,体现了中文的语言多样性,在同义词的使用上也可以体现作者的语言功底和对词语的驾驭能力。对同义词的使用习惯也可以看出一个作者的用词习惯,从同义词的使用情况出发,对同义词的使用习惯进行统计,总结出作者的写作习惯[7]。同义词算法如算法 2 所示。

算法 2　同义词特征统计

输入:同义词林,两段计算文档 *D*1 和 *D*2。

输出:***SynVec***。

1. BEGIN
2. 同义词林预处理,加载同义词林;
3. 对滑动窗口中的中文语句进行分词;
3. 分裂查找文档 D1 和 D2 中同义词,找到同义词表相应的位置,词频加 1;
4. 同义词和词频组成一个同义词对,更新同义词表,删除词频为 0 的同义词;
5. 比较两个滑动窗口同义词表,对位,0 补位;
6. 删除词频相同的同义词,降维;
7. 同义词向量集合作为最后同义词的参数;
8. 输出 ***SynVec***;
9. END

本文对哈工大同义词词林精减,通过测试文本集中的文档,将其中从未出现的同义词删掉,如果一个同义词词组中一个也没有出现,就将这组同义词删掉,形成新的同义词词林,最后根据测试结果将同义词词林精减到只有 2 200 组。获得剩余同义词和词频组成同义词向量,其中 *D*1 和 *D*2 代表两片测试文档,***SynVec*** 代表同义词向量结果。

本文选择了同义词林中的一组同义词“人,士,人物,人士”,并统计其在 10 篇文档中的使用分布情况,结果如表 2 所示。

表 2　同义词结果示例　　%

文档	“人”使用比例	“士”使用比例	“人物”使用比例	“人士”使用比例
古龙文档 1	0.99	0.00	0.01	0.00
古龙文档 2	0.98	0.00	0.02	0.00
古龙文档 3	0.99	0.00	0.01	0.00
古龙文档 4	0.98	0.00	0.01	0.00
古龙文档 5	0.99	0.00	0.01	0.00

续表 2

文档	“人”使用比例	“士”使用比例	“人物”使用比例	“人士”使用比例
琼瑶文档 1	0.85	0.00	0.15	0.00
琼瑶文档 2	0.89	0.00	0.11	0.00
琼瑶文档 3	0.90	0.00	0.10	0.00
琼瑶文档 4	0.88	0.00	0.12	0.00
琼瑶文档 5	0.81	0.02	0.17	0.00

根据表格中的数据可以看出,古龙的文档在“人,士,人物,人士”这一同义词组中,基本上只使用“人”这个词语,而琼瑶除了使用“人”这一词语外,还少量地使用了“人物”这一词语。因此可以通过同义词词组中同义词的使用偏好,来观察到作者的写作特点。

2.2.4 虚 词

虚词本身是没有意义的,它的意义是它在句子中的地位,虚词的数量有限,出现的频率没有实词高,大约占词汇使用率的 1/3 左右。可见虚词在整个文章的占比还是很大的,并且它数量有限,容易统计,根据这个特性可以表示作者的风格特征。

本文增加虚词的数量,通过自定义虚词表作为基准,对虚词表的虚词使用情况进行计算。首先制作虚词表,虚词表来源是《现在汉语虚词词典》,虚词表中一共有 840 个虚词,和同义词表相同,虚词表较大,含有一些生僻和不常用的虚词,会影响结果的计算。以搜狗新闻数据集为基准,对虚词表的虚词进行 TF-IDF 统计,删除 TF-IDF 过低的词。通过多次清洗,最后精简到 230 个虚词。选用 230 个虚词首先能控制在一个合理的数量中,这 230 个虚词能体现虚词在新闻集中的重要程度,最后形成一个虚词 TF-IDF 词对表。在虚词表中随意选择“被”这个虚词,计算其在 20 篇文档中的出现频率。文档风格提取情况如图 5 所示。

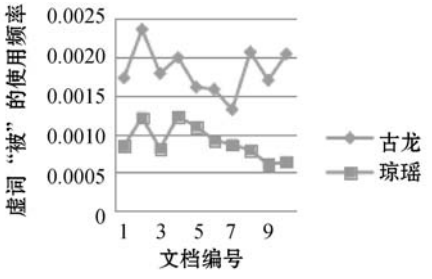


图 5 虚词特征

3 风格裂缝的识别

3.1 风格裂缝

风格裂缝指的是一个文本风格发生转变的位置,换句话说一个文章可能由不同作者共同完成,所以在作者识别之前进行基于作者分段技术变得尤为重要,即找出这篇文章每一个行文作者对应的行文部分,风格裂缝点如图 6 所示。通过写作风格分段,目标是找到风格裂缝点,即风格发生转变的位置。风格裂缝识别是通过风格的特征提取结合分类算法的技术,采用滑动窗口、降维等技术,找出风格裂缝点。风格裂缝是在多风格特征提取的基础之上提出的一个概念,通过风格裂缝的识别能更好地进行分段技术。

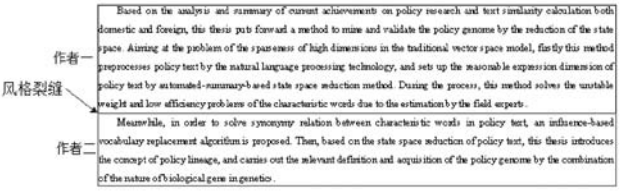


图 6 风格裂缝示例图

3.2 滑动窗口

滑动窗口以 5 个句子为一个整体,进行风格特征识别。每次向下滑动一个句子,对每个窗口进行风格统计,当风格发生转变时,每次风格和上一次发生的结果有逐渐的变化,直到风格相似度又趋近不变,则这个位置产生风格裂缝。

但是实际情况是有极少可能出现理想状态,为了更好地找到风格裂缝,需要在特征提取和分类算法上进行调优。因为论文篇幅较小,5 个句子所含的信息量较少,很大的可能性会出现偶然现象,假定每次的风格裂缝位置有很大的可能性发生在每个段的段尾。为了提高准确率,只能牺牲召回率。假定每次风格裂缝必然会发生在段尾,即假定文章中的每一段有且仅有一个作者。滑动窗口示例如图 7 所示。

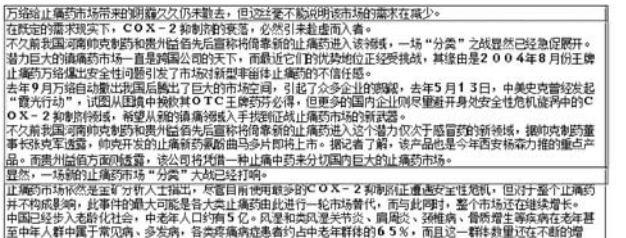

万络给止痛药市场带来的阴霾久久仍未散去,但这丝毫不能说明该市场的需求在减少。
在既定的需求现实下,COX-2抑制剂的衰落,必然引来赵虚而入者。
不久前我国河南帅克制药和贵州益佰先后宣称将倚靠新的止痛药进入该领域,一场“分羹”之战显然已经急促展开。
潜力巨大的镇痛药市场一直是跨国公司的天下,而最近它们的优势地位正经受挑战,其缘由是2004年8月份王牌止痛药万络爆出安全性问题引发了市场对新型非甾体止痛药的不信任感。
去年9月万络自动撤出我国后腾出了巨大的市场空间,引起了众多企业的觊觎,去年5月13日,中美史克曾经发起“霞光行动”,试图从国内申报欲其OTC王牌药芬必得,但更多的国内企业则尽量避开身处安全性危机旋涡中的COX-2抑制剂领域,希望从新的镇痛领域入手找到赶超止痛药市场的新武器。
不久前我国河南帅克制药和贵州益佰先后宣称将倚靠新的止痛药进入这个潜力仅次于感冒药的新领域,据帅克制药董事长张克军透露,帅克开发的止痛新药氯酚曲马多片即将上市。据记者了解,该产品也是今年西安杨森力推的重点产品。而贵州益佰方面则透露,该公司将凭借一种止痛中药来分切国内巨大的止痛药市场。
显然,一场新的止痛药市场“分羹”大战已经打响。
止痛药市场依然是宝矿分析人士指出,尽管目前使用最多的COX-2抑制剂止痛遇安全性危机,但对于整个止痛药并不构成影响,此事件的最大可能是各大卖止痛药由此进行一轮市场替代,而与此同时,整个市场还在继续增长。
中国已经步入老龄化社会,中老年人口约有5亿。风湿和类风湿关节炎、肩周炎、颈椎病、骨质增生等疾病在老年甚至中年人群中属于常见病、多发病,各类疼痛病症患者约占中老年群体的65%,而且这一群体数量还在不断的增加。

图 7 滑动窗口示例图

3.3 参数权重法

针对第 2 节中风格特征分析产生的特征冗余问题,每一个参数在风格裂缝识别过程中占用的权重不尽相同,所以在查找风格裂缝的时候需要找出每个参数的权重,然后通过参数调节的权重进行风格裂缝识别。

参数权重法首先对所有参数权重进行遍历,通过

多组新闻集进行遍历,对参数进行调优,最后找出每个特征的最优参数,虚词和同义词精减之后分别选用同一组参数作为权重,在训练过程中选中搜狗新闻集作为语料库。

算法描述:首先对数据集进行预处理,对数据集合进行特征提取,把数据集打乱顺序存到文件里,在其他特征参数权重不变的情况下,借用控制变量法的思想,控制词长度参数(WLP)从 0.01 到 0.99 进行计算,其他参数为 0.5,得到在其他参数不变的情况下参数 WLP 的最优值,最优值是以两篇文本相似度最低为标准。再在其他参数不变的情况下,以平均句子长度参数(ASLC)从 0.01 到 0.99 进行计算,得到 ASLC 的最优值,以此遍历所有的参数。然后都以上一次参数最优的结果为基准,继续上面的方法进行循环,直到参数最优值不变为止,得到参数权重组,目的是通过参数权重法发现每一个参数的有效性,删除无效参数。

得到的参数权重组,发现其中一些参数权重过小,这类特征对结果起到积极影响较小,但是会影响实验的效率,所以删除这些参数。

4　基于融合特征的风格聚类

4.1　实验思路

文本的特征提取是风格识别的主要方法,该方法对风格特征进行了层次分类,运用层次的角度进行特征的提取,加入特征与文章之间的一个映射关系。特征提取包括单维特征和多维特征两种,利用参数权重法对特征优化,把每个特征提取的结果作为最后 K-means ++ 分类器[8]的输入,通过滑动窗口找到风格裂缝,通过识别的风格裂缝点进行文章分段。

语料库选用自己构建的新闻语料库,语料集的主题包括利用爬虫技术在人民日报官网收集的关于时政、法制、旅游等方面的新闻,以及从虎扑新闻官网爬取关于体育的新闻。由于新闻集合中存在一些时间、图片、图片介绍和摄影师姓名等杂质。首先对新闻进行杂质处理,选取文中的正文。把新闻存成 .csv 文件,以作者姓名为新闻的标注,把 1 300 篇新闻分为 1 150个训练集和 150 个测试集,训练集和测试集的比例约为 9:1。为了验证小篇幅的准确性能,又把 150 篇测试集分为 100 篇,并按照篇幅存储,剩余 50 篇按照段落存储,大概是 215 段新闻。

4.2　单特征风格裂缝识别结果

对提取的 7 类风格特征进行单独实验,分别验证每一个风格特征对风格裂缝识别的效果,基于段落级别进行风格裂缝识别,取作者一时政编辑曹昆、作者二体育编辑郝帅、作者三法制编辑袁勃、作者四旅游编辑田虎、作者五时政编辑王政淇的新闻集融合作为测试集。实验结果如表 3 所示。

表 3　单特征实验结果展示表　　%

风格特征	准确率	召回率	F 评估值
词长度	35.9	65.8	46.45
平均句子长度	31.7	75.6	44.67
情感分析	21.7	55.5	31.20
词汇特征	45.6	82.3	56.68
特殊标点	43.8	85.3	57.98
同义词	55.5	81.7	66.10
虚词	57.2	87.9	69.30

本次实验放宽了召回率,这样准确率会随之减小,但是当前 F 值会相对增大。随着召回率的降低,召回结果的减小,准确率也会随之提升 5 ~ 10 个百分点。从结果可以看出,单维度特征维度偏少,效果不佳,情感分析结果较差,对风格裂缝识别作用较小;多维特征风格结果中,虚词、同义词和特殊标点的使用对风格影响较大,F 值偏高,相比而言词汇特征过于复杂,对风格裂缝识别成中性。

4.3　参数权重优化

在语料库方面首先随机抽取上述 5 名编辑作者的 100 篇文章形成一个小样本的训练集,用来对参数权重法进行训练。对训练集进行预处理,把每一个作者的文档集放到一个. txt 文件中,对每一个作者的文档集风格特征进行提取,形成风格特征向量。首先提取平均句子长度参数,进行分词处理,分词处理过后提取平均词长度、词汇特征、特殊标点符号;再提取虚词进行虚词 TF-IDF 算法,提取同义词填充同义词向量;最后计算训练集的情感偏向。

在计算平均句子长度时,以“。”“!”“?”作为评定句子结尾的三个标志,以每一个字作为一个长度计算。在分词过程中采用粗粒度分词系统,例如“北京大学”在粗粒度分词系统中不会被拆开,在细粒度分词中会被拆分成“北京”和“大学”两部分。本文在长度和词性特征上需要保证词汇的完整性。

在训练过程前预先设定 7 个参数权重,分别是平均句子长度、词长度、情感偏向、词汇特征、特殊标点、

同义词和虚词,所有同义词使用同一个特征权重,所有虚词也使用同一权重。实验结果如表 4 所示。

表 4 参数权重法结果

对比作者	词长度	平均句子长度	情感分析	词汇特征
作者一与作者二	0.27	0.27	0.12	0.52
作者一与作者三	0.18	0.32	0.13	0.62
作者一与作者四	0.45	0.17	0.07	0.87
作者一与作者五	0.31	0.31	0.21	0.62
作者二与作者三	0.52	0.12	0.12	0.43
作者二与作者四	0.64	0.08	0.14	0.76
作者二与作者五	0.12	0.07	0.08	0.44
作者三与作者四	0.54	0.34	0.12	0.45
作者三与作者五	0.21	0.31	0.11	0.77
作者四与作者五	0.41	0.10	0.12	0.79
作者一与作者二	特殊标符号	同义词	虚词	文本相似度结果
作者一与作者三	0.42	0.89	0.65	74.658 6
作者一与作者四	0.21	0.82	0.88	54.715 8
作者一与作者五	0.55	0.92	0.88	82.972 8
作者二与作者三	0.52	0.82	0.94	62.354 1
作者二与作者四	0.31	0.85	0.84	76.235 9
作者二与作者五	0.66	0.76	0.76	92.841 3
作者三与作者四	0.62	0.97	0.65	60.328 1
作者三与作者五	0.29	0.67	0.78	87.374 1
作者四与作者五	0.19	0.78	0.90	59.395 1
作者一与作者二	0.62	0.84	0.84	67.482 1

可以看出,作者四在词长度上明显与其他作者不同,经过实验发现作者四的平均词长度为 3.213 2,三字词、四字词占的比重较大,而其他作者都是在 2 ~ 3 之间徘徊。句子长度对结果影响较小,作者二、作者四和作者五句子长度较为相似,与其他句子差距也较小。情感分析对文章影响最小的原因是 5 名作者都是客观的新闻,对主观情感偏移较小。词汇特征的参数没有明显的规律说明多个特征影响权重不同,但是肯定是对风格裂缝识别实验有积极的影响。特殊标点符号的使用次数较为平均,只有作者三的特殊标点和大家较为相似。同义词和虚词上效果就比较明显,参数值较大,对结果影响也较大。从实验结果上看,同义词和虚词对结果影响较大,但是其他特征在特殊情况也能体现自己的写作风格。从写作风格提取结果进行风格相似度计算,发现作者一、作者三、作者五相似度较高。

4.4 风格裂缝识别实验

4.4.1 新闻集实验

风格裂缝识别数据集随机抽取上述 5 名编辑作者的 20 篇新闻,按照段落拆分,以段落为一个部分在实验开始的阶段使用滑动窗口技术,每次向下滑动一个句子,每次窗口句子数量为 5 个。随着实验的进行,在 K-means ++ 聚类的结果中聚类结果较差,因为每一次变化为一个句子,变化的幅度较小,每一次变化不明显导致聚类时邻近的窗口结果偏差较小。加上以段落结尾为风格裂缝出现的位置,准确率才会有所回升。在聚类过程中会导致 K 的结果不确定,是K-means ++ 算法中心点不准确导致的,所以许多聚类错误情况出现。实验结果如表 5 所示,可视化图如图 8 所示。

表 5 利用滑动窗口进行风格裂缝识别结果 %

作者	准确率	召回率	评估值
作者一	66.1	87.5	75.3
作者二	61.6	81.4	70.1
作者三	58.4	75.3	65.8
作者四	68.8	71.7	70.2
作者五	56.9	80.7	66.7

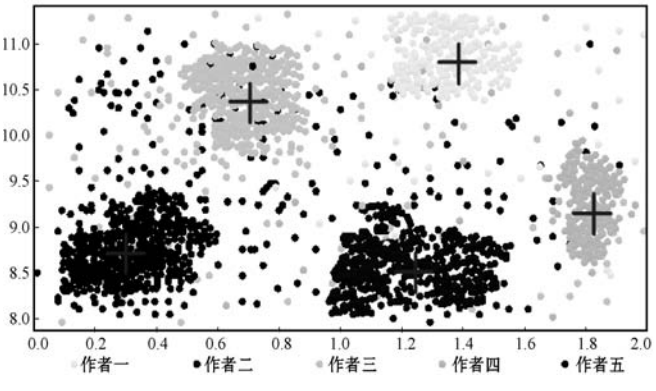


图 8 基于滑动窗口 K-means 可视化图

最后,本文放弃滑动窗口改用识别段落转换符,即把每一个段落视为一个作者完成的内容,以每个段落为单位进行风格特征提取,再根据提取的风格特征进行 K-means 聚类算法。实验结果如表 6 所示,可视化图如图 9 所示。

表 6 基于段落进行风格裂缝识别 %

作者	准确率	召回率	评估值
作者一	74.3	90.5	81.6
作者二	75.6	84.7	79.9
作者三	66.8	85.0	74.8

续表 6　　%

作者	准确率	召回率	评估值
作者四	71.5	81.7	76.3
作者五	67.9	92.7	78.4

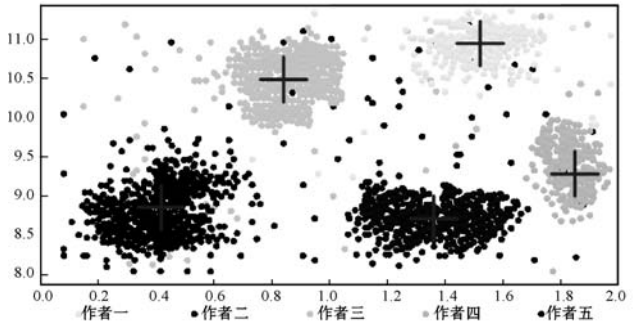


图 9　基于段落 K-means 可视化图

虽然滑动窗口的提出是为了尽可能全面地找出所有的风格裂缝点,但是由于每次变化一个句子,对结果变化不明显,风格聚类效果一般。基于段落进行风格裂缝识别效果要好于利用滑动窗口的实验,在准确率和召回率上都有所提升,在评估值上也能提升 10 个百分点。

4.4.2　小说集实验

《红楼梦》后 40 回原创性检测一直是文学家讨论的主要对象,本文对分割裂缝识别最后的实验就是以《红楼梦》为背景。使用《红楼梦》电子小说的网络资源进行全文下载,以每一回作为一个整体,进行风格特征提取,其中虚词不再使用本文的虚词表,而是使用 22 个文言文虚词表。对 120 回进行基于风格特征提取的风格聚类,结果统计分成前 40 回,中间 40 回,后 40 回,K-means算法的 K 值为 2,结果如表 7 所示,可视化图如图 10 所示。

表 7　红楼梦结果分析

章回	曹雪芹	高鹗	准确率/%
前 40 回	36	4	90.0
中 40 回	33	6	82.5
后 40 回	8	32	82.5

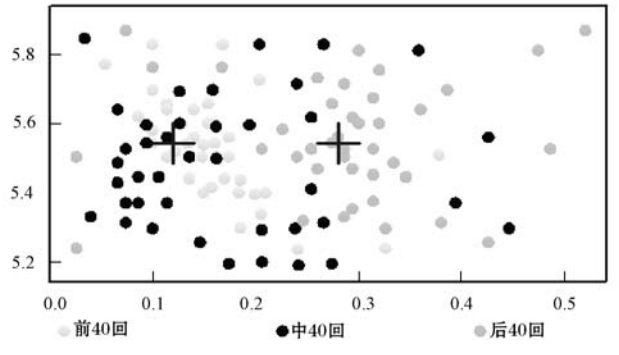


图 10　红楼梦结果分析可视化图

从结果可以看出,前 80 回的准确率较高,后 40 回相对偏低。通过对单独特征的实验结果发现,平均句子长度、情感分析和虚词对结果影响较大,前 80 回与后 40 回在句子长度上有明显的区别。情感分析影响较大的原因是,前 80 回偏积极,后 40 回偏消极,虚词对实验结果影响最大,22 个虚词表对实验结果影响较为积极。

5　结　语

本文提出一种多特征融合和无监督的机器学习算法相结合的方法进行风格裂缝识别。其中多特征融合是为了更好地提取作者的风格特征,而机器学习是以滑动窗口或段落为基准的,基于提取的特征进行分类,利用聚类算法对风格特征进行聚类,从而找到风格裂缝的位置。分别对新闻语料集和小说语料集进行实验,得出基于段落的裂缝识别比基于滑动窗口的实验效果在评估值上高出 10 个百分点,因此基于滑动窗口的实验思路仍需进一步改进。

参　考　文　献

[1] 刘悦. 基于语料库的红楼梦各部分写作风格研究[J]. 青年与社会,2014(3)(3):314 - 315.

[2] Tian S, Lu S, Su B, et al. Scene text segmentation with multi-level maximally stable extremal regions [C]//2014 22nd International Conference on Pattern Recognition, 2014.

[3] 刘耀,帅远华,龚幸伟,等. 基于领域本体的文本分割方法研究[J]. 计算机科学,2018,45(1):128 - 132,156.

[4] Xia F,Zhang Z. Study of text emotion analysis based on deep learning[C]//2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018.

[5] Sailunaz K, Alhajj R. Emotion and sentiment analysis from twitter text [J]. Journal of Computational Science, 2019, 36: 101003.

[6] 于涛,唐美华. 汉语小说中的词汇特征对比研究[J]. 重庆交通大学学报(社会科学版),2017,17(4):117 - 122.

[7] 李小涛,游树娟,陈维. 一种基于词义向量模型的词语语义相似度算法[J/OL]. 自动化学报:1 - 16 [2019 - 05 - 22]. https://doi.org/10.16383/j.aas.c180312.

[8] 傅彦铭,李振铎. 基于拉普拉斯机制的差分隐私保护 k-means ++ 聚类算法研究[J]. 信息网络安全,2019(2):43 - 52.