第 37 卷 第 9 期
Vol.37 No.9

计 算 机 工 程
Computer Engineering

2011 年 5 月
May 2011

• 博士论文 •

文章编号：1000—3428(2011)09—0004—02 文献标识码：A 中图分类号：TP311.13

# 基于语句节奏特征的作者身份识别研究

王少康[1,2]，董科军[1]，阎保平[1]

(1. 中国科学院计算机网络信息中心协同工作环境研究中心，北京 100190；2. 中国科学院研究生院，北京 100049)

**摘 要：**提出一种新的写作风格相似度评估方法，利用不同作者写作时在文章语句节奏控制方面的特点，鉴别作者的写作风格，从而达到作者身份识别的目的。该方法构建节奏特征矩阵模型来描述文本的语句节奏，利用点积相似度算法以及改进的 KL 距离算法来度量节奏特征矩阵之间的差异。实验表明，该方法在文学作品的作者识别方面具有较高的准确率。

**关键词：**文本挖掘；作者身份识别；文本相似度；节奏特征；多维矩阵

# Research on Authorship Identification Based on Sentence Rhythm Feature

WANG Shao-kang[1,2], DONG Ke-jun[1], YAN Bao-ping[1]

(1. Collaboration Environment Research Center of Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; 2. Graduate Universtiy of Chinese Academy of Sciences, Beijing 100049, China)

**【Abstract】**This paper proposes a new method of authorship similarity assessment, which identifies the authorship by sentence rhythm features of articles. The method constructs a rhythm feature matrix to describe the Sentence Rhythm Feature(SRF) of the text, and uses the inner product similarity algorithm and improves Kullback-Leibler(KL) divergence algorithm to measure the difference between the rhythm feature matrixes. Experiments show that it can make rather good results in literature authorship identification.

**【Key words】**text mining; authorship identification; text similarity; rhythm feature; multi-dimensional matrix

**DOI:** 10.3969/j.issn.1000-3428.2011.09.002

## 1 概述

作者的写作风格体现在其文章在语法、词汇、修辞以及句型等方面的特点上。类似于笔迹，不同的作者拥有各不相同的写作风格，利用这一点可以概率性地推断某篇特定文本的作者归属，做出作者身份识别判定，这在司法鉴定、信息安全以及文献考证等领域拥有广泛的应用前景，具有很高的实践意义。以往的研究者在写作风格识别领域做了大量的工作。文献[1]以平均词长和句长、词的搭配及词在句子中的位置等作为测量标准，利用卡方检验(Chi-square test)来判别 Shakespeare 和 Marlowe 作品的差别。文献[2]提出了一种“累积和(cumulative sum-cusum)”的判定方法来进行作者识别，达到了不错的效果。文献[3]在作者身份可视化方面做了很有意义的探索。国内方面，文献[4]利用语义网本体技术，探讨了论文本体的构建和论文相似度的计算。文献[5]使用虚词作为特征，从而构建向量空间模型来实现作者识别。文献[6]依据文体学理论，利用 HowNet 知识库，提出一种基于词汇语义分析的相似度评估方法。

本文的思路与以往研究者不同，将关注点放在一个较易被人忽视的写作特征——文章的语句节奏上，通过对不同作者的文章语句节奏进行分析，鉴别作者的写作风格，从而达到作者身份识别的目的。

## 2 语句节奏特征分析

### 2.1 文本的语句节奏

倪宝元主编的《大学修辞》中对“节奏”下了一个定义：语音的疾徐、高低、长短、轻重及音色的异同，在一定的时间内有规律地相间交替回环往复的组合形式。笔者有这样的体会，不同的文章在朗诵时往往会拥有其独特的节奏和韵律，这个节奏与朗诵者对文章的诠释无关，它是一种文章语句内在的奇妙韵律，就像音乐一样。正如高低音符的起伏排布形成了音乐，长短语句的交错运用以及分段的把握形成了文章的独特节奏，这种节奏虽然常常被人忽视，但它是一种客观的存在，特别是在文学作品中，语句节奏甚至是作品风格的重要组成部分。以往有研究者总结长短语句所代表的风格如表 1 所示。

**表 1 长短语句所表现的风格**

| 分类标准 | 具体分类 | 表现的风格 |
|---|---|---|
| 句子长度 | 长句 | 沉郁凝重、委曲细腻、繁丰华丽、热烈圆柔、气势雄浑 |
| | 短句 | 轻松活泼、明快粗率、简洁朴素、冷峻峭拔、慷慨激昂 |
| 句子结构 | 长短划一、结构整齐的整句 | 音节匀称、明畅流丽、简练经济、以少总多 |
| | 长短不等、结构自由的散句 | 音节参差、委婉迂徐、铺张扬厉、酣畅淋漓 |

笔者发现，文章的语句节奏具有强烈的作者个人色彩，它反映着作者表达文字的一种习惯和风格。也就是说，不同作者文章的语句节奏不尽相同，而同一作者的不同作品之间，其节奏却往往类似，正如当代著名作家王小波所说，优秀的

**基金项目：**国家“863”计划基金资助项目“科学数据网格及科研应用系统”(2006AA01A120)；中国科学院信息化基金资助项目“e-Science 虚拟科研平台研究与开发”(INFO-115-D01)；CNIC 青年基金资助项目“基于 Web 的传感器监控管理系统”(CNIC_QN_09005)

**作者简介：**王少康(1981－)，男，博士研究生，主研方向：智能搜索引擎，下一代互联网；董科军，副研究员、博士；阎保平，研究员、博士后、博士生导师

**收稿日期：**2010-10-25 **E-mail：**skwang@cnic.cn

作者所写的文章会“带有一种永难忘记的韵律”。因此，此特性可用来进行文章作者的识别。基于此假定，笔者构建数学模型，以多维概率矩阵的方式描述文本的语句节奏，并度量其间的差异，从而达到作者身份识别的目的。

**定义 1**(停顿性标点符号) 文章中产生停顿效果的标点符号，如逗号、句号、感叹号等；相应的，不产生停顿效果的标点符号为非停顿性标点符号，如引号、书名号等。

**定义 2**(语句节奏) 停顿性标点符号将文章分割成一个个的短句，这些短句的交错运用和起伏变化所产生的韵律称为文章的语句节奏。

为了更接近语句的朗读效果，本文以音节而不是字数来衡量语句的长度。中文是单音节的语言，所以可以认为一个中文字符的音节长度为 1；西文以单词作为基本语言单位，每个单词的音节数不确定，在实践中，为了计算简便，本文近似认为每个西文单词的音节长度为 2；短句中的非停顿性标点符号不计音节长度。文本的自然段分段对文本节奏会产生不同于标点符号的影响，这里加入“分段符”的概念，并认为一个分段符的音节长度为 0，分段符自身构成一个短句。这样可将一篇文本分割成不同音节长度的短句序列，从而建立节奏特征矩阵来描述该文本的语句节奏。

**定义 3**(节奏特征矩阵) 一个 $m$ 维的矩阵，$a_{i_1i_2\cdots i_p}$ 为 $\boldsymbol{A}$ 的矩阵空间中($i_1,i_2,\cdots,i_p$)处元素的值，如果矩阵中的每个元素值 $a_{i_1i_2\cdots i_p}$ 等于文档 $D$ 的短句序列中音节长度序列 $i_1i_2\cdots i_p$ 的出现频率，则该矩阵 $\boldsymbol{A}$ 称为文档 $D$ 的节奏特征矩阵。其定义公式如下：

$$\boldsymbol{A}=\left\|a_{i_1i_2\cdots i_p}\right\|$$

$$a_{i_1i_2\cdots i_p}=\frac{c_{i_1i_2\cdots i_p}}{\sum_{c\in\boldsymbol{A}}c_{i_1i_2\cdots i_p}} \tag{1}$$

其中，$c_{i_1i_2\cdots i_p}$ 为文档 $D$ 的短句序列中音节长度序列 $i_1i_2\cdots i_p$ 出现的次数。

显而易见，节奏特征矩阵之中的各元素之和为 1，可以认为，节奏特征矩阵是一种概率分布矩阵，该矩阵以音节长度序列出现概率的形式描述文档的语句节奏特征。

### 2.2 节奏差别度量

衡量两篇文本的节奏相似性，首先将每篇文本都按照 2.1 节的定义构建各自的节奏特征矩阵，然后计算节奏特征矩阵之间的相似性。可以认为，节奏特征矩阵的相似性反映了文本之间的节奏相似性，进而可以用来识别作者。

点积法常被用来计算矩阵的相似度，Kullback-Leibler Divergence(简称 KL 距离)算法常被用来测量同一事件空间中的 2 个概率分布之间的差别，本文将分别利用这 2 种算法来度量不同的节奏特征矩阵之间的差异。

#### 2.2.1 点积相似度算法(点积法)

矩阵之间的数值点积可做作为矩阵相似度的衡量算法，定义如下：

$$sim(\boldsymbol{A},\boldsymbol{B})=\sum_{a\in\boldsymbol{A},b\in\boldsymbol{B}}a_{i_1i_2\cdots i_p}\times b_{i_1i_2\cdots i_p} \tag{2}$$

其中，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 为 2 个节奏特征矩阵；$a_{i_1i_2\cdots i_p}$ 为 $\boldsymbol{A}$ 的矩阵空间中($i_1,i_2,\cdots,i_p$)处元素的值；$b_{i_1i_2\cdots i_p}$ 为 $\boldsymbol{B}$ 的矩阵空间中($i_1,i_2,\cdots,i_p$)处元素的值。函数 $sim(\boldsymbol{A},\boldsymbol{B})$的值越大，说明 $\boldsymbol{A},\boldsymbol{B}$ 矩阵的相似程度越高。

#### 2.2.2 改进的 KL 距离算法(KL 法)

在 Kullback-Leibler Divergence 算法的定义中，2 个概率分布 $P$、$Q$ 在事件空间 $E$ 中 KL 距离的定义如下：

$$D(P\|Q)=\sum_{x\in E}P(x)\mathrm{lb}\frac{P(x)}{Q(x)} \tag{3}$$

对其进行对称平滑化后得到 Jensen-Shannon Divergence：

$$JSD(P\|Q)=\frac{1}{2}D(P\|M)+\frac{1}{2}D(Q\|M) \tag{4}$$

$$M=\frac{1}{2}(P+Q) \tag{5}$$

节奏特征矩阵是一个音节长度的概率分布矩阵，将以上算法针对节奏特征矩阵进行套用，则矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 之间的差别度量 $RD(\boldsymbol{A},\boldsymbol{B})$为：

$$RD(\boldsymbol{A},\boldsymbol{B})=\sum_{a\in\boldsymbol{A},b\in\boldsymbol{B}}\Big(\frac{1}{2}a_{i_1i_2\cdots i_p}\times\mathrm{lb}\frac{a_{i_1i_2\cdots i_p}}{\frac{1}{2}(a_{i_1i_2\cdots i_p}+b_{i_1i_2\cdots i_p})}+\frac{1}{2}b_{i_1i_2\cdots i_p}\times\mathrm{lb}\frac{b_{i_1i_2\cdots i_p}}{\frac{1}{2}(a_{i_1i_2\cdots i_p}+b_{i_1i_2\cdots i_p})}\Big) \tag{6}$$

在文档 $D$ 的节奏特征矩阵 $\boldsymbol{A}$ 中，其元素 $a_{i_1i_2\cdots i_p}$ 为 0 的情况经常出现，它意味着 $D$ 中不存在音节长度序列为 $i_1i_2\cdots i_p$ 的短句序列。这种情况会引发式(6)中出现除数为 0 的计算，于是，本文引入极小值 $\varepsilon$ 来代替 0 元素。为保证节奏特征矩阵中的所有音节长度序列出现的概率之和为 1，须将矩阵中的非 0 元素折换为原来的 $\beta$ 倍(0< $\beta$ <1)。

于是改进的节奏特征矩阵重新定义为：

$$\boldsymbol{A}=\left\|a_{i_1i_2\cdots i_p}\right\|$$

$$a_{i_1i_2\cdots i_p}=\begin{cases}\dfrac{\beta\times c_{i_1i_2\cdots i_p}}{\sum_{c\in\boldsymbol{A}}c_{i_1i_2\cdots i_p}},\ c_{i_1i_2\cdots i_p}>0\\ \varepsilon,\ c_{i_1i_2\cdots i_p}=0\end{cases}$$

$$\beta=1-\sum_{c\in\boldsymbol{A},c_{i_1i_2\cdots i_p}=0}\varepsilon \tag{7}$$

## 3 实验分析

本文选取 10 位现代著名作家的作品进行实验，这 10 位作家分别是鲁迅、老舍、巴金、茅盾、金庸、古龙、梁羽生、张爱玲、王朔、余秋雨。因为作家的每部作品的长短不同，为了公平起见，按照文本长度而不是作品篇数来构建实验语料库。本文为每个作家选取 50 万字的作品文本构建语料库，分别建立起他们自己的 $n$ 维节奏特征矩阵($n\in[1.10]$)。

本文为每个作家在语料库之外随机选取了 20 段文本，每段 5 000 字，共 200 段文本 ；分别建立这 200 段文本的 $n$ 维节奏特征矩阵($n\in[1.10]$)；根据本文提供的方法将这些节奏特征矩阵与语料库中的矩阵进行比对，进而识别作者身份。实验结果如表 2 所示。

**表 2 点积法以及 KL 法的识别结果**

| 节奏特征矩阵维度 | 点积法 | | KL 法 | |
|---|---|---|---|---|
| | 识别篇数 | 识别率 | 识别篇数 | 识别率 |
| 1 | 84 | 0.420 | 148 | 0.74 |
| 2 | 84 | 0.420 | 152 | 0.76 |
| 3 | 93 | 0.465 | 164 | 0.82 |
| 4 | 104 | 0.520 | 164 | 0.82 |
| 5 | 141 | 0.705 | 132 | 0.66 |
| 6 | 153 | 0.765 | 108 | 0.54 |
| 7 | 150 | 0.750 | 112 | 0.56 |
| 8 | 150 | 0.750 | 104 | 0.52 |
| 9 | 142 | 0.710 | 96 | 0.48 |
| 10 | 121 | 0.605 | 102 | 0.51 |

识别率以折线图的方式表现如图 1 所示。可见，节奏特征矩阵维度数量的选择对识别结果有较大的影响。在利用点积法计算相似度的情况下，识别率的峰值出现在维度等于 6~8

(下转第 8 页)

新的目标中轴线。

### 4.4 迭代判断

重复上述步骤，则矩形拟合区域与目标真实形状逐渐趋于一致，中轴线计算精度也逐步提高。当中轴线角度计算结果趋于稳定时，退出迭代过程，输出最终计算结果。迭代终止条件可设置为 2 次计算目标倾角差小于某个阈值，或迭代次数超过最大限制。

## 5 实验结果

为验证本文方法的有效性和鲁棒性，在 CPU 为 Intel Core 2.33 GHz、内存 2 GB 的 PC 机上，采用 Matlab7.0 开发环境，分别对仿真图像与实际图像进行测试。

### 5.1 仿真图精度提取测试

利用仿真序列图像验证算法提取精度，目标由圆柱体加入高斯噪声构成，目标灰度沿中轴线垂线方向呈高斯分布。以 15°为间隔制作弹体倾角为 0°,15°,…,90°的仿真图 7 幅，弹体长度为 50 像素。采用本文方法提取目标面内倾角（中轴线与图像坐标系 $x$ 轴的夹角），迭代终止条件：角度变化量小于 0.05°。

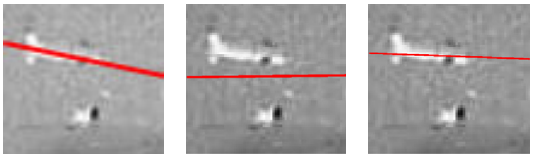
计算结果如表 1 所示，单位为度。提取误差标准差为 0.028 2，平均迭代次数为 10 次，平均用时 0.18 s，处理速度约为 5 帧/s。因靶场图像处理大多为事后处理工作，对精度的要求优先于计算速度的要求，本文算法满足判读要求。

**表 1 加噪声仿真图像的轴线提取结果**

| 理论值 | 计算结果 | 误差 | 迭代次数 |
|---|---|---|---|
| 0.000 | 0.058 | 0.058 | 4 |
| 15.000 | 14.975 | 0.025 | 16 |
| 30.000 | 30.016 | 0.016 | 12 |
| 45.000 | 44.913 | 0.087 | 14 |
| 60.000 | 60.047 | 0.047 | 12 |
| 75.000 | 75.003 | 0.003 | 9 |
| 90.000 | 89.968 | 0.032 | 4 |

### 5.2 实际图像测试

针对实际典型小目标图像，采用手工提取法、矩方法与本文方法分别提取目标中轴线进行比对。限于篇幅，选取一组图像为例，计算结果如图 4 所示。从图中可以看出，采用本文方法与手工标记方法结果一致，提取结果明显优于传统矩方法。



(a)手工提取方法 (b)传统矩方法 (c)本文方法

**图 4 提取结果比较**

## 6 结束语

本文提出了一种最优区域拟合的中轴线提取算法，较之于常规基于目标边缘的算法更适合于靶场小目标的中轴线提取，已在靶场目标三维姿态判读中得到了成功的应用。目前，算法对质心位置参数较为敏感，下一步可考虑利用跟踪算法自动确定质心位置。

### 参考文献


[1] 于起峰, 尚 洋. 摄像测量学原理与应用研究[M]. 北京: 科学出版社, 2009.

[2] 于起峰, 孙祥一, 陈国军. 用光测图像确定空间目标俯仰角和偏航角的中轴线法[J]. 国防科技大学学报, 2000, 22(2): 15-19.

[3] 徐秋平, 郭 敏. 基于变宽邻域图割和活动轮廓的目标分割方法[J]. 计算机工程, 2009, 35(8): 233-237.

[4] 魏 敏, 吴国华, 周 进, 等. 序列图像轴对称物体中轴线提取方法[J]. 半导体光电, 2007, 28(1): 143-146.

[5] 康文静, 丁雪梅, 崔继文, 等. 基于改进Hough变换的直线图形快速提取算法[J]. 光电工程, 2007, 34(3): 105-108.

[6] 姜泳水, 唐金辉, 陈学佺. 二值图像中物体几何主轴的提取方法[J]. 计算机工程, 2005, 31(18): 56-58.

[7] 孙即祥. 图像分析[M]. 北京: 科学出版社, 2005.


编辑 任吉慧

(上接第 5 页)

之间，随后缓慢下降；而在 KL 法的情况下，识别率的峰值出现在维度等于 3~4 之间。在峰值条件下，2 种算法的准确率都在 80%左右，KL 法略高于点积法。

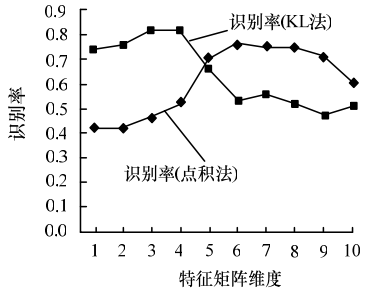



**图 1 点积法以及 KL 法的识别结果**

## 4 结束语

本文提出一种基于语句节奏特征分析的作者身份识别方法，重点关注于作者写作时内在的语言节奏风格。本方法构建节奏特征矩阵模型来描述文本的语句节奏，利用点积相似度算法以及改进的 KL 距离算法来度量节奏特征矩阵之间的差异。实验表明，在节奏特征矩阵维度选择得当时，本方法拥有相当不错的识别结果。

下一步工作将进一步研究节奏特征矩阵维度与识别结果的关系，以及实验文本长度对识别结果的影响，从而改进节奏特征矩阵的相似度算法，使其具有更高的准确性。

### 参考文献


[1] Smith M W. Recent Experience and New Developments of Methods for Determination of Authorship[J]. ALLC Bulletin, 1983, 11(3): 73-82.

[2] Farringdon J M, Morton A Q, Baker M D. Analysing for Authorship: A Guide to the Cusum Technique[M]. Cardiff, UK: University of Wales Press, 1996.

[3] Soboroff L M, Nicholas C K, Kukla J M, et al. Visualizing Document Authorship Using N-grams and Latent Semantic Indexing[C]//Proc. of Workshop on New Paradigms in Information Visualization and Manipulation. Las Vegas, USA: ACM Press, 1997.

[4] 聂规划, 付志超, 陈冬林, 等. 基于本体的论文复制检测系统[J]. 计算机工程, 2009, 35(6): 79-84.

[5] 孙晓明, 马少平. 基于写作风格的作者识别[C]//中国中文信息学会第五届全国会员代表大会暨成立二十周年学术会议论文集. 北京: 清华大学出版社, 2001.

[6] 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.


编辑 任吉慧