文章编号：#128

基于文本聚类的语言韵律和节奏风格特征挖掘 *

贺湘情[1]，刘颖[1]

（1. 清华大学人文学院中国语言文学系，北京 100084）

摘要：本文以朱自清、汪曾祺和刘亮程的散文作品为语料，旨在从文本的韵律和节奏出发，采用文本聚类的方法来挖掘出新的能够代表作品风格的特征。实验表明以句末用字韵母的 n 元组合、分句句长的 n 元组合、标点符号和整句句长作为风格特征，能成功地将这三位作家的作品区分开来。其中刘亮程句尾韵的舌位高于汪、朱二人，朱自清对韵脚的选择不如刘、汪二人丰富。汪曾祺的分句长最短，且最为讲究句式长短的对齐；刘亮程兼顾长短句的交错，节奏更富于变化；朱自清的句长变化最为平稳。

关键词：特征挖掘；韵律；节奏；文本聚类

中图分类号：TP391　　**文献标识码：**A

Mining Stylistic Features of Rhythm and Tempo Based on Text Clustering

Xiangqing He[1], Ying Liu[1]

(1.Department of Chinese Language and Literature, Beijing, 100084, China)

Abstract: We selected literary proses written by Ziqing Zhu, Zengqi Wang and Liangcheng Liu as corpora. Text clustering is used to mine new stylistic features from the perspective of rhythm and tempo. The experimental results show that n-grams based on the vowels of the last character of the sentence, n-grams based on the length of clauses, punctuations and length of sentences, all can successfully distinguish from the articles of the three authors. Specifically, Liangcheng Liu preferred to utilize the vowels of higher tongue position. Ziqing Zhu focused on some specific rhymes, but the rhymes used by Liu and Wang are more plentiful than those of Zhu. Wang's Clauses are the shortest, and he paid more attention to the order of sentence patterns. Long sentences and short sentences are alternatively used by Liu, and the tempos used by Liu are changeful. The sentence lengths used by Zhu are less changeful.

Keywords: Feature Mining; Rhythm; Tempo; Text Clustering

1 引言

计算风格学是数理语言学的一个分支，其所建立的基础是认为写作是个人将自己的思想用文字表达出来的一种活动，它隐含着作者个人独特的独立于表达内容的表达方式，即编排自己思想的方式，其在文本中表现为可衡量的语言特征，即遣词造句的习惯，且这种习惯常常是潜意识的，未被作者本人所察觉。

因此我们可以通过量化文本中的语言结构单位来刻画、研究语体、作品或作家的风格，也就是说，与传统的风格学（又称修辞学）靠体悟式的归纳、自我内省体验去意会作品风格所不同，计算风格学建立的基础是可供量化的语言学特征。对计算风格学来说，其关键点在于提取出能代表或区分不同风格的特征项，并且这些特征是可被量化统计且稳定出现的。

计算风格学的研究方法经历了从简单地统计某些特定语言结构单位的出现频率，到引

* 收稿日期：2014 年 7 月 18 日　　定稿日期：

基金项目：国家自然科学基金“基于语用信息的交互行为与语言特征的建模研究”（61171114）；教育部自主科研项目“基于大规模语料库的社会语用信息网的构建”（20111081010）

入 t 检验[1]、卡方检验[2]等假设检验统计量，再到使用典型相关分析[3]、主成分分析、因子分析[4]等多元统计方法的发展历程，目前最前沿的研究方法是利用机器学习领域中的文本聚类和文本分类模型来实现计算机基于作品风格的自动文本区分。

由于计算机无法直接处理文本数据，因此利用机器学习的方法来进行文本聚类首先需要将原始文本转化为由特征表示的结构化数据，以便于计算机存储、读取和计算。目前使用最广泛的文本表示模型是向量空间模型，其背后的概念即为将文本表示成向量空间中的向量，而能够代表文本的特征项则为该文本向量的分量，因此挖掘出能代表作品风格的特征项是最终实现计算机自动、准确地区分出不同风格作品的关键，这些特征的特点是在同一风格的一系列作品中稳定出现，且能使其与其他作品区分开来。

目前已提出并证实能代表作品风格的可计量的语言结构特征可分为词汇、句子、段落、语法、语义等五个层面。其中对词汇层面的挖掘最为成熟、深入而透彻，包括高频词、虚词[5]、单现词、关联词、词长、词的形符类符比等；句子层面包括句长、标点符号比例[6]、句类分布[7]、句式分布和句型分布[8]等；段落层面包括段落长、开端词、末端词、起始句、终结句、段落格式等；语法层面包括词性[9]、依存语法关系[10]等；语义层面包括基于 HowNet 的语义类[11]、基于概念词典的概念[12]以及基于哈工大同义词词林（扩展版）的同义词词集[13]等。

综上所述，我们不难发现目前基于风格的特征挖掘在语音层面上还比较薄弱，然而语音是语言的物质外壳，韦勒克和沃伦在其合著《文学原理》中就曾指出：“一件文学作品首先是一套声音的系统”，人们在阅读或写作的时候，即使没有读出声来，也必定会在心中“默念”。得益于语言的音乐美，每个文学作品都有着自身的节奏和韵律，它们是构成作品风格的不可忽视的重要因素，尤其是在诗歌和散文这两类文体中，和谐的语音、抑扬顿挫的语调以及语句变化的节奏本身就是审美对象。

本研究基于北大计算语言研究所所研发的《现代汉语语法信息词典》开发了一个拼音转化分析软件，用于将文本转换为现代汉语拼音，有效地解决了多音字标音的问题，准确率可达到 98%以上。此外该软件还能灵活地提取全文本以及位于词首、词尾、分句首、分句尾、句首、句尾的声母、韵母和声调，以及分句句长和整句句长，从而作为向量特征来构建文本向量，再利用基于 KL 散度（Kullback-Leibler Divergence, KLD）的层次聚类法对三位作家的文本进行聚类以检验特征的有效性，并进一步分析特征项能够代表作品风格的原因。

2 层次聚类

文本聚类是一种无监督的机器学习方法，与文本分类不同的是，它无需预先对文本进行人工类别标注，也不需要训练过程，而是通过计算文本之间的距离来作为衡量它们之间相似度的标准，最终将文本集合分组成多个类或簇，使得同一个簇中的文本具有较高的相似度，而不同簇之间的文本内容差异较大。

层次聚类是文本聚类的算法之一，按照分类原理的不同，又可划分为凝聚和分裂两种方法[14]，本文所采用的是凝聚型层次聚类法。该算法由树状结构的底部开始逐层向上进行聚合[15]，即最初将每个样本看作一个簇，然后根据距离最终将其合并为一个簇。

本研究采用 KL 散度（Kullback-Leibler Divergence, KLD），又称相对熵来度量算任意两个样本之间的距离，同时采用离差平方和法来度量类与类之间的距离。对于归一化的文本向量 P（X_1，X_2，……Xn）和 Q（Y_1，Y_2，……Yn），向量特征值的总和均为 1，且对于任何 n 皆满足 Xn>0 及 Yn>0，因此文本 P 和 Q 之间的 KL 散度可定义为：

$$KL(P \| Q) = \sum_{i=1}^{n} P(x) \log \frac{P(x)}{Q(y)} \qquad (1)$$

金明哲对其进行优化后获得了更好的聚类效果，因此本文采用其所提出的公式来来计算文本之间的 KL 散度，KL 散度越小，则代表样本之间的相似度越高。

$$KLD(P,Q)=\frac{1}{2}\sum_{i=1}^{n}[P(x_i)\log\frac{2P(x_i)}{Q(x_i)+P(y_i)}+Q(y_i)\log\frac{2Q(y_i)}{Q(x_i)+P(y_i)}] \quad (2)$$

3 实验及结果分析

3.1 语料资源

朱光潜曾在《谈美·谈文学》[16]中指出文学的情趣大半要靠声音节奏来表现，因此本研究定位于散文文体，从朱自清、汪曾祺和刘亮程三位作家的散文集中各选取了 40 篇作品。由于单个作品的字数参差不齐，为确保语料大小的同一，我们根据字数将每位作家的作品进行组合，以确保合并后的文本字数均保持在 10000 字左右。合并之后刘、汪、朱三人的文本数分别为 6、7、7。语料库规模的详细信息如表 1 所示：

表 1 语料库规模统计

作家	总字数	总词数	总句数	总分句数
朱自清	74,072	65,655	3153	9452
汪曾祺	72,841	64,917	3651	9733
刘亮程	60,717	51,943	2694	6644

3.2 以句末字韵母的 n 元组合为特征的韵律分析

中国古代的诗歌讲究押韵以达到音韵的和谐，朗读的顺畅，而优美的散文作品虽然并不苛责严格的押韵，却同样在韵脚的选择上有所讲究与经营。本文所选取的三位作家分别是现代、现当代和当代散文家的杰出代表，其中朱自清和汪曾祺受过良好的传统文化熏陶，古文功底深厚，因此其在散文创作中能有意识地传承古典诗歌中的声律元素，重视字句的声音。刘亮程虽然没有接受过系统的古文训练，但诗人出身的他曾在《对一个村庄的认识》中自白道：“我努力让自己像写诗一样写每一篇散文”，其近乎诗化的散文语言也同样是很讲究韵律和节奏感的。

因此我们以句末字韵母出现的频率及其 n 元组合作为特征来进行文本聚类，其中前者可反映出三位作家在句中偏好于以何韵母结尾，从而推断出其在韵脚的选用上有何异同；而后者则可折射出文本押韵的情况。

需要说明的是，所谓句末字韵母的 n 元组合是指基于文本每句的最后一个字的韵母所构建 n 元语法模型，为了全面考察文本连续押韵和隔句押韵的情况，我们分别提取了 2 元、3 元和 4 元语法模型中出现频率排名前 100 位的组合来共同组成特征项，向量空间总维度为 373 维。文本聚类的结果如图 1 所示，其中“z”指代朱自清，“w”指代汪曾祺，“l”指代刘亮程（下同）。

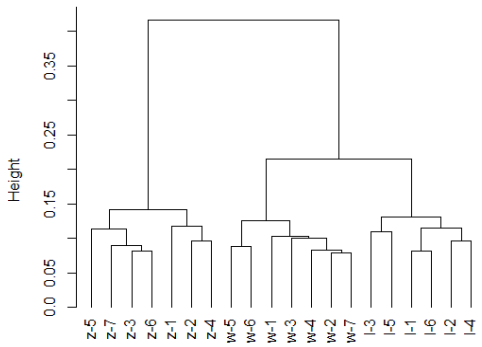

图 1 基于句末字韵母的 n 元组合的聚类结果

由图 1 我们可以发现以文本句末的用韵作为特征能很好地将三位作家的作品区分开来，也就是说句末韵母的使用和押韵情况能成为区分作者风格的有效特征。

3.2.1 句末字用韵

为了进一步探明三位作家在句末韵母的使用和押韵上有何异同之处，其句末用字的韵母分布频率图如图 2 所示：

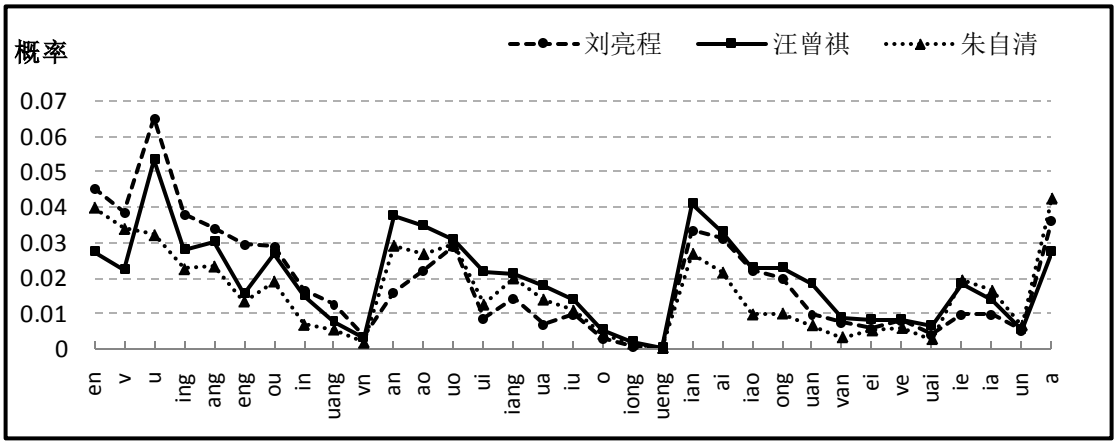

图 2 三位作家句末用字韵母分布频率图

图 2 是三位作家句末用字的韵母分布频率图，需要说明的是，由于韵母 e 和 i 的出现频率远远高于其它，它们共同在刘、汪、朱三位作家句尾所使用的韵母中分别占据了 38%、35%和 48%的比例，因此为了作图清楚，图 2 便将其省略。

由图 2 可以看出，三位作家句末字韵母的分布趋势大致相同，与汤云航[17]对汉语本身韵母的出现频率的统计结果相比，uo、iu、ong、ui、ei、ie、iong 在句末的使用明显少于其在普通话韵母中的出现概率，体现了句末用韵的独特性。此外我们还可以看到，总的来说刘亮程较常在句尾使用 u、ü、en、ing、eng 等舌位较高的韵；而汪曾祺除了 u 之外，更多地使用 an、ao、ian、ai、iao、uan 等舌位较低的韵，朱自清则相比更倾向于使用 a、ia 这类开口度较大的韵，反映出不同作家的句末用韵偏好。

由于在汉语中不同的字可能共享一个韵，因此我们试图将韵还原成字来进一步探讨其句尾用韵差异的原因。由于看单个的字可能没有意义，因此我们首先以词为单位将三位作家出现频率位于前 100 位的句尾词提取出来，再按照句末字的韵母加以归类。以元音 a、e、i、ü、u 所对应的字为例，结果如下表所示：

表 2 基于句末字元音的前 100 个句尾词归类表

a	刘亮程	吧、吗、啊、啥、它、大、马
	汪曾祺	吧、吗、大、茶
	朱自清	吧、吗、啊、罢、啦、哪、他、她、它、大
e	刘亮程	的、地、了、什么、呢、着、者
	汪曾祺	的、似的、得、极了、了、什么、么、呢、着、喝
	朱自清	的、似的、罢了、了、儿、什么、么、呢、着
i	刘亮程	子、村子、房子、麦子、日子、样子、叶子、一辈子、院子、里、村里、家里、那里、事、你、死、自己、意义
	汪曾祺	子、孩子、栗子、里、笔、字、意思、吃、事、而已、次、东西、关系、记、问题
	朱自清	子、孩子、影子、样子、里、心里、这里、哩、意思、而已、你、

		次、故事、记、名字、气、如此、是、为止、正义、自己、字
ü	刘亮程	去、过去、死去、下去、驴
	汪曾祺	去、兴趣
	朱自清	去、过去、回去、绿
u	刘亮程	处、动物、路、清楚、土、住
	汪曾祺	豆腐、湖、建筑、书、舒服
	朱自清	出、处、妇、路、书

由表 2 我们可以看出，朱自清在元音 a 使用上的凸显主要归根于其丰富的句末语气词的使用以及第三人称代词。此外就元音在发音上的响度而言，a 的响度最高，最能抒发强烈的感情，这也正好契合了下文对其文中多用感叹号的发现。

元音 e 不但在三位作家句尾韵中的出现频率均最高,而且其所对应的用字也大体一致，多为常用的助词和语气词。元音 i 的出现仅次于 e，不过其所对应的用字在三位作家中存在明显差异。刘亮程 i 韵的来源主要是粘附在词根后面的词缀“子”和表地点的“里”，且意象多为农村日常生活中常见的实物；而汪、朱的来源则相对较广，多为抽象名词。元音ü所对应的用字主要是加在动词后面表示趋向的“去”，而 u 则多对应“处、路、书”，特别的是汪文中的“豆腐”和“舒服”都是轻声，而刘、朱则没有。

3.2.2 基于句末字韵母的 n 元组合

韵母由韵首、韵腹、韵尾三部分，一般而言只要韵尾相同，韵腹相同或者相近即可视为押韵。我们根据《中华新韵（十四韵）简表》，分别提取了 2 元组合前 100 位，3 元组合前 50 位和 4 元组合前 50 位中押韵的组合，表 3 列出了主要押韵组合：

表 3 主要押韵组合

二元			三元			四元		
刘亮程	汪曾祺	朱自清	刘亮程	汪曾祺	朱自清	刘亮程	汪曾祺	朱自清
i i	e e	e e	i i i	e e e	e e e	e e e e	e e e e	e e e e
e e	i i	i i	e e e	e i e	e i e	e i e i	i e i e	e i i e
uo e	e uo	e uo	e i e	i e i	e a e	i i i i	e i i e	i e e i
u u	u u	uo e	i e i	i i i	i e i	i e i e	i e ou i	i e i e
e uo	uo e	a a	i u i	e ai e	e en e	e e uo e	e i e i	e i e i
ing ing	an an	en en	i a i	e an e	e v e	uo i i e	i e e i	e e uo e
ian ian	ian ian	uo uo	i en i	e u e	e u e	i e e i	e e e uo	
a a	uo uo		e u e	e ian e	i i i	uo e e uo	i i i i	
ang ang	a a		i ou i	e e ou	e ao e	u i i u		
ao ao	an ian		e uo e	i ao i	e ia e			
uo uo	ao ao		e e uo	e ing e	e an e			
ai ai	an uan		i v i		e ou e			
			i eng i		e ing e			
			i uo i		e ai e			
			i ao i		e ang e			
					e uo e			

					a e a			

从表 3 我们可以看出，在相邻两句的押韵上，朱自清对韵脚的选择较为单一，种类明显不如刘、汪二人丰富，且一半都属于“波”韵部（e、uo）。而刘、汪二人虽然韵脚的数量一致，但刘亮程除了反复使用“波”韵部外，其它韵脚所属的韵部则较为分散，而汪曾祺则集中于“波”和“寒”（an、ian、uan）这两个韵部上。

在相邻三句的押韵上，主要统计了“一韵到底”（如“i i i”）和“隔句押韵”（如“e i e”这两种形式。我们不难发现一韵到底只发生在“齐”韵部（i）和“波”韵部中，其余均为隔句押韵。而在隔句押韵中，刘亮程青睐押“i”韵，朱自清和汪曾祺则偏好押“e”韵，且朱自清中间插入的其它韵的形式更为丰富。

在相邻四句的押韵上，主要统计了“一韵到底”（如“e e e e”）、“隔句押韵”（如“e i e i”和“回环押韵”（如“e i i e”）这三种形式。由上表可知虽然刘亮程回环押韵的形式较另两位更多，但在数量的排名上比较靠后，因此与其对一韵到底和隔句押韵的使用相比，汪曾祺和朱自清在回环押韵上的使用更为突出。

3.3 以分句句长的 n 元组合为特征的节奏分析

散文的美除了韵律之外，还离不开节奏。所谓节奏，即是在一定时间间隔里的某种对立形式的反复。文章的节奏主要体现在顿歇上，从语言结构特征出发，可通过统计分句长及其前后连续变化来抽取隐藏的节奏。所谓分句是指复句中相对独立的单句形式，在本研究通过检测分号、逗号、冒号、顿号和破折号来加以识别。我们提取了所有单个分句长（以字数为单位）以及分句长的 2、3、4、5 元组合的前 196、117、55、46 位为特征对文本进行聚类，空间维度共 443 维，结果如图 3 所示：

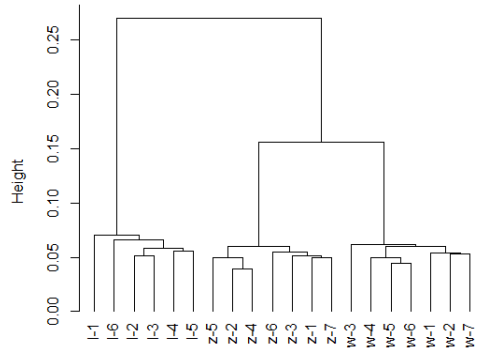

图 3 基于分句长的 n 元组合的聚类结果

由图 3 可知，以分句长及其 n 元组合为特征能成功地将三位作家的文本区分开来，说明它们是行之有效的能够反映作品风格的特征。

3.3.1 分句句长

为进一步比较三位作家在分句句长的连续使用上有何异同，其分句句长的频率分布图如图 4 所示：

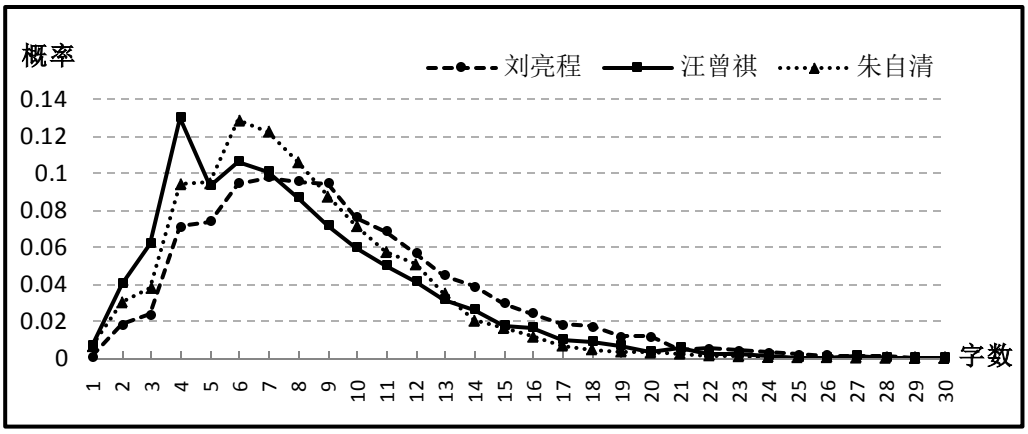

图 4 三位作家分句句长频率分布图

由图 4 我们可以看出，总体来说，汪曾祺的分句长最短，并大量使用四字分句；其次是朱自清，主要集中地使用 6~7 字分句；刘亮程的分句最长，由 6~9 个字组成的分句最多，且 9~20 字长的分句也明显多于其它两位作家。

究其原因主要有以下几个方面：第一，汪曾祺受明清散文小品的影响较深，偏爱文言句式和四字格，并且并不受现代汉语主谓宾必须完整的语法限制，经常省略主语和介词，颇有古代文人写意式白描手法的风韵，从而流露出典雅简洁的语言风格；第二，汪曾祺有意识地追求“精准”的表达手法，他曾在《小说笔谈》中指出“语言的唯一标准，是准确”，因此致力于使用最合适的词而较少使用修饰语。而朱自清处于“文言文”和“白话文”的交替时期，注重挑选和提炼口语；与此同时因在新文学运动中大量接触欧美文学，受到“翻译体”的影响，又使其某些句子变长。故其分句长介于汪曾祺和刘亮程之间。而刘亮程接受的是规范的现代汉语语法训练，句子大多主谓宾齐全，虽然也力求简洁的表达，但相较另外两位作者而言，仍不可避免得被拉长。

3.3.2 相邻分句句长变化

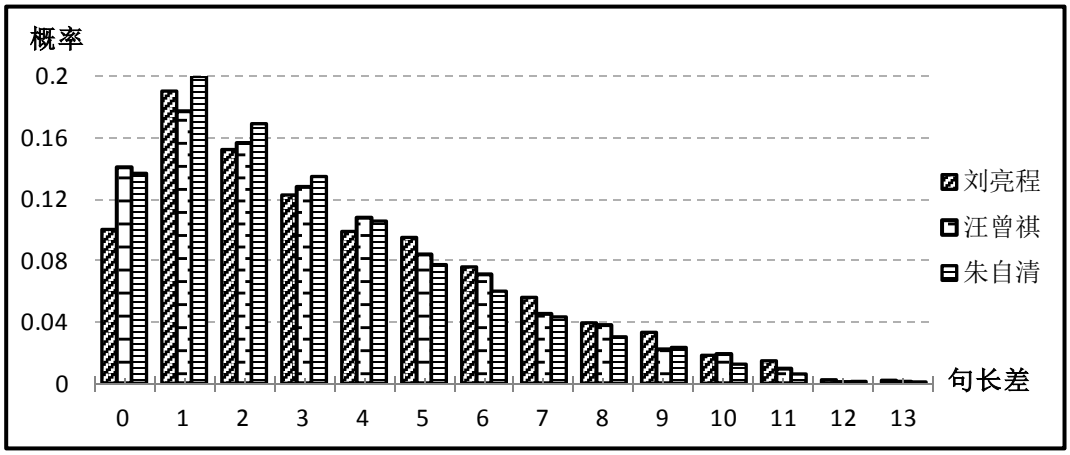

图 5 三位作家相邻分句句长变化图

从相邻分句的句长变化图来看，就主体而言三位作家的句长变化都较小，大多控制在 4 字以内，句长差小于等于 4 的情况分别在刘、汪、朱三人的文本中达到了 66%、71%和 75%的比例。然而细心观察之后我们仍能发现同中存异，相对而言朱自清的句长变化最为平稳，节奏最为舒缓，多近似对偶句。刘亮程在字数对仗工整方面（即句长差为 0）的把握明显不如汪、曾二人，尤其是汪曾祺最为讲究句式长短的对齐。此外从句长差大于等于 5 字以上的频率均高于汪、曾二人还可看出，刘亮程兼顾使用长短句的交错，其文本的节奏更富

于变化。

3.4 以整句句长和标点符号为特征的节奏分析

整句是指能表达一个完整的意思的语言运用的基本单位，它由词和短语构成，在本研究中通过检测句号、感叹号、问号和省略号来加以识别。标点符号除了用于标明语气之外，还可以起到标识不同时间长短的停顿，从而实现不同的节奏感，反映不同的文本风格。总的来说标点符号可分为点号、标号和符号三大类，其中可用来表示停顿的包括所有点号和一部分标号，它们是句号、问号、感叹号、逗号、顿号、分号、冒号，以及引号、破折号、省略号。由于标点符号还从另一方面折射了句子的编排结构，影响着句子的长度，因此我们以上文提到的十个标点符号和整句句长为特征来对三位作家的文本进行了文本聚类，向量空间维度共 114 维，结果如图 6 所示：

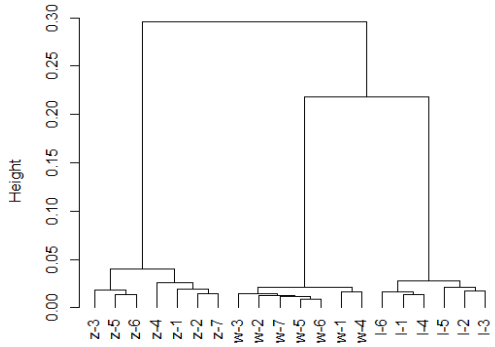

图 6 基于标点符号和整句句长的聚类结果

由图 6 我们可知以标点符号和整句句长为特征能较好地将三位作家的文本区分开来，它们是有效的特征。下面我们将具体地来看三位作家在句长和标点符号方面有何差异。

3.4.1 整句句长

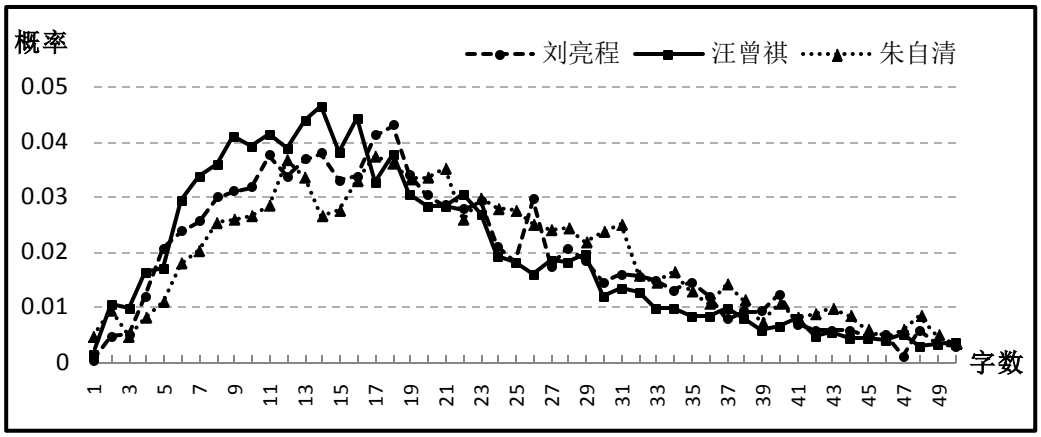

图 7 三位作家整句句长（分布 50 字以内）频率分布图

图 7 是三位作家的整句句长分布比例图，为了显示清楚，句长分布图中我们只选取了句长在 50 字以内（含 50）的句子。由图可知，和分句句长一样，汪曾祺的整句句长也最短，主要集中在 8~17 字。不同的是，朱自清反超刘亮程更多地使用长句，24 字以上的句子所占的比例均大致处于领先地位，刘亮程的句长分布则介于两者之间。

3.4.2 标点符号

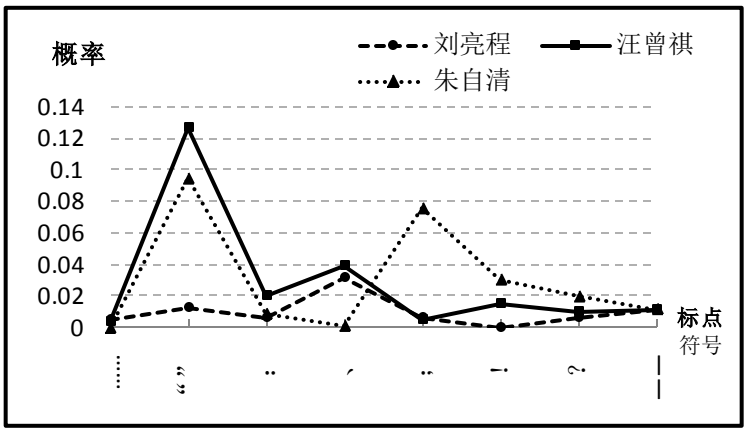

图 8 三位作家标点符号使用频率图

图 8 是三位作家的标点符号分布比例图，为了显示清楚，我们剔除了逗号和句号。总的来说，逗号和句号的出现频率占据着主导地位，它们共同在刘、汪、朱三位作家所使用的所有标点符号中分别达到了 91.3%、71.8%和 73.9%的比重，由此也可看出刘亮程的标点符号使用较为单一，主要集中在逗号、句号和顿号这三种上，语气情感较为平和舒缓。

相比之下，汪曾祺对引号和冒号的使用更为突出，这与其在散文中善于运用“对话”还原出日常生活的真实场景的写作风格有关，而这一家常式的语言风格也很好地解释了其句子最短的原因。

朱自清则表现出特别偏好使用分号，而分号主要用于复句内部并列分句之间的停顿，由此可反映出朱自清的句子结构较为缜密，层级分明，故长句颇多。此外其在感叹号、问号的使用上也有着更高的频率，说明其在情感的表达上更为外露和激烈。

4 结论

本文以朱自清、汪曾祺和刘亮程三位作家的散文作品为实验对象，利用基于 KLD 的层次聚类法挖掘出了新的能够代表作品风格的特征。实验结果表明句末用字韵母及其 n 元组合能够反映文本韵律，分句句长的 n 元组合、标点符号和整句句长能够反映文本节奏，以其为特征能成功地将这三位作家的作品区分开来，证明了它们是能代表作品风格的行之有效的特征，有望应用于作者归属、基于作品风格的自动文本分类等领域。

韵律方面，大体上刘亮程较常在句尾使用舌位较高的韵；而汪曾祺倾向于使用舌位较低的韵，朱自清则多使用开口度较大的韵。在相邻两句的押韵上，朱自清对韵脚的选择较为单一，种类明显不如刘、汪二人丰富。在相邻三句的隔句押韵形式上，刘亮程青睐押“i”韵，朱自清和汪曾祺则偏好押“e”韵，且朱自清中间插入的其它韵的形式更为丰富。在相邻四句的押韵上，汪曾祺和朱自清在回环押韵上的使用较为突出。

节奏方面，汪曾祺的分句长最短，并大量使用四字分句；其次是朱自清，刘亮程的分句最长。从相邻分句的句长来看，朱自清的句长变化最为平稳，节奏最为舒缓，多近似对偶句；汪曾祺最为讲究句式长短的对齐，对仗工整；刘亮程则兼顾使用长短句的交错，其文本的节奏更富于变化。和分句句长一样，汪曾祺的整句句长也最短，不同的是，朱自清反超刘亮程更多地使用长句。此外，刘亮程的标点符号使用较为单一，主要集中在逗号、句号和顿号这三种上，语气情感较为平和舒缓。相比之下，汪曾祺对引号和冒号的使用更为突出，折射出其口语化的特点。朱自清因偏好使用分号，故长句颇多，同时也多用感叹号和问号，说明其在情感的表达上更为外露和激烈。

本研究的不足之处在于，只研究了散文文体，并未将其与一般文体进行比较，未来还可进一步探寻其它文体的相关规律。

参考文献

[1] 赵冈，陈钟毅.红楼梦新探[M].北京：文化艺术出版社，1991.
[2] 李贤平.《红楼梦》成书新说[J].复旦学报（社会科学版），1987，5：3-16.
[3] 陈芯莹，李雯雯，王燕.计量特征在语言风格比较及作家判定中的应用——以韩寒《三重门》与郭敬明《梦里花落知多少》为例[J].计算机工程与应用，2012，48(3)：137-208.
[4] 万凯. 基于因子分析法的中文文本降维[D].广州：华南理工大学，2012.
[5] 金奕江，孙晓明，马少平.因特网上的写作风格鉴别[J].广西师范大学学报（自然科学版），2003，21(1)：62-66.
[6] 常淑慧.基于写作风格的中文邮件作者身份识别技术研究[D].天津：河北农业大学，2005 年.
[7] 李小凤.疑问句在报道语体与艺术语体中的对比研究[D].广州：暨南大学，2005.
[8] 邵长超.文艺语体和科技语体形谓句对比研究[D].广州：暨南大学，2007.
[9] 于灵子.科技语体和艺术语体定语位置上的形容词研究[D].广州：暨南大学，2006.
[10] 万晶.中文作者识别方法研究[D].长沙：湖南大学，2012.
[11] 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.
[12] 陈龙，范瑞霞，高琪.基于概念的文本表示模型[J].计算机工程与应用.2008，44(20)：162-164.
[13] 朱牧.基于写作风格特征的论文剽窃检查优化方法研究[D].上海：复旦大学，2011.
[14] 冯晓蒲，张铁峰.四种聚类方法之比较[J].微型机与应用，2010，29(16)：1-3.
[15] 孙吉贵，刘杰，赵连宇.聚类算法研究[J].软件学报，2008，19(1)：48-61.
[16] 朱光潜.谈美·谈文学[M].北京：人民大学出版社，2013：193-200.
[17] 汤云航. 普通话语音的统计分析[J].承德民族师专学报，1995，1：66-76.

作者简介：贺湘情（1991——），女，硕士研究生，主要研究领域为语料库语言学。Email: he_xiangqing@163.com；刘颖（1969——），女，副教授，主要研究领域为语料库语言学、计算语言学、自然语言处理和机器翻译。Email: yingliu@mail.tsinghua.edu.cn。