基于多维度分析法的鲁迅三种文体比较研究

范楚琳　刘颖

（清华大学 人文学院 中国语言文学系，北京 100084）

摘要： 该文从鲁迅书信、小说和杂文中提取出 387 个语言特征，采用随机森林和 k-means 聚类算法筛选出 58 个能够对三种文体取得较好区别效果的特征。该文采用比伯的多维度分析法对这些语言特征进行因子分析，得到 7 个比较重要的因子。该文根据每个因子中具有显著负荷值的语言特征，将 7 个因子解释为 4 个能够体现文体在写作角度、叙述视角、形式、语言系统等方面差异的维度，和 3 个能够体现文体存在某种特点的特征组合。书信和小说在互动性上相似，然而书信更具议论性、文言性和详细的写作特征，小说更具描写性、白话性和简短的写作特征；书信和杂文在议论性和详细的写作特征上相似，而书信互动性较强，杂文互动性较弱；小说和杂文则没有相似的维度。

关键词： 语体；文体；鲁迅；多维度分析法

A Comparative Study on Three Genres of Lu Xun Based on Multi-dimensional Analysis

Abstract: We collected 387 linguistic features from Lu Xun's letters, novels and essays, and used Random Forests and K-Means Clustering to select 58 features that could effectively distinguish three genres. We used Biber's Multi-dimensional Analysis to perform factor analysis on these features, and extracted 7 important factors. Based on linguistic features with significant factor loading values, we interpreted 4 factors as dimensions and 3 factors as feature combinations. The results show that letters and novels are similar in interactivity; while letters tend to be more argumentative, classical, and detailed, novels tend to be more descriptive, colloquial, and brief. Letters and essays are similar in argumentation and detailed structure; while letters tend to be more interactive. Novels and essays lack similar dimensions.

Key words: Register; Genre; Lu Xun; Multi-dimensional Analysis

0　引言

冯胜利[1]认为，语体是“实现人类直接交际中最原始、最本质属性的语言手段和机制”，而在交际过程中，语言是通过语体的正式与非正式产生距离的远和近，典雅与通俗产生距离的高和低效果的。按照冯胜利[2-3]，书面语属于典雅体和正式体，口语属于非正式体。口语和书面语之间的对立体现在词汇、句法等多种语法功能上。语体是“文体产生的源泉”，如“叙述故事一般都用非正式的语体，论证时事多与正式语体结伴而行”。故本文在对鲁迅的书信、小说和杂文这三种文本类型进行研究时，在研究文体特征的同时也探讨它们的语体属性。

曾枣庄[4]在《中国古代文体学》中提出，中国古代文体分类学的研究对象分别为体裁、体格和体类。体裁为每种文体所应具备的写作要求，如语言结构形态、表述方法等；体格则近似文体的表达效果或给人带来的主观感受。不同文体在内容、形式、角度、表达效果等方面存在差异。以书信为例，碑文主要记事，书信则在内容上可以无所不包；在形式上，论说文中的“说”较为短小，书信则可长可短，然措词须遵循上下等级关系；在写作角度上论说文或说理、或解释说明，书信则可叙事、可说理、可言情；在表达效果上，碑文典雅，书信则更为亲切。

语言学界有大量探讨现代汉语不同的语体机制所表现的具体语法差异的研究。王灿龙[5]通过考察指示代词“这”、“那”在大量语料中的用法发现：口语常常会在该用远指代词的情况下使用近指代词，或在该用近指代词的情况下使用远指代词，将小句所指事件“拉近”或“推远”，以表达对指称对象的亲密或疏远。宋文辉[6]等考察了现代汉语在不同语体中的有标志被动句、意念被动句和施事显现

基金项目： 2018 年度哲学社会科学基金重大项目“基于大数据技术的古代文学经典文本分析与研究”（18ZDA238）；教育部人文与社科一般项目，“语体特征的自动提取和研究”（17YJAZH056）

的被动句的分布情况，得出：“叫”字句、“给”字句和“让”字句的口语性较强，“为”字句和“于”字句的书面语性较强；在被动句中，书面语性越强，有标志被动句越多、意念被动句越少，被动句施事不显现的频率越高。亦有探讨不同文体所具备的语体特征的研究成果。冯胜利、王永娜[7]在选取四篇叙事文和两篇论说文，并对其中的通用体词汇、口语非正式语体词汇、书面正式语体词汇和庄典体词汇的数量进行标注和统计后发现：两种文体中“通用体要素均为主体要素”；论说体中书面正式语体词汇的数量要显著高于口语非正式语体，叙事体中则差异不明显。

语体是文体产生的源泉，故不同文体亦存在正式体、非正式体、典雅体等不同语体机制的差异，后者具体体现在词汇、句法等方面的差异上。同时，文体在形式、写作角度、表达效果等方面亦存在着不同。然而，语言学研究对语体和文体的探讨常常局限于列举微观而孤立的特征而缺乏对所有特征宏观综合的考察，文体学研究的成果大多基于主观的感受而缺乏较为可靠的语言学依据。因此，本文试图结合语言学研究的可靠性与文体学研究的综合性，以及二者在语体机制和文体特点等方面的成果，力求运用大量的语言特征来解释不同文体在语体机制、形式、角度、表达效果等方面的异同。

面对海量的文本和大量的语言特征，本文采用近年来数字人文领域提出的文本细读与计算机“远读”相结合的文学研究方法[8]，以及比伯在研究不同语域差异时提出的多维度分析法[9-14]来进行研究。比伯[9]采用因子分析算法，对语料库中的 481 篇文本、每篇文本 67 个语言特征的频率进行统计后，得出了五维“共现特征集合”。基于“共现特征的交际功能”，他分别将五个维度解释为“角色参与型与信息提供型”、“叙事型与非叙事型”、“清晰型与情景依赖型”、“显性的劝导型”和“客观型与非客观型”。计算不同文本的“维度分数”得以分析不同语域在不同维度上的相似性和相异性。运用多维度分析法，比伯先后对英语的口语和书面语域[10]、英语语域的历史演变[11]、大学口语和书面语域[12]、网络语域[13]等进行过不同程度的研究。近年来，多维度分析法还加入了通过直接计算文本在已知维度上的分数来确定新文本所属的语域，以及运用判别函数分析对文本进行分类以验证维度划分的合理性等内容。[14]

本文拟采用比伯的多维度分析法，和文本细读与计算机远读相结合的文学研究方法，对鲁迅书信、小说和杂文中可能具有区别性的语言特征（以词特征为主）进行统计和因子分析，以得出具有不同功能的共现特征集合。这些特征集合或代表了文体在叙事、议论、抒情和描写等方面的差异，或代表了文体在语体机制、形式、表达效果等方面的差异。本文还采用了随机森林和 k-means 聚类等算法，以得出对文本按文体分类和聚类准确率影响较大的特征集合，作为进行因子分析的数据集。

本文试图对文学文体的语言学功能进行解释，基于大量共现而非孤立的特征进行研究，并为传统文体学对文体特征所下的判断提供依据。尽管目前的探讨仅局限于同一作者所写的不同文体的文本，将来我们仍然可以继续进行同一文体不同时代、不同流派、不同作者的写作差异等方面的研究，以期为语体和文体研究、文学作品的文本风格分析、修辞和语用研究等提供更多的依据和成果。

1　实验方法

1.1 多维度分析法

比伯[9]提出的多维度分析法采用因子分析来确定变量即语言特征之间的相关性，把分布相似的语言特征分成一组：每组语言特征即一个因子，“从功能上被解释为”维度；每个维度包括一组“在文本中频繁共现”的特征，和另一组“极少在文本中出现”的特征。随后，根据不同因子中具有显著负荷值的语言特征，计算每一文本在不同维度上的得分。只有因子负荷绝对值大于 0.35 的变量才可用于计算，公式为：所有因子负荷值为正的语言特征在文中出现的标准化频率之和，减去所有负荷值为负的语言特征的标准化频率之和后得到的差。

1.2 随机森林

随机森林是“以决策树为基本分类器的一个集成学习模型”，“包含多个由 Bagging 集成学习技术训练得到的决策树,当输入待分类的样本时,最终的分类结果由单个决策树的输出结果投票决定”。[15]其优点有：在目前的算法中具有较高的准确率；能够高效地运行于大型数据集；能够处理数以千计的输入变量而无需降维；能够评估变量对分类的重要性程度；在运行的过程中能够产生泛化误差的内部

无偏估计——袋外误差估计（Out of Bag Error Estimate），而无需进行交叉验证等等。[①]

随机森林能够在模型训练结束后，给出每个变量对分类结果准确率的重要程度估计。故我们用其来衡量和筛选所有收集到的语言特征，并不断选取不同的特征集合进行训练，根据每次训练得到的袋外分数高低来确定最终用于因子分析的语言特征。

1.3 k-means 聚类

本文采用 k-means 聚类的方法来验证特征选取的合理性。我们通过随机森林提取出对分类结果准确率影响较大的特征集合，根据这些特征来对不同文体的文本进行聚类，倘若聚类结果的准确率较高，则说明通过随机森林提取的特征的确对文体具有较好的区分度。

2 实验过程和结果

本文总体的实验过程如下：

1. 建立语料库并提取所有与文本功能相联系的语言特征。
2. 将文本与语言特征输入随机森林训练，生成所有语言特征对分类准确度的重要性排名。
3. 选取重要性排名较高的特征进行多次组合，分别输入随机森林和 k-means 模型进行训练。
4. 选择能够取得较好分类和聚类结果的特征集合进行因子分析。
5. 根据因子分析的结果，计算所有文本在每个因子上的维度分数。
6. 根据每个因子中具有显著负荷值的语言特征，和每一文体所有文本的维度分数，解释不同维度的功能。

2.1 语料的收集和预处理

本文收集的语料来自人民文学出版社 2005 最新修订版本的《鲁迅全集》。[16]其中，书信收录于第十一卷至第十四卷，小说收录于第一卷的《呐喊》、第二卷的《彷徨》和《故事新编》中，杂文收录于第一卷的《热风》和第三卷至第六卷。第一卷的《坟》虽然被命名为“论文集”，在鲁迅写给友人的信中，亦曾将《坟》称作“杂文集”。[17]相较于文体特征较为明显的小说和书信，杂文往往带有鲜明的议论性与现实指向性，故我们将带有相同特征的《坟》亦收录在杂文一类中。序跋文是对一部书或作品的评论或介绍，与杂文中的其他文章差异较大，同时我们又能根据标题较为明确地判断出该文体，故语料库中删除了杂文集中所有的序跋。

由于本文的研究对象限定于鲁迅本人所写的白话文，我们删除了语料库中的存在大量引用他人段落和带有浓重文言色彩的文章。此外，鲁迅文章中所附的他人信件与文章，与文章中含有的大段文言摘抄也一并删除。为了统一语料库中文章的规模，我们将所有文章切分为一个个字数在 1000 字左右的片段。最终得到的语料库规模如表 1。

表 1 语料库规模

	书信	小说	杂文
片段数量	78	155	402
字数均值	1032.62	1049.17	1072.28
字数标准差	110.60	48.92	97.30

2.2 语言特征的提取和过滤

本文采用 NLPIR-ICTCLAS 汉语分词系统的 Python 接口对文本进行分词和词性标注。[②]表 2 为初步提取的 387 个语言特征，所有频率均以一千字为单位计算。

按朱德熙[18]，实词属于“难于在语法书里一一列举其成员”的开放类，虚词则属于“可以穷尽地列举其成员”的封闭类。因此，我们在高频词中选择提取的词以虚词为主，因为虚词的数量较为有限，在总量一定的情况下更容易比较相互之间在使用过程中的差异。

本文整句划分的标志为标点——“。”、“？”、“！”、“……”；分句划分的标志为——“，”、“；”、“：”、“——”。按李秀明[19]，叙事语体是“为了表达人物在一个连续时间内的动作行为”，具备“时间的连续性”和“施事的凸显性”两个基本特征，“动词的功能必须强化”；描写语体的一个重要特征为“场景凸显性”，动词的功能“表现出弱化的倾向”。故本文统计了每篇文本中不含动词的句子占所有句子数量的比例，以供文本叙事性程度衡量的参考。

我们首先在随机森林中输入从每个文本中统计的 387 个语言特征频率与文本文体类型进行训练，结果显示袋外分数为 0.9039。随后，提取在所有特征中分类重要性程度排名较高的 58 个特征重新训

[①] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#inter

[②] https://pynlpir.readthedocs.io/en/latest/api.html#module-pynlpir.nlpir

练，得到袋外分数 0.9354。图 1 显示了这 58 个特征的分类重要度排名，其中重要度排名最高的前 10 位特征分别为：平均词长、助词“着”、连词、地名、副词“已”、趋向动词、段落数量、括号、助词“的”、词的形符数。

表 2 提取的语言特征

特征	特征统计
词性	计算所汉语词性标记集[①]中所有的一类、二类、三类词性的频率
人称代词	第一人称代词（吾、我们、我、我辈、咱、俺、咱们）、第二人称代词（您、你、你们）、第三人称代词（他们、她们、它们、他、它、她、伊）的频率
词	语料库中出现频率较高的 242 个可能具有语体区别作用的词的频率，词的类符数，词的形符数，词汇丰富度，平均词长，词长离散度，单音词比例，双音词比例，多音词比例
单音词	名词、动词、形容词、代词、副词、介词、连词、助词中含单音词的比例
句首词	名词、时间词、处所词、方位词、动词、形容词、区别词、状态词、代词、数词、量词、副词、介词、连词、助词、叹词、语气词、拟声词、前缀、后缀作句首词的比例
句	句数、平均句长、句长离散度、分句数、平均分句长、分句长离散度、句中动词省略比例
段	段落数量、平均段长、段长离散度

我们将包含这 58 个特征频率的所有文本进行 k-means 聚类，得到纯度 0.9244，聚类结果如表 3。由表可见，78 个书信文本中，71 个被归为了类 1；155 个小说文本中，144 个被归为了类 0；402 个杂文文本中，372 个被归为了类 2。

2.3 因子分析结果

我们将包含 58 个标准化特征频率的所有文本导入 R，KMO 检验显示总体 MSA 值为 0.83，说明数据变量之间的相关性较强、适宜进行因子分析。对数据进行平行分析后显示：因子的数量为 10，主成分的数量为 8。平行分析碎石图如图 2。

图 1 选定特征的分类重要性程度排名

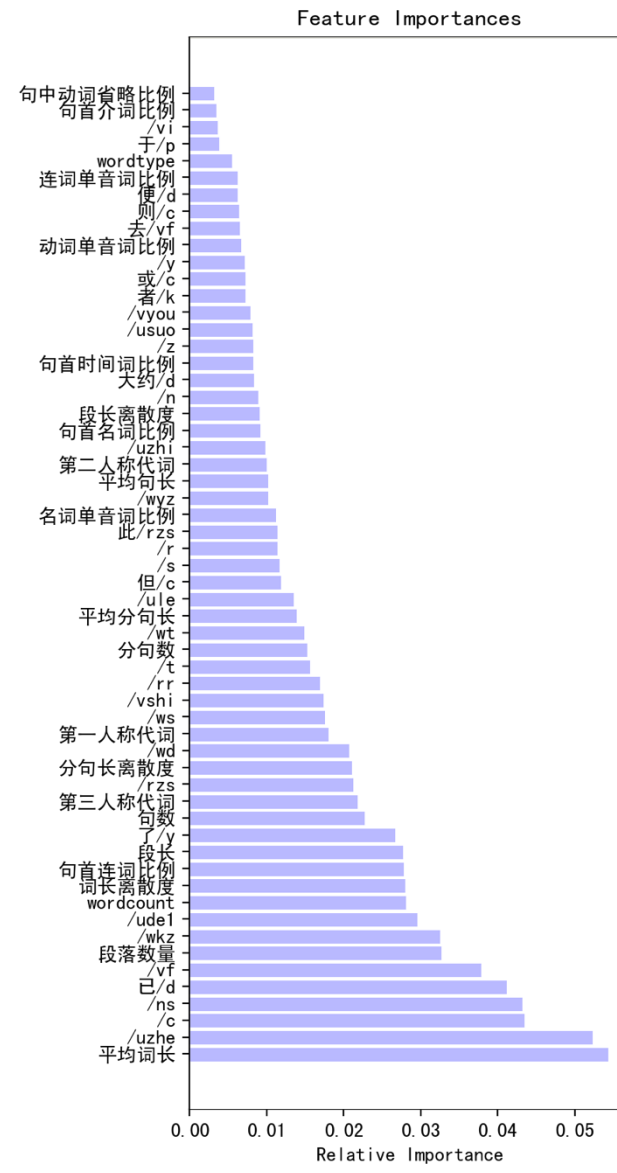

表 3 基于 58 个语言特征的 k-means 聚类结果

	类 0	类 1	类 2	合计
书信	0	71	7	78
小说	144	1	10	155
杂文	17	13	372	402
合计	161	85	389	

我们将因子数设定为 10，进行因子的提取和旋转。我们按最大似然法提取因子，并采取斜交的方法旋转以获得更好的解释性。[②]因子分析结果显示，10 个因子的累计方差贡献率为 53%。所有因子的方差贡献率如表 4。不同因子下每一变量的负荷值显

① http://ictclas.nlpir.org/nlpir/html/readme.htm#_Toc34628488

② Factor Analysis in R: https://www.uwo.ca/fhs/tc/

示，ML3、ML7 和 ML2 因子均只存在一个具有显著负荷值的变量，且这三个因子的方差贡献率均较低，故我们最终取前 7 个因子作为文本的维度，用于进行维度分数的计算和文体差异的解释。

图 2 包含 58 个变量的数据集平行分析碎石图

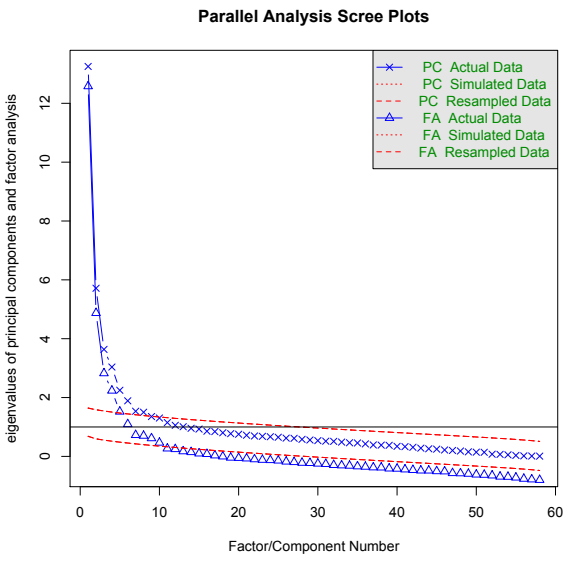

表 4 因子方差贡献率

	方差贡献率	累计方差贡献率
ML10	0.08	0.08
ML5	0.06	0.13
ML6	0.07	0.2
ML9	0.06	0.26
ML4	0.06	0.32
ML1	0.06	0.38
ML8	0.06	0.43
ML3	0.03	0.47
ML7	0.03	0.5
ML2	0.03	0.53

3 文本维度分析

3.1 维度一：描写与议论

按比伯[9]，因子负荷值为正的特征属于“积极”特征组，表示该组特征“在文本中频繁共现”；因子负荷值为负的特征属于“消极”特征组，表示该组特征“极少在文本中出现”。表 5 为第一个维度“ML10”下所有具有显著负荷值的语言特征。表 6 为我们根据 1.1 节中的多维度分析法计算出所有文体在维度一上的得分均值。

积极特征组显示：该维度出现了较多的助词“着”、处所词和状态词。按《现代汉语》[20]，动态助词“着”用于动词、形容词后，有时表示“动作正在进行”即“动作开始后、终结前的进行情况”，如“张着帆”；有时表示“状态在持续”即“动作完成后的存在形态”，如“门开着”。状态词，按《现代汉语语法信息词典》[21]，是“从形容词中分化出来的一类谓词”，即朱德熙所说的“状态形容词”，“描述事物或动作的状态”；而词典所界定的形容词实际上是他所说的“性质形容词”，“表示事物的性质”。按朱德熙[18]，“性质形容词单纯表示属性，状态形容词带有明显的描写性”。处所词，按《语法信息词典》表示地点，但区别于被划入名词的“地理名称”和“表示机关、单位、部门的词或专有名词”。由被划分为处所词的“路上”、“地下”、“海里”等词可见，处所词表示事物所处或动作发生所在的场所。因此，从维度一中频繁共现的助词“着”、状态词和处所词可知，该维度既体现了事物存在形态的持续或动作的进行，又表现了事物或动作的状态和处所，具有明显的描写性。

消极特征组表示：该维度中连词、句首连词、动词“是”和连词“但”出现得较少。按《现代汉语》[20]，连词“起连接作用，连接词、短语、分句和句子等，表示并列、选择、递进、转折、条件、因果等关系；判断动词“是”表示“事物等于什么或属于什么”，“事物的特征、质料、情况”或“事物的存在”，表示“一般的肯定”。因此，在维度得分较低的文本中连词与判断动词“是”出现的频率较高，且连词主要出现在句首、“但”出现得较多，意味着连词表示并列的可能性较低、以表示转折关系为主。当一个文本中出现了较多的表示转折和判断的成分时，常常表示该文本的议论性较强。综上，我们将维度一界定为描写与议论的维度。

我们对三种文体在该维度上所有文本得分的均值进行单因素方差分析，得到 p 值 1.0370e-115，说明书信、小说和杂文在维度一上的得分存在显著差异。Welch[22]对 T 检验进行了改进，使其能够在两组数据样本容量和方差不一致的情况下检验两组数据均值之间的差异。Zimmerman 认为，当两组数据样本容量不一致时，即使方差相等，Welch T 检验要比学生 T 检验的效果更好。[23]故本文采用 Welch T 检验比较三种文体两两之间均值的差异，所得 p 值如表 7。表 6、7 显示书信和杂文在维度一的得分差异较

小，小说则要明显高于它们。小说中动态助词“着”、状态词和处所词中共现的频率较高，如下：

老栓走到家，店面早经收拾干净，一排一排的茶桌，滑溜溜的发光。但是没有客人；只有小栓坐在里排的[桌前/s]吃饭，大粒的汗，从额上滚下，夹袄也帖住了脊心，两块肩胛骨[高高/z]凸出，印成一个阳文的“八”字。老栓见这样子，不免皱一皱展开的眉心。他的女人，从灶下急急走出，睁[着/uzhe]眼睛，嘴唇有些发抖。……店里坐[着/uzhe]许多人，老栓也忙了，提[着/uzhe]大铜壶，一趟一趟的给客人冲茶；两个眼眶，都围[着/uzhe]一圈黑线。（鲁迅小说，《药》）

处所词“桌前”表示动作“吃饭”的场所，状态词“高高”表示事物“肩胛骨”的状态，“睁着眼睛”、“坐着许多人”和“围着一圈黑线”均表示事物或人物存在形态的持续，“提着大铜壶”则表示动作的进行。以上两段展现了小说对人物的体态、动作、存在形态的描写，符合前文对该维度富于描写性的总结。小说在该维度上的平均分值最高，与相较于书信、杂文，小说更具描写性的一般认知相符。

表 5 维度一（ML10）中具有显著负荷值的特征

积极特征组		消极特征组	
助词“着”	0.52	连词	-0.71
处所词	0.41	句首连词比例	-0.58
状态词	0.39	动词“是”	-0.46
		连词“但”	-0.41

表 6 不同文体在维度一上的平均得分

	维度得分均值
书信	-1.97
小说	6.34
杂文	-2.04

表 7 维度一得分均值的 Welch T 检验

	p 值
书信、小说	1.6690e-46
书信、杂文	0.8543
小说、杂文	6.6957e-64

书信和杂文中连词、句首连词、判断动词“是”和连词“但”共现的频率较高。如下段：

木刻[是/vshi]一种作某用的工具，[是/vshi]不错的，[但/c]万不要忘记它[是/vshi]艺术。它[之所以/c][是/vshi]工具，就因为它[是/vshi]艺术的缘故。斧[是/vshi]木匠的工具，[但/c]也要它锋利，[如果/c]不锋利，则斧形[虽/c]存，即非工具，[但/c]有人仍称之为斧，看作工具，那[是/vshi]因为他自己并非木匠，不知作工之故。……（鲁迅书信 10，《致希仁斯基等》）

上文通过连词“但”、“之所以”、“如果”、“虽”和判断动词“是”的配合，展现了句间的逻辑关系与判断、得出结论的过程，是一个典型的议论段落。正如前文所述，维度一用于区别描写和议论的篇章。书信和杂文在该维度上的得分最低，说明与小说相比，两者均存在显著的议论性。

3.2 维度二：互动与非互动

表 8-10 分别为第二个维度“ML5”下所有具有显著负荷值的语言特征，三种文体在该维度上的平均分值与两两之间进行 Welch T 检验所得的 p 值。

表 8 维度二（ML5）中具有显著负荷值的特征

积极特征组		消极特征组	
第一人称代词	0.91	名词	-0.42
人称代词	0.88		
代词	0.85		
第二人称代词	0.51		

表 9 不同文体在维度二上的平均得分

	维度得分均值
书信	2.54
小说	2.12
杂文	-1.21

表 10 维度二得分均值的 Welch T 检验

	p 值
书信、小说	0.4070
书信、杂文	3.0977e-18
小说、杂文	2.3481e-13

表 8 显示该维度的积极特征包括了第一人称和第二人称代词，消极特征为名词。当第一人称和第二人称代词在文本中频繁共现时，常常意味着文中出现了较多主体和对方的交际互动，例如在对话中谈话者对自身和听话者的来回关注，书信中写信者对自身和收信人联系的建立和反复确认。名词通常在文本中出现的频率较高，然而，当一个文本中出现了远远超过正常频率的名词时，常常意味着作者对抽象事物的讨论或信息的提供，而在这个过程中作者关注的重点在于被提及的一系列的名词，对自身和对方的关注是不明显的。两者在叙述视角上存在着差异，故我们将维度二界定为互动与非互动。

对维度得分的均值进行方差分析得到的 p 值为 5.1124e-27，说明三种文体在维度二上存在显著差

异。表 9、10 显示，书信和小说在该维度上的得分差异较小，杂文则要明显低于它们。书信和小说中存在着较多的第一和第二人称代词，如以下两段：

[你/rr]什么时候可以毕业回国？[我/rr]自憾[我/rr]没有什么话可以寄赠[你/rr]，但以为使精神堕落下去，是不好的，因为这能使[自己/rr]受苦。第一着须大吃牛肉，将[自己/rr]养胖，这才能做一切事。[我/rr]近来的思想，倒比先前乐观些，并不怎样颓唐。[你/rr]如有工夫，望常给[我/rr]消息。（鲁迅书信 260617，《致李秉中》）

“[你/rr]以为[我/rr]发了疯么？[你/rr]以为[我/rr]成了英雄或伟人了么？不，不的。这事情很简单；[我/rr]近来已经做了杜师长的顾问，每月的薪水就有现洋八十元了。”（鲁迅小说，《孤独者》）

按曾枣庄[4]，书信是“人与人之间的交际工具，最具有实用价值”，而小说中维度得分较高的篇章也常常出现于对话和复述信件的场景中。篇章通过反复交替出现的第一人称代词和第二人称代词，体现了主体与对方的交流互动。与书信和小说相比，杂文中名词出现的频率较高，如以下一段：

[商品/n]固然是做不下去的，独立也活不下去。创造[社/n]的[人们/n]的[去路/n]，自然是在较有[希望/n]的“革命[策源地/n]”的[广东/ns]。在[广东/ns]，于是也有“革命[文学/n]”这[名词/n]的出现，然而并无什么[作品/n]，在[上海/ns]，则并且还没有这[名词/n]。（鲁迅杂文，《上海文艺之一瞥——八月十二日在社会科学研究会讲》）

杂文的篇章中名词使用得较为频繁，可以看出作者较少地关注自身与他人的交际互动与情感的传达，而将关注点着眼于抽象事物的叙述与思想理念的表达。与比伯[9]通过多维度分析法得出的第一个维度相同，其维度的积极特征组亦包含了第一人称代词和第二人称代词，消极特征组亦包含了名词。比伯将该维度总结为“角色参与型与信息提供型”，对话文本在维度上具有较高正值，因其“具有很高的交互性和参与性”，“参与者传递信息的倾向较少，也没有时间仔细地推敲提供的信息”；新闻社论、学术文章和官方文件在该维度上具有较高负值，因其“主要目的是传达信息，语境是高度受控的”，与我们总结出的第二维度具有相似性。

3.3 维度三：简短与详细

表 11 为第三个维度“ML6”下所有具有显著负荷值的语言特征。该维度的积极特征为句数、段落数和省略动词的句子比例，消极特征为平均句长、逗号频率和平均段长。当文本在该维度上得分较高意味着其句子、段落数量较多而句长、段长均较短，句中常常没有逗号和动词。当一个文本句子短而少动词、少停顿以及段落亦短时，常常给人一种以较少甚至是空白的笔墨传达出言外之意的感觉；反之当文本的句子和段落均长，句子很少省略动词且停顿较多时，则产生一种笔墨较多、叙述甚详的效果。两者在写法上存在着直观的差异，故我们将维度三界定为简短与详细。

表 11 维度三（ML6）中具有显著负荷值的特征

积极特征组		消极特征组	
句数	0.85	平均句长	-0.86
段落数量	0.52	逗号	-0.51
句中动词省略比	0.42	段长	-0.36

维度得分显示三种文体在维度三上存在显著差异；书信和杂文在该维度上的得分差异较小，小说则要明显高于它们。小说文本一般具有较多的句子和段落，句长和段长较短，句中常常没有停顿和动词的存在。譬如以下几段：

然而他终于去请白问山。

白问山却毫不介意，立刻戴起玳瑁边墨晶眼镜，同到靖甫的房里来。他诊过脉，在脸上端详一回，又翻开衣服看了胸部，便从从容容地告辞。沛君跟在后面，一直到他的房里。

他请沛君坐下，却是不开口。

“问山兄，舍弟究竟是……？”他忍不住发问了。

“红斑痧。你看他已经‘见点’了。”（鲁迅小说，《弟兄》）

在以上几段中，段落和句子的规模均较小。不含逗号的句子如：“然而他终于去请白问山。”和“他忍不住发问了。”不含动词的句子如：“红斑痧。”小说中较为简短的句子和段落出现得较为频繁，但这并不意味着小说表达的意义单薄，因为简短、有所省略的句子常常能够传达更为深远的内容，如出现《狂人日记》中的这一段：“狮子似的凶心，兔子的怯弱，狐狸的狡猾，……”几个偏正短语并列单成一段，具有诗意与尖锐的警示效果。较为简短的句子有时不含逗号，以上段落中的两句；有时为不含动词的名词性谓语句，如《狂人日记》中“今天晚上，很好的月光。”也蕴涵着言外之意。小说中不含动词、逗号的短句也常常出现于对话句中，而在对话中我们本身就倾向于使用不完

整的句子来表意，以达到省力与直截了当的效果。综上，小说中的非对话句、段倾向于采取简短的笔法以表达深远的含义，对话句倾向于简短以符合自然口语的特征；书信和杂文则句子较完整、停顿较多，句子和段落都很长，体现为一种详细的写法。

3.4 维度四：文言与白话

表 12 为第四个维度“ML9”下所有具有显著负荷值的语言特征。该维度的积极特征为助词“之”、处所指示代词“此”、副词“已”、连词“则”、时间词、词的类符数和介词“于”，消极特征为助词“的”和动词“是”。“之”、“此”、“已”、“则”、“于”属于典型的文言用词，“的”和“是”则是典型的白话用词。虽然分词系统将结构助词“的”和语气词“的”都标注成了结构助词，但不影响消极特征“的”的白话性。时间词虽为提供背景信息的表现之一，然而由于其它信息呈现的不明显，我们还是将维度四界定为文言与白话。

表 12 维度四（ML9）中具有显著负荷值的特征

积极特征组			
助词“之”	0.62	处所指示代词“此”	0.55
副词“已”	0.55	连词“则”	0.46
时间词	0.45	词的类符数	0.44
处所指示代词	0.42	介词“于”	0.35
消极特征组			
助词“的”	-0.55	动词“是”	-0.36

维度得分显示三种文体在维度四上存在显著差异；书信、小说和杂文两两之间均存在显著的差异，三种文体在该维度上的得分从高到低依次为书信、杂文、小说。书信中“之”、“此”、“已”、“则”、“于”倾向于频繁共现，如下：

慨自二十三日[之/uzhi]信发出之后，几乎大不得了，伟大[之/uzhi]钉子，迎面碰来，幸而上帝保佑，早有廿九日[之/uzhi]信发出，声明前[此/rzs]一函，实属大逆不道，合该取消，于是始蒙褒为“傻子”，赐以“命令”，作善者降[之/uzhi]百祥，幸何如之。现在对于校事，一切不问，但编讲义，拟至汉末为止，作一结束，授课[已/d]只有五星期，此后便是考试了。但离开此地，恐当在二月初，因为一月薪水，是要等着拿走的。

朱家骅又有信来，催我速去，且云教员薪水，当设法加增。但我还是只能[于/p]二月初出发。至于伏园，却[于/p]二十左右要走了，大约先至粤，再从陆路入武汉。……（鲁迅书信 261216，《致许广平》）

书信的用词较为古雅，可能是出于书信用于交际的实用性，而在文字交际的过程中人们倾向于使用较为传统的语言系统（如唐宋散文家用骈文写作公文），而区别用于发表的文学作品、议论文章等等。对比具有较多“的”、“是”的小说段落：

吃人[的/ude1][是/vshi]我哥哥！

我[是/vshi]吃人[的/ude1]人[的/ude1]兄弟！

我自己被人吃了，可仍然[是/vshi]吃人[的/ude1]人[的/ude1]兄弟！（鲁迅小说，《狂人日记》）

从书信、杂文到小说，文体的文言成分逐渐减少、白话成分逐渐增多。

3.5 特征组合

由于 ML4 因子、ML1 因子和 ML8 因子下的显著特征数量较少，不足够构成一个完整的维度体现文体在形式、角度、表达效果等方面的差异，而只能作为特征组合表现文体在某一方面的特点。故我们不以“维度”来命名这三个因子，而仅将其界定为“特征组合”。

1 特征组合一：分句长短结合 / 短而均衡

表 13 为特征组合一所有具有显著负荷值的语言特征。该特征组合的积极特征为平均分句长和分句长离散度，消极特征为分句数和逗号频率。特征组合得分较高的文本分句长较长且长短结合，句中逗号较少；反之则分句长短而均衡，句中逗号较多。

表 13 特征组合一（ML4）中具有显著负荷值的特征

积极特征组		消极特征组	
平均分句长	0.85	分句数	-0.86
分句长离散度	0.62	逗号	-0.84

特征组合得分显示三种文体在该特征组合上存在显著差异；书信、小说和杂文两两之间均存在显著的差异，三种文体在该特征组合上的得分从高到低依次为杂文、小说、书信。杂文的平均分句长较长，且呈现出长短结合的分布特征；书信中则分句长短而平衡，一个完整的句子中常常出现较多的停顿，如以下段落：

……例如罢，田军早早的来做小说了，却“不够真实”，狄克先生一听到“有人”的话，立刻同意，责别人不来指出“许多问题”了，也等不及“丰富了自己以后”，再来做“正确的批评”。但我以为这是不错的，我们有投枪就用投枪，正不必等候刚在制造或

将要制造的坦克车和烧夷弹。……（鲁迅杂文，《三月的租界》）

此地天气很好，已穿纱衫。我是好的，能食能睡，加以小刺猬报告她的近状，知道非常之乖，更令我放心。……（鲁迅书信 290525，《致许广平》）

同样是表达自己的观点，短而均衡的分句长显得较为亲切、富于人情，分句较长而夹杂着短句的段落则有一种陌生、疏离的效果。

2 特征组合二：单音词为主 / 单与非单音词结合

表 14 为特征组合二所有具有显著负荷值的语言特征。该特征组合的积极特征为词的形符数、名词单音词比例和动词单音词比例，消极特征为平均词长和词长离散度。特征组合得分较高的文本词长较短且变化较小，单音词比例较高；反之则词长较长，单音词与非单音词均有一定的比例。

表 14 特征组合二（ML1）中具有显著负荷值的特征

积极特征组		消极特征组	
词的形符数	0.80	平均词长	-0.89
名词单音词比例	0.55	词长离散度	-0.65
动词单音词比例	0.52		

特征组合得分显示三种文体在该特征组合上存在显著差异；书信、小说和杂文两两之间均存在显著的差异，三种文体在该特征组合上的得分从高到低依次为小说、书信、杂文。小说中具有较多的单音词，既可以出现在带有文言特征的句子中，如《铸剑》中的“[烟/n][消/v][火/n][灭/v]；[水波/n][不/d][兴/v]”；也可以出现在带有自然口语特征的句子中，如《奔月》中的“[拿/v][我/rr][的/ude1][射/v][日/b][弓/n][来/vf]！[和/cc][三/m][枝/q][箭/n]！”。杂文则呈现出单音词、双音词、多音词互现的特征，尤其是有较多音译词外来词的段落。然而在小说的对话中，也出现了具有以上特征的段落，如：

“[屋子/n]？”[四爷/n][仰/v][了/ule][脸/n]，[想/v][了/ule][一会/m]，[说/v]，“[舍间/s][可是/d][没有/v][这样/rzv][的/ude1][闲房/n]。[他/rr][也/d][说不定/vl][什么/ry][时候/n][才/d][会/v][好]……”（鲁迅小说，《长明灯》）

3 特征组合三：趋向动词和语气词“了”

表 15 为特征组合三下所有具有显著负荷值的语言特征。该特征组合只含有积极特征：趋向动词，表示“移动的趋向”，如“领来”、“沉静下来”、“看出”、“宋元以来”；语气词“了”，可以表示陈述和祈使的语气。[20]特征组合得分显示三种文体在特征组合三上存在显著差异；书信、小说和杂文两两之间均存在显著的差异，三种文体在该特征组合上的得分从高到低依次为小说、书信、杂文。小说中趋向动词和语气词“了”倾向于频繁共现，杂文则分布的较少。

表 15 特征组合三（ML8）中具有显著负荷值的特征

积极特征组	
语气词“了”	0.93
语气词	0.77
趋向动词	0.41

4 结论与展望

本文通过随机森林算法，提取出能够对鲁迅的书信、小说和杂文取得较好区别效果的 58 个语言特征，并对其进行因子分析得到 7 个具有功能解释性的因子。在这 7 个因子中，前 4 个为能够体现文体在语言系统、形式、叙述视角、写作角度等方面差异的维度，后 3 个仅为能够体现文体存在某种特点的特征组合。我们将不同文体在 4 个维度和 3 个特征组合上的异同总结为下表：

表 16 不同文体所具有的维度与特征组合

	维度一		维度二		维度三		维度四		特征组合一		特征组合二		特征组合三	
	描写	议论	互动	非互动	简短	详细	文言	白话	分句短而均衡	分句长且长短结合	单音词为主	单、非单音词结合	趋向动词、语气词“了”共现	趋向动词、语气词“了”极少出现
书信		+	+			+	+		+					
小说	+		+		+			+			+		+	
杂文		+		+		+				+		+		+

从表 16 可以看出，鲁迅的书信具备了带有较多文言用词和议论成分，笔法详细，第一和第二人称互动明显的特点；小说具备了带有较多白话用词和描写成分，笔法简短，第一、二人称互动明显的特点；杂文具备了带有较多议论成分，笔法详细，第一、二人称互动较少的特点。

书信和小说在互动性上相似，然而书信更具议论性、文言性和详细的写作特征，小说更具描写性、白话性和简短的写作特征；书信和杂文在议论性和详细的写作特征上相似，而书信互动性较强，杂文互动性较弱；小说和杂文没有相似的维度，小说更具描写性、互动性和简短的写作特征，杂文则更具议论性、非互动性和详细的写作特征。从维度特征上看，书信与杂文较为接近、与小说具有一定距离，小说与杂文的距离最远。

从特征组合上来看，书信的分句短而均衡，小说以单音词为主、趋向动词和语气词“了”出现得较多，杂文的分句长且长短结合、单音词与非单音词结合出现、趋向动词和语气词“了”出现得较少。特征组合虽不足以构成具有完整功能的维度，却启发我们在今后对这些特征进行进一步的研究，譬如书信的分句短而均衡、给人以亲切的效果，是否与其互动性有关？杂文的分句长且长短结合、给人以疏离感，是否与其非互动性有关？小说以单音词为主，是否与其具有较多白话用词有关？这些都可在今后继续考察。

本文将不同的语言特征划分为不同的维度与特征组合，以解释文体在语言系统、形式、叙述视角、写作角度等方面的差异，以期为基于主观印象的传统文体学与基于孤立特征的语体学提供新的思路。当然，本文所提取的语言特征以词特征为主，对维度的解释也较为粗略，希望能在此拍砖引玉，今后不断地修正与完善基于多维度分析法的文体研究。同时，算法也有改进的空间，譬如加入基于因子分析的结果对文本进行逻辑回归以预测文体，基于文本维度分数来预测文本类型或探讨文体内部的差异等等。此外，还可对被误分为其它文体的文本进行进一步的研究。

参考文献

[1] 冯胜利.语体语法及其文学功能[J].当代修辞学,2011,(第 4 期):2-8.

[2] 冯胜利.论语体的机制及其语法属性[J].中国语文,2010,(第 5 期):401,407.

[3] 冯胜利.语体语法的逻辑体系及语体特征的鉴定[J].汉语应用语言学研究,2015(00):1-21.

[4] 曾枣庄著. 中国古代文体学 下 中国古代文体分类学[M]. 上海：上海人民出版社；上海：上海书店出版社,2012.12.

[5] 王灿龙.试论“这”“那”指称事件的照应功能[J].语言研究,2006,(第 2 期):60-62.

[6] 宋文辉,罗政静,于景超.现代汉语被动句施事隐现的计量分析[J].中国语文,2007,(第 2 期):118-122.

[7] 冯胜利,王永娜.语体标注对语体语法和叙事、论说体的考察与发现[J].汉语应用语言学研究,2017,(第 0 期):21-27.

[8] 戴安德,姜文涛,赵薇.数字人文作为一种方法:西方研究现状及展望[J].山东社会科学,2016(11):26-33.

[9] （美）比伯,（美）康拉德,（美）瑞潘著. 语料库语言学[M]. 北京：清华大学出版社, 2012.10:84-106.

[10] Biber, D. Variation across speech and writing[M]. Cambridge University Press, 1988.

[11] Biber D, Finegan E. Drift and the evolution of English style: A history of three genres[J]. Language, 1989, 65(3):487-517.

[12] Biber D. University language: A corpus-based study of spoken and written registers[M]. Philadelphia: John Benjamins Publishing Company, 2006.

[13] Biber, D., Egbert, J. Register variation on the searchable web: A multi-dimensional analysis. Journal of English Linguistics[J], 2016, 44(2):95-137.

[14] Sardinha, T. B., Marcia V. P. Multi-dimensional analysis: Research methods and current issues [M]. London: Bloomsbury Publishing Plc, 2019.

[15] 董师师,黄哲学.随机森林理论浅析[J].集成技术,2013,2(01):1-7.

[16] 鲁迅. 鲁迅全集 2005 最新修订版[M]. 北京：人民文学出版社, 2005.11.

[17] 仲济强.从“论说”到“杂感”再到“杂文”：鲁迅文体意识脉络的钩沉[J].中国现代文学研究丛刊,2013,(第 1 期):159.

[18] 朱德熙著. 语法讲义[M]. 北京：商务印书馆,1982.09:40,86.

[19] 李秀明.语体特征与句型选择——以叙事语体和描写语体为例[J].绍兴文理学院学报,2013,(第 6 期):91-93.

[20] 黄伯荣，廖序东主编. 现代汉语 下[M]. 北京：高等教育出版社, 2002:14-15,38,42,45

[21] 俞士汶，朱学锋. 现代汉语语法信息词典[EB/OL].[2019-06-15]. http://dx.doi.org/10.18170/DVN/EDQWIL

[22] Welch, B. L. The generalization of student's problem when several different population variances are involved [J]. Biometrika, 1947, 34: 28－35.

[23] Zimmerman, D. W. A note on preliminary tests of equality of variances[J]. British Journal of Mathematical and Statistical Psychology, 2004, 57: 173-181.