Hans Journal of Data Mining 数据挖掘, 2022, 12(1), 73-79
Published Online January 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.121008

Hans汉斯

基于词频分析的K-Means特征聚类算法的《红楼梦》作者分析

郑佳莉，柯小玲*，江晓莹，陈淑悦

闽江学院，数学与数据科学学院(软件学院)，福建 福州

收稿日期：2021年12月17日；录用日期：2022年1月17日；发布日期：2022年1月27日

摘 要

本文提出一种“基于词频分析的K-means特征聚类算法”来分析存疑文献的作者信息。以《红楼梦》为例，根据在前80回和后40回中确定的特征汉字的出现频率，用基于词频分析的K-means特征聚类算法对其分析。以每10回为一个文本，研究前、中、后四十回的相似度，从而得出《红楼梦》的前八十回与后四十回很可能并非一人所作的论断。

关键词

词频，K-Means特征聚类算法，相似度

Analysis of the Author of *A Dream of Red Mansions* Based on K-Means Feature Clustering Algorithm with Word Frequency

Jiali Zheng, Xiaoling Ke*, Xiaoying Jiang, Shuyue Chen

College of Mathematics and Data Science (Software College), Minjiang University, Fuzhou Fujian

Received: Dec. 17th, 2021; accepted: Jan. 17th, 2022; published: Jan. 27th, 2022

Abstract

In this paper, “a K-means feature clustering algorithm based on word frequency analysis” is proposed to analyze the author information of doubtful documents. Taking *A Dream of Red Mansions* as an example, the K-means feature clustering algorithm based on word frequency is used to ana-

*通讯作者。

文章引用: 郑佳莉, 柯小玲, 江晓莹, 陈淑悦. 基于词频分析的 K-Means 特征聚类算法的《红楼梦》作者分析[J]. 数据挖掘, 2022, 12(1): 73-79. DOI: 10.12677/hjdm.2022.121008

lyze it according to the occurrence frequency of characteristic Chinese characters determined in the first 80 chapters and the last 40 chapters. Taking every 10 chapters as a text, by studying the similarity of the first, middle and last 40 chapters, it is concluded that the first 80 chapters and the last 40 chapters of *A Dream of Red Mansions* are probably not made by one person.

Keywords

Word Frequency, K-Means Feature Clustering Algorithm, Similarity

1. 引言

《红楼梦》创作于清朝的康熙、雍正和乾隆时期，以荣国府的日常生活为中心，主线是贾宝玉、林黛玉、薛宝钗的恋情和婚姻悲剧，以及大观园的小事，以贾宝玉和金陵十二钗为核心，流露出人类之美和悲剧之美。家庭的悲剧、女儿的悲剧以及主人公毕生的悲剧都揭露出一场封建末世危机。《红楼梦》生动地凸现了繁荣时代背后的严重危机，深入反映了那个时代的个体、家庭、社会和国家的命运，并表现出强烈的焦虑、遗憾、可惜、怀旧和无助。然而 1917 年北京大学开设了小说课程，标志着《红楼梦》作为古代小说研究的组成部分获得了学术制度的认可[1]，为其日后的发展奠定了基础。随后出现了李鹏飞开始关注红楼梦饮食描写，认为其“体现了曹雪芹高超的烹饪美学和贾府的南方生活习惯；在精神文化层面上包含着很深的文化、艺术和哲学意蕴”[2]，渐渐地开展了对《红楼梦》一系列的研究。

由于某些历史原因，《红楼梦》在其传播和保存过程中，不幸存在一定的缺失，就目前大众的认知，在 120 回中的前 80 回是由曹雪芹撰写的，而后 40 回则由高鹗续写。从文学作品的角度出发，不同作者的用词习惯、语言风格都大相径庭，研究《红楼梦》是否由同一作者撰写，对研究小说历史背景和文学发展历程有着深刻意义。古人的论点在文学术语和数理统计两大类中重复出现，得出的结论：《红楼梦》的前 80 回和后 40 回不是由同一个人所撰写。近年来，随着计算机技术的发展，学界出现了若干采用计算机技术来分析《红楼梦》作者问题的工作。王世海和施政对这些工作进行了总结[3] [4]。

词频表示一个词在文本中出现的频率，是一种用于情报检索与文本挖掘的常用加权技术，用于评估一个词对于一个文件或者一个语料库中的一个领域文件集的重复程度。而词频分析法是利用能够揭示或表达文献核心内容的关键词或主题词在某一研究领域文献中出现的频次高低来确定该领域研究热点和发展动向的文献计量学算法。由于一篇文回的关键词是文回内容的浓缩和提炼，因此，如果某一关键词在其所在领域的文献中反复出现，则可反映出该关键词所表征的研究主题是该领域的研究热点。这一算法在实际应用中，可以根据某一研究领域内所有关键词在相应领域的文献中出现的频次来辨认文献作者。因此，本文基于词频分析，了解并辨认可疑文献的真实作者。两部文献之间的言语格调差别，不但体现在某些单词类型的重复程度上，而且还体现在频繁呈现的单词的频率上。假设两部作品是由同一作者撰写的，则它们的相关系数会很高；假设这两部作品是由不同的作者撰写的，它们的相关系数将会很低[5]。

2. 相关研究进展

早在 1952 年，瑞典汉学家高本汉(B. Karlgren)就使用统计方法剖析了 32 种语法和京话与口语用法[6] [7]，并得出结论，所有 120 回都是曹雪芹写的，多数研究者提出了异议。赵冈、陈钟毅在《红楼梦新探》

中谈到高本汉存在的错误，一是前八十回与后四十回并非两部独立的小说，后者是前者的延续，后出者一定会对前者进行有意模仿；第二，高本汉的分类过于粗糙，只分出了“出现”、“不出现”两类，无法分辨出细微差别。研究表明，前 40 回和 67 回，除了 67 回，其余与前 80 回明显不同。可以断定，最后的 40 回和第 67 回不是前 80 回的原始作者所写；和第 67 回同样在一些脂本中缺失的第 64 回却和前 80 回有很高的相似性，可判定第 64 回为前 80 回的作者原作，而第 67 回可能是后 40 回作者补作。进一步剖析得出，第 105 回与后 40 回的其余各回有显著不同，可判定为不同作者所撰写[8]。

应用统计措施的主要研究如下：李国强等[9]依据《红楼梦》的词频和相关性对其作者进行了研究，将《红楼梦》的前中后各 40 回分为三个模块，得到这三个模块的相关度很高，但是相关性非常低。施建军[10] [11]通过两种算法进行了研究：K-均值聚类和支持向量机。其中，聚类技术没有解决施建军自己质疑聚类分析的方法存在的困难，那就是多大的风格差别算是不同的作者所作；支持向量机算法得出的结论：在前 80 回和后 40 回之间存在明显的风格差异，但对于参数和核函数选择敏感，目前比较成熟的核函数及其参数的选择都是人为的，根据经验来选取的，带有一定的随意性。叶雷[12]基于计量的文体特征，使用 K-means 特征聚类算法研究了《红楼梦》的作者，总结出最后 40 回和 67 回不是前 80 回作者写的。

本文主要使用基于词频分析的 K-means 特征聚类算法分析“红楼梦”的前 80 回和后 40 回的文本。分析《红楼梦》的 120 回的词量和词频，看它们是否是同一作者所写。

3. 基于词频分析的 K-Means 特征聚类算法

3.1. 理论依据

特征提取和特征选择算法均提高了学习性能，减少了计算开销，并提供了更通用的模型。然而，特征选择优于特征提取，因为它保留了原始特征并删除了一些冗余特征，从而使特征选择更具可读性和可解释性。特征提取将特征从原始空间映射到新的低维空间，并且变换后的特征没有物理意义。通过特征选择可以辨别不同类型样品的特征。

K-means 特征聚类算法用于计算相关程度，它对聚类所得的簇划分 k 组：$C=\{C_1,C_2,C_3,\cdots,C_k\}$，最小化平方误差为：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|_2^2,$$

其中 $\mu_i=\frac{1}{C_i}\sum_{x\in C_i}x$ 是簇 C_i 的均值向量。

该算法的主体思想是使用迭代算法：首先，随机初始化簇中心，将每个数据点分配给最近的簇中心；然后，将每个簇中心设置为所有分配的数据点的平均值；如此往复迭代，当簇中心保持不变时，该算法终止。

K-means 特征聚类算法应用广泛、速度快，虽对于离群点和孤立点敏感，但通过去除离群点后再聚类，减少离群点和孤立点对于聚类效果的影响。K-means 特征聚类算法的缺点在于 k 值的选择有困难。k 值的选择问题，在安徽大学李芳的硕士论文中提到了 K-means 算法的 k 值自适应优化算法[13]，并进行了改进：

1) 必须首先给出 k (要生成的簇的数目)，k 值很难选择。事先并不知道给定的数据应该被分成什么类别才是最优的；

2) 初始聚类中心的选择是 K-means 的一个问题。

李芳[13]设计的算法思路是这样的：可以通过在一开始给定一个适合的数值给 k，通过一次 K-means

聚类算法得到一次聚类中心。对于得到的聚类中心，根据得到的 k 个聚类的距离情况，合并距离最近的类，因此聚类中心数减小，当将其用于下次聚类时，相应的聚类数目也减小了，最终得到合适数目的聚类数。可以通过一个评判值 E 来确定聚类数，得到一个合适的位置停下来，而不继续合并聚类中心。重复上述循环，直至评判函数收敛为止，最终得到较优聚类数的聚类结果。本文对这种 K-means 算法加以改进，并加入了 jieba 分词，能准确地判断出是前八十回与后四十回之间存在差异。

即使处理大量数据，K-means 特征聚类算法的稳定性、效率和准确性(与实际标签识别相比)也非常好。该算法的时间复杂度的上限为 $O(nkt)$，其中 n 是样本大小，k 是划分的聚类数，t 是迭代次数。假定聚类数和迭代数保持相同，则 K-means 特征聚类算法所需的时间仅与样本大小有关，因此显示线性增长趋势。

为了避免仅以《红楼梦》一部作品作为样本进行聚类分析，不能够判别《红楼梦》作者的所属问题，本文对该算法进行了改进，并利用改进后的基于词频分析的 K-means 特征聚类算法对研究《红楼梦》作者的所属问题进行了分析研究。新算法的具体步骤如下：

1) 对《红楼梦》语料库出现的词汇进行频率降序排列；

2) 选择已排序的词汇(除去语气助词)，根据词频由高到低依次选择五个字；

3) 利用基于词频分析的 K-means 特征聚类算法进行计算，用 K-means 特征聚类来检验每个文本之间高频词的相关性，并用 jieba 分词对前八十回与后四十回进行进一步分析。

3.2. 模型假设

1) 假定标题给出的数据是真实可靠的；

2) 假定 TXT 格式的《红楼梦》完整文本中没有出现错误；

3) 假定每个样本都是独立的(不受影响或不受其余样本影响)；

4) 假定样本选择是随机且通用的；

5) 假定所选词汇使用频率的异同能够绝大部分地区分作者的写作格调；

6) 假定实验过程中的出现误差可忽略不计。

3.3. 实验与分析

《红楼梦》的正文共有 874,592 字，使用手动统计算法不仅工作量大，而且容易出现错误。但计算机具有快速和准确的计算这两个特征，并且使用计算机进行统计是很自然的结果。

李瑞芳等[14]使用 Java 编程对《红楼梦》中的字进行计数，得出《红楼梦》共有 189 个字出现频率超过 700 次，这 189 个字共被使用 498,630 次，这些字涵盖了全 56.0887%的内容，在 4401 个单字中，这 189 个字是最常见的。本文最终敲定研究“宝、贾、姐、老、红”这五个字在书中前八十回和后四十回中出现的频率，这五个字的选出是在删去了语气助词之后，又依据词频从高到低排序，筛选出的前五个字。

3.3.1. 利用 K-means 特征聚类算法检验

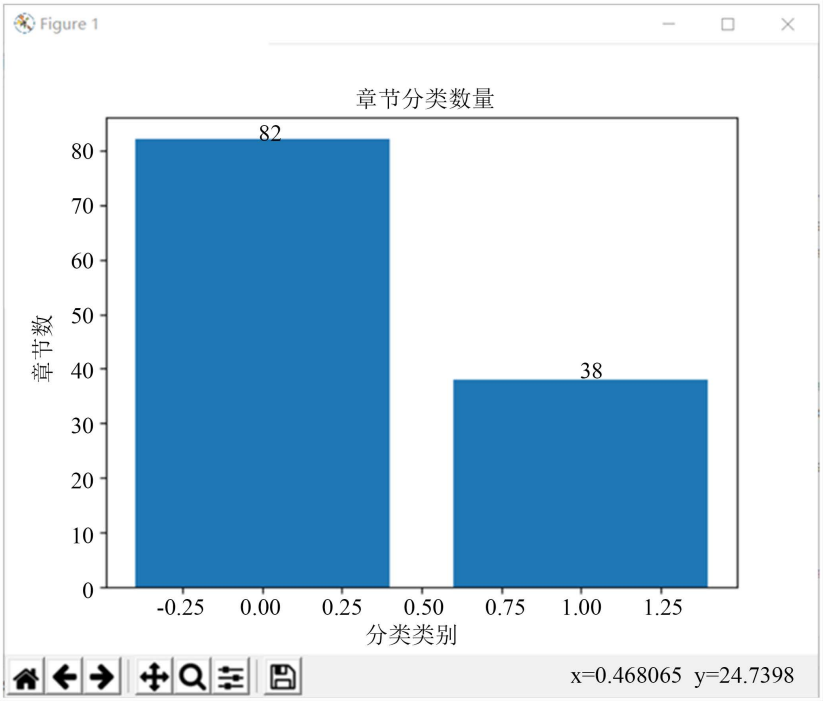
1) 实验过程

首先，根据每回的分词得到 TF-IDF 矩阵(各个词在每回的所有分词中出现的频率，也就是各个词在每回中出现的频率)；然后，以 TF-IDF 矩阵作为数据，通过余弦相似的 K-means 特征聚类算法进行分类。

K-means 特征聚类算法的演算算法：对指定的样本，根据样本间距离的大小，将样本划分为 2 个簇(2 个类别)，某样本离哪个簇的簇中心最近，这个样本就属于这个簇(属于这个分类)。余弦相似是通过测量两个向量的夹角余弦值来度量他们的相似性，为 1 则完全重复，0 则完全不相关。

2) 实验结果与分析

通过计算机检验可以得到的每个种类下的分类数量，得到两个结果(如图 1 所示)，对实验得出的结果取平均值，可以得出有八十回的样本被归为第一簇，有四十回的样本被归为第二簇。我们探讨的是《红楼梦》前八十回和后四十回的作者问题，结合实验结果可以推断出来前八十回和后四十回确实存在差异，但为了更验证这一想法，我们利用 jieba 工具对文本进行进一步处理。



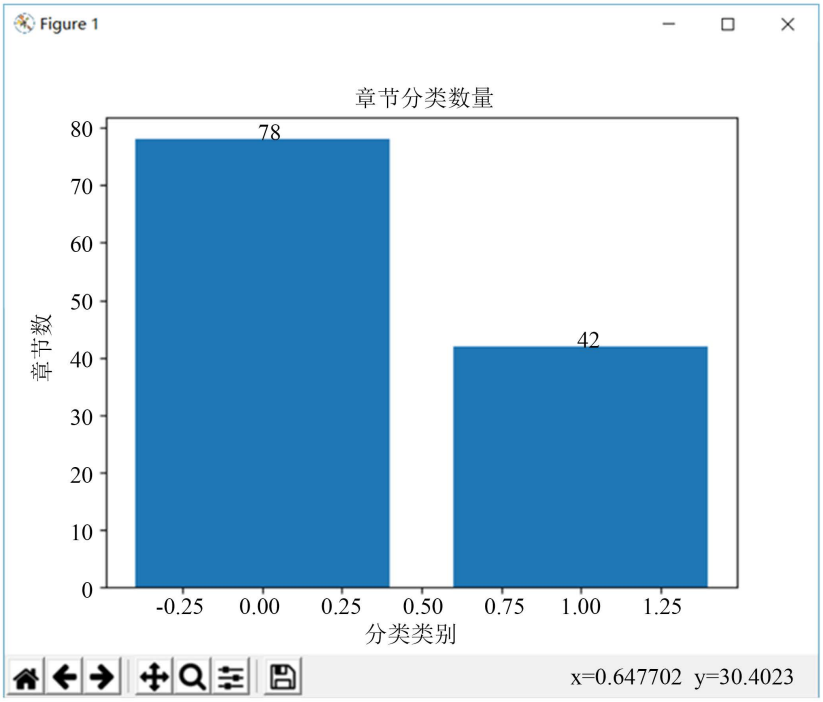


Figure 1. K-means algorithm test results of text data
图 1. K-means 算法对文本数据检验结果

3.3.2. 利用 Jieba 分词进行数据检验

以《红楼梦》作为实验文本，使用 jieba 工具对文本作分词处理。将《红楼梦》的 120 回分为 12 个语料库，每个语料库包含 10 回，因此第一个语料库包括第 1~10 回，第二个语料库包括第 11~20 回，……，第十二个语料库包括第 111~120 回。以每 10 回为一个样本，分别计算特征词的出现频率，并选择替换词进行统计。首先，数据附件每十回返回一个样本，导入 python；然后，应用 python 的 jieba 模块，按照词性(代词)进行切分，统计选定五个词汇出现的词频，结果示例如表 1 所示。

Table 1. Jieba word segmentation processing results of text data
表 1. Jieba 分词对文本数据进行处理结果

	1~10 章回	11~20 章回	21~30 章回	31~40 章回	41~50 章回	51~60 章回	61~70 章回	71~80 章回	81~90 章回	91~100 章回	101~110 章回	111~120 章回
“红”字	52	28	132	54	51	54	55	53	65	30	30	27
“宝”字	334	423	688	590	411	480	364	347	510	516	388	584
“贾”字	253	527	438	244	312	355	457	391	434	479	645	513
“老”字	160	153	113	177	264	225	193	199	257	271	326	212
“姐”字	166	301	259	210	305	185	623	322	285	198	237	273
合计	955	1432	1630	1275	1343	1299	1692	1312	1551	1494	1628	1609
章回总数	67,253	60,589	67,816	68,252	71,007	75,466	75,966	82,722	70,046	61,131	65,588	72,015
“红”字占比	0.07732%	0.04621%	0.19464%	0.07912%	0.07182%	0.07156%	0.07240%	0.06407%	0.09280%	0.04907%	0.04574%	0.03749%
“宝”字占比	0.49663%	0.69815%	1.01451%	0.86444%	0.57882%	0.63605%	0.47916%	0.41948%	0.72809%	0.84409%	0.59157%	0.81094%
“贾”字占比	0.37619%	0.86979%	0.64587%	0.35750%	0.43939%	0.47041%	0.60158%	0.47267%	0.61959%	0.78356%	0.98341%	0.71235%
“老”字占比	0.22304%	0.25252%	0.16663%	0.25933%	0.37179%	0.29815%	0.25406%	0.24056%	0.36690%	0.44331%	0.50009%	0.29438%
“姐”字占比	0.24683%	0.49679%	0.38192%	0.30768%	0.42954%	0.24514%	0.82010%	0.38926%	0.40688%	0.32389%	0.36135%	0.37909%

3.3.3. 结论分析

K-means 不足之处在于：只能分出红楼梦中有四十回和另外八十回不同，并不能分辨出是前八十回与后四十回不同，所以增加了 jieba 分词这一检验过程，更加明确的看出《红楼梦》的前八十回与后四十

回存在差异。通过实验，剔除异常数据后，可以看出：“红”字的占比在前八十回中平均数为0.0846425%，且众数为0.07%；但在后四十回中“红”字的占比平均数为0.056275%，众数为0.04%。“宝”字的占比在前八十回中平均数为0.648405%，且众数为0.5%；但在后四十回中“宝”字的占比平均数为0.7436725%，众数为0.8%，依次对“贾”“老”“姐”三个字进行类推可以得出，就平均利用率来讲，这五个词汇的使用情况在一到八的模块中差异不大，九到十二的差异也不大，但前八个模块与后四个模块却存在较大的差异，并且在众数的对比上更能明显地看出来《红楼梦》的前八十回与后四十回的不同。

4. 总结

为了剖析前八十回与后四十回产生差异的原因，本文在这里仅作一些初步的推断：一个是表达需求上的差异，另一个是创作风格上的差异，即文字的书写方式上的差异。平均用法的差异可以通过作者的措辞习惯来解释；导致语法动能和语言功能不同的原因相对复杂，不容易得出结论，初步猜测需要不同的表达方式和不同的作者创作风格来解释；除此之外，写作时间顺序也有一定影响。

基于词频分析算法的作者分析结果是数值统计的结果，是可能性的结果，可以作为文献和文献学中文本进一步分析和验证的参考[15]。本文中改进后的 K-means 特征聚类算法分析的结果支持前八十回和后四十回是由不同作者所创作的结论。

基金项目

福建省自然科学基金项目(2020J01844)，福建省大学生创新创业项目(S202010395034)。

参考文献

[1] 苗怀明. 红楼梦研究史论集[M]. 沈阳: 辽宁人民出版社, 2019.

[2] 李鹏飞. 人莫不饮食也, 鲜能知味也——谈《红楼梦》与饮食文化[J]. 红楼梦学刊, 2020(4): 84-120.

[3] 王世海. 论数理统计方法研究《红楼梦》作者问题的得与失[J]. 宜春学院学报, 2019, 41(4): 105-109.

[4] 施政. 《红楼梦》研究中的统计方法综述[J]. 吉林省教育学院学报, 2019, 35(1): 151-156.

[5] 马创新, 陈小荷. 从高频词等级相关角度探析《红楼梦》作者[J]. 中文信息学报, 2018, 32(11): 97-102.

[6] 胡适. 《红楼梦考证》(改定稿)[M]. 北京: 北京出版社, 2015.

[7] Koppel, M., Schler, J. and Argamon, S. (2009) Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, **60**, 9-26.

[8] 程东波, 柯小玲, 林施鑫. 基于等价性检验和特征聚类的《红楼梦》作者分析[J]. 理论数学, 2020, 10(5): 549-555.

[9] 陆尚辉. 基于 R 软件和 KNN 算法的《红楼梦》作者分析[J]. 魅力中国, 2017(7): 81+63.

[10] 施建军. 关于以《红楼梦》120 回为样本进行其作者聚类分析的可信度问题研究[J]. 红楼梦学刊, 2010(5): 318-335.

[11] 施建军. 基于支持向量机技术的《红楼梦》作者研究[J]. 红楼梦学刊, 2011(5): 35-52.

[12] 叶雷. 基于计量文体特征聚类的《红楼梦》作者分析[J]. 红楼梦学刊, 2016(5): 312-324.

[13] 李芳. K-Means 算法的 k 值自适应优化算法研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2015.

[14] 李瑞芳, 孙军波, 常诗珧. 基于计算机的《红楼梦》字词浅探[J]. 电脑知识与技术, 2009(5): 753-755.

[15] 叶雷. 基于计量文体特征聚类的《红楼梦》作者分析[J]. 红楼梦学刊, 2016(5): 312-324.