DOI:10. 13471/j. cnki. acta. snus. 2019. 07. 08. 2019A052

基于信息熵的语言风格分析方法初探*

王泓，方艳梅，黄方军
（中山大学数据科学与计算机学院，广东 广州 510006）

摘　要：提出一种对于词汇丰富程度的量化标准——信息熵，并验证信息熵的确可以反映文本的词汇丰富程度。先将英文小说分成四类，分别是魔幻/科幻小说，推理小说，幽默讽刺小说，儿童文学。并计算每一类中的每一本英文小说作品的信息熵，然后通过图表的方式对这四类小说的信息熵进行对比，并且根据以往的对于小说风格的研究和平时的阅读经验，观察四类小说的信息熵差别是否如同预期所猜想的一致。通过验证发现，儿童文学的信息熵普遍偏低，而魔幻/科幻小说的信息熵普遍较高，而根据以往的研究和平时的阅读体验来看，魔幻/科幻小说词汇丰富程度确实较高，儿童文学词汇丰富程度的确较低。之后用假设检验的方法验证不同类型作品信息熵的差异。由此说明信息熵可以作为反应词汇丰富程度的一个指标。

关键词：信息熵；词汇丰富程度；计量风格学；统计假设检验
中图分类号：TP391　　**文献标志码：**A　　**文章编号：**0529-6579（2020）06 -0113 -13

A preliminary study on text style analysis based on information entropy

WANG Hong, *FANG Yanmei*, *HUANG Fangjun*
（School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China）

Abstract: It is proposed and verified that the information entropy is a quantitative standard for lexical richness. Firstly, the English novels are categorized into four groups, namely, magic/science fiction, mystery novels, humorous satirical novels, and children's literature. Then the authors calculate the information entropy of each English novel, compare the information entropy of the four groups by means of graphs, and observe whether the difference of information entropy among these four categories consists with what the authors' expectation. Through verification, the authors find that the information entropy of children's literature is averagely the lowest, and the information entropy of magic/science fiction is generally higher. According to previous studies and our usual reading experience, the magic/science fiction indeed has higher vocabulary richness, and the vocabulary richness in children's literature is lower. Finally, the authors use hypothesis testing to verify the difference of entropy among the categories. Then, the authors conclude that information entropy can be used as an indicator of the vocabulary richness.

Key words: information entropy; vocabulary richness; stylometry; hypothesis testing

* **收稿日期：**2019 - 07 - 08
基金项目：国家自然科学基金（62072481，61772572）；中韩国际交流合作（61811540409）
作者简介：王泓（1996 年生），男；**研究方向：**机器学习与深度学习、数字媒体信息安全；E-mail：13719175117@163. com
通信作者：方艳梅（1966 年生），女；**研究方向：**机器学习与深度学习、数字媒体信息安全；E-mail：fangym@mail. sysu. edu. cn

信息熵（entropy）是来源于信息论中的一个概念，它描述了信息的不确定程度。信息熵越大，

信息熵在很多方面都有一定的应用。在关键帧的提取技术[1]中，可以利用互信息熵和互信息量对视频序列提取关键帧。在应用支持向量机（supported vector machine，SVM）进行网络入侵检测[2]时，可以利用信息熵选取特征，然后再进行入侵检测。对于相似性度量，信息熵也有一定的应用。使用聚类算法对样本进行聚类[3]时，可以使用叠加信息熵场对样本之间的距离进行度量。类条件概率在相似性度量中也有相应的应用[4]。

文本的语言特征可以表现作者在写作时的语言特点，是作者个人风格的深刻反映[5]。词汇丰富程度是文本的一个重要的语言特征。一本小说的词汇丰富程度越大，那么这本小说在表达上可能更加丰富，但同时阅读起来也可能会比较困难。目前在对文本进行聚类分析和分类中，语言风格是一个应用较多的分类依据[6]。说明语言风格在文本聚类中有着很好的应用。在一些文学作品的风格分析上，语料分析也有着重要的应用。

计量风格学（stylometry）是一门基于语料库，利用统计分析的方法，对文本的语言特征进行研究分析的学科[7-8]。目前已有很多学者对文学作品进行计量风格的分析，通过对不同作者的作品的语言风格特征进行对比，得到不同作者之间的语言风格差异，以此对不同作者进行简单的识别。文献［9］对苏童和毕飞宇两位作者的作品进行分析，分别选取了两位作者的四本小说作品，对小说中的标点符号，语气助词以及实词词类等可量化的语言特征进行比较，发现这些语言特征能够比较明显地区分这两位作者。在对两位作者的实词词类的分析上也可以看到，苏童在小说作品中使用实词的频次要高于毕飞宇在小说作品中的实词使用频次的。实词是用来表达意义的，具有很强的信息传递能力[10]。因此这个结果也可以表明，苏童的小说作品信息性比毕飞宇的小说作品信息性要强。因而实词的词类也可以成为区分两位作者的一个指标。在文献［11］中，对余华的小说作品和格非的小说作品进行了对比。在词汇丰富程度这一角度上，作者提出了词汇独特性，词汇多样性和词汇密度三个参数对词汇丰富程度进行描述。词汇独特性是指文本中只出现一次的词汇的数量占文本总词汇数量的比例。词汇多样性作者提出使用型例比（Type-Token Ratio，TTR）对其进行描述，即词型与词例的比值，并对其求对数得到对数TTR，对数TTR的值越高说明其词汇丰富程度越大。词汇密度是实词数量与整本小说的词汇数量的比值。这三个参数在对余华的小说作品和格非的小说进行对比分析后发现，与预期的两位作者的小说的词汇丰富程度是一致的。

本文希望对更多的作者进行分析，并提出一个比上述三个参数更加简洁的指标，对文学作品的词汇丰富程度进行分析。

在文学作品中可以用每个单词出现的频率作为概率，进而求出整本小说的信息熵。小说作品中的信息熵越高，说明这本小说用到的词汇重复性比较小，每个单词出现的频次较低。反之，小说作品中的信息熵越小，说明这本小说用到的词汇的重复性比较大，很多词汇出现了多次。从信息熵的定义和词汇丰富程度的概念来看，两者之间是很有可能有一定的联系的。信息熵可能能够反映一本小说作品的词汇丰富程度，成为衡量词汇丰富程度的一个参考性的数据指标。

1 信息熵

事件x的自信息表示的是事件发生之前，事件的不确定性，概率大的事件比较容易发生，预测其何时发生比较容易，因此不确定性比较小。同时也可以表示事件发生之后，事件所包含的信息量[12]，概率大的事件不仅容易预测，发生后所提供的信息量也小。将事件x的信息量简记为$I(x)$。

$$I(x) = -\log p(x) \tag{1}$$

$I(x)$是$p(x)$的单调递减函数，意味着概率越大，自信息会越小。

离散随机变量X的信息熵的定义表示为自信息的平均值[13-14]，记为$H(X)$。在信息论中，信息熵有如下的含义[15]。

在信源输出后，表示为每个新源符号所提供的的平均信息量。在信源输出之前，可以表示该信源的平均的不确定性。也可以表示信源的随机性大小，信息熵大的表示信源的随机性大。当信源输出之后，信息熵亦可视作接触不确定性所需要的信息量。

$$H(X) = \underset{p(x)}{E}[I(x)] \tag{2}$$

由式（1）可知，$I(x)$是事件x的自信息，E表示对随机变量用$p(x)$进行取平均运算，即可得到信

息熵。将式（1）代入式（2）可得到如下的式（3）。

$$H(X) = -\sum_{x} p(x)\log p(x) \tag{3}$$

本文希望通过计算每本小说作品的信息熵，把每一个英文实词当做是随机变量，根据信息熵的公式进行计算，通过分析进一步得到信息熵和小说词汇丰富程度的关系，从而验证信息熵是可以反映文学作品的词汇丰富程度的。

为了能够让实词出现的频率代替概率，这里引入了一个定理，即贝努力大数定律。贝努力大数定律[16]阐述了当试验次数很大的时候，事件发生的频率可以近似地当做事件发生的概率。

设$N(A)$表示事件A在N次独立重复事件中发生的次数，p表示事件A发生的概率，对任意的ε有，

$$\lim_{N\to\infty} P\left\{\left|\frac{N(A)}{N} - p\right| < \varepsilon\right\} = 1 \tag{4}$$

$N(A)/N$表示的就是事件A发生的频率，当N的数量足够大的时候，频率和概率的差值小于一个极小的数的概率趋近于1。这意味着此时可以将频率近似地当做概率。

在信息熵的计算中，根据贝努力大数定律式（4），当样本容量很大时某个事件出现的频率几乎接近于事件发生的概率。对于每本小说作品，单词的总数相对于每个单词出现的次数来说，是比较大的。因此可以把每本小说作品的单词数量近似当做是无穷大。在计算每一个单词出现的概率时，只要计算出单词的频率，将单词出现的频率当做是概率，带入信息熵的计算公式进行计算，即可得到每本小说作品的信息熵。

2　实验方案设计与实现

为了验证信息熵能够反映文学作品的词汇丰富程度，需要对实验的方案进行设计。首先需要进行样本的收集和处理，然后考虑分析的对象，接下来是进行分词处理并统计词汇数量，计算信息熵，最后处理并分析数据得到结论。

本文主要探讨的是信息熵与词汇丰富程度的关系。正如上述提到的，文学作品中的实词有很强的信息性，因此本文在信息熵的计算上主要考虑的是实词。计算每一个实词在所有实词中出现的频率，并将这个频率作为概率，根据上述提到的信息熵的式（3）计算信息熵。对于词汇丰富程度主要是基于对于不同类型的小说的定性分析以及平常的一些阅读经验进行分析。针对实词这一分析对象，设计流程图图1。

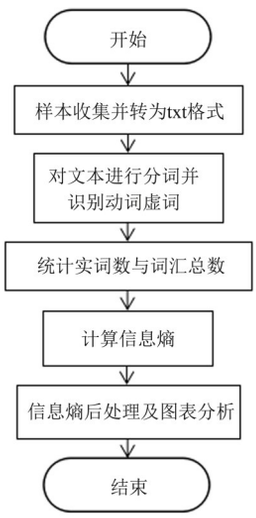

图1　方案设计流程图

Fig. 1　Flow chart of scheme design

如图1所示，分析信息熵与词汇丰富程度的关系，首先需要收集样本，然后把样本转化为text文档，对于每一本小说，即每一个文档，进行分词，然后识别出实词与虚词，对实词以及整本小说的词汇数量进行统计，计算出信息熵。接下来把得到的信息熵数据制成表格并且绘出图线，对表格数图进行分析。

2.1　样本收集

本文主要的研究对象是英文小说作品，需要收集小说作品主要是英国和美国两国的小说作品。为了便于研究，在一开始收集小说作品的时候，就按照不同的类别去收集。本文把英文小说作品分为四类，分别是魔幻/科幻小说，推理小说、幽默讽刺小说、儿童文学。对于每一类，需要收集多本小说作品，使得数据比较有说服力。每一类小说作品再按照作者进行分类，例如魔幻/科幻小说，《哈利波特》系列小说的作者成为一类，《魔戒》系列小说的作者也单独成为一类。这样也是为了探究写不同类型的小说的作者之间在信息熵上是否有什么区别。

由于一些小说作品受到版权的保护，在网上很难找到英文版的小说原著，因此有些作者只能找到两三本小说作品，甚至有些作者只能找到一本小说作品。但只要每一类的小说作品的数量足够，单个作者的小说作品数量少对研究的结果不会产生太大的影响。

有一些小说作品在网上只找到了PDF格式，

因此需要将PDF转化为TXT文本文档的格式。在网络上找到了可以在线转换的工具，因此就不需要自己编程实现PDF转换为文本文档了。

最终总共收集110本小说作品，其中魔幻/科幻小说共计34本，推理小说共39本，幽默讽刺小说共计16本，儿童文学共计21本。

2.2 预处理与分词

根据上述的分析对象，本次主要分析的是实词。因此在分词的时候，除了把每一行进行单词的划分之外，还需要考虑如何识别实词和虚词。

在统计实词的时候，相同单词的不同形态不应该进行区分，例如，do，does，did和done应该视为同一个单词，如果文本中出现了do的这四种形态，应当视为do这一单词出现了四次，而不是考虑成四个单词。

因此，在分词的同时应当考虑如何区分实词和虚词，并且对于每一个实词要将其还原为原形，这样才能比较准确地统计数据。但是如果自己实现对实词和虚词的识别，可能需要考虑使用机器学习的算法进行实现，但是这样需要花很多功夫在识别上，而本次探究的重点并不在于如何识别虚词和实词。同理，对于把实词还原为其原形，也可能需要采用机器学习的方法进行学习训练，才能有效地识别实词的各种形态并将其还原为原形，本次探究的重点也不在于此。

Python在数据分析上面有着很强大的优势，在文本处理上它也提供了很多很强大的工具。本文在分词和预处理上考虑使用Python中的NLTK库[17]，NLTK库中提供了分词的工具，并且会对于分好的每一个单词进行标记，能够有效地区分该单词是名词、动词、形容词或副词，也能很方便地标记出是形容词的原形还是比较级或是最高级等。

同时Python的NLTK库中还提供了将实词的各种形式还原为原形的工具，例如能有效地把动词的第三人称单数还原为原形，有利于减少计算重复单词的可能性，提高统计的准确性。因此，在分词和识别虚词实词时使用Python的NLTK库。

2.3 信息熵计算

在信息熵的计算中，每本小说作品，单词的总数相对于每个单词出现的次数来说，是较大的，可以近似地认为单词总数是无穷的。根据贝努力大数定律，可以将词频作为单词出现的概率进行计算。

求出每个单词出现的概率后，再进行求对数的计算，此时的对数是以2为底的对数。把每个单词出现的概率以及概率的对数相乘得到一个数值。对该数值进行求和计算，然后加上负号即可得到小说的信息熵数值。

对于每本小说，每遍历一个单词，都需要判断其为实词还是虚词，才能计算出实词的总数。如果是实词，那么实词的总数需要自增，实词的总数初始化为0，这样就能统计出整本小说作品的实词的总数有多少。

对于每一个单词，当判断为实词时，需要计算出该单词在文本中出现的次数，得到次数后，用次数除以实词的总数，即可得到单词出现的频率。利用贝努力大数定律，将频率近似为概率。根据式（3），将概率代入公式进行计算，从而得到整本小说的信息熵。

在实际实现的时候，由于实词中的名词、动词、形容词和副词最能够反映信息，因此在这里只考虑了这四种实词。

实现时，在完成分词的过程中就已经计算得到实词的总数（count），并且把所有实词（名词、动词、形容词和副词）存放与一个列表中。然后利用Python的字典这一数据结构，字典的每一个元素是一个二元组（key，value），在这里可以用key记录单词，用value记录单词出现的次数，字典中的key是不允许重复的。遍历列表中的每一个单词，判断是否已经出现在字典中，如果没有出现则以该单词为key对应的value值为1。如果已经出现过，则对应的value值自增。

然后遍历这个字典，对于字典中的每一值对（key，value），用value除以小说的实词总数（count），得到的就是该单词的词频。用这个词频作为该单词出现的概率，根据信息熵的计算式（3），求得整本小说的信息熵的值。

2.4 利用双曲正切函数处理数据

双曲正切函数在深度学习中有着重要的应用，可以应用于深度学习的激活函数，其本身能将数据映射到−1到+1之间。由于实验的数据基本上都落在7到12之间，因此可以将双曲正切函数做一个平移，使得实验的数据都能映射到−1到+1之间，这样数据之间的差别就会更加明显。

2.4.1 双曲正切函数

双曲正切函数：

$$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{5}$$

其定义域为（−∞，+∞），值域为（−1，+1）。该函

数能将任意实数映射到-1到+1之间。图2表示的是双曲正切的函数图像，横坐标的范围为-10到+10之间。

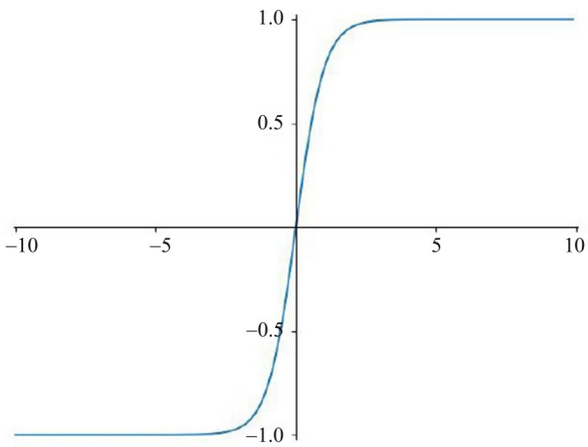

图2　双曲正切函数图像

Fig. 2　The figure of tanh-function

2.4.2　用双曲正切处理数据　实验中的信息熵的数据均落在7到12之间，因此可以通过将双曲正切的函数进行平移，将对称中心从（0，0）移到（9.5，0）。其平移后的双曲正切函数：

$$\tanh(x)=\frac{e^{x-9.5}-e^{-(x-9.5)}}{e^{x-9.5}+e^{-(x-9.5)}} \tag{6}$$

其图像如图3。从图中可以看出，该函数的确能将范围比较大的数据映射到-1到+1之间，使得数据更加紧凑，这样不同类别之间的差别会更加明显。

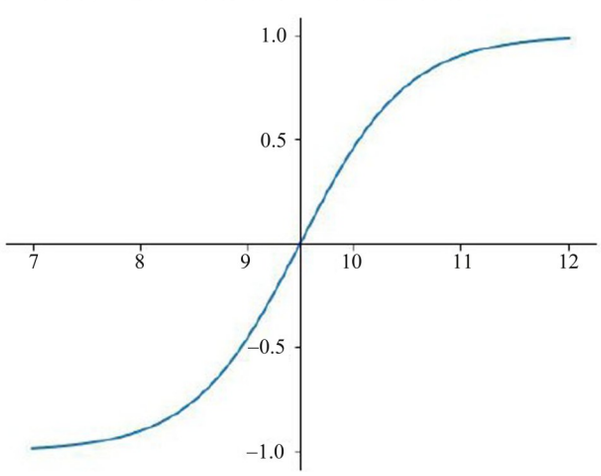

图3　平移后的tanh函数图像

Fig. 3　The figure of tanh-function after translation

2.5　绘制图表

得到的每本小说作品的信息熵会比较混乱，因此需要进行后处理，以便于后续的分析得到结果。

首先需要将所有的信息熵数据，按照不同小说的类型整理，放入四个不同的表格中，每个表格又要按照不同的作者再进行划分。对于不同的作者，可以对收集到的小说作品的信息熵求平均值，同样也记录在同一个表格中，这样就形成了四个表格。

在分析之前，把小说作品的信息熵数据用图表的方式呈现出来，会更方便进行观察和分析。所以需要把每本小说的信息熵用折线图的方式呈现出来。按照四类小说，画出四条折线段，这样能够方便地看出四种类型的小说作品有什么区别。

对于刚才求出来的平均值，也要按照不同类型的小说画成四条不同的折线段，这样两个折线图可以互相印证，共同得出规律。

在绘图之前将数据进行排序，使得四条折线段的变化趋势都是统一的由低到高，这样便于观察。Python中同样也提供了很方便的画图工具matplotlib库，可以很方便地对数据进行刻画。

3　结果分析

对于计算得到的四类小说作品的信息熵，整理成表格，将数据绘制成图并对其进行分析。在绘制折线图之前先利用双曲正切函数，即式（6），将信息熵映射到-1到+1之间。

3.1　魔幻/科幻小说分析

魔幻科幻小说信息熵如表1所示。表1的第一列是作者，第二列是作者对应的小说作品，第三列对应的是每本小说作品的信息熵，最后一列表示的是每位作者的所有小说作品的信息熵的平均值（之后的表格都表示相同的含义）。

从信息熵的数据来看，整体的魔幻/科幻小说的信息熵的数据都偏高，基本上都在10.5以上，有比较多的信息熵的值已经超过了11，从信息熵的含义上看，说明魔幻/科幻小说词汇的“不确定程度”比较大。根据以往的阅读经验，由于魔幻/科幻小说构造了一个新的与现实生活不一样的世界，因此需要用到的词汇会比较多，比较丰富，这类小说的词汇丰富程度通常都会比较大。在以往的研究中，魔幻小说通常会营造出一个“架空世界”，所谓架空世界是指魔幻文学作品中所营造的一个与现实世界完全不同的别样世界。这个世界有自己独特的地理环境、特殊的风俗习惯、奇异的种族生物；甚至具体的每一个种族都有自己独特的外形、历史、文化和语言等等[18]。这种架空世界的营造使得小说的词汇丰富程度会比较高。这与我们得到的信息熵的数据反映出来的趋势是

表 1 魔幻/科幻小说信息熵

Table 1 Entropy of magic/science fiction

作者	书名	信息熵	平均值
J. K 罗琳	Harry Potter and The Chamber Of Secrets	10. 669	10. 701
	Harry Potter and Deathly Hallows	10. 556	
	Harry Potter and the Goblet of Fire	10. 805	
	Harry Potter and The Half-Blood Prince	10. 843	
	Harry Potter and the Order of the Phoenix	10. 950	
	Harry Potter and the Prisoner of Azkaban	10. 636	
	Harry Potter and the Sorcerer's Stone	10. 450	
约翰. 托尔金	The Fellowship of the Ring	10. 811	10. 765
	The Return of the King	10. 740	
	The Two Towers	10. 744	
斯蒂芬妮. 梅尔	Twilight	11. 223	10. 807
	Breaking Dawn	11. 139	
	New Moon	11. 146	
	Eclipse	9. 972	
	Midnight Sun	10. 553	
罗伯特. 乔丹	The Gathering Storm	11. 175	11. 070
	Crossroads of Twilight	11. 098	
	Winter's Heart	11. 208	
	The Path of Daggers	11. 252	
	A Crown of Swords	11. 241	
	The Lord of chaos	11. 115	
	The Fires of Heaven	11. 030	
	The Shadow Rising	11. 061	
	The Great Hunt	10. 812	
	The Dragon Reborn	10. 524	
乔. 霍尔曼	Forever War	11. 127	11. 045
	Forever Peace	11. 068	
	Forever Free	10. 849	
乔治. 马丁	A Game of Thrones	10. 978	10. 877
	A Dance with Dragons	10. 939	
	A Storm of Swords	10. 778	
	A Clash of Kings	10. 832	
	A Feast for Crows	10. 858	
	The Eye of The World	10. 958	

基本上一致的。

对信息熵求平均值可以弱化同个小说作者不同小说的信息熵的差距，得到一个相对的信息熵数据。平均值可用来刻画每位小说作者的小说作品的总体情况。

3. 2 推理小说分析

推理小说是基于现实创作出来的小说，因此相较于魔幻/科幻小说来说，用到的词汇没有那么多样，因此其词汇丰富程度也会低于魔幻/科幻小说。而表 2 中反映出来的信息熵整体来说也会低于

表2　推理小说信息熵

Table 2　Entropy of mystery novel

作者	书名	信息熵	平均值
柯南道尔	The Adventure of Sherlock Holmes	10. 464	10. 546
	The Memoirs of Sherlock Holmes	10. 317	
	The Return of Sherlock Holmes	10. 849	
阿加莎	Curtain – Poirot′s Last Case	10. 267	10. 172
	Cat Among the Pigeons	10. 176	
	And Then There Were None	10. 406	
	Murder on the Orient Express	10. 243	
	The Murder of Roger Ackroyd	9. 770	
多萝西. 塞耶斯	The Nine Tailors	10. 448	10. 521
	Clouds of Witness	10. 965	
	Murder Must Advertise	10. 620	
	Strong Poison	10. 134	
	Gaudy Night	10. 440	
雷蒙德. 钱德勒	The Lady in the Lake	10. 097	10. 301
	Farewell， My Lovely	9. 970	
	The Long Goodbye	10. 608	
	The Big Sleep	10. 527	
达尔希. 哈米特	The Glass Key	9. 935	10. 191
	Red Harvest	10. 274	
	The Maltese Falcon	10. 364	
	The Thin Man	9. 352	
托马斯. 哈里斯	The Silence of the Lambs	11. 105	11. 074
	Red Dragon	11. 043	
丹. 布朗	The Lost Symbol	10. 493	11. 113
	Deception Point	10. 902	
	The Da Vinci Code	12. 321	
	Angels and Demons	10. 909	
	Digital Fortress	10. 941	
范. 达因	The Gracie Allen Murder Case	10. 472	10. 768
	The Winter Murder Case	10. 570	
	The Scarab Murder Case	10. 744	
	The Kidnap Murder Case	10. 687	
	The Kennel Murder Case	10. 720	
	The Greene Murder Case	11. 069	
	The Casino Murder Case	10. 825	
	The Bishop Murder Case	10. 943	
	The Canary Murder Case	10. 993	

魔幻/科幻小说。

从表2中的平均值也可以看出，整体来说，推理小说的信息熵的平均值会低于魔幻/科幻小说的信息熵的平均值，这也可以在一定程度上验证信

息熵的确可以反映词汇的丰富程度。

推理小说中会用到一些相对专业的词汇，为了增加阅读的体验，在辞藻上也会加以修饰。因此，虽然没有魔幻/科幻小说的词汇丰富程度那么大，但也不会很小。从表2的数据上看，多数信息熵数据落在10到10.5之间，也有部分数据落在10.5到11之间。

3.3 幽默讽刺小说分析

表3反映了幽默讽刺小说的信息熵。从整体上看，幽默讽刺小说的信息熵不如魔幻/科幻小说那么高。从往常的阅读经验来说，幽默讽刺小说大多篇幅不长，并且用到的词汇大多数是比较平实的，词汇的丰富程度不高。这和表3中反映出来的是相符的。

3.4 儿童文学分析

表4反映了儿童文学的信息熵。共收集了5位作者的儿童文学小说共计21本。儿童文学主要是一些童话故事以及其他的一些充满想象力的小说。

从整体的数据上看，儿童文学的信息熵都偏低，尤其是一些童话故事，例如《The Devoted Friend》，根据文献［21］研究，王尔德的《The Devoted Friend》这一童话故事多以口语化的对话呈现，因此可以推断出这一童话故事所用的词汇不会很丰富，口语化的语言的词汇丰富程度大都比较低。从表格中的信息熵数据来看也是如此，是不高的，而表3所呈现出来的信息熵数据也不高，从这里也可以说明信息熵的确能够反映出词汇的丰富程度的。

表3 幽默讽刺小说信息熵

Table 3 Entropy of humorous satirical novel

作者	书名	信息熵	平均值
狄更斯	A Tale of Two Cities	11.149	10.732
	Great Expectations	11.028	
	Oliver Twist	10.020	
	David Copperfield	11.060	
欧. 亨利	A Bird Of Bagdad	9.543	9.356
	A Blackjack Bargainer	9.980	
	A Chaparral Christmas Gift	8.798	
	A Call Loan	8.824	
	A Cosmopolite in a Cafe	9.206	
	A Little Talk About Mobs	8.341	
	A Matter of Mean Elevation	10.009	
	A Municipal Report	10.038	
	A Chaparral Prince	9.464	
乔治. 奥威尔	Animal Farm	10.164	10.575
	1984	10.986	
萨克雷	Vanity Fair	11.423	11.423
马克. 吐温	The Million Pound Note	9.509	10.157
	The Adventures of Tom Sawyer	10.804	

表3中提到的欧亨利的小说，根据已有的文献［19］研究，欧亨利的小说语言简洁优美，能够用最简单的语言把人物刻画得栩栩如生。他的小说并不对社会生活进行宏观的描述，而是通过他对生活细致入微的洞察力，在社会生活中选取素材，以小见大，通过不同的角度来折射那个五彩斑斓的社会[20]。因此欧亨利的小说在词汇丰富程度上只有8.572。

儿童文学主要的读者是儿童，因此这就决定了儿童文学不会用到太多丰富华丽的辞藻，这样才能方便儿童读者阅读。所以这类作品的词汇丰富程度偏低，这与信息熵反映出的现象也是相符合的，这也可以从一定程度上说明信息熵是可以反映词汇丰富程度的。

表4　儿童文学信息熵

Table 4　Entropy of children's literature

作者	书名	信息熵	平均值
刘易斯·卡罗尔	Alice's Adventures in Wonderland	9. 702	9. 702
罗伯特·斯蒂文森	Treasure Island	10. 194	10. 497
	New Arabian Nights	11. 276	
	Dr Jekyll and Mr Hyde	10. 021	
奥斯卡·王尔德	The Nightingale and the Rose	8. 359	9. 005
	The Happy Prince	8. 807	
	The Remarkable Rocket	9. 003	
	The Selfish Giant	7. 817	
	The Young King	9. 679	
	Birthday of The Infanta	9. 990	
	The Fisherman and His Soul	9. 756	
	The Star Child	9. 058	
	The Devoted Friend	8. 572	
克蕾西达·考威尔	How to train our dragon	10. 505	10. 520
	How to Speak Dragonese	10. 566	
	How to Cheat a Dragon's Curse	10. 484	
	How to Twist a Dragon's Tale	10. 595	
	How to be a pirate	10. 450	
E•B•怀特	Charlotte's Web	10. 086	10. 186
	Stuart Little	10. 132	
	The Trumpet of the Swan	10. 341	

3. 5　折线图分析

通过双曲正切函数将信息熵映射到−1 至+1 之间，将不同折线的差别拉大，使得分析更有说服力。

图 4 反映的是四类小说信息熵的范围以及差别。图中每一条曲线代表一类作品，每一个点代表一本作品。同一类作品用虚线连起来。绘图时使用 matplotlib 中的函数 ylim，将纵坐标范围限制在一定的范围之间，同时还使用了 plot 函数，对折线的颜色、标识等进行区分。另外还是用了 FontProperties 函数来设置图例样式，让图例能正确显示中文。图 4 同样是如此设置的。

大多数推理小说着重于对逻辑的叙述，让读者能够清晰地明白推理的逻辑和案件的过程，因此不同的小说对应的词汇丰富程度变化不会有特别大的起伏。从图 4 中可以看到其曲线是相对比较平稳的，与词汇的丰富程度的变化是相吻合的。这也侧面说明了信息熵可以一定程度反映词汇的丰富程度。

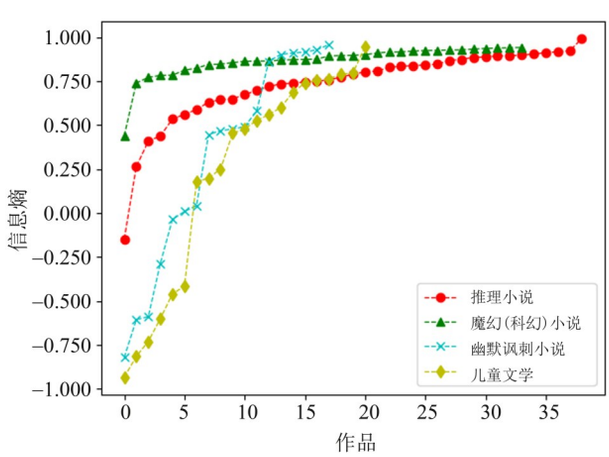

图 4　四类作品的信息熵比较分析

Fig. 4　Analysis of entropy in four kinds of works

魔幻/科幻小说在刻画时，更着重于描写刻画一个魔幻的世界，不同的魔幻/科幻小说所描写的魔幻世界多数是相似的，大多跟外太空，魔法世界相关，因此在词汇的丰富程度上不会有太大的变化。图 4 所反映出来的不同的魔幻/科幻小说的

信息熵的曲线也是相对平缓的，这与魔幻/科幻小说的词汇丰富程度的变化是相似的。这也可以从侧面反映出信息熵能够一定程度地反映词汇丰富程度。

图4中儿童文学和幽默讽刺小说的信息熵的曲线变化比较大。对于儿童文学，对于不同年龄层的读者，其词汇丰富程度也会有所不同。因此儿童文学的词汇丰富程度的变化会较大。幽默讽刺小说更多基于现实，在不同的时代，所描绘的内容也是不同的，因此幽默讽刺小说的词汇丰富程度的变化也会比较大。这与图4中儿童文学和幽默讽刺小说的信息熵的曲线变化是相吻合的。

从图4中可以看出，儿童文学的信息熵整体是最低的，而魔幻/科幻小说的信息熵整体是最高的，其他两类小说整体是落在两者之间的。

对于这四类小说根据刚才的分析以及从以往的阅读经验来说，儿童文学的可读性的确会较高，所用到的词汇比较简单，词汇的丰富程度不高，魔幻/科幻小说的可读性会差一些，词汇会比较复杂多变，词汇丰富程度会较高。这与图4中反映出了的情况是基本一致的，这可以说明信息熵的确可以反映词汇的丰富程度。

图5反映的是每一类小说中，不同作者的小说的信息熵的平均值，即表1到表4中的平均值一列的数据。每一条线代表不同类型的作品，每一个点代表一位作者，表示这位作者其所有小说作品的信息熵的平均值。图5是用于辅助图4，从平均信息熵的角度进行说明。

图5曲线的平缓程度与图4的是一致的，对其的解释也与图4类似。

从图5中也可以看出，儿童文学的信息熵平均值整体是最低的，而魔幻/科幻小说的信息熵平均值整体是最高的，其他两类小说整体是落在两者之间的。

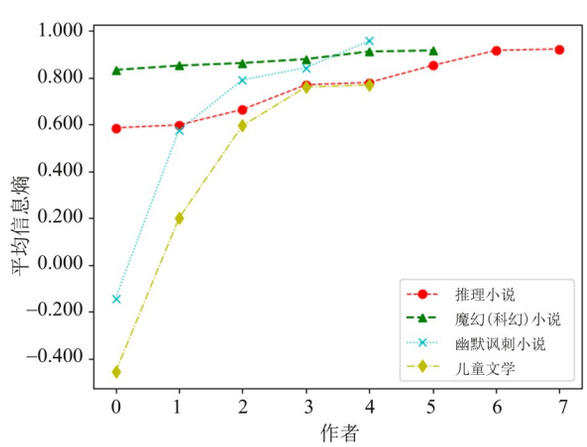

图5 四类作品不同作者平均信息熵比较分析

Fig. 5 Analysis of average entropy in four kinds of works

图5通过信息熵的平均值再次说明了，信息熵这一参数的确是可以在一定程度上反映词汇丰富程度的。

根据上述的结果以及分析，信息熵的变化趋势与小说的词汇丰富程度的变化趋势大致上是相吻合的。因此可以得出结论，信息熵这一个指标的确是可以反映词汇的丰富程度的。

4 假设检验验证差异

在验证样本和样本之间是否有差别的时候，通常会考虑假设检验的方法。在本文的第3节中已经就实验得到的图表进行了分析，从中的确能够看到不同的小说类型的信息熵是有差别的。

为了更加科学地分析其差异，本文在这里引入了假设检验，对每一类小说的均值进行假设检验，从而验证不同类型小说信息熵的差异性。

4.1 假设检验介绍

假设检验又称统计假设检验，常用的检验又 Z 检验，t 检验等等。假设检验利用的原理是小概率事件在一次实验中基本不会发生。假设检验的步骤大致可以分为以下几步。首先根据问题，提出原假设 H_0，以及备选假设 H_1。接下来给定显著性水平 α 和样本容量，确定检验统计量以及拒绝域。最后根据样本做出决策。

本文中的假设检验都是基于样本分布服从正态分布这一先验知识，并且假设不同类型的小说的方差均不同。

4.2 假设检验验证结论

根据得到的信息熵的数据，我们计算出每一种类型的小说的均值与方差，得到的结果如表5所示。本次假设检验要分别验证魔幻科幻小说的信息熵的均值是要比其他三类的小说的要高的，并且对于推理小说和幽默讽刺小说在一定的显著性水平下，它们的信息熵均值是要比儿童文学的要高的。

首先我们验证魔幻科幻小说的信息熵的均值是要比其他三类的要高的。以下我们以魔幻科幻小说和儿童文学为例进行阐述。

设 μ_1 表示魔幻科幻小说的信息熵的均值，μ_2 表示儿童文学的信息熵的均值，原假设为 H_0，H_1 为备选假设。这里采用的是 Z 检验，设统计量为 Z，

表5　不同类型小说信息熵均值与方差

Table 5　The mean and varience of entropy of different novels

小说类型	信息熵均值	信息熵方差	样本容量
魔幻科幻小说	10.886	0.232	34
推理小说	10.574	0.801	39
幽默讽刺小说	10.019	1.235	18
儿童文学	9.781	1.471	21

设拒绝域为 W。

$$H_0:\mu_1 \leqslant \mu_2 \tag{7}$$

$$H_1:\mu_1 > \mu_2 \tag{8}$$

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \tag{9}$$

$$W = \{Z \geqslant Z_\alpha\} \tag{10}$$

式（9）中 n_1 和 n_2 表示样本容量 X 和 Y 表示样本的均值，σ_1 和 σ_2 分别表示样本的标准差。将表5中的样本容量，均值和方差分别代入式（9），得到检验统计量的值为3.985。在显著性水平 α 为0.05的条件下，通过查表可得，Z_α=1.645，由于3.985大于1.645，因此我们有理由拒绝原假设。所以我们有理由相信，魔幻科幻小说的信息熵的均值是大于儿童文学的。

同样我们按照上述过程，分别检验魔幻科幻小说的信息熵的均值比推理小说的和幽默讽刺小说的要大。我们采用相同的检验统计量，得到的统计量的值分别为1.886和3.156，在显著性水平为0.05的条件下，我们能得到以下结论，检验魔幻科幻小说的信息熵的均值比推理小说的和幽默讽刺小说的要大。

类似地，我们也对推理小说的信息熵均值和儿童文学的信息熵均值进行假设检验。

我们采用相同的统计量，得到的统计量的值为2.635，同样在显著性水平为0.05的条件下，也能够说明推理小说的信息熵均值要比儿童文学的信息熵均值大。

从第3节的图表中可以看到，幽默讽刺小说的信息熵和儿童文学的信息熵有一定的差异，但差异没有那么明显，这里也利用假设检验进行说明。

我们同样也是对两种类型小说信息熵的均值进行假设检验，设 μ_1 表示幽默讽刺小说的信息熵的均值，μ_2 表示儿童文学的信息熵的均值。原假设和备选假设同式（7）~（8），采用同样的检验统计量。

经过计算得到的统计量的值为0.639，在显著性水平为0.27的条件下能够拒绝原假设。从这里也能够看出这两类小说的信息熵有一定的差异，但差异不明显。

上述的假设检验进一步验证了从图表中得到的结论，魔幻科幻小说的信息熵与其他三类小说的信息熵是有差异的，而且从总体的角度上魔幻科幻小说的信息熵看是要比其他三类的小说的要大的。同时儿童文学的信息熵从总体上看，是要比其他三类的小说的要低的。

5　信息熵在文本分类中的应用

信息熵这一个指标比较简单，因此具有可延展性。在使用长短期记忆神经网络（Long-Short Term Memory）或是循环神经网络（Recurrent Neural Network）[22-23] 对文本进行分类时，如果考虑使用词汇丰富程度作为一个特征，那么可以用直接使用信息熵，把信息熵作为表示词汇丰富程度的一个特征。在用其他的如BP神经网络 [24] 也可以考虑用信息熵作为一个输入。在使用一些简单的分类器，如逻辑回归 [25]，Fisher线性判别 [26]，支持向量机 [27-28] 等进行分类的时候，使用信息熵可以作为分类的一个较好的依据。对于聚类分析，如K均值聚类算法 [29-30]，也可将信息熵纳入考量。

6　结　语

本文通过对不同小说的信息熵的分析，探究信息熵与小说词汇丰富程度的关系。得到了信息熵可以反映小说的词汇丰富程度的这一结论，并提出信息熵在文本分类算法中的一些可能性的应用。

在以往的对文学作品的研究中，很少有基于信息熵这一个特点进行研究的。本文通过信息熵这一个指标，可以很简洁地反映出小说作品的词汇丰富程度这一个特点。信息熵的计算简单但反映的能力很强。相较于前面提到的词汇独特性，

词汇密度，词汇多样性（TTR），计算上更简单，反映能力也更强，能够更好地消除文本长度对于词汇丰富程度的影响。

参考文献：

[1] 郝晓丽,高永. CUDA框架下的视频关键帧互信息熵多级提取算法[J]. 电子科技大学学报，2018，47(5)：726－732.
HAO X L，GAO Y. Multi-level extraction algorithm of video key frame mutual information entropy under CUDA framework [J]. Journal of University of Electronic Science and Technology，2018，47(5)：726-732.

[2] 续拓,李洁,王颖. 叠加信息熵游走数据聚类算法[J]. 西安电子科技大学学报(自然科学版)，2018，45(4)：75-79.
XU T，LI J，WANG Y. Superimposed Information entropy walking data clustering algorithm [J]. Journal of Xidian University(Natural Science)，2018，45(4)：75-79.

[3] 宋勇,蔡志平. 一种基于信息论模型的入侵检测特征提取方法[J]. 电子科技大学学报，2018，47(2)：267-271.
SONG Y，CAI Z P. An intrusion detection feature extraction method based on information theory model [J]. Journal of University of Electronic Science and Technology，2018，47(2)：267-271.

[4] 郑碧如,吴广潮. 基于信息论方法的分类数据相似性度量[J]. 计算机与现代化，2018(5)：30-34.
ZHENG B R，WU G C. Similarity measure of classification data based on information theory[J]. Computer and Modernization，2018(5)：30-34.

[5] 刘颖，肖天久. 金庸与古龙小说计量风格学研究[J]. 清华大学学报(哲学社会科学版)，2014，29(5)：135-147.
LIU Y，XIAO T J. Study on the metrology and style of Jin Yong and Gu Long's novels [J]. Journal of Tsinghua University：(Philosophy and Social Sciences)，2014，29(5)：135-147.

[6] 贺湘情,刘颖. 基于文本聚类的语言韵律和节奏风格特征挖掘[J]. 中文信息学报，2014，28(6)：194-207.
HE X Q，LIU Y. Language rhythm and rhythm style feature mining based on text clustering [J]. Chinese Journal of Information，2014，28(6)：194-207.

[7] 李艳丽,李宛蓉,廖欣,等. 基于计量风格学的小说质量分析[J]. 计算机与现代化,2019，285(5)：23-28.
LI Y L，LI W R，LIAO X，et. al. Stylometry-based analysis of literature texts [J]. Computer and Modernization，2019，285(5)：23-28.

[8] STRUNK W Jr,WHITE E B. The elements of style[M]，Fourth Edition,MA：Allyn & Bacon,2000.

[9] 时季. 聚类分析方法在文学作品风格比较中的应用—以毕飞宇、苏童小说的比较分析为例[J]. 文教资料,2017,773(33)：19-22.
SHI J. The application of cluster analysis method in the comparison of literary styles-A comparative analysis of Bi Feiyu and Su Tong's novels[J]. Data of Culture and Education，2017，773(33)：19-22.

[10] 刘颖,肖天久.《红楼梦》计量风格学研究[J]. 红楼梦学刊，2014(4)：260-281.
LIU Y，XIAO T J. Research on the measurement style of "a Dream of Red Mansions" [J]. Studies on "a Dream of Red Mansions"，2014(4)：260-281.

[11] 金迪. 基于语料库的格非、余华小说计量风格学研究[D]. 南京:南京师范大学,2018.
JIN D. A corpus-based study on geometry and style of Ge Fei and Yu Hua's novels [D]. Nan Jing：Nanjing Normal University，2018.

[12] 田宝玉,杨洁,贺志强,等. 基础信息论[M]. 2版. 北京:人民邮电出版社，2008：18-26.

[13] COVER T M，THOMAS J A. Elements of information theory [M]. Second Edition，Beijing：Mechanical Industry Press，2008：13-16.

[14] SHANNON C E. A mathematical theory of communication [J]. The Bell System Technical Journal. 1948，27：379-423，623-656.

[15] 曹雪虹. 信息论与编码[M]. 2版. 北京:清华大学出版社，2009：6-15.

[16] 盛骤,谢式千,潘承毅. 概率论与数理统计[M]. 4版. 北京:高等教育出版社，2008：19-121.

[17] PERKINS J. Python 3 text processing with NLTK 3 codebook [DB/OL]. http://www. allitebooks. com/py-

thon-3-text-processing-with-nltk-3-codebook, August 2014.

[18] 鞠训科. 魔幻主义小说初探——以《魔戒》和《哈利·波特》为个案[J]. 广西教育学院学报,2007,2:121-123.
JU X K. A preliminary study of magical novels: a case study of The Lord of the Rings and Harry Potter[J]. Journal of Guangxi College of Education, 2007, 2: 121-123.

[19] 王军礼. 论欧·亨利小说的语言表达艺术[J]. 语文建设, 2016(23): 85-86.
WANG J L. The language expression art of O. Henry's novels[J]. Language Construction, 2016(23): 85-86.

[20] 陈丽. 欧·亨利小说的语言艺术探析[J]. 湖南科技学院学报, 2017, 38(07): 33-34.
CHEN L. An analysis of the language art of O Henry's novels[J]. Journal of Hunan Institute of Science and Technology, 2017, 38(7): 33-34.

[21] 徐海云.《忠实的朋友》语料库与文学的结合[J]. 海外英语, 2015(15): 177-179.
XU H Y. Corpus and literature of The Devoted Friend [J]. Overseas English, 2015(15): 177-179.

[22] GRAVES A. Supervised sequence labelling with recurrent neural networks[D]. München: Technische Universität München, 2006.

[23] 杨丽,吴雨茜,王俊丽,等. 循环神经网络研究综述[J]. 计算机应用, 2018,3 8(S2): 1-6,26.
YANG L, WU Y X, WANG J L, et al. A review of research on recurrent neural networks[J]. Computer Application, 2018, 38(S2): 1-6,26.

[24] 李友坤. BP神经网络的研究分析及改进应用[D]. 安徽:安徽理工大学,2012.
LI Y K. Research and analysis of BP neural network and its application[D]. Anhui: Anhui University of Science and Technology, 2012.

[25] 尹建杰. Logistic 回归模型分析综述及应用研究[D]. 哈尔滨: 黑龙江大学, 2011.
YIN J J. Summary and applied research of Logistic regression model analysis [D]. Harbin: Heilongjiang University, 2011.

[26] ABUZEINA D, AL-ANZI F S. Employing fisher discriminant analysis for Arabic text classification [J]. Computers & Electrical Engineering, 2017: S0045790617334845.

[27] 张浩然,韩正之,李昌刚. 支持向量机[J]. 计算机科学,2002, 29(12): 135-137.
ZHANG H R, HAN Z Z, LI C G. Supported vector machine [J]. Computer Science, 2002, 29 (12): 135-137.

[28] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程,2004, 30(10): 6-9.
QI X N. A survey of support vector machines and their applications [J]. Computer Engineering, 2004, 30 (10): 6-9.

[29] 王千,王成,冯振元,等. K-means聚类算法研究综述[J]. 电子设计工程,2012, 20(7): 21-24.
WANG Q, WANG C, FENG Z Y, et al. A survey of K-means clustering algorithms [J]. Electronic Design Engineering, 2012, 20(7): 21-24.

[30] 吴夙慧,成颖,郑彦宁,等. K-means算法研究综述[J]. 数据分析与知识发现,2011, 27(5): 28-35.
WU S F, CHENG Y, ZHENG Y N, et al. A survey of K-means algorithm research [J]. Data Analysis and Knowledge Discovery, 2011, 27(5): 28-35.

（责任编辑　冯兆永）