文章编号：1003-0077(2019)12-0129-06

先秦诸家学派的相关系数与特征词研究

马创新[1]，梁社会[2]，陈小荷[3]

（1. 江苏师范大学 语言科学与艺术学院，江苏 徐州 221009；
2. 南京师范大学 国际文化教育学院，江苏 南京 210097；
3. 南京师范大学 文学院，江苏 南京 210097）

摘　要：为了发现先秦诸家学派之间的相关度，找出能够代表各学派主题特征的特征词，该文首次对诸家学派之间的相关关系作量化考察，对诸家思想的主题特征作统计分析。通过研究发现，儒家与道家之间的相关度最高，兵家与墨家之间的相关度最低，道家与其他各学派之间的相关系数的均值最大。该文还通过分析特定学派中各个词型与其他各学派中相同词型的等级之间差额大小，筛选出能够代表学派主题的特征词。

关键词：数字人文；诸子百家；相关度；主题

中图分类号：TP391　　**文献标识码**：A

Study on the Correlation Coefficient and Characteristic Words of the Pre-Qin Schools

MA Chuangxin[1], LIANG Shehui[2], CHEN Xiaohe[3]

(1. Linguistic Sciences and Arts School, Jiangsu Normal University, Xuzhou, Jiangsu 221009, China;
2. International College for Chinese Studies, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
3. College of Liberal Arts, Nanjing Normal University, Nanjing, Jiangsu 210097, China)

Abstract: In order to determine the correlation between the pre-Qin schools, to detect the characteristic words that can represent the theme characteristics of each school, this paper makes a quantitative investigation of the relations among the various schools, together with the subject characteristics of the various ideas of each school. It is revealed that the correlation between the school of Confucianism and the school of Taoism is the highest, the correlation between the school of Soldiers and the school of Mo is the lowest, and the mean value of the correlation between Taoism and other schools ranks at the top. This paper also selects the characteristic words which can represent the subject of the school by analyzing the difference in the rank of word types between schools.

Keywords: digital humanities; all classes of authors; relevance; topics

0　引言

中华文化博大精深、源远流长，先秦时期，老子、孔子、韩非子、孙子、墨子等诸子开创了中国历史上第一次学术文化的繁盛时期。在此阶段，出现了道家、儒家、法家、兵家、墨家等学派。这些学派的学术思想是中华民族精神文明的结晶，它们对后世的政治、经济、文化等各个领域都产生了深远的影响[1]。

历代学者都很重视对于诸子百家学术思想和学术渊源的研究。儒家希望通过人生修养、克已复礼、实行仁政，以实现博施济众、老安少怀的理想社会[2]，据《淮南子·要略》，墨子本来是学习儒家思想的，因不满于儒家主张而创立墨家，提出了“天志”、“明鬼”、“非攻”、“节用”、“节葬”、“兼爱”等主张，反对儒家思想中的一些主张[3]。法家倡导在政治中推行法治，不同与儒家主张的德治和礼治，但儒法两家都强调人治，这又不同于道家的“自然无为而治”。儒墨道学派的共同点是依据他们各自的人生理想发展出人生修养理论，目标就是创造一个和平的社会

收稿日期：2019-05-06　　**定稿日期**：2019-07-02

基金项目：江苏省社会科学基金(15YYC001)；国家社会科学基金(15BYY096)

环境[4]。兵家思想与老子《道德经》之间在军事思想方面也具有相同之处，有些学者甚至把《道德经》视为兵书。道儒法墨兵诸家思想之间有差异之处，但他们的最终目标都是在寻找救世治世的良方[5]。

如今信息时代，计算机技术、信息技术与人文研究融合而形成了一个新的研究领域，即“数字人文”研究，数字人文的出现对人文学术的研究方法与过程产生深远影响[6]。以往学界对于先秦诸家思想的研究仅限于定性的评价和讨论，那么在数字人文兴起的今天，能否利用计算语言学方法，对先秦诸家学派之间的相关关系作量化考察，对诸家思想的主题内容作统计分析呢？本文首次提出通过计算不同学派之间在高频词型等级方面的相关系数来获得学派之间的相关系数，以评价学派的影响力和学派之间的亲疏远近，通过计算学派之间在词型等级上的差异度来获取学派的主题特征词[7]。

1 先秦诸家学派的相关度分析

1.1 实验语料和实验方法

为了分析先秦诸家学派的相关度并且筛选出各学派的特征词，我们分别从儒家、道家、法家、兵家、墨家等学派的作品中选取出具有代表性的文献作为实验语料。本文选取的儒家代表文献是《论语》和《孟子》，道家代表文献是《老子》和《庄子》，法家代表文献是《韩非子》和《商君书》，兵家代表文献是《孙子》和《吴子》，墨家代表文献是《墨子》，由于墨家的文献比较少，所以只从墨家中选取一部代表性文献。

本研究所采用的实验方法是：

(1) 从儒家、道家、法家、兵家等每个学派选出代表性文献；

(2) 对每部文献的词型分别统计词频并且按照频次降序排列，然后使用并列法确定各部文献的词型等级并做等值化处理[8]；

(3) 筛选出每个学派的共有词型，并且取该词型的“转化等级”的均值作为“最终等级”；

(4) 使用斯皮尔曼等级相关系数公式，计算学派之间共有的高频词型等级序列的相关系数。

1.2 确定各部文献的词型等级

“词型等级”是按词型在文献中的出现频次(即词型的词例数)递减排序，把出现频次最高的词型等级定为1，次高的词型等级定为2，依次类推。但对于如何确定同频词型的等级，国内外学者提出最大值法、最小值法、平均值法、并列法四种方法[9]。本文使用并列法确定同频词型的等级，即把出现频次最高的词型等级定为1，次高的词型等级定为2，依次类推，频次相等的词型为一个等级，以其在语料中词频序值为等级。

通过分析表1发现由于各部文献的词型数、词例数差异都较大，使得各部文献的词型等级数差异明显，其中《韩非子》的最大词型等级为228，《吴子》的最大词型等级为47，两部文献的词型等级数量相差约五倍。

表1 先秦文献的等级数和等级系数

	词例数	词型数	等级数	等级系数
《论语》	14 608	1 622	90	1.11
《孟子》	32 079	2 723	125	0.80
《老子》	6 005	986	58	1.72
《庄子》	59 586	4 838	165	0.61
《韩非子》	98 415	4 640	228	0.44
《商君书》	19 559	1 399	105	0.95
《孙子》	6 620	814	58	1.72
《吴子》	4 595	896	47	2.13
《墨子》	73 779	3 920	203	0.49

这种情况就使得各部文献之间的共有词型难以做等级差异比较和加减运算。例如，《老子》和《庄子》同为道家文献，《老子》的词型等级数为58，《庄子》的词型等级数为165。“聖人”这个词型在《老子》中的词型等级为27，在《庄子》中的词型等级为70。单从词型等级的大小差异来看，“聖人”在《老子》应该比在《庄子》更常见，而实际上“聖人”在《老子》中出现33次，在《庄子》出现108次。

由于上述原因，我们提出要对各部文献中的词型等级做等值化处理。方法是给每部文献设定一个等级系数，特定文献中每个词型的等级都要乘以它的等级系数，从而将由并列法确定的“原始等级”转变为“转化等级”。各部文献的等级系数是不同的，特定文献的等级系数等于100除以该文献的最大词型等级。例如，《老子》中的最大词型等级为58，它的等级系数就约等于1.72；《庄子》中的最大词型等级为165，它的等级系数就约等于0.61。表1的第五列给出了每部文献的等级系数。

设定好等级系数之后，就容易比较各部文献共

有词型的等级差异，例如，“聖人”在《老子》中的原始等级为 27，乘以等级系数 1.72，转化等级为 46.44；“聖人”在《庄子》中的原始等级为 70，乘以它的等级系数 0.61，转化等级为 42.70。可见“聖人”在《老子》和《庄子》中的转化等级差异很小，在这两部文献中的重要性相差不大。表 2 中给出《老子》中出现频次排前 20 位的词型的出现频次、原始等级和转化等级。

1.3　确定各学派的共有词型和等级

我们统计了各个学派代表文献的共有词型数及共现率，以及共有词例数及共现率，如表 3 所示。例如，在《论语》和《孟子》中都出现的词型有 978 个，占《论语》和《孟子》总词型数（3 367 个）的 29%。《论语》和《孟子》总词型数等于“《论语》词型数＋《孟子》词型数－《论语》与《孟子》的共现词型数”，因共现词型在两部文献中都出现，故不能重复计算。这些共现词型在两部文献中共出现 41 118 次，占《论语》和《孟子》总词例数（46 686 次）的 88%。

通过分析表 3 发现，除了墨家之外，其他四个学派的词型共现率都比较低，在 14%至 42%之间，而词例共现率较高，在 78%至 89%之间，这说明共现词型的出现频次相对较多，大多属于高频词。

在同一流派内，我们以共现词型在两部文献中的转化等级的均值作为最终等级。表 4 选取儒家学派的 10 个共现词型，展示它们计算最终等级的方法。我们使用这种方法，计算出各个学派的共现词型的最终等级。

表 2　转化等级的计算方法示例（等级系数为 1.72）

词型	出现频次	原始等级	转化等级	词型	出现频次	原始等级	转化等级
之	276	1	2	為	125	6	10
不	252	2	3	者	110	7	12
其	157	3	5	也	103	8	14
以	150	4	7	無	100	9	16
而	139	5	9	有	88	10	17

表 3　先秦诸家学派的共有词型数、共有词例数及共现率

学派	代表文献一	词型数	词例数	代表文献二	词型数	词例数	共有词型数	词型共现率%	共有词例数	词例共现率%
儒家	《论语》	1 622	14 608	《孟子》	2 723	32 079	978	29	41 118	88
道家	《老子》	986	6 005	《庄子》	4 842	59 600	714	14	51 169	78
法家	《韩非子》	4 638	98 414	《商君书》	1 399	19 559	1 129	23	104 797	89
兵家	《孙子》	814	6 620	《吴子》	896	4 595	503	42	9 843	88
墨家	《墨子》	3 920	73 788				3 920	100	73 788	100

表 4　最终等级的计算方法示例

词型	《论语》中转化等级	《孟子》中转化等级	最终等级
曰	1	3	2
之	2	1	2
不	3	2	3
也	4	2	3
子	6	26	16
而	7	4	6
其	8	6	7
者	9	6	8
以	10	5	8
有	11	9	10

1.4　学派之间的相关度计算

学派之间相关度的计算方法采用“斯皮尔曼等级相关”系数，计算如式（1）所示。

$$R_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \tag{1}$$

其中，D_i 表示每一对词型相应的两个等级之差，n 表示样本数。

斯皮尔曼等级相关系数适用于研究数据是具有等级性质的成对数据。但是，两个学派出现的词型数据并不是成对的，所以采用这种计算方法所得到的相关系数是一个近似值。我们用 ARs 来表示“以学派 A 中特定数量词型为样本”与学派 B 中全部词

型比较所得到的相关系数，对于在学派A中出现而学派B中没有出现的词型，就假定该词型在学派B中的最终等级为101。同样，以BRs来表示“以学派B中特定数量词型为样本”与学派A中全部词型比较所得到的相关系数，对于在学派B中出现而学派A中没有出现的词型，就假定该词型在学派A中的最终等级为101。学派A与B的相关度用ABRs来表示，ABRs等于ARs与BRs的均值，即：ABRs=(ARs+BRs)/2。也就是说，学派A与B的相关度就等于“以学派A中特定数量词型为样本”与学派B的全部词型比较所得到的相关系数，加上“以学派B中特定数量词型为样本”与学派A的全部词型比较所得到的相关系数，两个系数之和再除以2所得到的商[10-11]。

我们选取各个学派中最终等级排在前120位的词型作为样本，计算各学派之间的相关系数。汇总相关数据，如表5所示。

表5 先秦各学派之间的相关系数

	儒家/%	道家/%	法家/%	兵家/%	墨家/%	均值/%
儒家	—	86.16	71.12	70.34	66.45	73.52
道家	86.16	—	76.85	76.58	65.99	76.40
法家	71.12	76.85	—	59.11	77.08	71.04
兵家	70.34	76.58	59.11	—	35.56	60.40
墨家	66.45	65.99	77.08	35.56	—	61.27
均值	73.52	76.40	71.04	60.40	61.27	—

通过分析表5，能够发现以下两点：

(1) 儒家和道家的相关系数最大，这是因为这两个学派的思想接近，都认同人性本善，提倡以道德为基础的治国理念，具有“民本”思想。兵家和墨家的相关系数最小，这是因为两学派的思想差异较大，墨家宣扬仁政，主张兼爱和非攻，而兵家文献主要谈论用兵之道，分析取得军事胜利的策略。

(2) 道家与其他各家学派相关系数的均值最大，这说明道家思想在先秦时期的影响力最强。在当时，儒家思想的影响力也弱于道家，其他各家学派都在较大程度上受到道家思想的影响。兵家思想与其他各家学派相关系数的均值最小，这是因为兵家思想毕竟是在极特殊的战争时期才使用的策略，其他各家思想对战争都持有谨慎态度。

2 先秦诸家学派的特征词研究

2.1 实验方法

先秦诸家学派的特征词研究所使用的语料，与各学派之间的相关度分析所使用的语料完全相同。诸家学派的特征词研究实验方法共分四步，其中的前三步与各学派之间相关度分析方法的前三步完全相同，即都是要先选取各学派代表作，然后统计各代表作的词频，并依此确定词型等级和做等值化处理，再确定各学派的共有词型和“最终等级”。

第四步是通过分析某学派各个词型与其他各学派中相同词型的等级之间差额大小，筛选出该学派特征系数较高的词型。

第五步是筛选掉其中的虚词，只保留特征系数较高的实词作为该学派的特征词。

2.2 各流派的特征词计算

我们通过计算学派中各个词型的特征系数，来筛选出学派中的特征词。特征系数的取值范围是正值、负值或零。计算词型特征系数的方法如式(2)所示。

$$D_j = \frac{\sum_{i=1}^{n} D_i}{n} \tag{2}$$

其中，D_j 表示某个词型在特定学派中的特征系数，D_i 表示“某词型在‘对比学派’中的最终等级”减去“该词型在特定学派中的最终等级”所得到的差[12]，n 表示“对比学派”的数量，在本研究中，n 的取值应该大于或等于1，小于或等于4。

本研究中共有五家学派，在对比时，我们是用一家学派与其他四家学派相比较，所以在本研究中 n 不会大于4。在用特定学派中的某个词型与其他多家学派的词型对比时，有时该词型不能在其他多家学派中也都出现。如果该词型在其他学派中都没有出现，就无法计算其特征系数，所以我们规定该词型必须在另外至少1家学派中出现，才把该词型放在特征词统计范围之内[13-14]。所以在本研究中，式(2)中的 n 的取值应该大于或等于1，小于或等于4。经过统计，我们发现各个学派中符合此条件的词型数量是：儒家882个、道家663个、法家1 016个、兵家489个、墨家1 377个。从表3中可以看到，儒家、道家、法家、兵家和墨家的共有词型数分别为978个、714个、1 129个、503个、3 920个，除墨家以

外，其他各家学派符合此条件的词型数均占各家共有词型数的 89％以上。

由式(2)可以看出：

(1) 词型的特征系数是与特定学派联系在一起的，是在特定学派中的特征系数，同一词型在不同学派中的特征系数是不同的。

(2) 当特定学派中某词型的特征系数为正值时，表示该词型在特定学派中所处的等级位置是比较靠前的，高于该词型在多家“对比学派”中的等级均值；当特征系数为负值时，表示该词型在特定学派中所处的等级位置是比较靠后的，低于该词型在多家“对比学派”中的等级均值；当特征系数为零时，表示该词型的在特定学派的等级值等同于该词型在多家“对比学派”中的等级均值。

使用该方法，计算各家学派中所有词型的特征系数，按照特征系数的大小降序排列，并且筛除虚词只保留实词。表 6 中列举出各学派特征系数最大的 30 个词型。

表 6　先秦各学派的特征词及其特征系数

学派	特征词及其特征系数
儒家	孔子(56)、仁(44)、子(42)、君子(40)、問(39)、禮(37)、由(32)、好(26)、夫子(26)、子路(25)、子貢(24)、邦(22)、聞(21)、學(20)、吾(20)、言(20)、見(19)、我(19)、友(19)、舜(18)、食(18)、何(18)、詩(18)、曰(17)、思(17)、志(17)、云(17)、求(16)、受(15)、小人(14)
道家	德(42)、聖人(40)、物(30)、生(29)、始(26)、天下(26)、大(26)、常(23)、名(22)、化(21)、道(21)、形(20)、終(20)、成(19)、天(16)、真(15)、知(15)、身(14)、心(14)、窮(14)、復(14)、失(13)、觀(13)、和(11)、靜(11)、精(10)、果(10)、吾(9)、柔(9)、虚(9)
法家	主(62)、法(61)、姦(53)、私(52)、力(51)、重(50)、官(50)、功(48)、刑(48)、明(46)、臣(46)、治(44)、爵(43)、國(41)、賞(41)、農(40)、令(40)、王(39)、勢(39)、亂(38)、強(38)、兵(38)、弱(37)、民(35)、禁(35)、戰(34)、亡(34)、秦(34)、世(33)、輕(32)
兵家	軍(55)、戰(52)、敵(47)、兵(46)、擊(43)、勝(32)、地(25)、進(25)、眾(24)、起(23)、卒(21)、間(20)、陳(20)、險(15)、水(15)、凡(13)、車(13)、將(13)、旌(11)、動(11)、輕(10)、退(10)、山(10)、形(9)、右(9)、絕(8)、機(7)、備(7)、合(7)、先(7)
墨家	尺(76)、城(63)、愛(57)、兼(54)、天(54)、鬼(54)、義(54)、政(53)、厚(52)、步(51)、長(50)、二(48)、中(47)、上(46)、當(46)、利(45)、攻(45)、說(44)、守(44)、意(43)、治(41)、令(41)、家(40)、同(40)、命(40)、天下(40)、非(39)、百姓(38)、從(38)、子(37)

通过分析表 6，能够发现以下两点：

(1) 使用本方法所筛选出的各学派特征词与各学派的思想主题是相符的，能够代表各学派的思想特征。例如，儒家学派的特征词有“孔子”“仁”“君子”“問”“禮”“學”等，这与儒家主题特征是相符的[15]；道家的特征词有“德”“聖人”“物”“生”“始”“天下”“道”等，与道家主题特征是相符的；法家学派的特征词有“主”“法”“私”“官”“刑”“明”“治”“令”“賞”等，与法家的主题特征相符[16]；兵家的特征词有“軍”“戰”“敵”“兵”“擊”“勝”“地”“進”等，与兵家的主题特征相符[17]；墨家的特征词有“尺”“城”“愛”“兼”“天”“鬼”“義”等等，与墨家的主题特征相符[18]。

(2) 由以上分析发现，我们所提出的特征词计算方法是完全可行的，可以在同类研究中推广使用。

3　结语

先秦诸家思想对后世影响深远，历来研究者众多，但以往的研究全都是对诸家思想的异同作定性地分析和评论。本文首次采用计算语言学方法对先秦诸家思想的异同做定量的统计和比较，先分别找出各家学派中的共有词型序列，然后计算各学派高频词型等级之间斯皮尔曼等级相关系数，经过数据统计之后发现儒道两学派之间的相关度最高，兵墨两学派之间的相关度最低；道家与其他各学派之间的相关系数的均值最大，说明先秦时期道家对其他学派的影响力最大。本文还通过计算各学派词型等级之间的差异度，来获取各个学派的主题特征词，这些特征词能够反映出各个学派的主要思想特征。

参考文献

[1] 陈小荷,冯敏萱,徐润华. 先秦文献信息处理[M]. 北京:世界图书出版公司,2013:5-8.

[2] 王启发. 荀子与儒墨道法名诸家[J]. 中国史研究,2000(3):40-57.

[3] 洪恩赐. 论先秦儒墨道人生哲学的终极理想[J]. 江西社会科学,2001(11):5-8.

[4] 王成儒. 论先秦儒墨道法之间的边际思想[J]. 学术论坛,1991(3):46-51.

[5] 赵小雷. 法家与兵家[J]. 西北大学学报(哲学社会科学版),2009,39(1):44-49.

[6] 柯平,宫平. 数字人文研究演化路径与热点领域分析[J]. 中国图书馆学报,2016(6):13-30.

[7] 马创新,梁社会. 面向语言分析的语料库技术平台建设[J]. 智能计算机与应用,2019,9(4):100-103.

[8] Zipf,G. K. Selected Studies of the Principle of Relative Frequency in Language[M]. Cambridge: Harvard University Press,1932: 2-8.

[9] Zipf,G. K. Human Behavior and the Principle of Least Effort[M],Cambridge: Addison-Wesley,1949: 25-30.

[10] 马创新,陈小荷. 文献中的词语分布、词型等级和风格计算[J]. 中文信息学报,2017,31(4):20-27.

[11] 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.

[12] 马创新,陈小荷. 文献中的词型分区规律与高频特征词的发现[J]. 语言文字应用,2018(3):124-133.

[13] 冯志伟. 用计量方法研究语言[J]. 外语教学与研究,2012,44(2):256-269.

[14] 靖继鹏,马费成,张向先. 情报科学理论[M]. 北京:科学出版社,2009:35-39.

[15] 胡泽. 墨、道、法三家学说与儒家学说渊源关系释证[J]. 中北大学学报(社会科学版),2014(6):5-8.

[16] 蒋重跃. 试论道法两家历史观的异同[J]. 文史哲,2004(4):73-80.

[17] 张文儒. 中国兵家与儒、道、法各家的兼容与互补[J]. 江汉论坛,1998(6):9-13.

[18] 李存山. 诸子百家与儒道佛三教的社会文化功能[J]. 中国哲学史,1998(1):21-26.

马创新(1980—),通信作者,博士,讲师,主要研究领域为计算语言学、信息计量学。
E-mail: machxin@126.com

陈小荷(1952—),博士,教授,博士生导师,主要研究领域为计算语言学。
E-mail: chenxiaohe5209@126.com

梁社会(1979—),博士,副教授,硕士生导师,主要研究领域为计算语言学。
E-mail: liangshehui@njnu.edu.cn