·信息管理·

数字技术下《老子》文本与先秦两汉典籍的关系挖掘*

高瑞卿[1] 董启文[1] 方 达[2] 王弘治[3] 方 勇[2]

(1. 华东师范大学数据科学与工程学院 上海 200062;
2. 华东师范大学中文系 上海 200062;
3. 上海师范大学人文学院 上海 200234)

摘 要:[**目的/意义**]理解老子思想关乎理解中国早期文化,结合数字人文的方法,开展实证研究。利用大数据计算的方式,通过定量统计、定性分析,解决老子研究领域长期存在的疑而难决的源头、影响等方面的问题,发掘依靠阅读经验难以发现的文本组织特征及相互关系。[**方法/过程**]统计河上公版《老子》语料的字频;进行相似度分析和典籍引用情况分析;最后训练出古汉语的BERT模型,利用生成的字嵌入计算典籍句子之间的相似程度,在《老子》之前的典籍上进行相关性研究。[**结果/结论**]使用TF-IDF进行文本向量化,得出《老子》与其后世的作品中的《淮南子》最为相似;使用BERT模型的自监督学习训练,达到在完形填空任务上52.11%的精度和在预测是否是下一个句子上98.45%的精度,相似度计算结果显示出《墨子》与《老子》密切相关。这种方法引起了我们对《老子》和《墨子》间论说思想关系的一番新思考。

关键词:BERT;数字人文;相似度;关系挖掘;先秦;老子

中图分类号:TP393;G251 **文献标识码**:A **文章编号**:1002-1965(2021)10-0099-09

引用格式:高瑞卿,董启文,方 达,等.数字技术下《老子》文本与先秦两汉典籍的关系挖掘[J].情报杂志,2021,40(10):99-107.

Research on the Relationship Between the Text of "Laozi" and the Classics of the Pre-Qin and Han Dynasties Based on Digital Technologies

Gao Ruiqing[1] Dong Qiwen[1] Fang Da[2] Wang Hongzhi[3] Fang Yong[2]

(1. School of Data Science and Engineering, East China Normal University, Shanghai 200062;
2. Department of Chinese Language and Literature, East China Normal University, Shanghai 200062;
3. School of Humanities, Shanghai Normal University, Shanghai 200234)

Abstract:[**Purpose/Significance**]Understanding the Laozi's thoughts relates to comprehend the early culture of Chinese. In this study, digital humanities methods were applied to empirical research. By using the method of big data calculation, including quantitative statistics and qualitative analysis, many long-standing problems in the field of Laozi's research were deeply explored, such as the source, influences and other aspects of difficulties, mainly about the text organization characteristics and interrelationships which are difficult to find by read-

收稿日期:2021-02-23 修回日期:2021-04-02

基金项目:国家社会科学重大基金项目"中国诸子学通史"(编号:19ZDA244)研究成果之一;国家社会科学基金项目"《经典释文》音义辞典"(编号:19FYYB008)研究成果之一;华东师大幸福之花先导基金重大研究专项"'幸福之花'先导研究基金项目——大数据视野下的老子思想源头与涵义研究"(编号:44300-19312-542500/005)的研究成果之一。

作者简介:高瑞卿,女,1997年生,硕士,研究方向:自然语言处理和文本挖掘;董启文,男,1977年生,博士,教授,研究方向:数据科学应用技术、包括网络信息学、机器学习和计算广告等;方 达,男,1987年生,博士,助理研究员,研究方向:诸子学研究;王弘治,男,1977年生,博士,副教授,研究方向:汉语史;方 勇,男,1956年生,博士,教授,研究方向:诸子学研究。

通信作者:王弘治

ing. [**Method/Process**] The word frequencies were counted on the "Laozi" corpus of Heshanggong's version. Similarity analysis were conducted and the citation of classics were analyzed. The BERT model were trained on ancient Chinese, and the generated word embeddings were used to calculate the similarity between classic sentences. [**Result/Conclusion**] By using TF-IDF for text vectorization, we found that "Huainanzi" is the most similar work with "Laozi" among its later works. By training the self-supervised learning model, BERT, a model whose accuracy reached 52.11% on the cloze task and 98.45% on predicting whether it's the next sentence task was got. The result of similarity calculation indicates the close relevance of "Laozi" and "Mozi". The proposed method could help us to rethink about the theoretical and ideological relationship between "Laozi" and "Mozi".

Key words: BERT; digital humanities; similarity; relationship mining; Pre-Qin; Laozi

《老子》是春秋时期老子(李耳)的哲学作品,又称《道德真经》《道德经》《五千言》《老子五千文》,是中国古代先秦诸子分家前的一部著作,是道家哲学思想的重要来源。老子思想对中国现代文明建设[1]、当代教育[2]、生态幸福观的建构[3]、我国行政管理建设[4],甚至对日本近现代名家[5]都产生了深刻影响。据联合国教科文组织统计,《老子》一书是除《圣经》外被译成外国文字发布量最多的文化名著。

从古至今解老注老者很多,在老学思想文化史上较有影响的,从最早的列子、庄子、文子、稷下黄老学的宗老,韩非子解老喻老,河上公、严君平、王弼诸家注老,唐代傅奕,宋代王安石、苏辙、吕惠卿,明末王夫之,元代吴澄等,直到晚清经世学者魏源,无不发表心得,增益老学,为后人探究老子思想留下了丰厚的文化资源[6]。因而《老子》版本众多,历史上比较流行且重要的版本有:马王堆帛书(甲本为 5344 字。乙本为 5342 字,外加重文 124 字);今本,河上公《道德经章句》(5201 字,外加重文 94 字),王弼《老子道德经注》(5162 字,外加重文 106 字),傅奕《道德经古本》(5450 字,外加重文 106 字)。

老子思想是中国文化早期发展的一个典型,如何理解其思想,关乎如何理解中国早期文化。关于老子的起源、内涵及影响问题历史上以及当代已经有了大量的研究成果。本研究主要针对利用大数据计算的方式,对先秦老子思想的源头及涵义进行实证性研究并得出相应结论,利用文本分析技术、人文数字处理手段,通过定量统计、定性分析,尝试解决老子研究领域长期存在的疑而难决的问题,发掘依靠阅读经验难以发现的文本组织特征及相互关系;在完成上述研究的同时,获得古汉语语料训练的 BERT 预训练语言模型并开源在 GitHub 上(https://github.com/RuiqingGao/ancient_Chinese.git)。

1 古文信息处理进展

从概念的内涵和外延上看,古文信息处理是一个交叉的研究领域,涉及了数学、计算机科学、语言学和图书情报学等多个学科的理论、方法、知识和技术[7]。近年来,古籍自动录入技术令古籍数字化工作取得了丰富的实际成果[8]。在此基础上,众多学者在古代典籍的自动分词[9-11],命名实体自动识别[12-15],句子或术语对齐[16-17]研究,古文断句研究[18-20]以及文本分类[21]方面取得了一系列研究成果。

从研究方法上看,仅在统计层面上有简单的词频统计[22],互文度量[23]等方法;更深入地,N-gram 模型[24],贝叶斯模型[9],支持向量机[10],条件随机场等传统机器学习模型被大规模应用在对古文的信息挖掘中,这与关于古代汉语文本的信息处理研究集中在分词、命名实体识别的任务上有密切关系。近些年来,随着运算力的快速提升和深度学习的再度兴起,循环神经网络[19,20]、BERT[18]模型也被应用到古文断句中。

从研究对象上看,先秦古汉语典籍是研究所用的主要语料,特别是早些年,大部分研究学者仅用少量典籍划分出训练集和测试集进行模型训练与应用;近年来,深度学习模型在各种人工智能研究领域取得了显著的成功后,研究者使用的语料才丰富起来,在一个研究项目中就涉及到史藏、诗藏、儒藏、集藏和子藏等文献。在研究对象内容上,早期应用传统机器学习模型时,使用的语料较为单一,但在近些年,研究所用语料涉及众多方面,包括地理,历史,小说,宗教,物产方方面面,以混合语料研究居多,以追求大而通用的模型。

尽管关乎古代文献典籍的研究已经有丰富的研究成果,但是从人文学科研究的深度来看,上述的数据化处理还停留在较为基本与浅显的层面,并不能深入到思想的阐释领域,而思想研究与阐释的主观性也正是人文学科的关键所在。从人文学科研究的广度来看,前述的古文信息处理呈现出散点状,没有完整的体系,缺乏不同典籍在时间维度上关联与逐步发展的研究。

总而言之,如何利用大数据得出具有洞见性的结论,还需要进一步的探讨。本文所开展的研究,力求找出不同典籍之间的联系。

2 研究方法

2.1 研究框架 本研究以大数据计算为依托,以《老子》典籍为研究对象,通过数字人文技术,考察了老子思想与其他典籍之间的相似程度与关系。所使用的技术手段如图 1 所示。

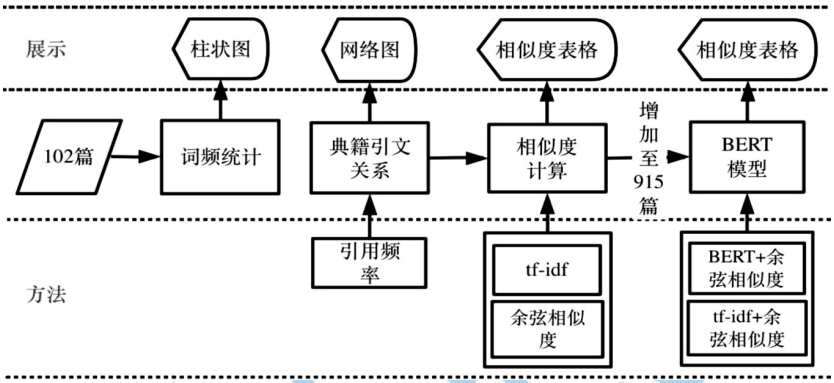

图 1　本文研究框架

首先,选取河上公《道德经章句》这一版本作为本研究所用《老子》语料,下文提到的《老子》均为这一版本。河上公本大致成书于两汉之际[25],文句简古,近民间系统,是历史中流传最广的《老子》版本之一,与另一流传广泛的王弼《老子道德经注》相比,一些学者认为河上公注本有的地方胜过了王弼注本,保存不少精义。

研究过程中以字为单位统计老子语料中的语词出现状况,从宏观上了解《老子》的用字特点;统计典籍之间的引用情况并绘出网络图;精选出先秦到东汉之间的典籍,基于统计进行相似度分析;使用先进的深度学习方法,训练 BERT 模型,按照 BERT 模型生成的字嵌入计算典籍内部句子之间的相似程度,并与使用 TF-IDF 的方法作为对比,将《老子》中所有句子和其他典籍中句子的相似度计算情况汇总,得出创作于《老子》之前的、与其最相似的典籍。本研究以《老子》为着力点,分别研究了其对后世的影响和之前作品与《老子》的关系,广泛研究了先秦两汉时期的典籍特征,力图从数据角度为人文学科提供实证。

2.2　**语料库介绍**　本研究使用 102 篇典籍所构成的语料库,这些典籍涵盖了先秦诸子百家的法、道、墨、儒、兵等学术派别的代表作品;作品的时代性大致包含了传统六艺经典,一般战国典籍,从战国到西汉的过渡性质的典籍,确定的西汉典籍,东汉典籍,汉后典籍(有引用早期内容),共计 8686320 字。

之后在训练 BERT 模型时,本研究将语料库范围扩大到 915 部作品,加入了三国两晋南北朝时期的汉达古籍。考虑到先秦两汉时期流传的作品较少,只是中国古代文学的一小部分,因此加入先秦两汉之后朝代的语料,可以更加有效地训练字向量,能足够开展文本数字分析。

2.3　**关键技术**

2.3.1　统计分析　词频分析是对文章中重要词汇出现的次数进行统计与分析,是文本挖掘的重要手段,是文献计量学中传统的、具有代表性的一种内容分析方法。在分析《老子》内容环节,本研究首先想到的方法就是基于字的层面,从宏观上查看老子的内容情况,按照《老子》中每个字出现的次数统计后,以频次从大到小排列。

引文分析(Citation Analysis)利用数学及统计学的方法和比较、归纳、抽象、概括等逻辑方法,通常是对科学期刊、论文、著者等各种分析对象的引证与被引证现象进行分析,进而揭示其中的数量特征和内在规律的一种文献计量分析方法[26]。本研究所用语料中存在表 1 中的引用情况,例如《中论》中使用《诗》的句子“高山仰止,景行行止”。如果典籍 A 引用了典籍 B 的句子,则绘制出一条由节点 A 指向节点 B 的有向线段,并且通过引用次数的高低控制有向线段的粗细,引用次数越多,线段越粗。该部分的结果将借助 Gephi 软件进行典籍节点布局和作力导向图。

表 1　语料库中典籍的引用示例

典籍	典籍中句子
中論	《詩》云:「高山仰止,景行行止」
周易	《象》曰:「履霜堅冰」、陰始凝也
呂氏春秋	《詩》曰:「執轡如組」
孟子	王說,曰:「《詩》云:『他人有心,予忖度之』」

2.3.2　文本向量化　仅统计字频无法表示文本丰富的语义信息,因此需要语言知识表示。在自然语言处理技术中,一般将词义信息编码到词语的向量化表示中,目前常用的文本表示方式分为离散式表示和分布式表示,本研究在进行分析《老子》与其他典籍的关系时,采取了这两种向量化手段。离散式表示选用了容易理解且解释性较强的 TF-IDF 方法,将每一篇典籍表示成向量;分布式表示采用深度神经网络的预训练嵌入式表示方法,具体选取了流行的 BERT 模型,将典籍中的句子投影到高维空间中,为后续应用数学上计算向量间的夹角余弦值做铺垫。

TF-IDF 方法简单快速,是最常用的基于统计的向量化方法,本研究使用该方法简述如下。在一个给定的典籍或句子里,字频指的是某一个给定的字在该典籍或句子中出现的次数,用 TF(Term frequency)表

示先进行正规化,以防止它偏向文本长的典籍。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

以上式子中 $n_{i,j}$ 是该字在典籍 d_j 中的出现次数,而分母则是在典籍 d_j 中所有字的出现次数。

逆向文件频率(inverse document frequency,IDF)是一个词语普遍重要性的度量。

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \tag{2}$$

分子 $|D|$ 表示本语料库中的典籍总数,分母表示含字 t_i 的典籍数目,如果存在字未出现在本研究所用的语料库中,就会导致被除数为零,因此在计算时进行了拉普拉斯平滑。

$$tfidf_{i,j} = tf_{i,j} * idf_i \tag{3}$$

某一典籍内的高频率字,以及该字在整个语料典籍集合中的低文件频率,可以产生出高权重的 TF-IDF。因此,TF-IDF 倾向于过滤掉常见的字,保留重要的词语。该方法的缺点是有时候用字频来衡量文章中的一个词的重要性不够全面,且这种计算无法体现位置信息,无法体现字的上下文的关系。因此之后加入了 BERT 模型进行字的向量化表示。

BERT[27](Bidirectional Encoder Representation from Transformers)是由 Google AI 于 2018 年 10 月提出的一种基于深度学习的语言表示模型,是 NLP 领域近期重要的研究成果。BERT 模型可以视为一种自监督的文本向量化手段,它充分地描述了字符级、词级、句子级甚至句间关系特征。

BERT 模型的主要结构是真正双向的 Transformer 编码器。Transformer 是 2017 年谷歌《Attention is all you need》[28]论文中提到的。论文中提出了 transformer 一种新的结构,包括编码器与解码器两部分,有很强大的文本编码能力,应用于机器翻译领域上时,取得了很好的效果。该模型的训练任务有两个,一个是完形填空任务(Masked Language Model),随机遮盖典籍语料中 15% 的字(本研究的粒度在字层面),将经过双向编码器的隐层向量送入 softmax,来预测被遮盖的字;第二个任务是训练模型捕捉句子联系的能力(Next Sentence Prediction),即给出两个句子 A 和 B,B 有一半的可能性是 A 的下一句话,训练模型来预测 B 是否为 A 的下一句话。

通常情况下,基于 BERT 模型的自然语言处理任务需要经过预训练与微调两个阶段。在本研究中的模型中,预训练阶段如图 2 所示,首先利用大规模未标注过的典籍文本,也就是 915 部作品进行充分的自监督学习,有效学习文本的语言特征得到深层次的文本向量表示,使用 12 个注意力头,12 层隐藏层,在 18 144 个字符基础上训练出 784 维的语境向量表示。鉴于本研究不存在有标注的下游任务,因此没有进行第二阶段的微调工作。

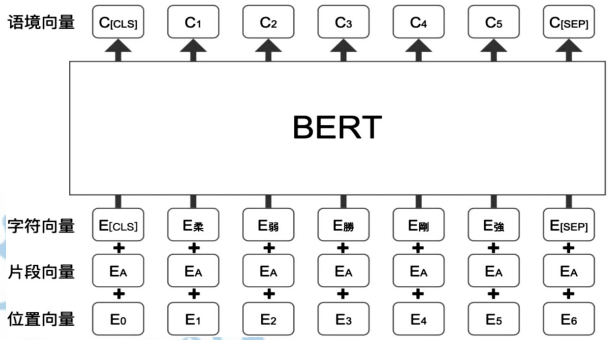

图 2 本研究 BERT 模型的输入输出示例

2.3.3 相似度计算 在信息检索、网页判重、推荐系统中,都涉及到对象之间或者对象和对象集合的相似性计算,本研究通过相似度计算比较两个典籍或者典籍间句子的相似性。相似度计算中的关键技术主要是两个部分,对象的特征表示,特征集合之间的相似关系。对象的特征表示已在“文本向量化”中介绍;本文使用了余弦相似度进行特征集合之间的相似关系计算,其简单,广泛使用且效果很好。余弦相似度是基于向量空间模型(Vector space model)的,使用 A、B 两个语境向量夹角的余弦值作为衡量两个典籍或句子间差异的大小,余弦相似度体现的是每个向量的方向关系(角度),而非幅度,计算方法如公式 4。

$$\cos\theta = \frac{A \cdot B}{\|A\| * \|B\|} \tag{4}$$

3 实验结果与分析

3.1 **老子内容整体分析** 河上公版本的《老子》典籍中,有八十一章,去除标点符号、河上公的注以及章名,有字 5 665 个。根据统计“之”“不”“以”“其”“而”“為”“者”“無”“也”“天”是出现次数最多的字。根据史学知识可知,“之”“以”“其”“而”“者”“也”是常用的虚词,“不”“無”是否定副词,不过“無”同时又是《老子》思想中的核心概念,后代魏晋玄学又进一步标榜老子的“贵无”思想。在《老子》的文本当中,“無”也经常作为抽象提炼的概念名词使用,如《老子》著名的第一章中“無名天地之始;有名萬物之母。故常無欲,以觀其妙;常有欲,以觀其徼。”一般断句都把“無”作为否定性的存在动词。但是《庄子·天下篇》说:“老聃闻其风而悦之,建之以常无有。”,似以“常无”“常有”为断句处,后世王安石说“道之本出于无,故常无,所以自观其妙。”在这种解读中,“無”就被作为单独的哲学概念使用了。除此以外,出现次数最多的前 10 个字中,另有“为”“天”是实词。在文本中出现次数超过 30 次的字已经展示在图 3 中,可以看出,

除去实词、方位词、形容词,主要名词的出现频次由高到低有“为”“無”“天”“人”“道”“善”“德”“物”“民”。《老子》主要围绕“道”和“无为”的宗旨,将“道”运用在治国理政中的“德”来进行论述其思想观念[28]。在天、人关系上,老子尊天更敬人,外王论是“贵以贱为本”“无为而无不为”,社会理想是“小国寡民”“至德之世”[29]。本研究中抽取出的高频词基本能够凸显出《老子》的这些哲学思想。

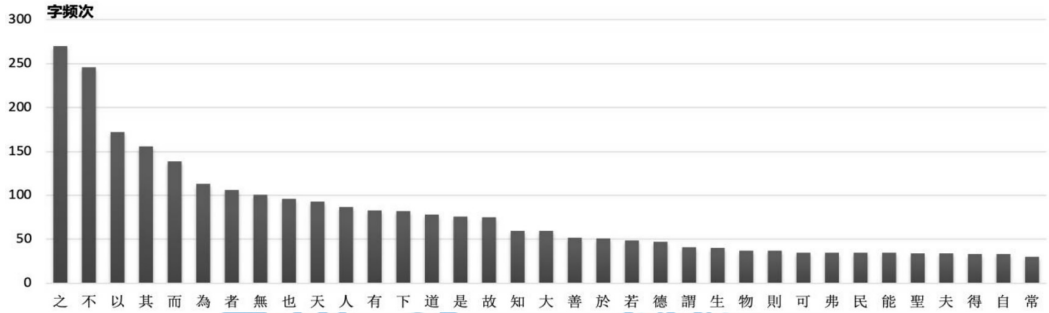

图 3　河上公版本《老子》的字频次统计

3.2　**典籍引用分析**　通过对《老子》的引用统计得出《老子》中没有引用其他典籍,因此,本小节展开《老子》一书被先秦两汉时期哪些典籍所引用,也就是《老子》的后世影响。统计后,按照节点的大小与节点的入度的大小成正比,边粗细与引用次数高低成正比,做出图 4。

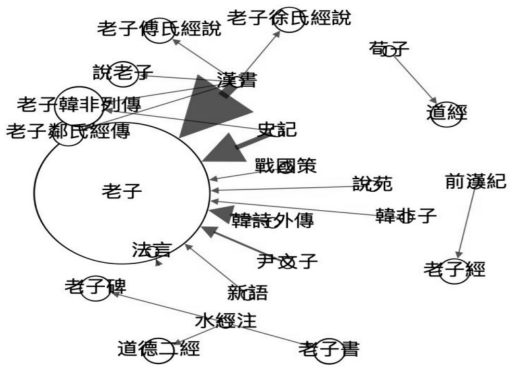

图 4　《老子》的被引情况网络图

由于本研究是以书名号中含有“老子”字段匹配,不免有所疏漏,因此在匹配书名的过程中还加入了《老子》的上下篇“德經”“道經”作为关键词,并加入《老子》的其他称谓“五千言”“五千文”“道德经”。通过图 4 可以清晰地看出,《汉书》对《老子》的引用最多,其次是《史记》和《韩诗外传》,他们的引用老子的次数分别是 9 次,5 次,3 次。另外《尹文子》引用《老子》2 次,图 4 中其他典籍之间的引用次数都是 1 次。《汉书》《史记》是两部纪传体史书,是“前四史”之二,其中提及”老子“主要是引用《老子》中语句和陈述历史人物“好《老子》书”。《韩诗外传》《尹文子》中都是直接引用《老子》中的典句。其他典籍关于对《老子》一书的引用,主要是引用其中的句子。

道家思想影响广而远,其它经典著作还有《文子》《庄子》《列子》《淮南子》等,可以从图 5 中寻找各个典籍之间的引用关系与思想继承发展脉络。

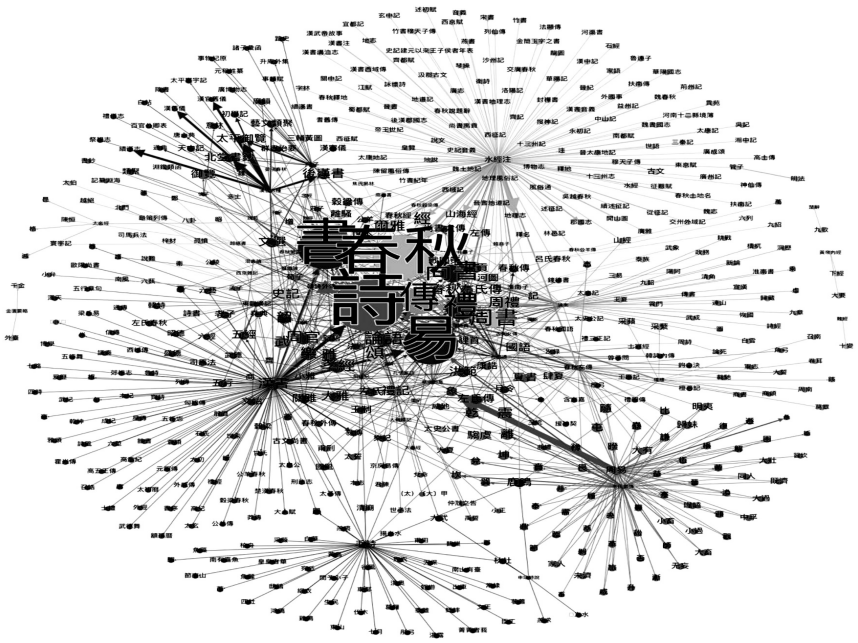

图 5　以入度大小绘制节点得到的先秦两汉典籍引用关系网络

道家学派是以老庄学说为中心的学术派别,代表人物有老子和庄子。接下来,本研究分析了庄子的思想精髓《庄子》一书的引用情况。《庄子》全书中引用了《诗》《书》《礼》《乐》《春秋》,并被《尸子》所引用。《诗》《书》《礼》《乐》《春秋》对《庄子》思想有所影响,在《庄子》中被引次数分别是 6、6、5、4、3。《尸子》一般认为是托名战国时期作品,兼容诸家之学,其中引用了《庄子》释文的内容,例如"天地四方曰宇,往古往今来曰宙。"一句,见于唐初陆德明所作的《庄子》释文引《三苍》之说《释文》,虽然晚出,但《三苍》乃秦时文献,因此并不能排除《尸子》中内含早期的文献来源的可能性。

本小节同时做出以出度大小绘制节点的网络关系图,通过图 6 可以清晰地看出,在所挖掘的文本中,《水经注》《汉书》《周易》《京氏易传》是最常引用其他典籍的先秦两汉时的史书。《汉书》篇幅较长,是综合型史籍,《水经注》原是对地理名著汉代《水经》的注解,大部分内容为北魏时郦道元所作,引用范围十分广泛,包括自然地理、人文地理、山川胜景、历史沿革、风俗习惯、人物掌故、神话故事等,因此多引用先秦两汉旧籍。《周易》《京氏易传》尽管篇幅相对很短,但是其非常尊重前贤的著作,开创了"引经据典"的先河,因此这两部典籍引用其他典籍的次数也较多。

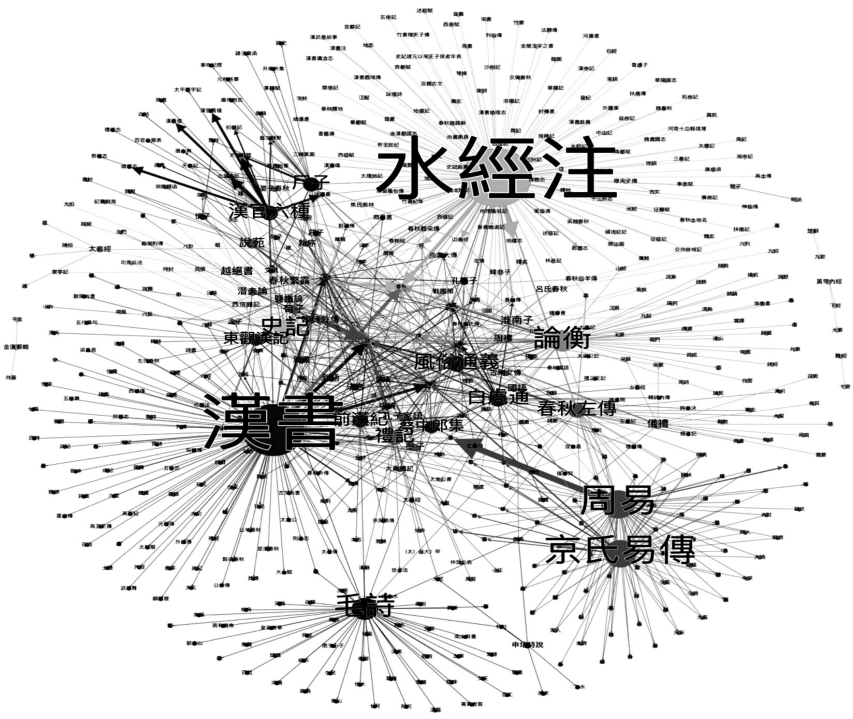

图 6 以出度大小绘制节点得到的先秦两汉典籍引用关系网络

3.3 《老子》与后世典籍的关系 借助 Python 的 Scikit-Learn 模块计算每一篇文档的 TF-IDF 的值,两两典籍之间计算余弦相似度,得到 102 篇典籍的两两相似度结果。经过分析其中的相似度情况,可以得出,《九章算术》《新语》《忠经》《楚辞》《佛说二十四章经》《尔雅》与其他典籍都不相似的,这与它们的主题内容有关,是相对于其他典籍所独特的。《九章算术》是数学专著;《新语》是西汉时期陆贾的政论散文集;《忠经》是系统总结忠德的专门经典;《楚辞》是中国文学史上第一部浪漫主义诗歌总集,是一种新诗体;《佛说二十四章经》一般认为是中国第一部汉译佛经;《尔雅》是辞书之祖。通过上面的分析,说明应用 TF-IDF 和余弦相似度的计算方法对分析典籍之间的关系一定程度上有效,因此在对《老子》的分析时,表 2 抽取出《老子》典籍的相似情况,进行详细说明。

表 2 《老子》与其他典籍相似度降序排名前 5

序号	典籍 A	典籍 B	相似度
1	老子	淮南子	0.333019
2	老子	山海经	0.328041
3	老子	列子	0.291855
4	老子	莊子	0.291179
5	老子	焦氏易林	0.250778

该小节中得到的《老子》与其他典籍相似度排名,体现了《老子》思想对后世的影响,正是因为有"道家"思想后,后世作品中才有包含"黄老思想"的典句著作。整体上看,与《老子》相似度的计算结果最高只有 33.30%,最相似的《淮南子》是一部集诸子百家思想于一体的著作,道家思想在其中占据了一定的优势,它的"道"既继承了先秦道家老庄思想,同时又染有汉初"黄老思想"的时代气息[30]。《列子》《庄子》是道家

重要典籍,《山海经》多神仙怪诞之风,《易林》同为玄学所宗,这两种文献在用字遣词的风格上与《老子》相近,似乎也不能算特别意外。除此以外,其他典籍与《老子》的相似程度较低,本部分不再分析。

3.4 《老子》与其之前典籍的关系　探索《老子》与其之前典籍的关系,主要是为了研究老子的思想是受哪些已有思想的影响。为了量化比较与《老子》的相似程度,该部分选取了两种方法,以获得不同方法的结果和对比差异,这两种方法是 TF-IDF 和余弦相似度、BERT 和余弦相似度。两种方法的原理和 BERT 的参数设置已在本文第二部分"研究方法"中说明。使用 915 部作品进行 BERT 模型的自监督学习,有效地学习到文本的语言特征得到深层次的文本向量表示,通过训练达到在完形填空任务上 52.11% 的精度和在预测是否是下一个句子上 98.45% 的精度,参考 Sentence-BERT[31],使用句子中各个字嵌入的平均后的向量作为句子的表示,之后使用了余弦相似度衡量句子相似程度。

该部分需要查找成书于《老子》之前的典籍,这类典籍难以考证确切成书时间且数量较少。《周易》《尚书》《诗经》《墨子》《论语》《逸周书》,这些基本都可以肯定在老子文本之前成书的[32]。《楚辞》与《老子》时代相对接近。本部分研究的基本理念是把《老子》当成一个混杂的文本,来源不单一,只要时代相近的都纳入范围。最终,本小节研究使用的典籍有《周易》《尚书》《诗经》《墨子》《论语》《逸周书》《楚辞》。

将《老子》中的每一个句子与前述 7 部典籍中的每一个句子计算相似度得分并由大到小排序,分别取前 20、10、5 条相似的句子,目的是防止出现因所划分的数目不同而导致的结果差异。以"此兩者同出而異名。"这句为例,使用 TF-IDF 与 BERT 生成的字嵌入,之后计算出的相似句子结果如表 3。需要说明的是,表 3 中重复出现的句子是典籍原始内容中就存在相同的句子。

表 3　与"此兩者同出而異名"最相似的前 20 个句子

方法	TF-IDF			BERT		
序号	句子	相似度	来源	句子	相似度	来源
1	同名。	0.536	墨子	故同處其體俱然。	0.739	墨子
2	君子以同而異。	0.519	周易	二名一實,重同也。	0.704	墨子
3	有其異也,為其同也,為其同也異。	0.475	墨子	夫物有以同而不率遂同。	0.680	墨子
4	異。	0.450	墨子	同、異而俱於之一也。	0.679	墨子
5	異。	0.450	墨子	所謂非同也,則異也。	0.671	墨子
6	同異交得。	0.419	墨子	此乃一是而一非者也。	0.658	墨子
7	長人之異短人之同,其貌同者也,故同。	0.401	墨子	是俱有,不偏有偏無有。	0.647	墨子
8	名。	0.396	墨子	則義不同也。	0.639	墨子
9	所謂非同也,則異也。	0.385	墨子	法同、則觀其同。	0.634	墨子
10	同、異而俱於之一也。	0.383	墨子	二必異,二也。	0.631	墨子
11	丘同,鮒同,是之同,然之同,同根之同。	0.370	墨子	然則義果自天出矣。	0.627	墨子
12	有非之異,有不然之異。	0.368	墨子	然則義果自天出矣。	0.627	墨子
13	同。	0.361	墨子	然則義果自天出也。	0.619	墨子
14	同。	0.361	墨子	不外於兼,體同也。	0.617	墨子
15	同極而異路兮,又何以為此援也?	0.339	楚辞	是類不同也。	0.616	墨子
16	同異交得,放有無。	0.328	墨子	君子以同而異。	0.611	周易
17	使者出。	0.321	论语	貴者公,賤者名,而俱有敬傻焉,等異論也。	0.610	墨子
18	民好惡其不同兮,惟此黨人其獨異。	0.321	楚辞	此言而非兼,擇即取兼,即此言兼費也。	0.606	墨子
19	節出,使所出門者,輒言節出時摻者名。	0.320	墨子	有非之異,有不然之異。	0.603	墨子
20	仗者,兩而勿偏。	0.320	墨子	同異交得。	0.603	墨子

表 3 是在句子层面上进行了相似度计算,为了衡量典籍之间的相似程度,本研究将排名前 20、10、5 个相似句子的相似程度相加,即每一个句子都为自身的来源典籍进行加权,以表 3 为例,如果只根据"此兩者同出而異名",那么《老子》之前的 7 部典籍的每一部得到一个分数,即根据这一个句子,可以得出每部典籍与《老子》相似的排序,表 3 的结果显然表现出《墨子》与《老子》最相似。依照此方法,将《老子》中所有的句子的权重相加再排序后得到与《老子》的典籍相似的结果,如表 4 与表 5 所示。

表 4　基于 TF-IDF 计算相似度之后的排序情况

序号	前 20 个句子	前 10 个句子	前 5 个句子
1	墨子	墨子	墨子
2	论语	论语	论语
3	周易	周易	周易
4	尚书	尚书	尚书
5	逸周书	逸周书	逸周书
6	楚辞	楚辞	楚辞
7	诗经	诗经	诗经

表 5 基于 BERT 计算相似度之后的排序情况

序号	前 20 个句子	前 10 个句子	前 5 个句子
1	墨子	墨子	墨子
2	周易	周易	周易
3	逸周书	逸周书	逸周书
4	尚书	尚书	尚书
5	论语	论语	论语
6	楚辞	楚辞	楚辞
7	诗经	诗经	诗经

量化方法固定时,无论是前 20 还是前 10、前 5 个最相似的句子统计的结果的排序都是没有差别的。两种方法都认为《墨子》是与《老子》最相似的典籍,这一现象看似打破了传统对于道家、墨家学说的分野。除却《墨子》是本小节研究所用的文本中篇幅最长的,因此其含有的句子最多,在一定程度上增加了从中找出与《老子》相似的文本的可能因素以外,一个比较意外的结果是,通过文本相似度的计算,计量结果进一步引导我们对两种文献的内容进行了比对和思考,令我们发现《老子》和《墨子》中反映出有关思想史与科技史相结合的近似观念。对此问题,我们拟另设专文讨论。

《周易》被誉为“大道之源”,《老子》思维方式与方法同样与《易经》有着内在的联系[33]。后世《易》《老》并称,同谓之玄这一结果,与前一小节通过余弦相似度分析,发现汉代《易林》与《老子》之间的相似关系恰可呼应。

《逸周书》与《尚书》都是记言史书,一说《逸周书》为孔子删书时剔落的部分。诸子皆出于六经王官之学,“书经”是诸子文献中的高频引用来源,相传老子本为周朝柱下之史,由此我们或可理解“书”类文献与《老子》文本间的联系。

表 4 表 5 中,两种方法对《论语》的判断是差距较大的,TF-IDF 方法中,《论语》排在第 2,而在 BERT 方法中《论语》排在第 5 位。《论语》儒道本身就是中华文明中不可缺少的两个思想宝库,从其发展历程来说,都是“同源一体”的[34]。

《楚辞》《诗经》在两种方法中都与《老子》最不相似,两者都是我国早期诗歌中的杰作;而《老子》的核心思想是哲学上的朴素辩证法。通过计算相似度,表现出渊源上较远的关系这一结果是合理的。

4 结论与展望

《老子》内容丰富,凝结着道家的思想与智慧。本文通过两种相似度计算的方式,以寻求与《老子》相似的文本进行文本关系研究,主要利用了文本相似度,用机器学习方法进行了探索。基于 TF-IDF 的方法更多在统计层面上,因此对一些专有名词的关联度比较敏感;利用 BERT 语言模型的方法得到的结果对整体语义的把握相对好;两者可以满足对文本使用的不同需求。为方便之后的学者进一步探索,本研究开源出所训练的 BERT 模型。

在统计《老子》字频情况时,本文从字层面捕捉到《老子》反复提到的“为”“無”“天”“人”“道”“善”“德”“物”“民”等概念;进行典籍引用分析后,《汉书》《史记》等作品提及《老子》的频次较高,多是描述古人“好《老子》书”,可以窥见《老子》一书的在后世广为流行。通过一系列相似度对比,本研究发现《淮南子》受《老子》影响颇深,《山海经》次之,《淮南子》本身是一部集诸子百家思想于一体的著作,先秦道家老庄思想在其中占据了一定的优势,《山海经》则是玄学所宗,与《老子》在遣词上的相近之处。在《老子》出现前的典籍上进行相似度计算时,本研究挖掘出《墨子》的思想史与科技史与其近似,例如两者都对自然有探究倾向。

本项研究与之前通过字词的字符文本检索的比较研究有很大不同,更多是基于语言的特征分析得出的数据统计。本文的分析,不是基于普通的文本阅读印象,而是利用字词在文本中的出现频率和文本的向量关系,初步建立文本之间的联系。在此基础上,再进行思想史研究的考量。定量分析尤同探矿,是进行深入挖掘前的可行性测试手段。这种方法,正在帮助我们对《老子》和《墨子》间的论说思想关系进行一番新的思考。

目前本文的研究较为粗略,关于《老子》、道家学说仍有很多亟待研究的方面。在今后的工作中,我们准备从文本的字段、词段、句法与虚词特征等方面进行更细致的研究,为一个完整的大数据视域下老学的起源、发展提供更加有针对性的、具体的实证结果。

参 考 文 献

[1] 黄承梁. 老子之道和现代文明建设研究[D]. 山东大学,2016.

[2] 李璐茜. 老子思想的当代教育功能[D]. 太原科技大学,2015.

[3] 王晓晓. 老子思想与当代生态幸福观的建构[D]. 山东师范大学,2014.

[4] 余 萍. 老子思想对我国行政管理建设的启示[D]. 电子科技大学,2012.

[5] 徐水生. 老子思想对日本近现代名家的影响[J]. 商丘师范学院学报,2015,31(1):24-30.

[6] 李振纲. 老子哲学的生命意识与价值关怀[J]. 河北师范大学学报(哲学社会科学版),2020,43(5):79-86.

[7] 黄水清,王东波. 古文信息处理研究的现状及趋势[J]. 图书情报工作,2017,61(12):43-49.

[8] 赵 阳,顾 磊. 基于中文信息处理的古籍整理研究评述[J]. 图书情报工作,2010,54(3):116-119,63.

[9] 俞敬松,魏 一,张永伟,等. 基于非参数贝叶斯模型和深度学

习的古文分词研究[J]. 中文信息学报,2020,34(6):1-8.
[10] 王姗姗,王东波,黄水清,等. 多维领域知识下的《诗经》自动分词研究[J]. 情报学报,2018,37(2):183-193.
[11] 黄水清,王东波,何　琳. 以《汉学引得丛刊》为领域词表的先秦典籍自动分词探讨[J]. 图书情报工作,2015,59(11):127-133.
[12] 徐晨飞,叶海影,包　平. 基于深度学习的方志物产资料实体自动识别模型构建研究[J]. 数据分析与知识发现,2020,4(8):86-97.
[13] 黄水清,王东波,何　琳. 基于先秦语料库的古汉语地名自动识别模型构建研究[J]. 图书情报工作,2015,59(12):135-140.
[14] 汤亚芬. 先秦古汉语典籍中的人名自动识别研究[J]. 现代图书情报技术,2013(Z1):63-68.
[15] 王东波,高瑞卿,沈　思,等. 面向先秦典籍的历史事件基本实体构件自动识别研究[J]. 国家图书馆学刊,2018,27(1):65-77.
[16] 梁继文,江　川,王东波. 基于多特征融合的先秦典籍汉英句子对齐研究[J]. 数据分析与知识发现,2020,4(9):123-132.
[17] 李秀英. 基于历史典籍双语平行语料库的术语对齐研究[D]. 大连理工大学,2010.
[18] 俞敬松,魏　一,张永伟. 基于 BERT 的古文断句研究与应用[J]. 中文信息学报,2019,33(11):57-63.
[19] 程　宁,李　斌,葛四嘉,等. 基于 BiLSTM-CRF 的古汉语自动断句与词法分析一体化研究[J]. 中文信息学报,2020,34(4):1-9.
[20] 王博立,史晓东,苏劲松. 一种基于循环神经网络的古文断句方法[J]. 北京大学学报(自然科学版),2017,53(2):255-261.
[21] 王东波,何　琳,黄水清. 基于支持向量机的先秦诸子典籍自动分类研究[J]. 图书情报工作,2017,61(12):71-76.
[22] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报,2016,42(2):66-80.
[23] 姜　欣,姜　怡,方　森. 文本翻译索引的互文度量方法[J]. 计算机应用,2010,30(7):1938-1940.
[24] 肖　磊,陈小荷. 古籍版本异文的自动发现[J]. 中文信息学报,2010,24(5):50-55.
[25] 王宝利. 从避讳现象谈《老子河上公章句》的成书时代[J]. 兰州学刊,2006(8):47-48.
[26] 王曰芬. 文献计量法与内容分析法综合研究的方法论来源与依据[J]. 情报理论与实践,2009,32(2):21-26.
[27] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[28] Correia G M, Niculae V, Martins A F T. Adaptively sparse transformers[J]. arXiv preprint arXiv: 1909.00015, 2019.
[29] 魏质方.《道德经》中的价值主张及其现代启示[J]. 美与时代(上),2020(5):73-74.
[30] 杨　栋,曹书杰. 二十世纪《淮南子》研究[J]. 古籍整理研究学刊,2008(1):78-88.
[31] Reimers Nils, Gurevych Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks [J]. arXiv preprint arXiv: 1908.10084, 2019.
[32] 钱　穆.《先秦诸子系年》[M]. 上海:商务印书馆, 2001.
[33] 解光宇,孙以楷. 老子与《周易》[J]. 国学,2013(12):20-23.
[34] 何大海. "儒道同源"的若干考证[J]. 学理论,2013(20):30-31.

(责编/校对:贺小利)