

基于机器学习的古籍目录互著与别裁探析^{*}

张力元 王 军

摘 要 目录是组织与利用古籍资源的重要工具,也是图书情报学科的重点研究对象。互著与别裁作为古典目录学中的两种辅助方法,能在深入剖析文献内容特征的基础上,根据内容的多元性将文献准确、完整地记载于目录体系中,达到“类例既分,学术自明”的效果。将互著与别裁映射为文本挖掘中的文本分类问题,提出基于机器学习以实现互著与别裁的方法框架,为古籍在目录体系中的多类目记载提供方法。首先利用 TextCNN 与 BERT 两种机器学习模型对先秦诸子六家十部典籍文本进行分类训练,结果显示 BERT 优于 TextCNN,可以达到 91.64% 的分类准确率;之后用微调训练后的 BERT 模型对《荀子》与《管子》进行篇、章粒度的分类判断,最终得出这两部图书各篇章互著与别裁的结果。本研究展现了在数字人文视域下,数字技术对古典目录学、古典文献学以及学术史研究的应用价值。图 5。表 7。参考文献 43。

关键词 古籍目录 互著 别裁 机器学习 数字人文

分类号 G257

Research on Inter Record and Analytic Record of Classical Bibliography Based on Machine Learning

ZHANG Liyuan & WANG Jun

ABSTRACT

Bibliography is an important tool to organize and utilize ancient books and is also the key research object of library and information science. As two auxiliary methods in classical bibliography, inter record and analytic record aim at recording the documents accurately and completely in the bibliography system according to the diversity of contents on the basis of the in-depth analysis of the content characteristics of the documents so as to achieve the function of “once classes are divided, academic ideas become clear”. However the traditional methods of inter record and analytic record are mainly completed by human beings, which leads to several issues such as low efficiency, high cost, weak scientificity, weak objectivity and poor reliability, etc.

This study aims at introducing machine learning to classical bibliography under the perspective of digital humanities to provide a new implementation strategy for inter record and analytic record. The study first proposes to map inter record and analytic record of classical bibliography to the problem of text classification

^{*} 本文系国家自然科学基金国际重点合作项目“中国儒家学术史知识图谱构建研究”(编号:72010107003)的研究成果之一。(This article is an outcome of the key international cooperation project “Research on the Construction of Chinese Confucian Academic History Knowledge Graph”(No. 72010107003) supported by the National Natural Science Foundation of China.)

通信作者:王军,Email: junwang@pku.edu.cn, ORCID: 0000-0003-0593-1034 (Correspondence should be addressed to WANG Jun, Email: junwang@pku.edu.cn, ORCID: 0000-0003-0593-1034)

and puts forward a method framework based on machine learning to contribute countermeasures for the multi-category record of ancient books in the bibliographic system. This study takes the pre-Qin schools and representative books as the object to verify the method. Two machine learning models, namely TextCNN and BERT are used to classify ten ancient books from six schools of pre-Qin dynasties, and each school has one or two books correspondingly. The classification result shows that fine-tuned BERT model outperformed TextCNN model and can achieve 91.64% of classification accuracy. The fine-tuned BERT model is then used to classify two controversial books, namely *Xunzi* and *Guanzi*, and the results of inter record and analytic record of two books are obtained. It is suggested that *Xunzi* should be classified to Legalism category and Confucianism category, while *Guanzi* should only be classified to Legalism category. In particular, this study also finds that the first 26 chapters of *Xunzi* are more inclined to Legalism and the last six chapters are more inclined to Confucianism.

The innovations of this paper are mainly reflected in the following aspects: it puts forward the method of inter record and analytic record based on machine learning, and puts forward the method of text classification to distinguish the categories of ancient books from the particle size of books, sections and chapters and to judge the ideological tendency of scholars. It shows the value of digital technology to the study of classical bibliography, classical philology and academic history from the perspective of digital humanities. 5 figs. 7 tabs. 43 refs.

KEY WORDS

Classical bibliography. Inter record. Analytic record. Machine learning. Digital humanities.

0 引言

我国古典目录学历史悠久,具有“辨章学术,考镜源流”的学术价值,至今仍是“学中要紧要事”^[1],对古籍资源的组织与利用、古籍数据库的设计与开发、古籍联合目录的建设与统一等具有重要意义。互著与别裁作为古典目录学的两种辅助方法,能在剖析文献内容的基础上,根据内容的多元性将文献准确、完整地记载于目录体系的多个类目下,达到“类例既分,学术自明”的效果,曾应用于《七略》《文献通考·经籍考》《澹生堂书目》《直斋书录解题》《书目例略》《百川书志》《艺文类聚》等书目中^[2-4]。但互著与别裁的传统实现策略主要依托于人工,这不仅对人力和学术素养有较高要求,而且存在主观性较强或难以准确判断的问题,以致无法在海量古籍上得到普遍运用。

本文在数字人文视角下引入机器学习方法,为互著与别裁提供新的实现策略。在古籍数字化的基础上智能化地实现互著与别裁,可以高效处理更多古籍资源与更多类目的对应关系,也可以深入到篇章做出细粒度的分类判断。这不仅能够拓展互著与别裁的理论与实践边界,而且能够提高古籍资源的组织与利用效率,为数字环境下的“即类求书,因书究学”提供量化保障。在新文科背景下,该方法还可以为人文研究提供新的分析维度,助力发现新问题。

本文首先提出基于机器学习以实现互著与别裁的方法框架,之后分别利用 TextCNN 和 BERT 两种机器学习模型对先秦诸子六家十部典籍文本进行分类训练,让机器学习各家类别与典籍文本的对应关系,再选取训练后分类效果更好的 BERT 模型对《荀子》和《管子》分别进行分类判断,最终依据分类结果提出对这两部典籍的互著与别裁建议。

1 研究综述

互著与别裁的思想早见于明代祁承燾《澹生堂藏书目录·庚申整书略例》的“互”与“通”,后于清代章学诚《校讎通义》中正式提出^[5,6]。互著与别裁虽然作用对象不同,但二者的原理和方法相通,即在分类编目时,将内容广泛的或是具有多个主题的图书及篇章全面、科学地记载于目录体系内的多个类目中,实现目录辨考学术和指导文献检阅的多重作用,对于实现古典目录学“辨章学术,考镜源流”具有重要价值^[3,7]。同时也要辩证认识到,互著与别裁的传统实现方法具有时代局限性,在客观性、准确性、效率等方面均存在问题。如徐召勋在1979年就指出:“假如一本书包括甲、乙两个类的内容,而且各占一半,这究竟是用互著法还是用别裁法呢?还有的书包括两个类的内容,不是一半对一半,而是一大半对一小半。这样的书又该如何处理呢?”对此他的回答是“结合具体情况而定”^[8]。实际上,在回答这个问题前,首先需要解决的是如何量化书中不同类别所占的比例,即何为“一半”“一大半”“一小半”。基于人工主观判定的方法很难回答这个问题,这意味着过去的互著与别裁判断可能存在错误。然而,这一问题在近三十年并没有得到学者们的充分重视,近年有关互著与别裁的研究多集中于再论其内涵^[9]、追溯其起源^[2,3]以及辨析书目应用实例^[4],几乎没有学者再次审视互著与别裁在实现方法上的问题,也没有将量化方法引入其中。

近五年来,随着“数字人文”文理交叉融合理念的推广,有学者逐步意识到可以利用数字技术来完善并丰富古典目录学的理论体系和实践方法。在理论方面,陈志新指出,目录学的未来要依托于大数据和人工智能等技术来创新和进步,实现过去依靠大学问家也无法有效完成的任务^[10]。在实践方面,李瑞龙和李明杰提出利用数字技术实现古典目录学辨考学术的思想^[11],但其在技术探索时更侧重于梳理学术源流,而跳过了更为基础的分类图书环节。李惠等构造了古籍提要

共现网络并探讨了提要推荐对于古籍知识发现的潜力^[12]。由此,已有数位学者通过理论探讨或案例分析证明数字技术对古典目录学的价值。

若从技术的角度将互著与别裁理解为文本分类问题,那么国内外均已较为丰富的研究成果可以借鉴。文本分类的核心任务是将文档自动划分到预先定义好的类别中,其研究可追溯至1960年代Maron根据受控词表辨识文档主题、自动分类文档并索引科学文献^[13]。目前机器学习是文本分类的主要方法,并且在英语和现代汉语等语境下有较为成熟的应用,例如用于垃圾邮件识别、网页主题分类、文本情感分类、文本风格分类等。在图书与情报领域,一个典型的应用场景是利用主题词等信息实现图书自动分类标引,如王昊等利用机器学习模型构建多层次图书自动分类系统,根据题名、关键字和文摘信息自动给出中图法分类号^[14]。随着数字化建设带来的古籍数字资源的生长,有学者开始将这一技术运用在古籍文本分类上,例如张馨怡利用词向量技术表示古诗文本,再通过TextCNN模型将古诗判断为爱国类别或其他类别^[15]。截至目前,尚未有研究利用新技术来解决古典目录学中互著与别裁存在的问题。

2 基于机器学习的互著与别裁方法框架

本文提出一套基于机器学习的互著与别裁方法框架,如图1所示。该框架整体逻辑为:在古典目录学已有研究经验与成果的基础上,由互著与别裁的内涵和原则来统领整个方法。根据具体的研究需求或目录需求确定文献范围,通过对多源异构语料的采集、清洗与融合,最终构建研究所需的语料库。利用语料库内结构化的文本和类别标签来训练分类器,再利用分类器来判别待分类文本的类别概率,并根据概率阈值为文本做出分类建议。若文本是以书为单位,则对应于互著标引,若文本是以篇章为单位,则对应于别裁标引。最终结果可应用于目录补充、学术史研究、图书编目以及古籍数据库按类浏览和索引等方面。

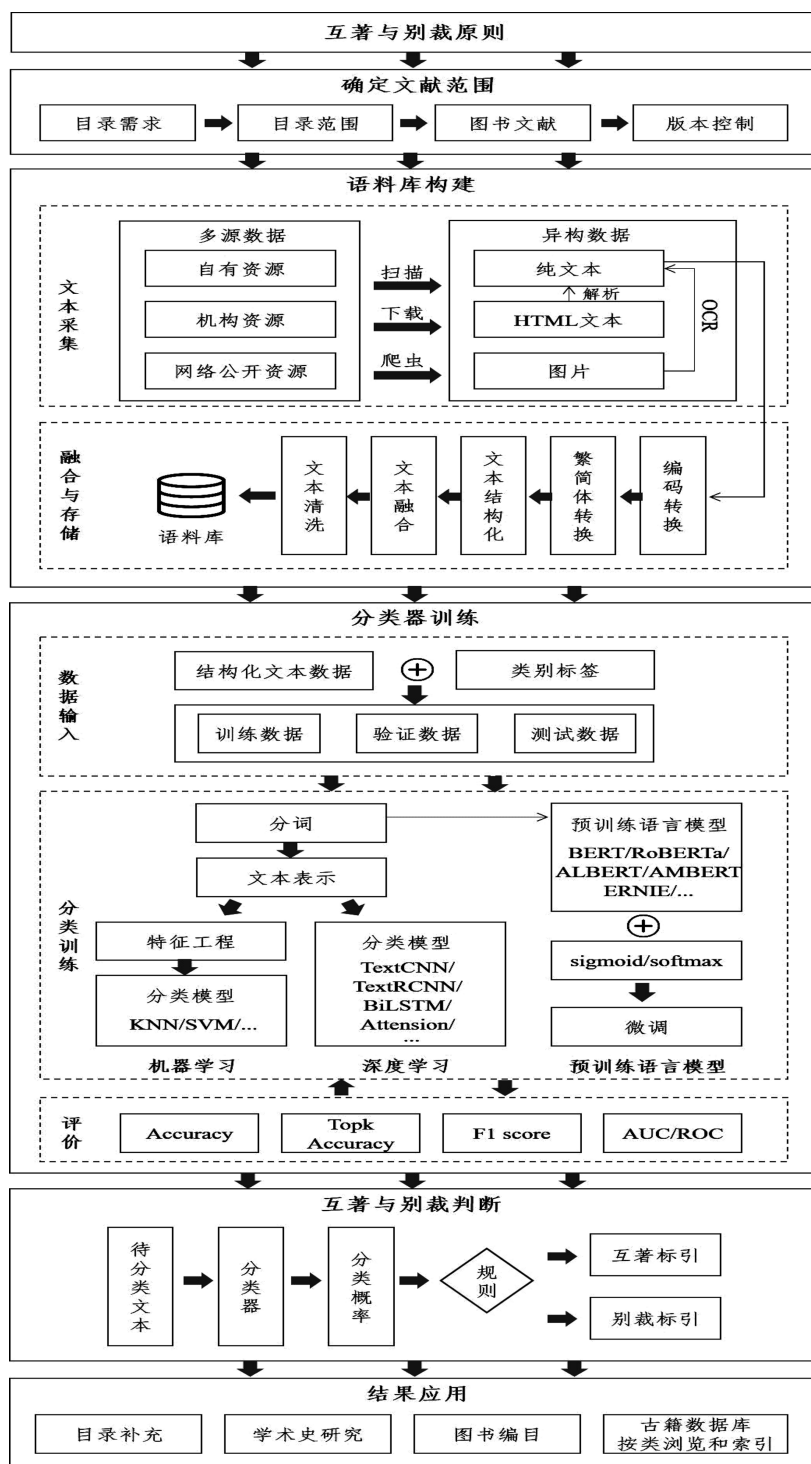


图1 基于机器学习的互著与别裁方法框架

3 分类模型训练实验

为检验机器学习框架是否可以有效判别古籍类目,本文以四部的子部为对象,以先秦诸子六家为类目,以各家代表典籍为文本训练分类模型,并分析分类结果的准确性。

3.1 实验数据

代表典籍的选定主要参考《汉书·艺文志》中记载的先秦诸子典籍,并参考梁启超在《汉

书·艺文志·诸子略考释》中对《诸子略》中学派、学者、著述的存、佚、伪考释结论^[16]。表1为最终选定的类目与典籍,将六家十部典籍作为训练语料,将《荀子》和《管子》两部典籍作为辨别语料。

本文在学衡数据网站^①下载上述典籍的HTML格式文本,用Python调用bs4中的Beautiful Soup 4.4.0解析文件,将其转换为按类目、书名、篇、章、正文组织的结构化文本。经过数据清洗后,将3220条训练数据按照8:2的比例随机划分训练集和验证集,详见表2。

表1 先秦诸子类目及对应典籍

语料类型	类目	典籍
训练语料	儒家	论语、孟子
	道家	老子、庄子
	法家	韩非子、商君书
	名家	公孙龙子
	墨家	墨子
	兵家	孙子、吴子
辨别语料	荀子、管子	

表2 清洗后的先秦典籍篇章数据统计

类别标签	学派	典籍	篇数	章数
0	儒家	论语	20	492
		孟子	7	246
1	道家	老子	81	81
		庄子	33	482
2	法家	商君书	26	146
		韩非子	55	765
3	名家	公孙龙子	6	9
4	墨家	墨子	53	836
5	兵家	孙子	14	96
		吴子	6	67
待分辨	待分辨	荀子	32	511
		管子	75	937

① 学衡数据中文核心典籍数据网址:<http://core.xueheng.net>

3.2 模型介绍

本文使用 TextCNN 和 BERT 两种模型进行分类实验。模型基于文本语义建模,将句子、段落向量化地表示,据此训练神经网络可以捕捉到丰富的语义及结构信息,是目前常用的文本分类方法。TextCNN 模型需要先分词再训练词向量,因古文分词技术尚不成熟,故本文分别尝试使用 Jieba 和单字两种分词方法,然后利用 word2vec 的 CBOW 模型训练词向量。BERT 模型是采用双向深度 Transformer 编码器的预训练

语言模型^[17,18],无需分词即可将句子或段落向量化表示,本文在 BERT-Base-Chinese 预训练模型的基础上继续微调训练。两种模型分别在原始语料、去标点语料、去标点和停用词语料上各重复运行 5 次,每次随机划分训练集与验证集,取 5 次中的 Train-accuracy、Val-accuracy、F1 的平均值和最优值进行比较。

3.3 分类评价

分类训练实验结果如表 3 所示。

表 3 分类训练结果

预处理策略	评价指标		TextCNN 单字 (%)	TextCNN Jieba (%)	BERT (%)
原始	Train-accuracy	mean	67.19	66.79	97.74
	Val-accuracy	max	68.01	68.25	97.86
	F1	mean	60.96	61.49	91.64
		max	61.18	62.58	92.70
		mean	58.52	59.52	91.91
		max	58.81	60.50	93.00
去标点	Train-accuracy	mean	61.40	39.65	96.57
		max	61.68	40.99	96.78
		mean	57.17	37.61	90.31
	Val-accuracy	max	58.39	40.37	91.15
		mean	53.79	31.48	90.43
		max	56.01	34.32	91.18
去标点停用词	Train-accuracy F1	mean	56.74	31.18	95.80
		max	57.22	35.25	95.96
	Val-accuracy	mean	53.73	31.64	89.54
		max	55.28	34.78	90.84
	F1	mean	49.98	19.55	89.06
		max	51.60	24.95	89.79

实验结果表明,在其他超参数相同的情况下,TextCNN 与 BERT 均在不过滤标点和停用词的原始语料上达到最优分类效果,预处理复杂度的提高反而会带来准确率和 F1 值较大幅度的下降,这一特点在利用 Jieba 分词训练 TextCNN 分类模型时尤为显著,在过滤掉标点和停用词后该分类模型平均准确率由 61.49% 下降到 31.64%。这一现象说明古文的预处理策略会在较大程度上影响分类模型的效果,对于本文选取的先秦诸子语料而言,保留标点与虚词可以有效提高分类模型的训练效果。这一发现验证了 Uysal 和 Gunal 此前的研究结论,他们利用 SVM 模型分别在英语和土耳其语语料上训练分类器,发现保留停用词的语料预处理策略可以获得最高的准确率^[19]。对比 BERT 和 TextCNN 可以发现,BERT 的分类效果明显好于 TextCNN,在原始语料上 BERT 的 F1 值为 91.91%,远高于 TextCNN-Jieba 的 59.52%。BERT 的另一个优势在于无需分词,以单字为单元编码训练就可以取得较高的分类准确率,规避了目前古文分词不成熟的问题。

综上,在原始文本上,BERT 分类训练可以达到最优的效果,其平均准确率可以达到

91.64%,这说明预训练语言模型可以基于古籍文本有效分辨先秦不同学派的典籍。下文将进一步利用在这十部典籍上微调训练后的 BERT 模型对《荀子》与《管子》做互著与别裁分析。

4 互著与别裁实验

4.1 《荀子》实验与讨论

《荀子》位于目录体系中的儒家类,但其学派归属问题始终是重要的研究内容。唐代韩愈指出荀子思想“不粹”,宋代二程及朱熹等理学家认为荀子思想“极偏颇”并将其归为法家,现在也仍有儒家说、法家说、杂家说等观点并存^[20]。本研究利用机器学习方法对《荀子》做互著与别裁探索,可以为该问题的解决提供新的视角和维度。

4.1.1 《荀子》互著结果及讨论

将《荀子》511 个章节的文本输入 BERT 模型后可以得到各条文本的分类结果,将各条文本划分到概率最大的类别中,得到表 4。若将各条文本对应于各学派的概率进行统计汇总,可以绘制如图 2 所示的分类概率箱型图。

表 4 《荀子》分类鉴别结果统计

学派	章节	章节占比 (%)	平均字符数	总字符数	总字符占比 (%)
儒家	134	26.22	119	15 978	17.24
道家	43	8.41	188	8 102	8.74
法家	259	50.68	208	53 780	58.02
名家	1	0.20	167	167	0.18
墨家	61	11.94	190	11 620	12.54
兵家	13	2.54	234	3 039	3.28

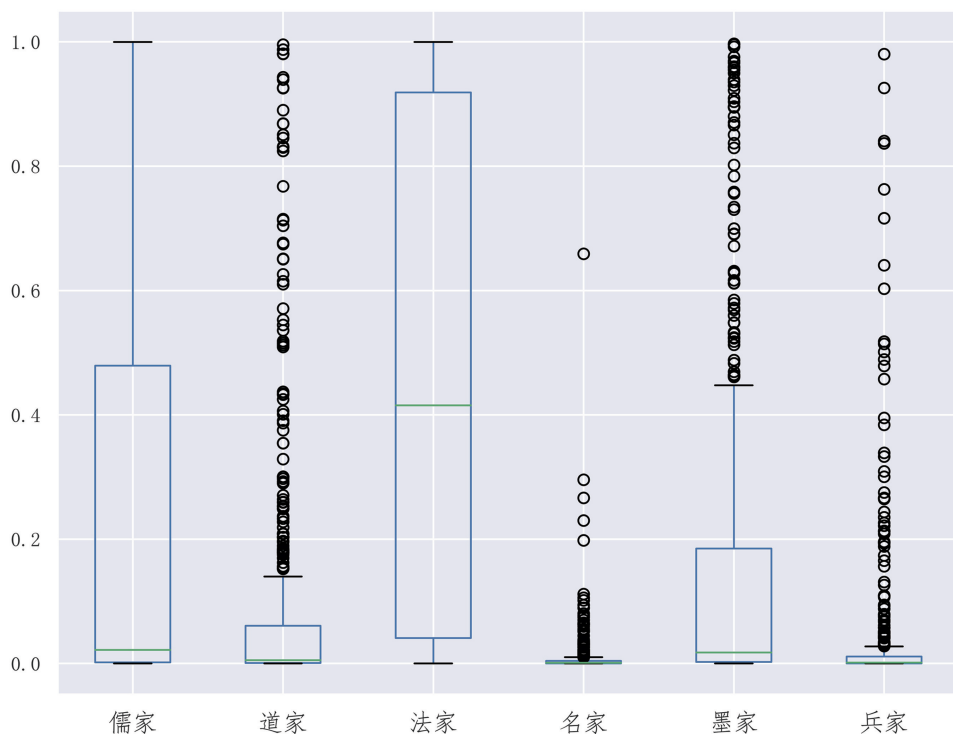


图2 《荀子》分类概率箱型图

根据表4和图2,《荀子》最接近法家与儒家,同时也兼及其他学派,这与传统研究结论相符。对于儒法之争的问题,分类模型将更多的条目判断为法家,揭示出《荀子》更接近以《韩非子》和《商君书》代表的法家而非以《论语》和《孟子》代表的儒家。本文以“均值比较”确定互著阈值,即若文本属于某一类的概率超过平均分类的概率,则可判断文本属于该类。经计算,法家与儒家的概率大于16.67%,分别为46.68%和24.95%,故将《荀子》在法家和儒家两个类目中做互著标记。

此互著结果体现了荀子“礼法并施”的思想特征。荀子在继承儒家的礼制思想的同时,也由性恶论的观点出发,认为刑、罚等法家手段是治国所必需,如《荀子·劝学》篇有“礼者,法之大分,类之纲纪也”^[21]。这一互著结果也呼应了荀子研究中一直存在的儒法之辨。吕思勉在《经子解题》中评价“语其宗旨,实与法家最近,

而又蒙儒家之面目者也”^[22];赵法生认为荀子属于受儒家影响的法家,即“儒法”^[23];王正提出荀子的“法”同法家的“法”有所不同,是带有法家因素的儒家,最多称为“法儒”^[24]。本文通过机器学习模型得到的量化结果也如实反映出这一特点。

4.1.2 《荀子》别裁结果及讨论

本文借助由机器学习模型得到的二维热力图(见图3)分析《荀子》内部篇章所属学派,由此做出别裁判断。每一篇对应六个学派的概率和为1,具体对应每个学派的概率等于该篇中各章节对应各学派概率的平均值,颜色越深说明概率越大,越接近该类目。

在《荀子》的绝大多数篇目中,法家对应颜色最深,说明这些篇目更接近于法家,而儒家、墨家、道家次之,与互著结果一致。以《荀子》互著于法家和儒家为基础,可以为《荀子》具体篇目做出别裁标注,结果如表5所示。

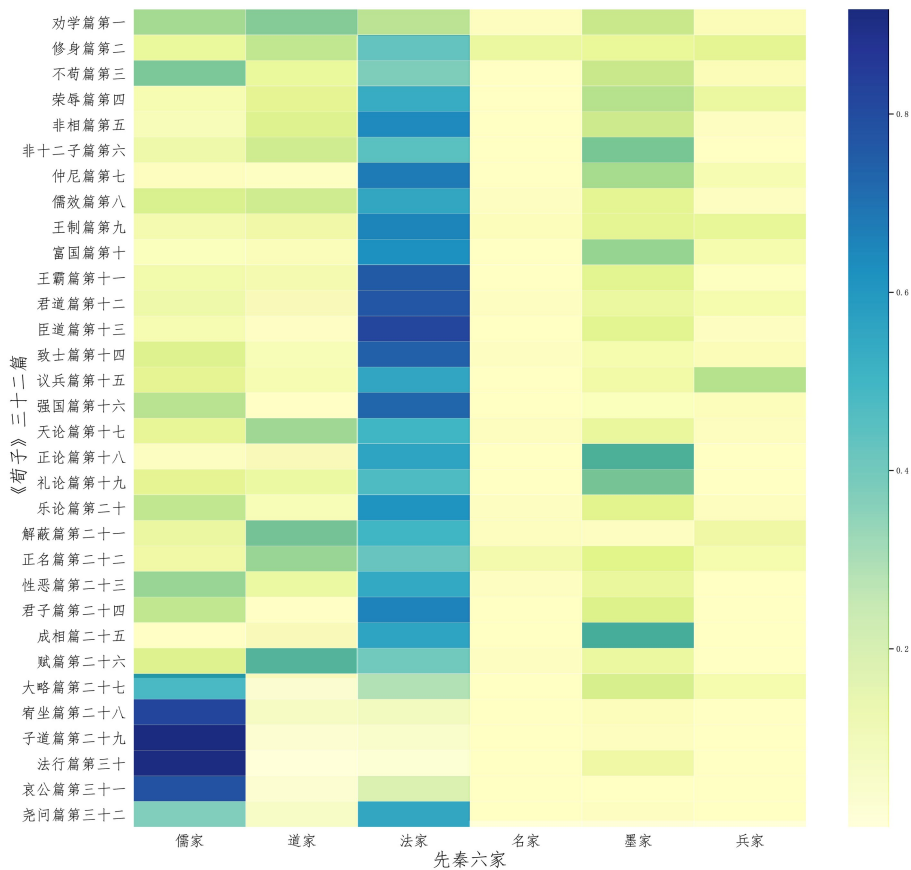


图3 《荀子》三十二篇对应六家概率分布热力图

表5 《荀子》别裁结果

类 目	《荀子》篇目
道家	劝学、修身、儒效、天论、解蔽、正名、赋
墨家	劝学、不苟、荣辱、非十二子、仲尼富国、正论、礼论、成相
兵家	议兵
名家	无

虽然此前没有学者对《荀子》三十二篇的每一篇都分析思想归属,而本文所给出的别裁建议也未必绝对准确,但若将传统与数字化两种方法得到的结论相比较,也许可以通过“三角论证”取得一些新的突破。对于本文所列举的道家篇,赵吉惠在《荀子非儒家辨》中列举《天论》

《劝学》《解蔽》《儒效》《正名》篇论述荀子的道家思想倾向^[25],李刚兴列举《天论》和《解蔽》论述荀子对道家思想的吸收与改造^[26],许倩撰文论述《解蔽》中的道家思想^[27],雷震和郭成杰从养生观的角度论述《修身》篇同老庄的相似之处^[28]。对于本文所列举的墨家篇,刘宝春指出

《大略》《不苟》《荣辱》《富国》《君德》《王霸》《儒效》《强国》《君子》《君道》《臣道》《正论》《劝学》《性恶论》《成相》中或多或少地带有墨子“利民”“贵义”“尚贤”“节用”“所染”思想^[29],赵吉惠也指出《荣辱》篇有墨家早期的义利观体现^[25]。对于本文所列举的兵家篇,《议兵》是荀子军事思想的主要体现,也是学者们研究荀子军事思想的主要依据^[30,31]。

此外,模型揭示出以《大略》为分界,《荀子》前后文本存在“断层”,即前二十六篇更侧重法家思想,后六篇侧重儒家思想。《荀子》后六篇大量引述孔子及弟子言论,内容涵盖仁、礼、义、信、孝道等儒家观念。唐代杨倞指出,《大略》《宥坐》《子道》《法行》《哀公》《尧问》并非荀子本人所作,认为《大略》篇为“弟子杂录荀卿之语”,《宥坐》《子道》《法行》《哀公》《尧问》五篇“皆荀卿及弟子所引记、传杂事”^[32]。董志安指出后五篇“与荀卿基本学说不尽相符,而多少带有其他儒家派别的思想痕迹”,由此判断后五篇既不是出自荀子本人,也不是出自荀子传授^[33]。廖名春认为,《荀子》后五篇包括荀子整理和纂集的资料及弟子之作,提出考察荀子思想应主要以前二十七篇为依据^[34]。本文模型发现的“断层”现象揭示出《荀子》文本前后思想的差异,与上述传统研究结论相符,能对此类研究形成有益补充。

4.2 《管子》实验与讨论

《管子》是对管仲学派思想的记载,书中记载的思想较为复杂,涉及多个学派,在学术史研

究中始终有学派分属的争议,并且这种争议也体现在目录类目的变化中^[35]。如《管子》在《汉书·艺文志·诸子略》中记载在道家类目,在《隋书·经籍志》《四库全书》中列为法家类目,李宗邺、潘俊杰等人主张《管子》是杂家的汇集^[36,37]。

本文利用机器学习方法对《管子》提出互著与别裁建议,助力于《管子》思想问题的讨论。

4.2.1 《管子》互著结果及讨论

将《管子》937个章节文本输入BERT模型后可以得到各条文本的分类结果,将各条文本划分到概率最大的类别中,得到表6,若将各条文本对应于各学派的概率进行统计汇总,可以绘制如图4所示的分类概率箱型图。

根据表6和图4,《管子》最接近法家,与墨家也较为密切。在“均值比较”规则下,《管子》中仅法家一类超过了16.67%,为70.07%。因此仅将《管子》在法家一个类目中做标记,不适合做互著标引。这说明《管子》内容虽兼容并包,但法家思想在其中占绝对主导,仅适合著于法家一类,这与《管子》在《隋书·经籍志》《四库全书》等目录中的分类一致。

4.2.2 《管子》别裁结果及讨论

本文借助由机器学习模型得到的二维热力图(见图5)分析《管子》内部篇章所属学派,由此做出别裁判断。每一篇对应六个学派的概率和为1,具体对应每个学派的概率等于该篇中各章节对应各学派概率的平均值,颜色越深即概率越大,越为接近该类目。

表6 《管子》分类鉴别结果统计

学 派	章 节	章节占比(%)	平均字符数	总字符数	总字符占比(%)
儒家	31	3.31	409	12 666	2.88
道家	69	7.36	406	28 035	6.37
法家	682	72.79	511	348 614	79.15
名家	0	0	0	0	0
墨家	113	12.06	366	41 305	9.38
兵家	42	4.48	234	9 812	2.23

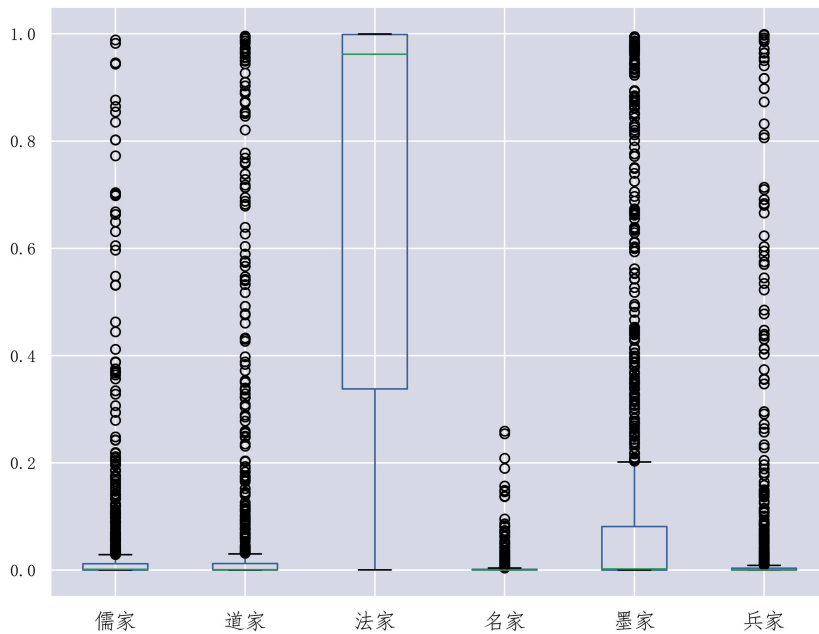


图4 《管子》分类概率四分位图

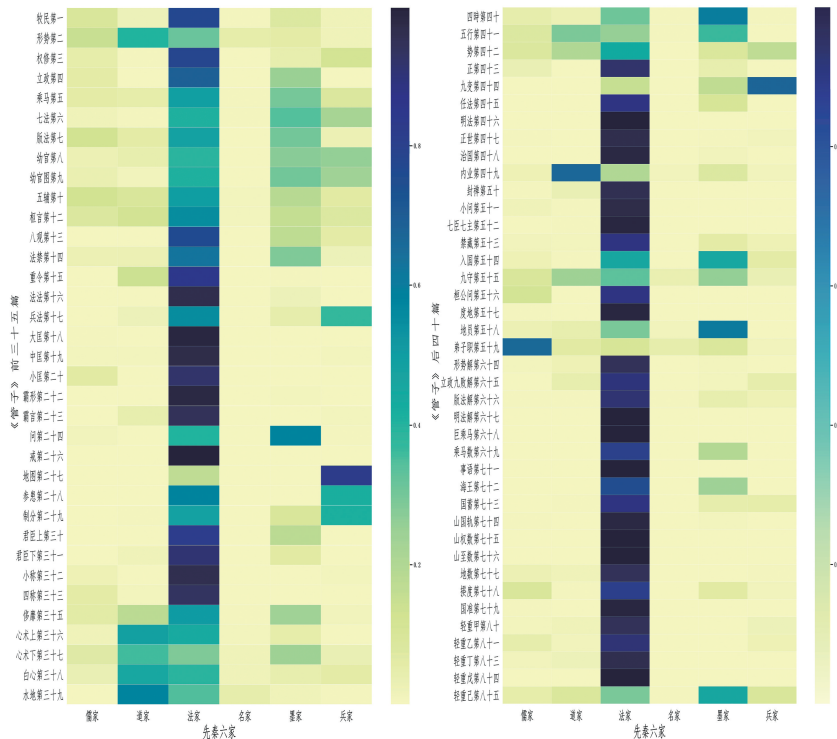


图5 《管子》七十五篇对应六家概率分布热力图

在《管子》绝大多数篇目中,法家对应颜色最深,说明这些篇目更接近于法家,而墨家、道家次之,这与《管子》互著结果一致。以《管子》

仅著录于法家类为基础,可以为《管子》具体篇目做出别裁标注,结果见表7。

表7 《管子》别裁结果

类目	《管子》篇目
儒家	弟子职
道家	形势、侈靡、心术上、心术下、白心、水地、五行、势、内业、九守
墨家	立政、乘马、七法、版法、幼官、幼官图、五辅、八观、法禁、问、君臣上、侈靡、心术下、四时、五行、入国、九守、地员、乘马数、海王、轻重己
兵家	七法、幼官、幼官图、兵法、地图、参患、制分、势、九变
名家	无

同样,虽然此前没有学者对《管子》七十五篇的每一篇都分析思想归属,而本文所给出的别裁建议也未必绝对准确,但若将两种方法得到的结论相比较,也许可以通过“三角论证”取得一些新的突破。对于本文所列举的儒家篇,冯友兰指出《弟子职》与《弟子规》实属一类^[38],罗根泽认为《弟子职》的作者疑是汉代儒家^[39],张连伟虽不赞同汉代儒家说,但同样认为《弟子职》体现了儒家所重视的学堂礼仪^[40]。对于本文所列举的道家篇,《管子》四篇《心术上》《心术下》《白心》《内业》所体现的道家思想已是一种共识。但是,模型也得到了一些不同于现有观点的结果,如罗贤龙判断《法法》《仁法》《明法》属于黄老作品^[41],而本文将这几篇判列为法家类,这种差异值得关注。这种差异同样存在于《管子》的墨家思想:模型判别《管子》中大量篇章都与墨家思想相近,但是传统研究对《管子》墨家思想的研究较少,本文通过机器学习模型揭示出《管子》与《墨子》思想之间的相似性,还值得未来更多的探讨。对于本文所列举的兵家篇,池万兴指出《幼官》《七法》《参患》《地图》《制分》《九变》《兵法》是兵家文字^[42],罗根泽指出《七法》《兵法》《制分》为兵家所出,《幼官》和《势》为兵阴阳(兵家四大流派之一)所出^[39],梁启超列举《七法》《九变》《兵法》《霸言》《小匡》中含有军政思想^[43]。由此可知,模型得到的量

化结论,既可与已有研究结论相印证,也存在不同之处,这些正是值得未来更多探讨及批判的靶向所在。

5 结语

古籍是承载中华优秀传统文化的重要载体,合理有效地保护与利用古籍资源是弘扬中华文化、树立民族自信的重要举措。本文借助机器学习技术,从古籍资源中挖掘出新的知识,能为古籍研究提供新的方法与论据,让古籍资源焕发出新的生机,让古籍“活起来”,并助力于科学研究和社会文化的进步。

本文提出利用机器学习来实现互著与别裁,从全书、篇、章等粒度辨析古籍类别,并对先秦诸子典籍进行了实验。实验结果不仅显示了机器学习对古籍文本分类的潜力,也对《荀子》及《管子》两本书给出了具体的互著与别裁建议,此外实验结果展现了在数字人文视域下,数字技术对古典目录学、古典文献学以及学术史研究的应用价值。

鉴于时间和语料等方面的限制,本文仍存在一定的不足和局限性,未来的研究可以从以下三个方面展开:在应用范围方面,进一步探究在更大规模的古籍目录及语料上的适用性;在机器学习方法方面,可以让BERT模型对大规模古籍语料做进一步的预训练,或结合更准确的

分词策略提高 TextCNN 的表现,也可以采用更为先进的机器学习模型;在互著与别裁判断策略上,进一步将此前依靠主观模糊性判断的习惯

例用阈值具体化、定量化表示出来,从而制定更合理的判别规则。

参考文献

- [1] 陈晓华. 传统目录学仍是“学中要紧事”[N]. 中国社会科学报, 2016-08-18(001). (Chen X H. Traditional bibliography is still “the most important thing in learning”[N]. Chinese Social Sciences Today, 2016-08-18(001).)
- [2] 李景文. “互著”、“别裁”起源时间考辨——读王重民先生《校讎通义通解》[J]. 图书情报工作, 2012, 56(7): 140-144. (Li J W. The origins of “exchange indexing” and “reindexing”: with a study of Wang Chongmin’s *Jiao Chou Tong Yi Tong Jie*[J]. Library and Information Service, 2012, 56(7): 140-144.)
- [3] 王国强. 中国古代书目著录中的互著法和别裁法[J]. 郑州大学学报(哲学社会科学版), 2002(4): 130-133. (Wang G Q. The inter-record method and analytic-record method in ancient Chinese booklist record[J]. Journal of Zhengzhou University (Philosophy and Social Sciences Edition), 2002(4): 130-133.)
- [4] 韩建立.《艺文类聚》中的“互著”与“别裁”[J]. 图书馆学刊, 2014, 36(4): 117-119. (Han J L. Inter record and analytic record in *Yi Wen Lei Ju*[J]. Journal of Library Science, 2014, 36(4): 117-119.)
- [5] 章学诚. 校讎通义[M]. 北京: 古籍出版社, 1956. (Zhang X C. *Jiao Chou Tong Yi*[M]. Beijing: Classics Publishing House, 1956.)
- [6] 罗友松, 朱浩. “互著”、“别裁”的理论探讨始于谁?——与徐召勋同志商榷[J]. 图书馆杂志, 1982(1): 12-13. (Luo Y S, Zhu H. Who started the theoretical discussion of “inter record” and “analytic record”? Discussing with comrade Xu Zhaoxun[J]. Library Journal, 1982(1): 12-13.)
- [7] 章学诚. 校讎通义通解[M]. 上海: 上海古籍出版社, 2009. (Zhang X C. *Jiaochou Tongyi Tongjie*[M]. Shanghai: Shanghai Classics Publishing House, 2009.)
- [8] 徐召勋. 论“互著”与“别裁”——章学诚目录学思想初探[J]. 安徽大学学报, 1979(3): 102-110. (Xu Z X. On “inter record” and “analytic record”: a preliminary study of Zhang Xuecheng’s bibliography thoughts[J]. Journal of Anhui University, 1979(3): 102-110.)
- [9] 子厚. “互著”与“别裁”[J]. 国家图书馆学刊, 2003(4): 26. (Zi H. “Inter record” and “analytic record”[J]. Journal of the National Library of China, 2003(4): 26.)
- [10] 陈志新. 中国目录学传统的当今表现——目录学去哪儿了?[J]. 图书馆, 2020(7): 7-9. (Chen Z X. The present performance of Chinese ancient bibliography: where has bibliography gone?[J]. Library, 2020(7): 7-9.)
- [11] 李瑞龙, 李明杰. 数字人文背景下古典目录学“辨考”思想的实现路径初探[J]. 山东图书馆学刊, 2019(1): 14-19, 40. (Li R L, Li M J. Probe on the realization path of the Biankao thought of Chinese classical bibliography under the background of digital humanities[J]. The Library Journal of Shandong, 2019(1): 14-19, 40.)
- [12] 李惠, 陈涛, 侯君明, 等. 钩玄提要——古籍目录智能分析工具构建[J]. 中国图书馆学报, 2021, 47(4): 97-112. (Li H, Chen T, Hou J M, et al. Noting the essential: an explorative tool for catalog annotations in Chinese rare-book collections[J]. Journal of Library Science in China, 2021, 47(4): 97-112.)
- [13] Maron M E. Automatic indexing: an experimental inquiry[J]. Journal of the ACM, 1961, 8(3): 404-417.
- [14] 王昊, 严明, 苏新宁. 基于机器学习的中文书目自动分类研究[J]. 中国图书馆学报, 2010, 36(6): 28-39.

- (Wang H, Yan M, Su X N. Automatic classification of Chinese bibliography based on machine learning[J]. Journal of Library Science in China, 2010, 36(6): 28-39.)
- [15] 张馨怡. 基于 TextCNN 的古典诗词爱国情怀研究[D]. 上海: 上海师范大学, 2020. (Zhang X Y. Research on patriotic feelings of classical poems based on TextCNN[D]. Shanghai: Shanghai Normal University, 2020.)
- [16] 梁启超. 汉书艺文志诸子略考释[M]. 北京: 中华书局, 2015. (Liang Q C. Hanshu Yiwenzhi Zhuzilue Kao shi[M]. Beijing: Zhonghua Book Company, 2015.)
- [17] 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.
- [18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, 2017: 5998-6008.
- [19] Uysal A K, Gunal S. The impact of preprocessing on text classification[J]. Information Processing and Management, 2014, 50(1): 104-112.
- [20] 郭士榕, 关键英. 荀子归属之争看荀子对孔孟儒学的转化与创新[J]. 学术交流, 2019(10): 70-78. (Guo S R, Guan J Y. Probe into Xunzi's transformation and innovation of Confucius and Mencius Confucianism in the dispute of his school attribution[J]. Academic Exchange, 2019(10): 70-78.)
- [21] 刘柱彬. 仁政理想与礼法并施的冲突与融合——孟子与荀子刑法思想之比较[J]. 法学评论, 2001(3): 124-131. (Liu Z B. The conflict and fusion between the ideal of benevolent government and the combination of rites and laws: a comparison between Mencius and Xunzi's thought of criminal law[J]. Law Review, 2001(3): 124-131.)
- [22] 吕思勉. 经子解题[M]. 北京: 中国书籍出版社, 2006. (Lv S M. Jingzi Jietyi[M]. Beijing: China Book Press, 2006.)
- [23] 赵法生. 荀子的政制设计与学派归属[J]. 哲学研究, 2016(5): 68-77, 129. (Zhao F S. Master Xun's political system design and the attribution of his school[J]. Philosophical Research, 2016(5): 68-77, 129.)
- [24] 王正. “法儒”还是“儒法”?——荀子与法家关系重估[J]. 哲学研究, 2017(2): 40-47, 129, 128. (Wang Z. Legalist-Confucian or Confucian-Legalist? Reconsidering the relationship between Xunzi and the Legalists[J]. Philosophical Research, 2017(2): 40-47, 129, 128.)
- [25] 赵吉惠. 荀子非儒家辨[J]. 管子学刊, 1990(4): 29-36, 19. (Zhao J H. Xunzi is not Confucian[J]. Guan Zi Journal, 1990(4): 29-36, 19.)
- [26] 李刚兴. 简析荀子对道家思想的吸收和改造[J]. 四川师范学院学报(哲学社会科学版), 1990(2): 77-80, 94. (Li G X. A brief analysis of Xunzi's absorption and transformation of Taoism[J]. Journal of China West Normal University (Philosophy & Social Sciences), 1990(2): 77-80, 94.)
- [27] 许倩. 从《解蔽》看荀子对道家思想的融会与发展[J]. 邢台学院学报, 2004(4): 31-34. (Xu Q. Jiebi: an interpretation of Xunzi's digestion and development of Taoism[J]. Journal of Xingtai University, 2004(4): 31-34.)
- [28] 雷震, 郭成杰. 荀子的养生哲理与体育教育思想[J]. 体育文化导刊, 2004(11): 72-73. (Lei Z, Guo C J. Xunzi's philosophy of health and physical education thoughts[J]. Sports Culture Guide, 2004(11): 72-73.)
- [29] 刘宝春. 谈墨学对荀子思想的影响[J]. 济宁学院学报, 2010, 31(2): 5-8. (Liu B C. Analysis on the influence of Mohism on Xunzi[J]. Journal of Jining University, 2010, 31(2): 5-8.)
- [30] 李桂生. 先秦兵家研究[D]. 杭州: 浙江大学, 2005. (Li G S. Research on the military strategists in pre-Qin [D]. Hangzhou: Zhejiang University, 2005.)

- [31] 黄朴民. 荀子军事思想简述[J]. 邯郸学院学报, 2013, 23(4): 9-12. (Huang P M. Brief introduction to Xunzi's military thought[J]. Journal of Handan University, 2013, 23(4): 9-12.)
- [32] 王先谦. 荀子集注[M]. 北京: 中华书局, 1988. (Wang X Q. Xunzi Jizhu[M]. Beijing: Zhonghua Book Company, 1988.)
- [33] 董治安. 先秦文献与先秦文学[M]. 济南: 齐鲁书社, 1994. (Dong Z A. Pre-Qin document and pre-Qin literature[M]. Jinan: Qilu Press, 1994.)
- [34] 廖名春. 《荀子》各篇写作年代考[J]. 吉林大学社会科学学报, 1994(6): 52-57. (Liao M C. An analysis of the writing age of Xunzi[J]. Jilin University Journal Social Sciences Edition, 1994(6): 52-57.)
- [35] 尹清忠. 《管子》研究[D]. 曲阜: 曲阜师范大学, 2009. (Yin Q Z. Research on Guanzi[D]. Qufu: Qufu Normal University, 2009.)
- [36] 李宗邨. 中国历史要籍介绍[M]. 上海: 上海古籍出版社, 1982. (Li Z Y. An introduction to important Chinese historical books[M]. Shanghai: Shanghai Classics Publishing House, 1982.)
- [37] 潘俊杰. 《管子》为杂家著作考[J]. 管子学刊, 2007(3): 19-22. (Pan J J. Making textual criticism of Guan zi be a masterpiece of pre-Qin eclectics[J]. Guanzi Journal, 2007(3): 19-22.)
- [38] 冯友兰. 冯友兰论管子[M]//司马琪. 十家论管. 上海: 上海人民出版社, 2008. (Feng Y L. Feng Youlan comments on Guanzi [M]//Sima Q. Ten reviews on Guanzi. Shanghai: Shanghai People's Publishing House, 2008.)
- [39] 罗根泽. 管子探源[M]. 长沙: 岳麓书社, 2010. (Luo G Z. Guanzi Tan Yuan[M]. Changsha: Yuelu Press, 2010.)
- [40] 张连伟. 《弟子职》与先秦的学校教育[J]. 中共石家庄市委党校学报, 2008(2): 29-31. (Zhang L W. Di Zi Zhi and pre-Qin school education[J]. Journal of the Party University of Shijiazhuang City Committee of CPC, 2008(2): 29-31.)
- [41] 罗贤龙. 《管子》黄老治道思想研究[D]. 合肥: 安徽大学, 2015. (Luo X L. A study on the governance of emperor Huang and Laozi in Guanzi[D]. Hefei: Anhui University, 2015.)
- [42] 池万兴. 《管子》研究[D]. 兰州: 西北师范大学, 2003. (Chi W X. The research of Guanzi[D]. Lanzhou: Northwest Normal University, 2003.)
- [43] 梁启超. 梁启超论管子[M]//司马琪. 十家论管. 上海: 上海人民出版社, 2008. (Liang Q C. Liang Qichao comments on Guanzi [M]//Sima Q. Ten reviews on Guanzi. Shanghai: Shanghai People's Publishing House, 2008.)

张力元 北京大学图书馆馆员。北京 100871。

王 军 北京大学信息管理系教授, 博士生导师。北京 100871。

(收稿日期: 2021-09-13; 修回日期: 2021-12-14)