计算人文视域下的《史记》三家注引书知识标注与计量分析初探*

□齐月　刘雏菲　李文祺　孟凯　王东波　刘浏

摘要　基于古籍文本知识挖掘和知识库构建、围绕数据分析与可视化呈现等视角展开的计算人文探索，已逐渐成为古籍保护和研究利用的重要方向。计算人文视域下的古籍引书研究能够为传统研究问题带来新思路与新技术，拓宽古籍研究视角，提供可靠数据支撑。本文以人工标注结合深度学习的方法，对《史记》三家注中的引书知识进行了标注研究，随后分别从分类视角和三家注对比的视角出发统计并呈现了引书和引用作者的分布规律。本研究以《史记》三家注为例，形成了一套完整的古籍引书知识标注技术流程和框架，并将统计计量和可视化分析方法引入了古籍引书研究，对于推动和完善《史记》研究和古籍引书研究均具有参考价值。

关键词　计算人文　数字人文　古籍引书　《史记》三家注　文本知识挖掘
分类号　G250
DOI　10.16603/j.issn1002－1027.2024.05.008

1　引言

随着古籍数字化工作的快速发展，以及相关自然语言处理技术的逐渐成熟，面向古籍的计算人文研究得到了广泛的关注。较之于更偏向主观思辨与内省方法的传统人文学科研究，计算人文能够从量化角度进行研究，引入更丰富的客观证据，以计算为方法，对数据资源进行重构，使其成为包含更丰富知识的结构化的知识库。同时，数据化阐释和可视化呈现的方式，让客观的数据更直观易懂、更容易验证[1]。

古籍引书是中国传统文献学的重要内容，古人撰书历来以旁征博引为豪，这些所征引文献来源于书籍或独立篇章，本文将其统称为“引书”。在引书中包含着丰富的文献资料和潜在信息。对古籍的引书进行研究，一方面有助于读者更好地理解与把握所注经典本身，另一方面，其涉及的大量古籍文献，对考察文献源流、辑录文献佚文[2]、校勘文献异文[3]，有着不可取代的学术价值。《史记》是我国第一部纪传体通史，历代为之作注者众多，百衲本收录的三家注版本（即南朝宋裴骃《史记集解》，唐朝司马贞的《史记索隐》和张守节的《史记正义》，以下简称《史记》三家注），在古籍引书研究中具有重要价值[4]。然而，《史记》三家注引书内容丰富，关联复杂，传统研究方法仅能窥其一角，难以观其全貌。计算人文视角从方法上弥补了这一缺憾，提供了一个新颖的考察思路，能够发挥数据和计算优势，其有效性也得到了相关研究的初步验证[5]。

本研究从计算人文的视角出发，以《史记》三家注引书为对象，结合人工和深度学习的方法，对《史记》三家注文本中引书知识进行标注、校对、消歧和补全，将古籍引书中的内容和知识进行结构化组织。在此基础上，从多个维度统计和分析了《史记》三家注中的引书知识，包括其在数量、类型等多方面的分布情况，用可视化的方式呈现了其特征。本文可以为计算人文视角的古籍引书研究和《史记》三家注研究提供有效的实证参考。

* 国家自然科学基金青年项目“基于深度学习的典籍引书知识图谱构建及应用研究”（项目编号：72004095）和国家社会科学基金重大项目“中国古代典籍跨语言知识库构建及应用”（项目编号：21&ZD331）的研究成果之一。
通讯作者：刘浏，邮箱：liuliu@njau.edu.cn。

2 相关研究进展

近年来，计算人文不断突破学科界限，向历史、艺术、宗教、哲学等人文学科蔓延，在广度和深度上增强了对人文学科研究内容的认知。目前国内计算人文以古籍为对象的研究最为显著，包括围绕唐诗的相关研究[6-7]，围绕方志物产的研究[8]，对春秋《左传》的相关研究[9-10]，面向四库全书的预训练模型的相关研究[11-13]。对于计算方法与人文研究交叉的这一新兴领域，学界使用较多的表述有计算人文、数字人文、人文计算等，本文认为这些表述指称对象并无差别，而“计算人文”的可解释性更好[14]，因此全文将使用“计算人文”一词，旨在探索计算方法在古籍引书研究中的可行性和有效性。

古籍引书的相关研究主要指的是古籍之间互相引用的情况，传统研究多基于文献学和历史学归纳分析古籍引书特点。通过体式梳理[15]、引用统计[16]、特点对比[17]等方法探索古籍引书规律。随着研究方法的跨学科应用，学者尝试使用计量学方法[18]和机器学习方法[19]来实现计算人文视角下的古籍引书研究，为该领域研究提供了新的思路与方法。

《史记》三家注是一部特殊类型的集注，三种书原来各自单行，宋人刻书，把“三家注”散入正文各句之下，合为一本。三种书先后递补传承，各具特色，相得益彰，从多方面对《史记》进行了诠释和补充，对后人研读有很大帮助。目前对《史记》三家注的研究主要聚焦于文献学、历史学、语言学、校勘学等领域，旨在探索其特点[20-21]、体例[15]、文字[22]、训诂[23]、音韵[24-25]、版本[26]、注解[27]等，并挖掘其在文献学和史料学[4]等领域的价值。

综上所述，传统的古籍引书研究需要人工进行长时间的阅读、整理和归纳，而引入计算人文研究视角，能够借助自然语言处理前沿技术，利用人工结合机器标注的方法，实现对古籍文献的自动化处理和智能化分析，实现对古籍多维度的研究及可视化的呈现。通过对数据的计量分析，能发现古籍引书的规律和趋势，提高古籍研究的效率和准确性。

3 引书知识标注的技术框架

3.1 数据来源

本研究所使用数据为《百衲本二十四史》中的《史记》，该版本《史记》包含了《史记集解》《史记索隐》和《史记正义》三家注解，即《史记》三家注，全书共131卷，约200万字。《史记》三家注的数字化语料来源于北京扫叶科技文化有限公司，该语料包含了与纸质版典籍严格对应的行数和页码，其中还为部分稀有字进行了造字编码处理，具有较高的质量。为保证引书知识标注的准确性和效率，本研究对其进行了初步的人工处理，具体按照《史记》原文和三家注解各占一行的格式进行划分，最终处理结果如图1所示。

治五氣．．↓
（【集解】王肅曰．五行之氣．．↓
【索隱】．謂春甲乙木氣．夏丙丁火氣之屬．是五氣也．）．↓
↓
藝五種．．↓
（【集解】藝．樹也．詩云．藝之荏菽．周禮曰．穀宜五種．鄭玄曰五種．黍．稷．菽．麥．稻也．．↓
【索隱】．藝音魚祭反．種也．樹也．五種即五穀也．音朱用反．此註所引見詩大雅生民之篇．爾雅云．荏菽．戎也．郭璞曰今之胡豆．鄭氏曰．豆之大者是也．
【正義】．藝音魚曳反．種音腫．）．↓

图1 《史记》三家注语料格式

3.2 引书知识的标注、校对和消歧

面向《史记》三家注中最常见的两类引书知识：引书名和引用作者，本研究参考中华书局点校本《史记》（含三家注）[28]进行了一轮标注、两轮校对的工作。首先对全文引书知识进行人工标注，然后基于深度学习下的命名实体识别对标注结果进行第一轮校对，最后再进行人工的第二轮校对工作。人工标注结果如图2所示，BOOK和AUTHOR标签分别对应引书名和引用作者。

司馬相如者．蜀郡成都人也．字長卿．少時好讀書．學擊劍．
（【索隱】/BOOK 呂氏春秋/．劍伎云持短入長．倏忽縱橫之術也．

故其親名之曰犬子．
（【索隱】/AUTHOR 孟康云/．愛而字之也．）

图2 《史记》三家注引书知识语料标注示例

人工标注完成后，本研究进行了两轮校对。第一轮校对工作使用计算机自动完成，该工作可以转换成一种命名实体识别任务，基于深度学习模型实现，以提高校对效率。命名实体识别任务还可以为后续相关的古籍引书自动标注和校对提供参考。本研究选择了 BERT-base-Chinese[29]、SikuBERT[13]、SikuRoBERTa[11]三种深度学习预训练模型，BERT-base-Chinese 是由谷歌基于 BERT 模型采用中文语料库训练而成，对于中文自然语言处理任务有较好的适用性。SikuBERT 和 SikuRoBERTa 是由南京农业大学基于《四库全书》语料推出的面向古籍的领域化预训练模型，对于繁体古籍文本的处理任务性能更佳。在命名实体识别任务中，以引书名和引用作者为实体对象，使用 BIOES 标记将人工标注的文本自动转换成序列化表示形式，如表 1 所示。

表 1　使用 BIOES 标记的引书知识序列化表示示例

实体类型	序列化表示
引用作者	徐 B-AUTHOR
	廣 E-AUTHOR
非实体	曰 O
引书名	春 B-BOOK
	秋 I-BOOK
	合 I-BOOK
	誠 I-BOOK
	圖 E-BOOK
引书名	禮 S-BOOK

命名实体识别的训练和测试语料来源于人工标注的《史记》三家注语料，模型的超参数设置如表 2 所示。

表 2　深度学习模型超参数表

超参数	含义	值
max_seq_length	最大输入序列长度	128
batch_size	每个批次训练的数据量	32
epochs	训练轮次	3
warmup_proportion	预热学习率	0.4

对命名实体识别结果的测试采用十折交叉验证的方法，评测指标包括：准确率 P(Precision)、召回率 R(Recall)和调和平均值 F1(F1-measure)，其结果如表 3 所示。三种预训练模型均表现不俗，SikuRoBERTa 模型的性能略优于 BERT-base-Chinese 和 SikuBERT 模型，引用作者的识别效果明显好于引书名，这也反映出两者之间存在的标注难度差异。根据命名实体识别结果，再结合人工的第二轮校对，可以完成对人工标注的查漏补缺，保证引书知识标注的准确性。

此外，在语料文本中标注得到的引书知识存在较多指称歧义问题，如同名异指、异名同指、缩略名等。面向古籍文本的命名实体歧义消解目前尚不成熟，很难作为自动校对的有效手段，因此本研究参考《史记索隐引书考实》[30]、《史记三家注引书索引》[31]等文献，人工消解了这些歧义，并实现引书名的统一编号，如表 4 所示，为后续的计量分析提供更丰富的参考知识。

表 3　命名实体识别结果

实体类型	模型	准确率(P,%)	召回率(R,%)	调和平均值(F1,%)
引用作者	SikuBERT	96.76	97.68	97.21
	SikuRoBERTa	97.02	97.99	97.49
	BERT-base-Chinese	92.42	94.70	93.49
引书名	SikuBERT	90.10	92.43	91.15
	SikuRoBERTa	90.89	93.13	91.97
	BERT-base-Chinese	86.11	88.62	87.28

表 4　引书名歧义对照表

引书 id	引书名	古籍异名
Z041	春秋左传	左传、春秋左传、春秋左氏、春秋左氏传、左传例、左氏传
S056	世本	世本、居篇、系本
G002	春秋公羊传	公羊传、春秋公羊传、公羊、公羊春秋、公羊文
D001	大戴礼	大戴礼、大戴、大戴记
D010	帝王世纪	帝王世纪、帝王代纪、帝王纪、帝纪
Z035	宗国都城记	宗国都城记、国都记、国都记、城记

4 《史记》三家注引书的统计分析

通过上文所述引书知识的标注、校对和消歧等工作，除去对《史记》自身的引用和其他无关引用，本研究共获取《史记》三家注16971条引书记录及相关引书知识。本研究重点对其中的引书名和引用作者的分布情况进行计量分析。从引用数量来看，《史记》三家注中，引书名实例为6816次，引用作者实例为10313次，其中，一条引书实例同时包含引书名和引用作者的情况并不常见，仅出现158次。本节先考察6816次引书名（以下简称为引书）的分布情况。

4.1 引书分布规律的分类考察

为了细致考察《史记》三家注引书的分布规律，本研究从多个角度分类对引书的频率、种类等情况进行统计分析。一方面，以《史记》三家注的“本纪”“表”“书”“世家”和“列传”五部分（以下简称为五部分）作为分类依据；另一方面，《史记》三家注以《补史记·三皇本纪》开篇，其后“本纪”有十二卷，“表”有十卷，“书”八卷，“世家”三十卷，“列传”七十卷，对应语料中的131卷，本研究还将以卷作为分类考察的另一个视角。

4.1.1 引书频次—种类规律

首先按照《史记》三家注五部分统计引书频次和种类，相关数据如表5所示。可以看出，引书主要集中于“列传”“本纪”和“世家”，“表”的引书频次最少。引书频次在五部分及其所属卷间分布差异明显，这体现出三家注在注释五部分以及各卷内容时，引书的目的或风格存在差异，因此可以借助更多维度的统计数据来考察引书的分布特点，探究其中可能蕴含的规律。本研究还统计了引书的种类数，并依此计算出平均每种引书的引用频率，如表5所示，藉此可以对比五部分引书的分散—集中程度，从数值的意义来看，引书频率可以理解为每种引书的平均被引次数，因此，对于五部分来说，引书频率低说明该部分的引书更分散，引书频率高则说明该部分的引书更集中。具体来看，五部分的引书频率与频次、种类的关系各不一样，比如“书”的引书种类相对较多，引书频次则相对较低，因而引书频率最低，表现出更分散的引书特点；而“世家”的引书频次较高，但引书种类较少，其引书频率最高，表现出更集中的引书特点。

五部分所含卷数各有区别，将这一因素考虑进去，计算出卷均引书频次和卷均引书种类有助于更准确地把握五部分引书的密集—稀疏程度。如表5所示，“本纪”部分的卷均引书频次和卷均引书种类都是最高，呈现出高度密集的引书分布特点，这一特点与引书频次和引书种类基本一致。“书”部分的情况较为特殊，引书频次和引书种类都不高，但其卷均引书种类和卷均引书频次都是第二高的，表现出较密集的引书分布。而“列传”部分虽然引书频次和种类都最高，但其卷均数相对都较低，因而引书分布较为稀疏。

表5 《史记》三家注引书统计（频次—种类）

	本纪	表	书	世家	列传	总计
引书频次	1917	271	578	1688	2362	6816
引书种类	240	33	140	198	307	525
引书频率	7.99	8.21	4.13	8.53	7.69	12.98
卷均引书频次	147.46	27.10	72.25	56.27	33.74	52.03
卷均引书种类	18.46	3.30	17.5	6.60	4.39	4.01

综合《史记》五部分的引书特征，通过象限图呈现其分布，如图3所示。横轴通过五部分的引书频率衡量不同部分引书的分散-集中程度，纵轴通过卷均引书频次和种类表示不同卷引书的密集-稀疏程度。各部分点的具体坐标为：$(x_i-\bar{x},(y_i-\bar{y})*(z_i-\bar{z}))$。其中$x_i$为第i部分的引书频率，$\bar{x}$为五部分引书频率的均值，$y_i$为第i部分的卷均引书频次，$z_i$为第i部分的卷均引书种类数，$\bar{y}$为五部分卷均引书频次均值，$\bar{z}$为五部分卷均引书种类均值。借助象限图不仅能更直观的观察《史记》五部分的引书分布特征，还可以通过散点在坐标轴上的映射反映该部分与均值的偏离程度，同时各部分间相对位置一目了然，有助于考察相近分布及相似规律，如“表”“世家”和“列传”三部分都集中在第四象限，在引书情况上可能存在相似的分布特征。

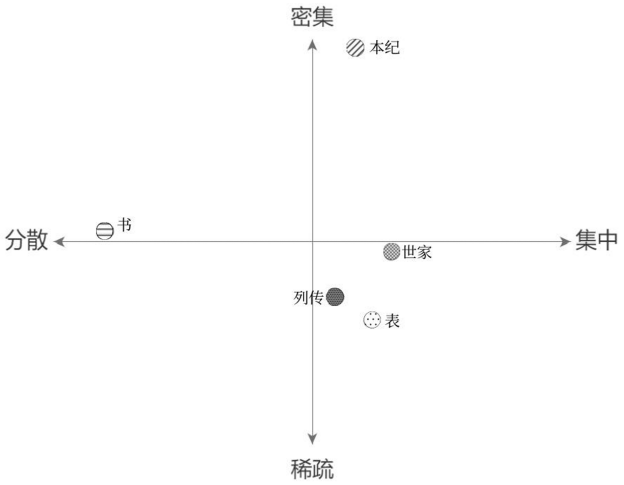

图 3 《史记》三家注引书的分散—集中/密集—稀疏四象限图

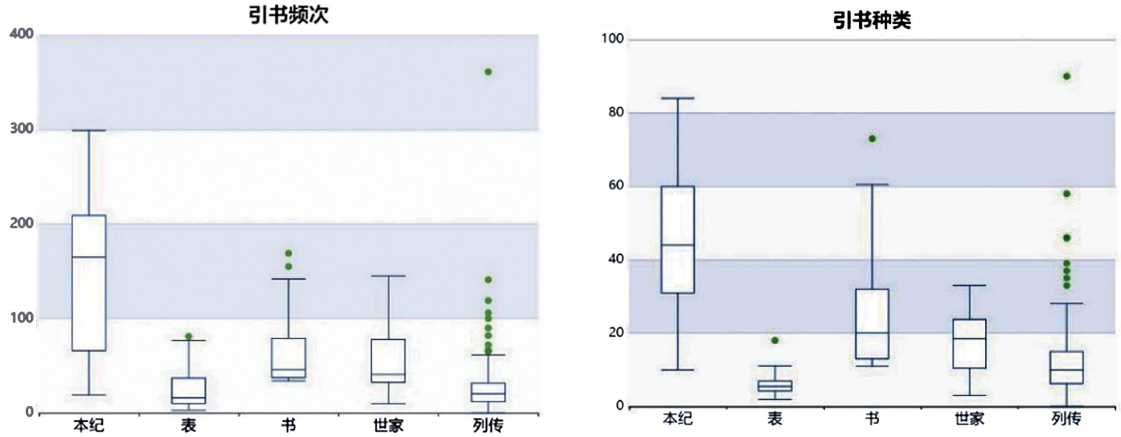

图 4 《史记》三家注引书的频次—种类箱线图

四象限图有助于快速分析出《史记》三家注引书的整体风格，但无法准确揭示出以卷为单位的具体分布情况和差异。将以卷为单位的引书频次和种类的最大值、最小值、中位数和四分位数，通过箱线图呈现出来，有助于直观地观察各组数据的分布范围，便于检测数据异常及数据偏态情况，如图 4 所示。通过图 4 能够清晰发现五部分引书频次和种类分布的彼此差异，“本纪”部分的分布区间最高，印证了其在五部分中最重要的地位。值得注意的是，其余四部分中“书”的区间高于“表”和“列传”，表明了其在引书视角的重要地位，而“列传”在排除异常值后整体区间都较低。

此外，“本纪”部分的引书频次呈左偏分布，种类呈对称分布，均无异常值，分布均较为离散，整体波动较大，这意味着“本纪”中各卷间引书情况差异较大，少数几卷引书较多，如《五帝本纪》《周本纪》等。其余四部分在引书频次上呈明显的右偏分布，且除“世家”外，均在上部分存在异常值，说明存在某一卷或多卷的引书频次特别多，比如“列传”部分的《司马相如列传》，“书”部分的《天官书》和《封禅书》，“表”部分的《惠景间侯者年表》。结合引书种类来看，“世家”部分的引书频次呈右偏分布，而引书种类则呈现左偏分布，说明在“世家”部分中，少部分卷引用了较多种类的书，而引书的频次则分散于大部分卷中。

最后，本节将五部分各卷的引书频次和种类分布情况呈现为二维散点图的形式，散点图可以从另一个视角看出五部分引书分布情况，与箱线图互为参考，用以对上文分析结果进行补充，并考察频次和种类之间的线性规律，如图 5 所示。可以发现，《史记》三家注引书频次和种类成正比关系，五个类别之

间的细微差异可以通过回归方程进行量化对比，并为更深层的引书规律分析提供数据支撑。图中还可以看出各卷引书极值情况，比如“列传”部分在右上角的特殊值对应了第 117 卷《司马相如列传》，这与箱线图的统计结果相一致。此外，图右下角总体部分的方形框线也对应了各部分引书的分布范围，其为引书分布提供了更清晰的数据参考。

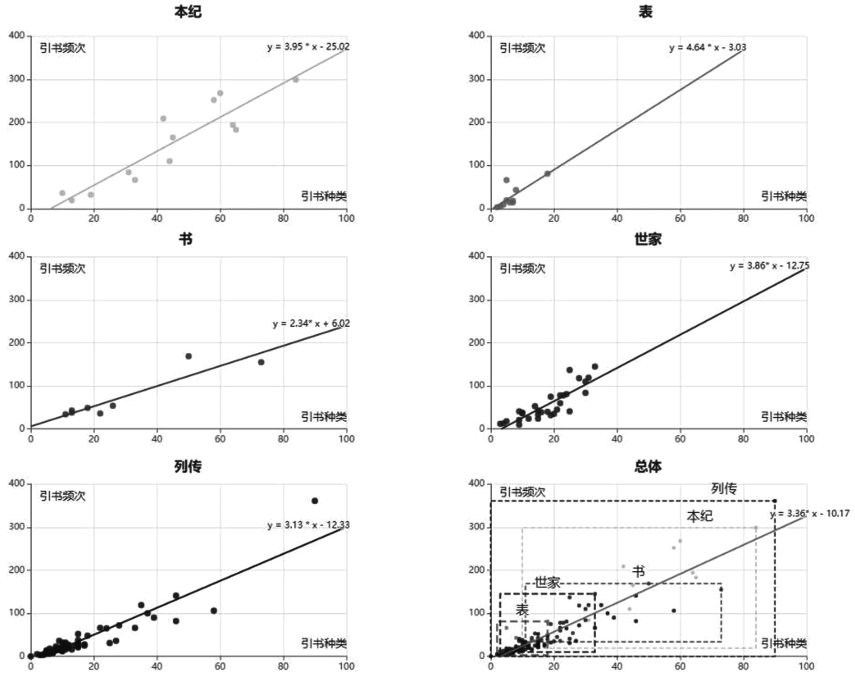

图 5 《史记》三家注引书的线性分布规律对比

4.1.2 高频引书

考察高频引书有助于探寻《史记》关联文献，构建相关引用知识网络，并为更深入进行古籍引书研究提供数据参考。本研究统计高频引书时，利用了引书名歧义消解的结果，同时将篇章名并入其所属书名，《史记》三家注前十高频引书见图 6。高频引书的一些特点值得关注，比如对《汉书》中大量篇名的引用，又比如对《春秋左传》和《汲冢纪年》各种别名的引用。具体举例如下：

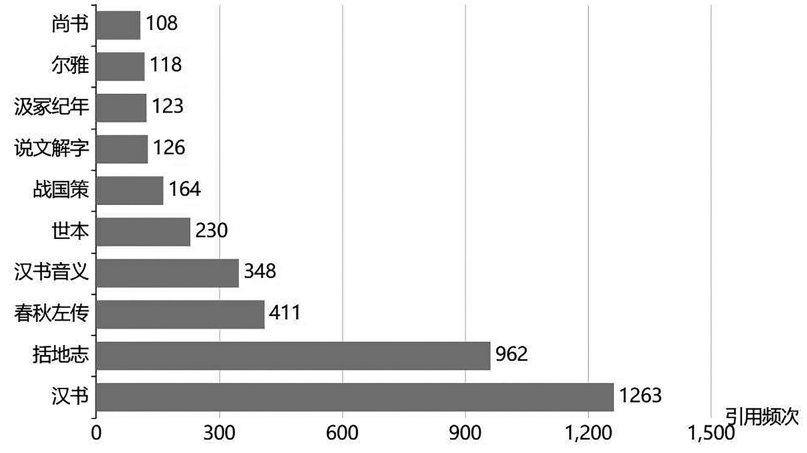

图 6 《史记》三家注的前十高频引书

(1)《汉书》的篇名引用分布。《史记》三家注引书中，《汉书》的被引用频次最高，它是中国第一部纪传体断代史，是继《史记》之后中国古代又一部重要史书。对于《汉书》的引用有两种方式，一是直接引

用《汉书》534次；第二种是引用《汉书》的篇章729次，其中引用“志”最多，共579次，包括《郊祀志》《礼乐志》《天文志》和《地理志》等，其中《地理志》出现了510次，占《汉书》总被引频次的40%左右。

(2)《春秋左传》和《汲冢纪年》的别名引用。《春秋左传》在三家注中共被引411次，以《左传》为主要被引形式，共331次，还包括《左氏》(66次)等其他别名的引用。《汲冢纪年》在三家注中被引形式更为特殊，如表6所示。

表6 《史记》三家注引用《汲冢纪年》的异名频次分布对比

引书名	集解	索隐	正义	总计
及冢古文	1	0	0	1
汲郡古文	1	0	0	1
汲冢	1	0	0	1
汲冢古文	1	1	3	5
汲冢紀年	6	2	2	10
汲冢書	0	0	1	1
汲冢竹書	1	1	0	2
紀年	11	81	4	96
明紀年	0	0	1	1
竹書	0	0	2	2
竹書紀	0	0	1	1
竹書紀年	0	0	2	2
总计	22	85	16	123

4.2 三家注对比下的引书分布规律考察

4.2.1 引书频次—种类规律

综合考察《史记》整体的引书频次与种类后，其引书分布特点已得到较好地呈现。在此基础上，分别统计《史记集解》《史记索隐》《史记正义》(以下分别简称《集解》《索隐》《正义》)三家注各自的引书频次和种类，观察三部注疏文献的引书情况，如图7和表7所示。《正义》在“本纪”部分的引书频次为三家最高，有973次，主要是对《括地志》《帝纪》《左传》的引用；与之相对的是在“表”部分，《正义》仅有一次引用。《集解》在除“表”外的四部分引书频次皆为最少，其引用特点是“数家兼列”，以徐广的《史记音义》为底本，参考并补充汉晋时期其他注家的注文，因此多数引用采用“作者+内容”的格式，而非直接提及具体引书。

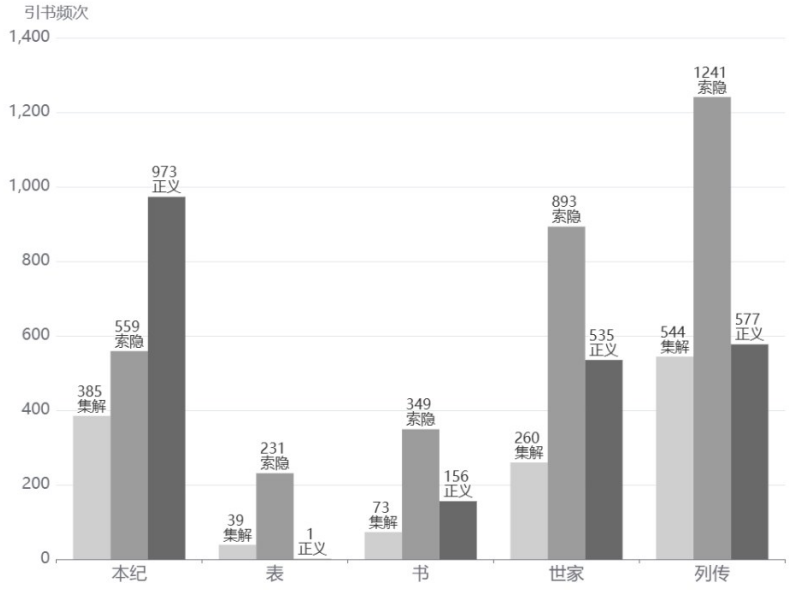

图7 三家注对比下的五部分引书频次

表 7　三家注对比下的引书统计

三家注	引书频次	引书种类	引书频率	卷均引书频次	卷均引书种类
集解	1301	161	8.08	9.93	1.23
索隐	3273	343	9.54	24.98	2.62
正义	2242	283	7.92	17.11	2.16
总计	6816	525	12.98	52.03	4.01

在引书频次的基础上，进一步通过引书频率、卷均引书频次、卷均引书种类三个特征考察三家注的引书分布规律，如图 8 所示，在五部分及各卷间引用差异明显的前提下，三家注视角下的各部分引书分布特征仍有相似之处。如“书”部分，在三家注的引用中都呈现出较为分散的特征；与之相对的是“世家”部分，引书分布较为集中。“本纪”和“列传”部分则分别对应了较为密集和较为稀疏的引书分布。

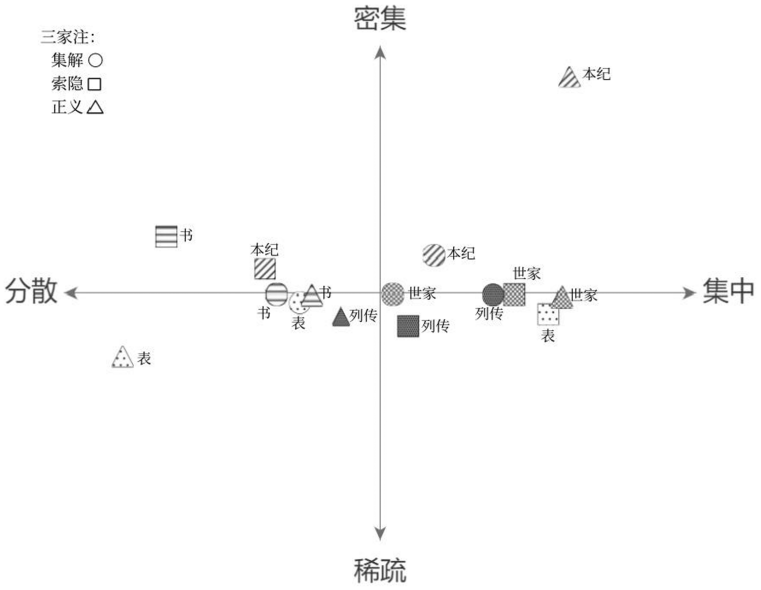

图 8　三家注对比下的引书分散—集中/密集—稀疏四象限图

结合三家注视角的引书分布散点图来看，不难发现三家在注解《史记》时引用习惯各不相同，受引书目的与关注内容的影响。考察三家注的共有引书能够挖掘与《史记》主题内容密切相关的古籍，而特有引书则反映了三家注各自针对《史记》的独特关注点。为探寻三家注间引书的联系与区别，统计共同引书及独有引书见表 8，其中前四类表示共有引书，后三类表示特有引书。

表 8　《史记》三家注共有和特有引书分类

分类	引书种类	引书举例
集解 ∩ 索隐 ∩ 正义	76	《吴地记》《春秋传》《孟子》《说文》
（集解 ∩ 索隐）－正义	29	《列士传》《史记音隐》《韩诗》《汉志》
（集解 ∩ 正义）－索隐	8	《海外经》《六艺》《毛诗序》《慎子》
（索隐 ∩ 正义）－集解	73	《古今注》《外国传》《释名》《河图》
集解－索隐－正义	48	《孔子三朝记》《瑞应图》《弈指》《韩诗章句》
索隐－集解－正义	165	《归藏易》《春秋纬》《诗纬》《冀州记》
正义－集解－索隐	126	《五经通义》《古今地名》《尚书考灵耀》《毛诗义疏》

4.2.2 高频引书

高频引书是反映三家注引书特点的重要参考因素，本研究对其各自所引用的高频引书进行统计汇总，如表9所示。不难发现，三家注的高频引书直接存在明显差异。《集解》对音义的重视可与下文考察中其引徐广《史记音义》相呼应，而《正义》对地理类书籍的引用特色则进一步得到体现。

表9 三家注对比下的高频引书

集解			索隐			正义		
引书id	引书名	频次	引书id	引书名	频次	引书id	引书名	频次
H029	汉书音义	345	H024	汉书	789	K010	括地志	961
H024	汉书	261	Z041	春秋左传	260	H024	汉书	207
Z041	春秋左传	65	S056	世本	185	Z041	春秋左传	86
H050	皇览	40	Z005	战国策	144	D010	帝王世纪	42
S056	世本	28	S072	说文解字	95	S021	尚书	32

5 《史记》三家注引用作者的统计分析

《史记》三家注中除了对书名的引用外，对作者的单独引用更多，同样是《史记》注解的重要组成部分。与上文引书的计量分析类似，本节分别从“本纪”“表”“书”“世家”“列传”五部分和三家注视角对引用作者的现象进行分类统计和计量分析。

5.1 引用作者分布规律的分类考察

5.1.1 引用作者频次—种类规律

《史记》三家注引用作者共10313次，通过描述频次分布特征能够探究《史记》三家注引用作者的规律，故本节首先根据篇章对应的五部分呈现作者的在各卷的被引频次—种类，从整体视角总览引用作者情况概貌，考察各部分间的引用规律，具体数据如表10所示。

从《史记》五部分来看，引用作者频次与引书频次情况类似，最高部分为“列传”部分；“本纪”和“书”部分引用作者卷均频次更高；“表”和“列传”部分引用作者卷均频次较低。但相较于引书，引用作者在各卷间的分布差异更加显著，单卷最高频次更高，在“列传”第117卷《司马相如列传》中达到668次。为完善引用作者情况全貌，在频次的基础上，补充统计种类、卷均种类等多维度的统计数据，借此挖掘引用作者特点和规律。

表10 《史记》三家注引用作者统计（频次—种类）

	本纪	表	书	世家	列传	总计
引用作者频次	2565	237	1133	2601	3777	10313
引用作者种类	124	31	78	118	164	247
平均引用频率	20.68	7.64	14.52	22.04	23.03	41.75
引用作者卷均频次	197.31	23.7	141.63	86.7	53.96	78.73
引用作者卷均种类	9.54	3.1	9.75	3.93	2.34	1.89

结合引用作者种类来看，在频次更高时，种类更少，导致引用作者的平均引用频率远高于引书，即更明显的集中程度。“表”部分的频次和种类都是最少，平均引用频率也最低，表现出分散的趋势。“本纪”与“书”的引用作者卷均频次远高于其他部分，呈现出高度密集的分布特征。“世家”与“列传”部分引用作者的情况与引书情况基本一致，虽频次和种类高，但由于卷数较多，卷均频次都相对较低，呈现出较为稀疏的引用作者分布，《史记》三家注引用作者的分散-集中以及密集-稀疏规律通过图9可以得到更清晰的把握。

此外，图9还呈现了分散—集中和密集—稀疏分布中，引用作者与引书的区别，可以看出“本纪”“书”与“世家”三部分的分布十分相近，在同一象限内，“列传”部分虽略有差异，整体上仍表现为集中但稀疏的特点；“表”部分差异稍大，但在卷均分布上都呈现出稀疏的特点。

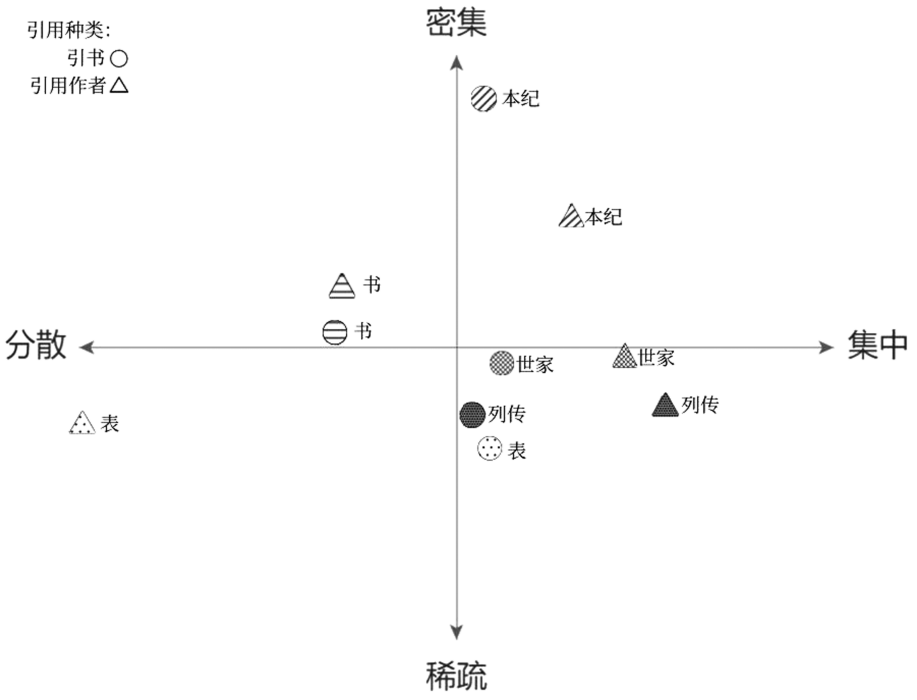

图 9 《史记》三家注引书和引用作者的分散—集中/密集—稀疏四象限图

与引书考察类似，图 10 呈现了各卷引用作者的集中离散趋势及数据偏态情况。引用作者的分布差异与引书情况基本一致，而数据细节表现更为明显。“本纪”的分布区间仍然最高，而“书”的区间则较之引书进一步提高。“本纪”各卷的引用作者频次和种类分布整体波动较大，且仍均为左偏态，没有异常值，说明“本纪”各卷间的引用作者差异仍然较大。其他四部分中，“书”部分在频次上的分布较之引书明显更加离散，而“列传”在种类上较之引书更离散。其余各部分的频次和种类分布较为集中，异常值仍然存在，如“列传”部分的《司马相如列传》等。“表”“世家”和“列传”部分各卷的频次与种类的偏态分布仍然相反。引用作者频次都为右偏分布，种类都为左偏分布，这与上文考察的引书分布情况相一致。

最后，图 11 呈现了《史记》三家注引用作者的线性分布情况，与引书的线性分布情况类似。引用作者频次与种类成正比关系，“表”部分的回归线斜率最小，这与图 9 呈现出的分散但稀疏的引用作者特征相一致。

5.1.2 高频引用作者

对《史记》三家注引用作者按频次进行排序，可以得到高频作者如表 11 所示。前五分别为徐广、郑玄、韦昭、服虔与孔安国，合计被引占比 1/3 以上。其中徐广的《史记音义》是目前所存最早的《史记》注本，记载了六朝时期的《史记》异本，对校正今本《史记》具有重要价值。前十高频引用作者里，大部分都由《集解》引用，裴骃征引徐广、孔安国、郑玄等人的注释成果来梳理《史记》的正文，体现了裴骃对不同版本、异文材料的重视。

表 11 《史记》三家注的前十高频引用作者

作者 ID	作者名	频次
x12	徐广	2364
z22	郑玄	628
w21	韦昭	594
f07	服虔	547
k01	孔安国	531
r01	如淳	435
d13	杜预	433
y21	应劭	327
z12	张晏	262
j02	贾逵	261

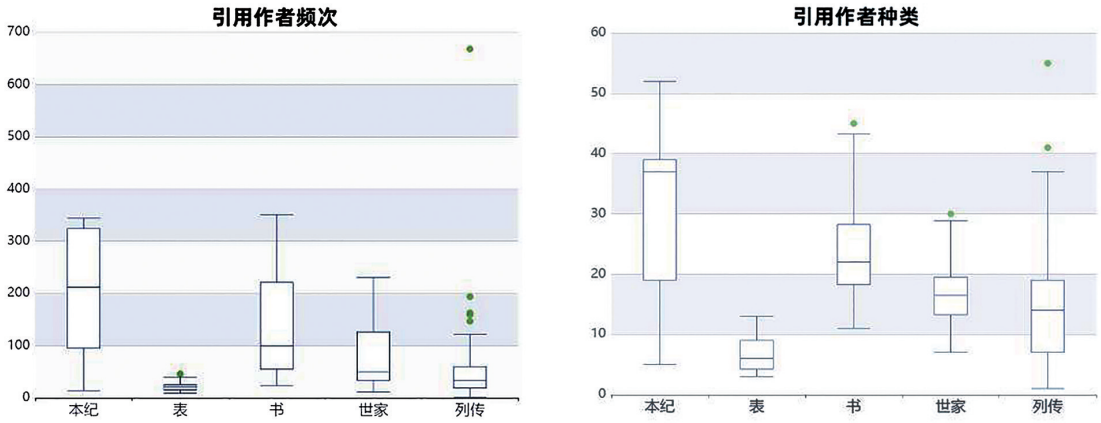

图 10 《史记》三家注引用作者的频次—种类箱线图

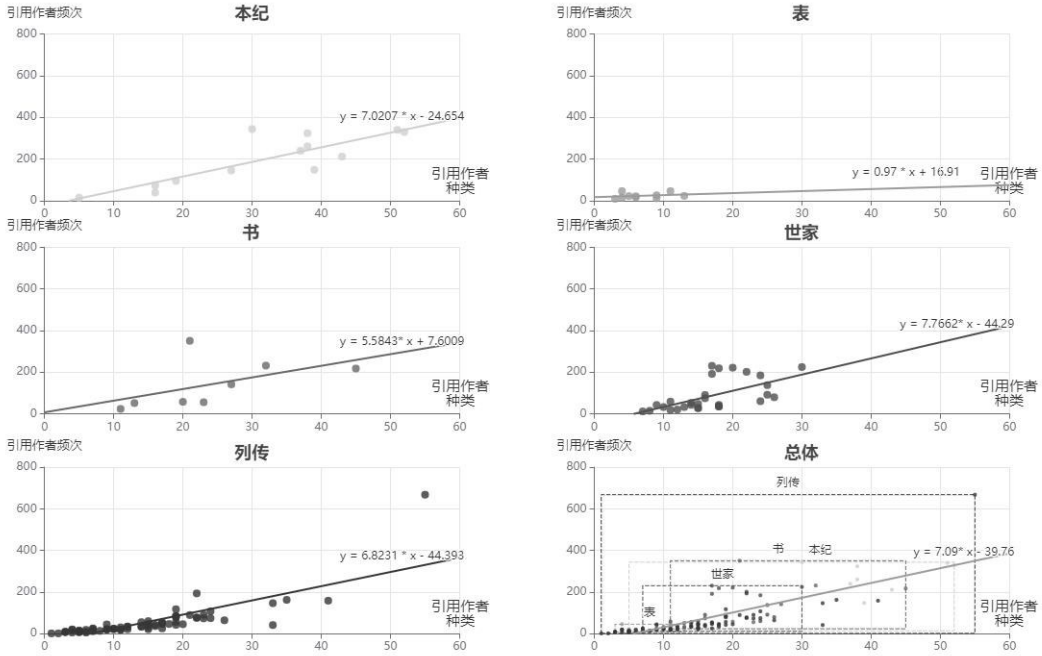

图 11 《史记》三家注引用作者的线性分布规律对比

5.2 三家注对比下的引用作者分布规律考察

5.2.1 引用作者的频次—种类规律

本节进一步从三家注对比视角下考察引用作者的特点，以完善对分布规律的理解，如图 12 所示。三家注对比下的引用作者频次分布与引书差异较大，尤其是《集解》，这表明其更多通过引用作者来注解，符合其“数家兼列”的引用特点。然而《集解》引用作者的种数仅有 90 位，为三家最低。造成这种差异的原因，可能是《集解》的作者裴骃所处时代为南朝宋，而另外两家作者身处唐代，由于时代限制，裴骃可选择引用作者的数量更为有限。此外，不排除作者之间不同学派、不同个人解读和史观的不同造成的引用差异。

分散—集中和密集—稀疏的分布考察与上文类似，如图 13 所示，聚焦于三家注的对比后可以发现，《集解》的引用作者的分布与另外两家差异明显，而这一特点在三家注引书中并不明显。这进一步验证了《集解》引用作者的独特性。此外，相较于三家注各自的引用风格，《史记》本身结构对三家注解的影响同样显著，基于内容展开的引用现象构成了引用频次、种类等引用特征，进一步呈现为《史记》整体的引用风格。

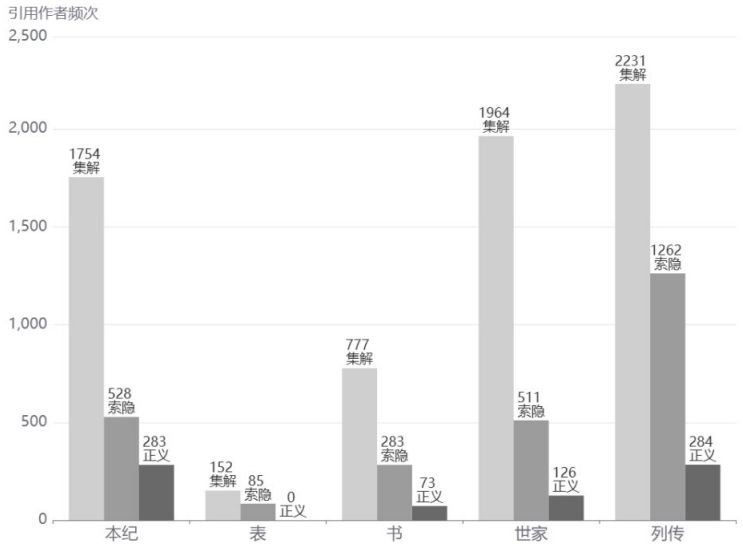

图 12　三家注对比下的引用作者频次

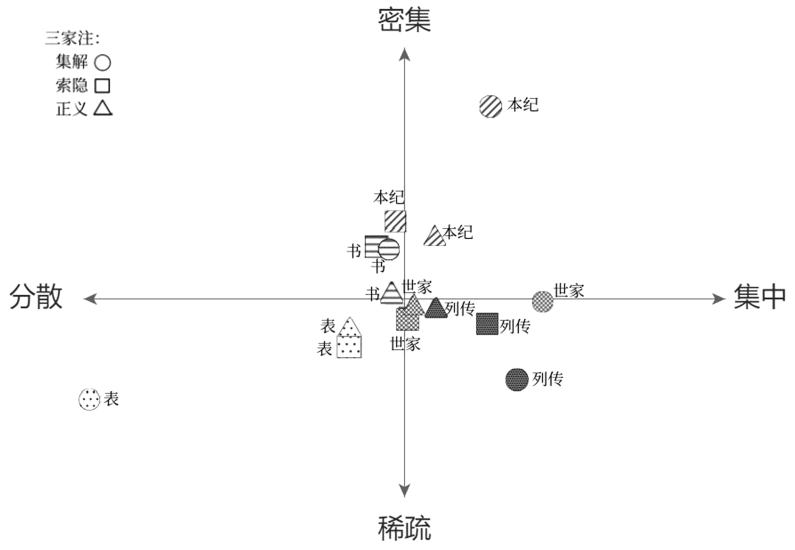

图 13　三家注对比下引用作者的分散—集中/密集—稀疏四象限图

5.2.2　高频引用作者

三家注对比下前五位高频引用作者如图 14 所示。三家对徐广的引用差异明显，《集解》引用徐广高达 2236 次，《索隐》对徐广的引用次数仅有 111 次，《正义》的前十高频作者则没有徐广，这与上文所述《集解》引用作者的独特性有关。此外，《集解》区别于其他两家的主要有东汉的马融和三国时期的贾逵和王肃。《索隐》与《正义》的高频作者引用比例均较为平均，且有六位作者共现，体现出二者引用作者的相似性，这与两者处于同一朝代也存在关联性。

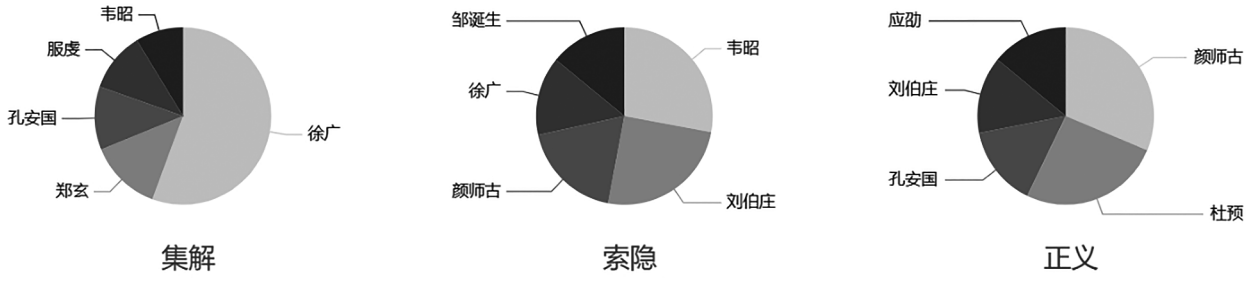

图 14　三家注对比下的高频引用作者(前五)

6 结论和未来展望

本文以《史记》三家注为语料，以其中的引书知识为对象，在人工标注及模型辅助校对后，结构化组织古籍中的引书内容。以此为基础，分别统计引书和引用作者分布特征并可视化呈现，以此考察《史记》三家注潜在的引书规律。从《史记》五部分和三家注对比两个视角出发，量化分析引用频次、种类、高频引书等统计数据，多维度描述《史记》引书分布特征，全面呈现引书知识全貌，为传统的古籍引书研究提供了可行思路与可靠结果。

相较于传统的人文研究，本文通过数据量化分析与可视化呈现方法计量分析古籍引书知识及引用特征，未来还将以引书原文内容为基础，对《史记》三家注重点关注的主题内容进一步挖掘和分析，同时围绕引书的异文现象和引用风格展开深入研究，以求全面完整的实现《史记》引书知识挖掘，为传统古籍引书研究提供新的参考。而从计算人文的视角来看，如何从数据出发提出更有价值的古籍引书研究问题，发现更有意义的古籍引书规律，还值得未来进行更加全面深入的探索。此外，在计算人文研究下，大规模高质量数据的获取至关重要，人工标注的成本和自动标注的准确性之间如何平衡成为一个不容忽视的问题。深度学习和大语言模型在解决这一问题中的有效性还值得进一步探索。

参考文献

1 黄水清，刘浏，王东波. 计算人文的发展及展望[J]. 科技情报研究，2021，3(4)：1－12.
2 曹书杰. 辑佚学的性质对象任务内容和意义[J]. 古籍整理研究学刊，1999(4)：38－43.
3 胡绍文. 论古典诗歌异文校勘的方法[J]. 湖南科技大学学报(社会科学版)，2010，13(2)：106－110.
4 应三玉. 试论《史记》三家注的价值及其影响[J]. 中国典籍与文化，2004(3)：15－22.
5 刘浏，齐月，刘雏菲，等. 计算人文下的古籍引书研究及全文本知识库的构建[J]. 情报学报，2023，42(12)：1498－512.
6 周莉娜，洪亮，高子阳. 唐诗知识图谱的构建及其智能知识服务设计[J]. 图书情报工作，2019，63(2)：24－33.
7 宋雪雁，霍晓楠，刘寅鹏，等. 数字人文视角下《全唐诗》贬谪诗人社会关系研究[J]. 现代情报，2022，42(2)：14－21.
8 朱锁玲，包平. 数字人文在中国农史研究中的实践与思考——以中华农业文明研究院数字人文项目为例[J]. 农业图书情报学报，2021，33(8)：79－87.
9 李斌，王璐，陈小荷，等. 数字人文视域下的古文献文本标注与可视化研究——以《左传》知识库为例[J]. 大学图书馆学报，2020，38(5)：72－80，90.
10 刘浏，黄水清，孟凯，等.《春秋》三传女性人物的人文计算研究[J]. 图书情报工作，2020，64(23)：109－23.
11 王东波，刘畅，朱子赫，等. SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究[J]. 图书馆论坛，2022，42(6)：31－43.
12 刘江峰，冯钰童，王东波，等. 数字人文视域下 SikuBERT 增强的史籍实体识别研究[J]. 图书馆论坛：，2022，42(10)：61－72.
13 刘畅，王东波，胡昊天，等. 面向数字人文的融合外部特征的典籍自动分词研究——以 SikuBERT 预训练模型为例[J]. 图书馆论坛，2022，42(6)：44－54.
14 黄水清，刘浏，王东波. 计算人文学科的内涵、体系及机遇[J]. 图书与情报，2023(1)：1－11，153，45.
15 牛巧红.《史记索隐》引书体例考辩、述补[J]. 古籍整理研究学刊，2017(5)：46－50.
16 张蔚虹. 郭庆藩与王先谦集解《庄子》引书比较[J]. 贺州学院学报，2021，37(2)：46－52.
17 王丽.《世说新语》刘孝标注与《三国志》裴松之注比较研究[D]. . 沈阳：辽宁大学，2021.
18 马创新，陈小荷. 基于引文分析的古籍文献影响力评估[J]. 大学图书馆学报，2016，34(1)：16－24.
19 周好. 引书的自动识别及分析[D]. 南京：南京农业大学，2019.
20 赵英翘.《史记》三家注体例略述[J]. 社会科学辑刊，1988(2)：69－71.
21 杨炜.《史记》三家注之特点比较[J]. 渭南师范学院学报，2018，33(9)：85－89.
22 方心棣.《史记》三家注通假琐议[J]. 安徽教育学院学报(哲学社会科学版)，1995(3)：60－63.
23 周振风.《史记》三家注研究[D]. 南昌：南昌大学，2005.
24 黄坤尧.《史记》三家注之开合现象[J]. 中国语文，1994(2)：121－124，38.
25 孙利政.《史记》三家注音注校议[J]. 励耘语言学刊，2020(2)：20－30.
26 张玉春. 明廖铠刊《史记》三家注本版本系统考[J]. 古籍整理研究学刊，1999(6)：38－40.
27 赵生群.《史记》三家注称引诸子考校[J]. 浙江师范大学学报(社会科学版)，2009，34(2)：20－23.
28 司马迁. 史记[M]. 北京：中华书局，1959.
29 Devlin J，Chang M W，Lee K，et al. BERT：Pre－training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Linguistics：Human Language Technologies. Minneapolis，Minnesota：Association for Computational Linguistics，2019：4171—4186.

30 程金造. 史记索隐引书考实[M]. 北京：中华书局，1998.

31 段书安. 史记三家注引书索引[M]. 北京：中华书局，1982.

作者单位：齐月、刘雏菲、李文祺、王东波、刘浏，南京农业大学信息管理学院，人文与社会计算江苏省高校哲学社会科学重点研究基地，南京农业大学领域知识关联研究中心，江苏南京，210095

孟凯，南京农业大学马克思主义学院，江苏南京，210095

收稿日期：2024 年 3 月 28 日

修回日期：2024 年 6 月 8 日

（责任编辑：李晓东）

Knowledge Annotation and Quantitative Analysis of Citations of the Three Commentaries on Records of the Grand Historian in the Perspective of Computational Humanities

QI Yue　LIU Chufei　LI Wenqi　MENG Kai　WANG Dongbo　LIU Liu

Abstract: This study aims to explore a new method for studying the *Three Commentaries on Records of the Grand Historian* from the perspective of computational humanities. Focusing on the phenomenon of citations in the *Three Commentaries on Records of the Grand Historian*, this study combined manual and deep learning methods from the perspective of digital humanities to organize the content and knowledge in the *Three Commentaries on Records of the Grand Historian* and constructed a knowledge base of ancient citations. By building a named entity recognition model based on deep learning, this study automatically proofread the manual annotation results. Combined with another round of manual proofreading and entity disambiguation, it constructed a knowledge base of citations in the *Three Commentaries on Records of the Grand Historian* to conduct a comprehensive statistical analysis. It started from two types of citations, namely, citing books and citing authors. With the five parts of the *Records of the Grand Historian* (Annals, Chronological Tables, Treatises, Hereditary Houses, and Biographies) as the types of classification, it carefully examined the frequency-type law of citations, especially the scattered—concentrated and dense-sparse distribution phenomena. It also used statistical data to construct a frequency-type box plot to examine the data anomalies and data skewness in the citations. The study further examined and compared the linear distribution law of citations. At the same time, examples were given to analyze high-frequency cited books and to explain the causes of statistical laws. In addition, with comparative study of the three commentaries, the above-mentioned citation phenomenon and distribution law were examined and analyzed in more detail. The research forms a complete set of technical processes and frameworks for the annotation of ancient book citations through annotation specifications, knowledge annotation, knowledge base construction, knowledge measurement and analysis. The knowledge base constructed has important resource value for the study of ancient book citations and the study of the *Records of the Grand Historian*. The exploration of ancient book citations using statistical measurement and visualization analysis methods provides a new idea for the traditional study of ancient book citations. The results of quantitative analysis are of reference value for the study of "Historical Records" and ancient book citations. It also has reference significance for the application of computational humanities perspective in the field of ancient Chinese classics.

Keywords: Computational Humanities; Digital Humanities; Ancient Book Citations; *Three Commentaries on Records of the Grand Historian*; Textual Knowledge Mining