DOI:10.13530/j.cnki.jlis.2023006

# 历代史志目录的数据集成与可视化

李文琦　王凤翔　孙显斌　黄芷欣　李芃蓓

**摘　要**　古籍目录及其分类体系具有重要的学术价值,数字学术的发展为古籍目录的数字化保存和利用以及开展数字工具支持的目录学研究提供了新的契机。本文以时间跨度两千多年的八种史志目录为数据源,以机器预处理与专家校对相结合的人机迭代方式对数据进行记录拆分和字段抽取、数据补全、规范化以及书目认同,最终完成11万余条书目记录的结构化、规范化集成。在此数据集的基础上,从领域专家的研究需求出发,结合统计、可视化、检索等方法,利用人机交互技术构建了一个历代古籍目录可视化分析系统。该系统包括书目统计以及分类演化分析两个主要部分:一方面可对书目数据进行细粒度统计和可视化呈现,以帮助学者清晰地比较、追踪类目的消长;另一方面可对所有典籍在历代目录中的分类演变轨迹以及各类目所收典籍的源流进行可视化分析,以更好地实现类目分合转化的模式识别。本研究为数字学术背景下的目录学研究提供了数据基础和分析工具,不仅为学者省去了大量数据收集、整理的时间,还通过新的技术和视角助力分析、比较等解释性研究。图8。表3。参考文献36。

**关键词**　古籍目录　数字学术　数字人文　目录学　数据集成　可视化

**分类号**　G257

# Diachronic Data Integration and Visualization of Ancient Book Catalogs Compiled for Imperial Collections

LI Wenqi, WANG Fengxiang, SUN Xianbin, HUANG Zhixin & LI Pengbei

**ABSTRACT**

Ancient book catalogs record and classify a large number of Chinese ancient books. They are of great academic value for studying both ancient literature and traditional knowledge organization. The development of digital scholarship shed new light on the digital preservation and reuse of these ancient book catalogs as well as the domain research supported by digital tools. Digital scholarship facilitates the digitization and datafication of ancient book catalogs. Moreover, new methods and computational tools are provided to enable the exploration of large collections, and new research questions can be raised from fresh perspectives. Recent studies have introduced computational methods to analyze the abstracts and classification systems of the ancient book catalogs. But these studies were based on only one catalog or a particular category. It is imperative to integrate the catalogs throughout the history and provide digital tools for scholars to explore and analyze them diachronically and holistically.

In this study, we selected eight representative catalogs, mostly from official histories, as data sources. They were *Hanshu Yiwenzhi*, *Suishu Jingjizhi*, *Jiutangshu Jingjizhi*, *Xintangshu Yiwenzhi*, *Songshi Yiwenzhi*, *Mingshi Yiwenzhi*, *Qingshigao Yiwenzhi* and *Siku Quanshu Zongmu*. These catalogs cover major

通信作者:王凤翔,Email:wfximer@pku.edu.cn,ORCID:0000-0002-7383-4634(Correspondence should be addressed to WANG Fengxiang, Email:wfximer@pku.edu.cn, ORCID:0000-0002-7383-4634)

dynasties in Chinese history with a time span of more than two thousand years. We adopted a semi-automated data processing approach to integrate the book entries in eight catalogs. The whole integration process was iterated by machine pre-processing and expert manual correction and contained three main steps—record splitting and field segmentation, field completion and normalization and book identification. Eventually we got more than 110 000 structured data records, and identified over 7 000 books that were recorded in at least two catalogs.

Based on the integrated data, we designed and developed an interactive visual analysis system that included features of statistics, visualization and record query. The system is designed to mainly meet two research requirements proposed by expert users. First, the system provides granular statistics and graphs that can help scholars to compare and trace the change of book volumes in different categories and catalogs. Second, it provides an interactive visualization tool that can be used to explore how different books are classified differently in each catalog, and thus manifests the changes of knowledge organization as well as the origin and evolution of academic thoughts.

In conclusion, this study provides data foundation and analytic tools for the studies of ancient book catalogs in the context of digital scholarship, which not only saves the effort on manual data collection and collation, but also provides new perspectives to identify and solve hermeneutics problems with new techniques. 8 figs. 3 tabs. 36 refs.

**KEY WORDS**

Ancient book catalogs.　Digital scholarship.　Digital humanities.　Bibliography.　Data integration.　Visualization.

## 0　引言

古籍目录是将一批书名和叙录依次编列的总聚,是目录学研究的主要对象之一[1],也是两千余年古籍整理工作留下的重要学术遗产。清代学者王鸣盛在《十七史商榷》中说“目录之学,学中第一要紧事,必从此问途,方能得其门而入”[2]。古籍目录在图书分类和学术史等方面都有宝贵的研究价值。一方面,古籍目录可以看作是我国最早的知识组织系统,是学者“即类求书,因书究学”的入学门径,广泛使用的四部分类法在当今的图书馆古籍分类工作中仍被应用。另一方面,古籍目录历来有“辨章学术,考镜源流”的学术史功用,其类目组织的分合转化体现了学术知识的流变,见证了中国历代知识生产、组织以及知识结构的发展变化。

目前,古典目录学研究仍以传统的人文范式为主,信息技术虽被广泛应用到古籍数字化实践中,却尚未与古典目录学研究融合,而数字学术的发展为这种融合带来了新的契机和方向[3]。数字学术是指基于某一学科领域,使用数字工具产生并以数字形式呈现的学术[4],不仅包含使用数字证据和方法的学术研究,还包括文献等学术成果的数字化保存和利用[5]。对于古典目录学这一领域,数字学术意味着通过对古籍目录的数字化和数据化,使其以更适合数字环境的形式供今人阅读和使用,并通过数字工具的使用为目录学研究提供识别问题、发现问题甚至是解决问题的新方式。目前在古籍和古籍目录的数据化方面已经有了一定数量的标准规范[6]和方法框架[7],但由于历代目录所收典籍数量庞大,著录不规范统一,所以一直以来都缺乏针对大规模、跨时代的古籍目录数据化处理及集成的实践。此外,目录学研

究,特别是对目录分类的研究多为定性研究,少量定量研究涉及的书目统计数据也是局部的、分散的。近期虽有研究[8,9]引入计算手段对目录的提要、分类等进行分析,但研究基础都是单一的目录或类目数据。因此,学界亟需在现有研究基础上,对历代目录进行数据整合,并引入数字方法和工具对长时段、全领域的目录分类演化进行分析,以探索数字学术环境下的新的研究范式。

在此背景下,本研究利用现有数字技术为目录学研究提供数据基础和分析工具。具体而言,以时间跨度两千余年的八种史志目录为数据源,以机器预处理与人工校对相结合的方式对数据进行结构化、规范化处理,结合统计、可视化、检索等分析方法和人机交互技术,构建一个古籍目录的可视化分析系统。

## 1 研究综述

古籍目录的分类作为"纲纪群籍簿属甲乙之学"[10],一直是古典目录学的重要分支。正如姚名达所言,"有书目而不分类,未得尽目录之用也",目录分类的研究价值可见一斑。古籍目录的四部分类法虽自《隋书·艺文志》后保持稳定,但"四部之界画并不严谨,各篇小类之内容并不单纯"[11]。姚名达的《四部分类源流一览表》将十三种书目的类目增删变化及类目内涵的对照关系梳理一清,充分发挥了目录学的考辨作用[12]。其他从目录学发展历史的角度进行全局梳理的研究[13]与之类似,都是从类目层面进行定性的分析讨论,没有全面深入到类目下具体收录的典籍。此外,还有一些针对古籍目录中特定子类的演变和子类间关系的研究[14-17],这类研究依据类目所收录的典籍对类目内涵进行界定,同时对典籍数量进行统计,进而分析类目的增减。由于没有数据化书目作为基础以及缺少统计分析工具的支持,这类研究目前还只局限于个案分析。

在图情领域,古籍目录作为两千多年以来的古典目录学成果对现代目录学有着重要的参考价值[3],其数字化保护和开发在数字人文这一新兴研究范式下也备受关注。从古籍整理和文献编目来看,古典目录学中的文献分类、叙录等方法可以视作最早的古籍整理范式[18]。而后随着计算机技术的发展,文献整理和编目也逐步从数字化转向数据化[19],特别是随着数据驱动的第四研究范式的兴起,数字学术背景下的古籍整理逐渐以数据和知识为对象[18]。在这一背景下,古籍目录作为古典文献研究的重要数据,在人工智能、语义网等技术的加持下,可以为古典人文研究提供数字学术的数据基础[7]。2009年,北京大学中文系与国家图书馆合作,基于结构化古籍目录数据开发了"历代典籍总目分析系统",具有古籍目录数据的浏览、检索和共现等功能。目前在古籍数据化方面已经有了一些标准规范,如《国家图书馆古籍元数据规范与著录规则》和《古籍元数据规范》以及各高校图书馆的古籍元数据规范[20]等。在此基础上,夏翠娟等设计了古籍术语词表并将不同来源、格式的古籍目录数据进行融合,构建了"中文古籍联合目录及循证平台"[7]。除构建数据基础的相关研究外,也有学者提出了利用数字技术对古籍目录进行分析研究的方法框架,如李惠等以《四库全书总目》为例,通过对目录中的古籍、人物和提要构建分析网络,实现古籍提要的知识发现[9]。

在古籍目录的分类研究方面,数字学术的发展使面向大量书目的数据化以及基于数据的全局量化分析成为可能。自从 Moretti 提出"远读"的概念以来[21],统计和可视化就成为分析人文数据的常用技术手段[22]。描述性统计、可视化以及文本分析等技术方法也同样被应用到分类系统及知识组织系统构建的研究中[23-26],已有研究论证了这些方法在分析类目体量变化、构成以及增长分布等方面的有效性[27]。李瑞龙等以易类典籍为例,探索桑基图这一可视

化图表对从数量上分析古籍目录分类演变的可行性[8]。然而这些研究中的研究对象在数据规模和时间跨度上都很有限,且应用的可视化方法还局限于统计数据静态描述性的展示,而对于古籍目录分类演变这种涉及历史、社会等人文方面的解释性分析,交互式可视化更能帮助学者从定量和定性两种视角进行探索[28]。

本文在现有的古籍数据化,特别是古籍目录的数据化研究的基础上,对具有代表性的历代古籍目录进行数据化、规范化处理,一方面可以对古籍目录进行数字化保存,另一方面也可作为数据基础支撑数字古典目录学研究。此外,本文还通过构建交互式可视化分析系统,为历时性的、全局性的、基于大规模数据的古籍目录分类演变分析提供有效的研究工具。

## 2　研究框架

数字学术环境的建立是一个持续的、迭代的设计过程,其建立基础是对用户及其交互需求的充分理解[29]。遵循以用户为中心的设计思想,本研究从最初的需求调研,到明确数据范围、确立设计目标、进行原型设计,再到最终的系统反馈,经历了多次迭代,邀请三位古典文献学领域的学者作为专家用户提供领域知识并密切参与到数据处理及系统设计的各个环节。首先,通过与专家用户的多次交流确认了他们对于古籍目录研究的两层需求,即数据层需求(R1)和分析层需求(R2、R3)。具体需求如下。

R1:形成一个跨时代的、统一规范的、结构化的、可供检索查询的代表性古籍目录数据集。

R2:考察历代典籍分布情况、不同时代新生产典籍的数量,以及前代典籍历经长时段留存亡佚情况,揭示知识生产、消亡、流变的趋势。

R3:全面考察和记录每部典籍在历代史志目录中归属的变化,以及每个典籍子类的分合流变,并进一步总结出各种知识分类演化的模式,然后分析其背后蕴含的规律和意义。

基于以上需求,本研究主要由数据集成(满足 R1)及交互式可视化系统(满足 R2、R3)两部分构成。在数据集成过程中,首先,笔者与专家用户共同确定项目初期需要覆盖的数据范围,即八种代表性史志目录,并由专家用户提供原始目录文本以确保数据的可靠性。其次,以尽可能保留最完整原始目录文本内容为原则设计数据结构,并通过机器预处理与人工校对相结合的半自动化方式,迭代进行数据的结构化与规范化集成。根据古籍目录的特性,这部分工作主要包括记录拆分与字段抽取、数据补全以及规范化与书目认同。最后,根据分析需求对数据进行转换,并设计交互式可视化分析系统。与静态可视化图表相比,交互式可视化更适合人文领域的探索性研究,更能帮助用户进行对比分析、发现规律模式[28]。可视化分析系统的设计遵循 Shneiderman 提出的视觉信息搜索“概览—缩放、筛选—按需提供细节”的整体流程,以支持用户逐步深入地进行可视化分析[30,31]。具体而言,为满足需求 R2,对各目录、各类目中的典籍数量以及留存情况进行统计,并对这些统计数据进行交互式的统计图表展示;为满足需求 R3,对规范化后的数据进行转换以得到每个典籍在历代目录中的类目归属,并进一步设计针对类目演化的可视化分析功能,支持筛选、检索、数据展示等交互。由此,基于集成数据构建了一个完整的交互式可视化分析系统,如图 1 所示。

## 3　数据集成

### 3.1　数据集范围

古籍目录虽有官修与私撰之分、综合与专科之别,但从目录学史的进程来看,不论是何种类,历代修撰的目录共同构成了古籍目录体系。官修目录是古籍目录的主干,正史所附史志目录是一种特殊的官修目录,因其自成系统,一般

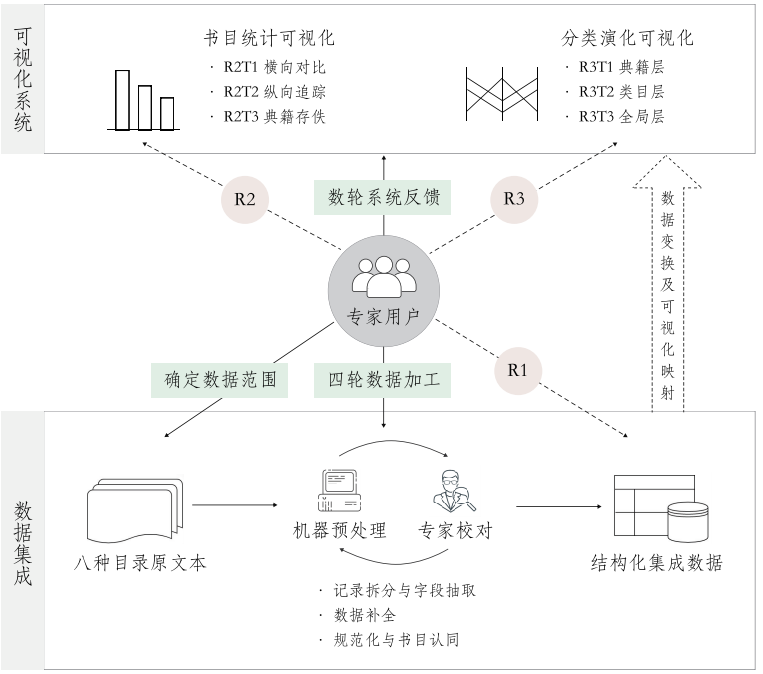



**图 1　研究设计框架**

将其单列。史志目录在很大程度上承接了官修目录的成果,并且贯通主要的历史时期;相较于私家目录,史志目录收书种类无疑更广泛全面,更能反映特定时代典籍存佚的整体情况。时间上的贯通性和收书范围的全面性决定了史志目录是研究和利用古籍目录不可或缺的、也是可获得的最基础的资料。正史目录的主体部分是正史的经籍志或艺文志,即《汉书·艺文志》《隋书·经籍志》《旧唐书·经籍志》《新唐书·艺文志》《宋史·艺文志》《明史·艺文志》以及《清史稿·艺文志》,其中《明史·艺文志》和《清史稿·艺文志》只收录本朝典籍。为了考察典籍的流传情况,本研究也将《四库全书总目》(包括补充禁毁书和未收书)和《清史稿艺文志补编》(以下简称《补编》)、《清史稿艺文志拾遗》(以下简称《拾遗》)纳入研究对象。以上构成了本研究的数据来源(见表 1)。从时间跨度上看,该数据集覆盖了从汉代到清代主要的历史时期。

从分类上看,《七略》将图书分为"六艺略""诸子略"等六略。《七略》虽早已亡佚,但《汉书·艺文志》是《七略》"删其要"而成,继承了古籍目录体系的源头——六略分类。此后,古籍目录分类经魏晋"甲乙丙丁"四部分类过渡,到《隋书·经籍志》形成"经史子集"四部分类法,此后一直沿用,成为古籍分类法的主流,乾隆时编纂的古代官修目录中的集大成之作《四库全书总目》使用的也是四部分类法。因而,由史志目录和《四库全书总目》所构成的数据集不仅具有时间上的贯通性、收录种类的广泛性,还具有分类方式的传承性和目录学史发展历程的代表性,足以支撑开展跨时段的研究。

### 3.2　数据处理与集成

除了前文所述目录分类方式的相续相承,史志目录之间的内在联系还表现在对同种典籍或相关典籍的著录上。典籍是知识和文化的主要载体,中国古代知识的传承亦体现在古代典籍的历代相传上,正因为如此,今人才能通过大量的古代文献回溯贯穿几千年的文化。理想情况下,在典籍存续的时段里所编写的史志目录

**表 1　数据集构成及时间覆盖范围**

| 目录名称 | 编纂时间 | 书目记载覆盖时期 |
| --- | --- | --- |
| 汉书·艺文志 | 东汉光武帝建武三十年(54 年)至和帝永元四年(92 年) | 收录先秦至西汉中期著作;对西汉后期的书亦有增补 |
| 隋书·经籍志 | 唐贞观十五年(641 年)至高宗显庆元年(656 年) | 收录先秦至隋代著作 |
| 旧唐书·经籍志 | 五代后晋高祖天福六年(941 年)至出帝开运二年(945 年) | 收录先秦至唐玄宗开元年间著作 |
| 新唐书·艺文志 | 宋仁宗至和元年(1054 年)至嘉祐五年(1060 年) | 收录先秦至唐代著作 |
| 宋史·艺文志 | 元顺帝至正三年(1343 年)至五年(1345 年) | 收录先秦至宋宁宗嘉定年间著作;对嘉定后的宋代书亦有增补 |
| 明史·艺文志 | 清康熙二十年(1681 年)至雍正十三年(1735 年) | 收录明代著作 |
| 四库全书总目 | 清乾隆三十七年(1772 年)至四十七年(1782 年),此后至乾隆六十年(1795 年)仍有修订 | 主要收录先秦至清乾隆时期著作 |
| 清史稿·艺文志 | 1914 年至 1927 年 | 主要收录清代著作,含清人辑佚的古佚书 |
| 四库禁毁书目 | 2000 年出版 | 主要收录《四库全书》禁毁的宋末至清乾隆时期著作 |
| 四库禁毁书目补编 | 2005 年出版 | 主要收录《四库全书》禁毁的明代至清乾隆时期著作 |
| 清史稿艺文志补编 | 1982 年出版,20 世纪 50 年代编成 | 主要收录清代著作,含清人辑佚的古佚书 |
| 清史稿艺文志拾遗 | 2000 年出版 | 主要收录清代著作,含清人辑佚的古佚书;部分民初书籍亦酌予收录 |

应将该种典籍收录其中,以求周全。实际情况是任何一种目录都不太可能囊括所有存世典籍,但史志目录还是很大程度上保留了当时存世典籍的面貌,这是其研究价值所在。另一方面,同一种书在不同的目录中收录于何种类目或者因某种原因在某个目录中缺收同样值得关注。在过去的目录学研究中,鲜有能够将众多目录的所有类别集中综合分析的,遑论条分缕析式地考察每一条书目记录,也就无法准确回答类似以上的诸多宏观问题,其主要原因在于目录类别和典籍之间的内在联系没有完全体现于既有目录的形式之上。对于人工而言,书目形式和内容上的微小差异对理解影响不大,但人工的局限在于不能方便地处理大量的数据;对于计算机而言,其长处在于可以进行大量数据的存储和运算,却对数据格式的规范要求很高。因此,为构建机器可操作的历代古籍目录数据集,首先要保证数据内容的准确性和数据格式的规范性。

由于这些目录陆续产生于过去两千多年的时间里,著录体例等方面的差异带来了数据规范处理的诸多困难,如需要大量的人工干预以解决异体字、同书异名、同名异书、字段缺失、称呼混用等问题。为解决目录数据的各种不规范

问题,本研究以人机迭代的方式对书目数据进行规范集成。所谓人机迭代,是指由计算机自动处理与人工校对和修改交替进行的数据加工方式,可以在提升效率的同时保证数据的质量。对于普遍存在的、规模较大的不规范数据,先由计算机按照规则,使用正则表达式等技术处理,对于不便自动处理的,转由人工处理,即由具有古典文献专业背景的专家逐条手工修改。此外,专家还需对机器处理的结果做校正,进而归纳可以自动处理的新规则,手工处理完成后再转入计算机处理流程,如此反复迭代。具体的数据处理过程主要包括以下三个步骤:记录拆分与字段抽取、数据补全、规范化与书目认同(见图 2)。

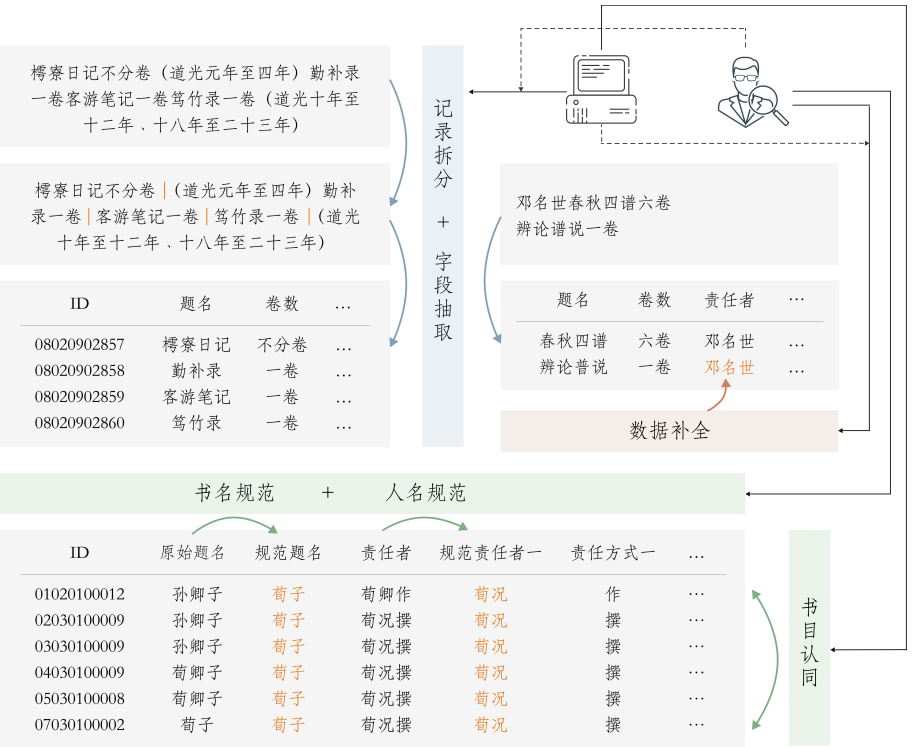



**图 2 数据规范与集成流程示例**

(1)记录拆分与字段抽取。这一过程实际是数据的结构化。在原始数据中,每种典籍的责任者、朝代、题名、卷数等信息几乎混杂在一起,甚至一条记录包含多种书而没有做显式区分,如《清史稿·艺文志》《补编》及《拾遗》史部传记类日记之属有“樗寮日记不分卷(道光元年至四年)勤补录一卷客游笔记一卷笃竹录一卷(道光十年至十二年、十八年至二十三年)”。根据括号标识和数字与“卷”字相连的模式可批量提取时间和卷数信息,并以卷数为分界标志拆分出四种书。但由于全部数据中括号的用处不仅仅用于标示时间信息,因此需要人工修改错误的识别记录。若记录中杂有责任者和责任方式,责任方式的有穷性便成为提取责任者的入口,即首先尽量穷举所有可能的责任方式并生成责任方式字典,再根据责任方式前为责任者的规则提取这两个字段信息。事实证明,这种操作方式是可行的,但也有部分错误需要手工纠正,如将“乾隆间纂”中表示修纂时间的“乾隆间”误提为责任者。在对所有书目记录进行拆分和字段抽取的同时,本文还为每一条结构化的数据赋予了 ID,该 ID 的前两位代表所属

目录,三、四位代表一级类目,五、六位代表二级类目,后五位则为该类目下书目依次编号。以#01020300004 为例,其中 01 代表《汉书·艺文志》,02 代表一级类目“诸子略”,03 代表二级类目“阴阳家”,00004 则为该类目下的第四本书《邹子》。如此便可以方便计算机识别、操作,特别是在处理后续分类问题时可以直接通过 ID 识别书目的分类。

(2)数据补全。在将所有记录拆分并抽取出字段数据后便得到了结构化数据,但很多字段的信息仍存在缺失或不明确的问题,需要将之补全。以责任者缺失为例,宋志的体例历来为学者所费解,有些前一条责任者系于题名前,后一条虽无责任者但承上条,为同一责任者,有些确无责任者,如“邓名世春秋四谱六卷”条后接“辨论谱说一卷”条,“春秋四谱六卷”的责任者为邓名世,“辨论谱说一卷”无责任者,但经查考文献,此条责任者亦为邓名世。然而这种处理方式只可用在经部,子部文献庞杂,多无责任者,便不能草率认为其承接上条,因此在加工数据时,只能尽可能查考补充。又如《四库全书总目》中含“坿”的纪录(如“易数钩隐图三卷坿遗论九事一卷”),实为多个题名共用一个责任者,在第一轮计算机自动提取责任者时忽略了这种情况,仅为之对应一个题名,在人工校对阶段发现可利用此规则进行自动抽取,则在第二轮计算机处理阶段一并补充。除了责任者字段,其余字段皆做相同的处理。

(3)规范化与书目认同。最后也是最重要且繁杂的一项工作是题名和责任者认同问题的处理,其目的是通过保证同种书在各处的记录有完全相同的表达形式,以便让计算机识别不同目录中的同种书。在假定题名相同并且责任者相同的两种书是同一种书的前提下,本文对题名和责任者做了规范处理以满足机器认同的要求,这一工作主要依赖于手工完成。题名认同和人名认同实际上紧密相连,既需要依据责任者识别出同一种书的不同题名,又需要依据题名判断其责任者是否为同一个人。因此,在处理时为了便捷地互为参考,本文将八种目录的所有数据放在一起,先后按照题名排序和人名排序,让同名书和同责任者书尽量靠近并逐条分析,以如上两次完整的数据清洗为一轮次,共进行四轮认同工作。为了保证同一种书和同一个责任者最后的规范表达相同,在每遇到一种特殊的表达形式时,需要事先规定对应表达的规范表达形式,如责任者为僧名的数据存在“法号”“释+法号”“僧+法号”三种表达形式,本文将其统一为“僧+法号”。其他如称字、号、俗名等不统一的情况也作类似处理,以保证数据的一致性。而对于“不知责任者”“不著撰人名氏”等责任者缺失的表达,则统一改为空值。题名的特例相对于人名要少,也仍根据专家判断选择一个最为通用的规范名。以《荀子》为例,在汉志、隋志及旧唐志中,此书题名为《孙卿子》,而在新唐志和宋志中题名为《荀卿子》,在四库提要中题名为《荀子》,在进行数据处理时由专家统一规范化为《荀子》。同时,该书的责任者在汉志中记录为“荀卿”,而后皆为“荀况”,经规范化处理后统一为“荀况”。此外,原始数据存在大量异体字,为了更好地进行认同,本文在数据加工时通过制作规范字表对异体字进行替换,尽可能地将规范题名和责任者中的异体字统一为规范字,如“溪—谿”“略—畧”“衮—袞”“考—攷”“德—悳”等(规范字—异体字)。至此,数据的结构化和规范化便已完成,基础数据字段包括“ID”“原始题名”“规范题名”“卷数”“责任者”“责任方式”“所属类别”以及“提要”等,其中对于存在多个责任者的情况则分别以“责任者二”“责任方式二”等来表示(具体数据示例见表 2,其中空值字段已省略)。在此数据的基础上,本文对所有书目记录做了一一匹配,不仅保证题名、责任者名相同,还要确保有多位责任者时责任者的顺序也是一致的,以尽可能减少错误匹配,由此完成书目认同的工作,保证了跨目录的书目数据具有一致性和贯通性,为后续的书目分类演化等分析奠定了基础。

表 2　结构化数据记录示例

| ID | 原始数据 | 原始题名 | 规范题名 | 卷数 | 所属类别 | 责任者一 | 责任方式一 |
|---|---|---|---|---|---|---|---|
| 01020100012 | 《孙卿子》三十三篇。名况,赵人,为齐稷下祭酒,有《列传》。 | 孙卿子 | 荀子 | 三十三篇 | 诸子略/儒家 | 荀况 | 作 |

以上论述了目录数据处理的主要问题及流程,另有许多繁琐的特例无需一一展开讨论。总之,数据处理的主要难点在于古代文字记录的诸多特殊表达现象,并且这些特殊表达现象具有类型多、难以规则化的特点。本研究以手工处理为主,手动和自动处理相结合的方式,尽可能在预处理和后处理上让计算机辅助手工加工数据,确保数据准确性并提高加工效率。最终获得了 11 万余条数据,经过书目认同后至少出现在两种目录中的书籍七千余种,各目录数据分布如表 3 所示。

表 3　数据处理前后书目记录数对比

| 目录名称 | 原始文本条目数 | 结构化数据记录数 |
|---|---|---|
| 汉书·艺文志 | 651 | 658 |
| 隋书·经籍志 | 5 208 | 5 208 |
| 旧唐书·经籍志 | 3 018 | 3 018 |
| 新唐书·艺文志 | 5 242 | 5 242 |
| 宋史·艺文志 | 9 779 | 9 779 |
| 明史·艺文志 | 4 673 | 4 673 |
| 四库全书总目(含补充禁毁书和未收书) | 11 212 | 11 255 |
| 清史稿·艺文志(含补编、拾遗) | 68 623 | 73 411 |
| 总计 | 108 406 | 113 244 |

## 4　可视化分析系统构建

### 4.1　书目统计及典籍存佚

官修目录中典籍的数量直接反映了当时的知识生产情况和知识结构概貌。现有研究所涉及的统计数据多为局部的、分散的,不便于研究者重复使用。本研究的集成数据集为大规模的机器统计提供了基础,不仅确保了统计数据的准确性,而且方便研究者集中调取不同粒度的统计数据用于分析比较。然而,大量的统计数据虽为量化分析提供了基础,但其本身难以清晰呈现其中蕴含的变化规律,数据可视化则可以通过视觉元素增强用户对数据的感知,帮助用户直观地进行数据比较和发现数据变化趋势[32,33]。根据专家用户提出的统计分析需求 R2,本研究进一步拆解出可视化分析系统需要支持的基于统计数据的分析任务。

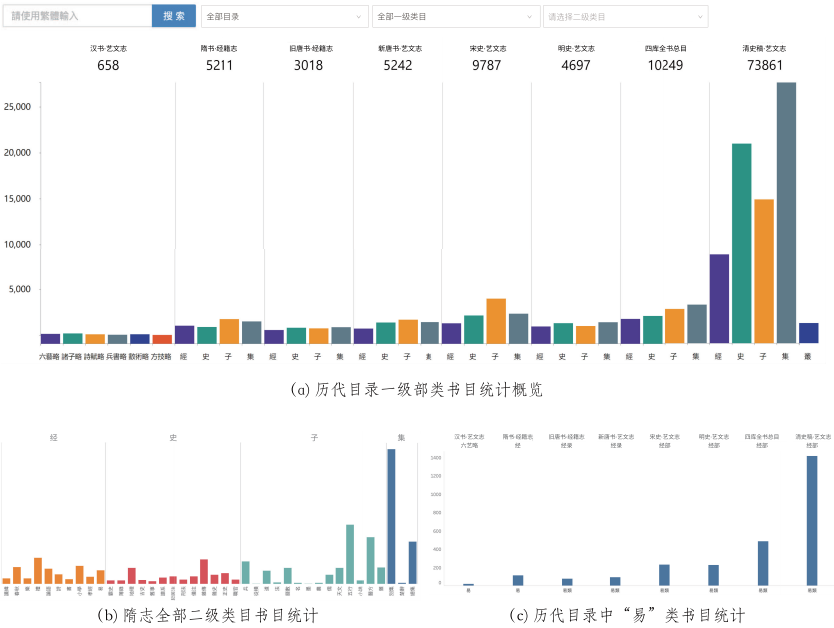
R2T1:横向对比。对特定目录中一级、二级类目下所收典籍的数量进行比较分析,以窥探当时的知识分布概貌。

R2T2:纵向追踪。对历代目录中特定类目

下的典籍数量进行对比分析,以揭示该思想、学说的发展趋势及兴衰演变。

R2T3:典籍存佚。考察历代目录对前代典籍的收录情况以及其所收典籍在后续目录中的留存情况。

以上三个任务的核心都是完成典籍数量的对比,因而选用柱状图进行可视化呈现,以最有效地通过位置和长度两种视觉元素来传达数量对比的信息。进入可视化分析系统后,用户可通过图 3(a)对历代目录一级部类书目形成总体概览性认知。图 3(a)反映的是对八种目录的整体书目统计,以及每部目录中一级类目的典籍数量。在此图的基础上,用户可以点击任意目录查看该目录所有二级类目的典籍统计,从而实现 R2T1 的分析任务(图 3(b)以隋志为例)。同时,该柱状图的上方是目录和类目的筛选区,用户可以通过筛选器选择所关注的目录和某一具体二级类目,从而对比该类目典籍数量的历代变化以支持 R2T2 的分析(图 3(c)以全部目录中的“易”类为例)。



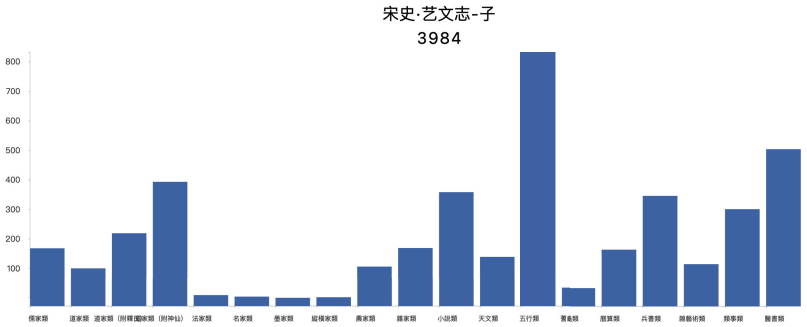

(a) 历代目录一级部类书目统计概览

(b) 隋志全部二级类目书目统计

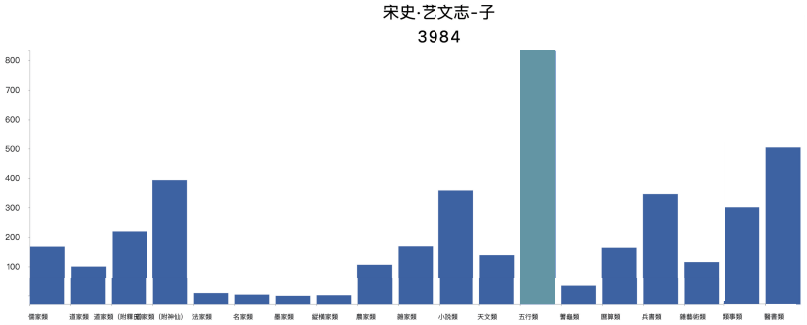
(c) 历代目录中“易”类书目统计

**图 3　书目统计可视化——横向及纵向对比分析示例**

书目统计可视化可以通过缩放、筛选、检索等交互帮助用户按照视觉信息搜索的流程进行 R2T1 和 R2T2 的可视化分析。例如,用户通过各目录一级类目的典籍统计概览图可以清晰地看出除了明志(只收录当朝之书)以外,各目录中的典籍数量是随朝代递增的。而在所有一级类目中,宋志的子部典籍数量明显高于清史稿以前的各目录的一级类目,点击宋志的子部便可以对该类目进行放大观察,查看其下的二级类目和对应的具体典籍条目(见图 4(a)),可以发现其中“五行”类书籍最多。按照用户的需求,还可以通过点击某一、二级类目查看该类目所收典籍条目的数据细节和原始文本(图 4(b)以“五行”类为例)。进一步地,用户还可以通过查询其他目录中的子部典籍分布进行对比,如可以发现旧唐志和新唐志中最多的是道家典籍,而明志中最多的则为儒家典籍,清史稿中医书最多,以此,人文学者便可以进一步推究其背后的学术文化背景。




(a) 宋志—子部下二级类目的书目统计




| 编号 | 书名 | 卷数 | 原始类目 | 作者信息 |
|---|---|---|---|---|
| #05031300001 | 三命通照神白經 | 三卷 | 子部/五行類 | ['(不詳-不詳-1-郭璞-撰-_責任人-_確定-_不詳--)'] |
| #05031300002 | 五行運氣 | 一卷 | 子部/五行類 | ['(不詳-不詳-1-陶弘景-撰-_責任人-_確定-_不詳--)'] |
| #05031300003 | 青子錄班氏經 | 一卷 | 子部/五行類 | [] |
| #05031300004 | 五行元統 | 一卷 | 子部/五行類 | ['(不詳-不詳-1-李淳風-撰-_責任人-_確定-_不詳--)'] |
| #05031300005 | 太一金鏡式經 | 十卷 | 子部/五行類 | ['(不詳-不詳-1-王希明-撰-_責任人-_確定-_不詳--)'] |

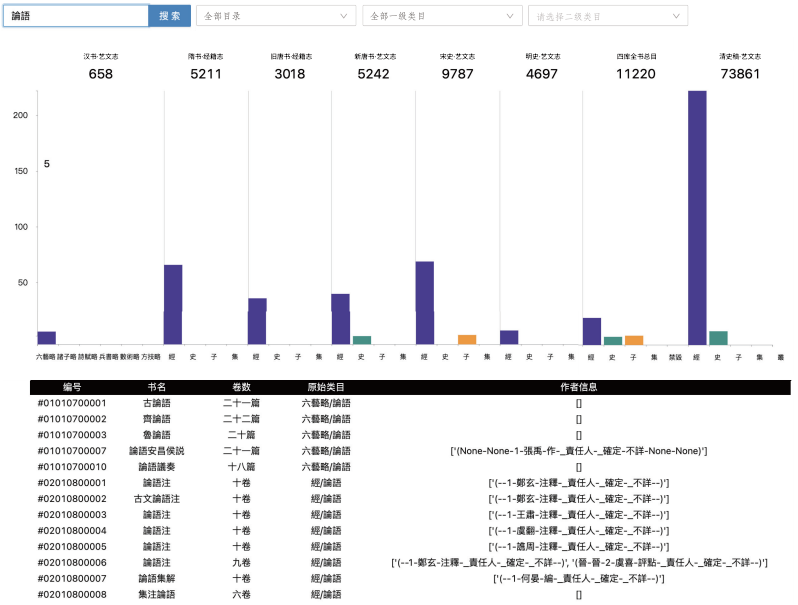
(b) 宋志—子部—“五行”类具体典籍数据详情

**图 4　书目统计可视化——以宋志子部为例逐步分析**

此外,为了支持用户更细致的书目数据查询需求,可视化分析系统还提供检索功能,可以支持对题名、责任者以及规范题名中含有该字段的典籍条目进行模糊匹配,从而查看其所属类目以及原始文本条目。以关键词“论语”为例,从检索结果中可以发现其在新唐志以前都只收录于经部,而自新唐志起在史部、子部也有收录(见图 5)。

从数据规模来看,R2T1 和 R2T2 的统计工作通过人工计数虽然耗时却也是可以实现的,而 R2T3 对典籍存佚分析所涉及的时间跨度和数据量统计则是难以通过人工方法完成的,甚至可以说该分析任务的提出是建立在由计算机辅助完成的大规模数据处理、规范化集成工作基础之上的,这体现了数字学术、数字人文不仅可以为传统的学术研究提供数字化的分析方法,还可以为传统的学术研究提供新的发现问题的视角[5]。在进行了书目认同工作之后,便可以判断某目录中的一部典籍是否在前代目录中被收录过,以及其是否在后续的目录中依然被收录。如果一部典籍在前代目录中没有出现过,则可以认定为当代的新生典籍,如果在后续目录中未被收录则可暂定为“亡佚典籍”。当然,此处的“亡佚”指的是在目前的数据集范围内该书在后续的目录中没有留存,并不意味着彻底亡佚。为了直观对比历代目录的新生典籍和典籍留存情况,本文以图 6 所示的可视化进行了呈现。从图 6 中可以清晰地看出,新唐志、旧

唐志对前代典籍的收录较多,而隋志、宋志和四库提要的新生典籍较多;从典籍的留存情况来看,隋志、旧唐志中的典籍在后代留存的相对较多,而汉志和宋志中的典籍留存较少。这种统计方式和可视化可以有效帮助人文学者对典籍的存佚进行分析,在后续研究中,通过对现存古籍总目的数据加以集成并接入该可视化分析系统,便可以反映各代典籍的留存亡佚情况。




| 编号 | 书名 | 卷数 | 原始类目 | 作者信息 |
|---|---|---|---|---|
| #01010700001 | 古論語 | 二十一篇 | 六藝略/論語 | [] |
| #01010700002 | 齊論語 | 二十二篇 | 六藝略/論語 | [] |
| #01010700003 | 魯論語 | 二十篇 | 六藝略/論語 | [] |
| #01010700007 | 論語安昌侯説 | 二十一篇 | 六藝略/論語 | ['(None-None-1-張禹-作-_責任人-_確定-不詳-None-None)'] |
| #01010700010 | 論語讲奏 | 十八篇 | 六藝略/論語 | [] |
| #02010800001 | 論語注 | 十卷 | 經/論語 | ['(--1-鄭玄-注釋-_責任人-_確定-_不詳--)'] |
| #02010800002 | 古文論語注 | 十卷 | 經/論語 | ['(--1-鄭玄-注釋-_責任人-_確定-_不詳--)'] |
| #02010800003 | 論語注 | 十卷 | 經/論語 | ['(--1-王肅-注釋-_責任人-_確定-_不詳--)'] |
| #02010800004 | 論語注 | 十卷 | 經/論語 | ['(--1-虞翻-注釋-_責任人-_確定-_不詳--)'] |
| #02010800005 | 論語注 | 十卷 | 經/論語 | ['(--1-譙周-注釋-_責任人-_確定-_不詳--)'] |
| #02010800006 | 論語注 | 九卷 | 經/論語 | ['(--1-鄭玄-注釋-_責任人-_確定-_不詳--)', '(晉-晉-2-虞喜-評點-_責任人-_確定-_不詳--)'] |
| #02010800007 | 論語集解 | 十卷 | 經/論語 | ['(--1-何晏-編-_責任人-_確定-_不詳--)'] |
| #02010800008 | 集注論語 | 六卷 | 經/論語 | [] |

**图 5　检索“论语”后相关书目分布展示**

**图 6　新生典籍及典籍留存情况**

## 4.2　目录分类演化

古代典籍的分类是古典目录学的重要研究内容。古籍目录分类的价值一方面体现在便于典籍的查找,“即类求书,因书究学”,另一方面由于古籍目录的分类始终以图书内容、学术门类为主要依据,因而能分门别类地总结学术源流,即郑樵所谓“类例既分,学术自明”[34]。虽自隋志以后四部分类法稳定下来,但四部中的细类划分是随着学术发展和典籍数量的增多而变化的,且即使在各目录中的细类名称相同的情况下,著录内容和划分标准也并不完全一致,因而要了解古籍目录的分类不能只看类目标

题,还需要考察各类著录了什么书,并比较各目录如何处理不同典籍的归属[1]。当前古籍目录分类研究多以定性的、局部的研究为主,而本研究所使用的经过规范化和书目认同处理后的集成目录数据则为跨时代的全局分类演化分析提供了数据支撑。相较于从类目本身及其小序来分析学术流派演变,此数据集可以支持从类目所收录典籍的角度自下而上研究类目的内涵及其变化。同上,在设计分类演化研究的可视化分析功能时,首先对专家用户提出的分类演化研究需求 R3 进行了分析任务拆解。

R3T1:典籍层。查询某一部典籍在历代目录中所归属类目的变化。

R3T2:类目层。考察分析每个目录子类的分合流变,即该类目所收典籍在历代目录中的来源去向以及所占比例。

R3T3:全局层。基于以上两层的总体分析,总结出各种知识分类演化的模式。

其中,R3T1 是基本分析单元,其他两个任务建立在实现 R3T1 任务的基础之上。为完成 R3T1,需要将每一部典籍分别映射到每一部收录过该典籍的目录的具体子类目,也就是说在本研究的数据集中,典籍分类记录可能涉及八个维度(目录),而每一个维度则有几十种子类作为其值域。从数据来看,由于分析的是分类变化,对于只被一部目录收录的典籍不纳入分析范围,因而在书目认同工作的基础上,该可视化以 7 046 条至少被两部目录收录过的典籍及其分类信息作为主要数据;从可视化设计来看,对于多维数据的可视化展示,最常用的可视化技术是平行坐标。虽然平行坐标最初被用于展示样本数据的不同定量变量的取值,但由于它在多元数据的分类、变化等探索性分析上具有优势[35],因而十分适用于展示大量典籍的分类演变。由此,本研究构建了基于典籍的历时性分类演变轨迹的交互式可视化分析功能(见图 7)。图中的每一个纵轴代表一部目录,纵轴上的每一个坐标点为该目录的二级子类目,每一条线从不同纵轴的某坐标点穿过即展示一部典籍在每个目录中分别归属于哪个类目,即该典籍的分类演变轨迹,轨迹的颜色与典籍第一次被收录的目录相对应。需要说明的是,由于不同典籍可能具有相同的分类演变轨迹,因此每一条线可能代表了不同典籍的重合轨迹。此外,由于在典籍的流传过程中,并不是所有目录都收录过该典籍,因此每个目录坐标的最下方都补入“未收录”这一坐标点以实现典籍流传轨迹的延续性。

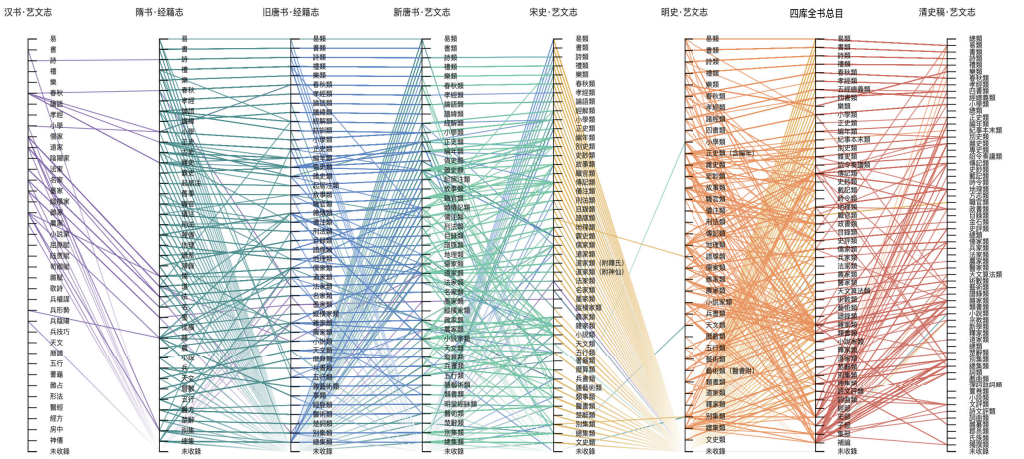
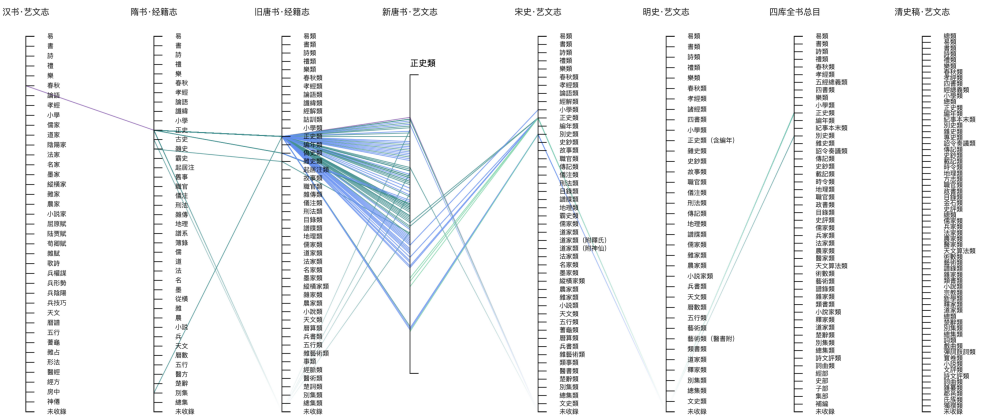



**图 7　目录分类演化概览**

对于任务 R3T1,一方面,用户可以通过检索一部典籍从而观察其分类演变在该可视化中的轨迹线条及相应的分类信息;另一方面,用户可以通过点击可视化中的一条具体的轨迹线来查看其相应的典籍书目信息。对于任务 R3T2,从微观层面看,用户可以通过点击任意纵轴上的坐标点来查看相应目录子类中所收录的具体典籍的分类演变(图 8(a)以新唐志中的“正史”类为例);从宏观层面看,该系统还提供弧形图来呈现某一类目中的典籍在前代和后代目录中的归属类目,其中线条的粗细反映对应典籍的数量。以新唐志中的“正史”类为例(见图 8(b)),从可视化中可以看出该类目中 60%的典籍在旧唐志中的“正史”类被收录过,36%的典籍在隋志中的“正史”类被收录过,还有很少一部分在前代目录中归属于“伪史”“杂史”等类目,而该类目下的典籍仅有不到 10%流传到后代目录中的“正史”类,也有少数流传到“别史”类。而对比宋志与四库提要,可以发现留存的典籍完全一致,这说明在宋代留存下来的“正史”类典籍,明代以后基本没有亡佚,究其原因,可能是“正史”类典籍非常重要,且数量也不多。以此全局和细节相结合的展示方式,用户可以通过此分类演化交互式可视化,从不同层面进行探索分析。




(a) 新唐志—正史类典籍分类演化轨迹

(b) 新唐志—正史类典籍源流及数量占比

**图 8　任一类目典籍的分类演变——以新唐志“正史”类为例**

与 R2T3 类似,R3T3 问题的提出也是对以往古典目录学分类研究的突破,在本研究所开发的基于平行坐标可视化技术的分析工具中,用户可以通过人工观察或者自动化图形抽象两种方式进行图像分析,即研究者通过对不同的类目演化模式进行遍历性的观察来对类目分合转化模式进行归纳分析,或者可以通过拓扑结构分析的方法对平行坐标进行视觉抽象,从而对类目分合转化模式进行自动化的全面总结。

## 5 总结

本研究以对古典文献及图情领域都有重要研究价值的古籍目录为研究对象,以探究古籍目录的分类演化及其背后的学术流衍为核心研究问题,对八种代表性的官修史志目录进行了数据集成,并构建了交互式的可视化分析系统以支持学者的检索、统计、分析等研究需求。数字学术背景下,人文学者的需求主要体现在数据集成和分析工具两方面[36],本研究的主要贡献便是从这两个方面为古籍目录的相关研究提供支持,为学者省去了大量进行重复性数据收集、整理及统计等基础工作的时间,并通过新的技术和视角助力解释性研究。

首先,数字学术可以通过信息技术将大量的学术文献数字化、数据化,使学者可以基于更大的信息量提出新的研究问题,并显著地从时间和空间上拓宽对问题的理解[4]。本研究通过机器预处理与专家校对的迭代处理模式对八种古籍目录数据进行集成,该数据集实现了对题名、责任者等项目的结构化和规范化,且书目选择具有时间跨度长、典籍覆盖范围广等特点,为未来古籍目录研究,特别是基于计算手段的数据分析及统计提供了数据基础。同时该数据处理方法及流程可以为后续更大范围的古籍目录数据集成提供程式化参考。在数字人文研究领域,目前主要的制约瓶颈是基础数据的加工,而本研究证明了人机迭代的方式既能保证数据的质量,又能极大地提升数据加工效率,为大规模人文数据加工提供了工程实施思路和成功的可行性验证。

其次,在文献数据集成的基础上,以重要的学科问题为框架来进行学术论证并以多种形式提供学术解释是数字学术未来发展中的必要组成部分[4]。本研究构建的交互式可视化分析系统实现了对古籍目录数据的细粒度统计和可视化呈现,可以帮助学者清晰地比较、追踪类目的消长。同时,该系统将所有典籍在历代目录中的分类演变轨迹以及各类目所收典籍的源流进行可视化呈现,可以更好地实现类目分合转化的模式识别,进而探究历代知识生产情况以及知识组织的流变,并对其背后蕴含的规律和意义进行诠释。在数字学术环境中,该系统作为分析工具并不是要提供最终的研究话语权,而是为学者识别问题、发现问题和解决问题提供新的视角和手段。

在以上数据集成和可视化分析系统的基础上,未来笔者还将对更多书目进行集成,包括更多公藏和私家目录以及《中国古籍总目》等,以丰富古籍书目数量,同时还可为官私目录对比、典籍存佚等问题提供数据基础。在数据处理上,将引入众包模式,鼓励更多的学者参与到数据处理及校对的过程中,使数字学术拥抱开放学术;在系统构建上,将不断吸收用户对系统的各类建议和反馈,以实现系统的持续优化,根据用户的研究需求构造新的分析和可视化功能,最终构建一个数据全面、功能强大的中国古籍总目分析平台,以支持数字学术背景下的古典文献和目录学研究。

**致谢**: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。

**参考文献**

[1] 来新夏. 古典目录学(修订本)[M]. 北京:中华书局,2013. (LAI X X. Traditional bibliographic studies (revised) [M]. Beijing:Zhonghua Book Company,2013.)

[2] 王鸣盛. 十七史商榷[M]. 上海:上海古籍出版社,2013. (WANG M S. Discussions on seventeen histories [M]. Shanghai:Shanghai Classics Publishing House,2013.)

[3] 柯平,刘旭青. 中国目录学七十年:发展回溯与评析[J]. 中国图书馆学报,2019,45(5):101-111. (KE P,

LIU X Q. 70 years of Chinese bibliography development:retrospect and analysis[J]. Journal of Library Science in China,2019,45(5):101-111.)

[4] AYERS E L. Does digital scholarship have a future?[J]. EDUCAUSE Review,2013,48(4):24-26.

[5] RUMSEY A S. New-model scholarly communication:road map for change[C]//Ninth Annual Meeting of Scholarly Communication Institute. Charlottesville,USA,2011.

[6] 中华人民共和国文化部. WH/T 66-2014古籍元数据规范[S]. 北京:国家图书馆出版社,2015.(Ministry of Culture of the PRC. WH/T 66-2014 The Chinese ancient books metadata specifications[S]. Beijing:National Library of China Publishing House,2015.)

[7] 夏翠娟,林海青,刘炜. 面向循证实践的中文古籍数据模型研究与设计[J]. 中国图书馆学报,2017,43(6):16-34.(XIA C J,LIN H Q,LIU W. Designing a data model of Chinese ancient books for evidence-based practice[J]. Journal of Library Science in China,2017,43(6):16-34.)

[8] 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 Bian Kao thought of Chinese classical bibliography under the background of digital humanities[J]. The Library Journal of Shandong,2019(1):14-19,40.)

[9] 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 essentials:an explorative tool for catalog annotations in Chinese rare-book collections[J]. Journal of Library Science in China,2021,47(4):97-112.)

[10] 汪辟疆. 目录学研究[M]. 上海:华东师范大学出版社,2000.(WANG P J. Studies of bibliographies[M]. Shanghai:East China Normal University Press,2000.)

[11] 姚名达. 中国目录学史[M]. 北京:商务印书馆,2014.(YAO M D. History of Chinese bibliographies[M]. Beijing:The Commercial Press,2014.)

[12] 马学良. 从《四部分类源流一览表》看目录学的考辨作用[J]. 图书馆理论与实践,2012(7):46-49.(MA X L. The Biankao function of bibiliographies:from *Four-division Classification Evolution Glance*[J]. Library Theory and Practice,2012(7):46-49.)

[13] TSIEN T-H. A history of bibliographic classification in China[J]. The Library Quarterly,1952,22(4):307-324.

[14] 耿素丽. 浅析古籍目录中子部农家类与谱录类之关系[J]. 文献,2002(1):158-169.(GENG S L. Brief analysis of the relation of Nongjia category and Pulu category in ancient book catalogs[J]. The Documentation,2002(1):158-169.)

[15] 万彩红. 史志目录中易学文献分类研究[J]. 南方论刊,2015(6):58-60.(WAN C H. Study of the classification of Yi category in the bibliographic sections of Chinese history books[J]. Nan Fang Lun Kan,2015(6):58-60.)

[16] 程有庆. "杂家"与"杂家类"浅说[J]. 文献,2002(3):233-241.(CHENG Y Q. Brief discussions of "Zajia" and "Zajialei"[J]. The Documentation,2002(3):233-241.)

[17] ZHAO X H. Xiaoshuo as a cataloguing term in traditional Chinese bibliography[J]. Sungkyun Journal of East Asian Studies,2005,5(2):157-181.

[18] 李明杰. 数字环境下古籍整理范式的传承与拓新[J]. 中国图书馆学报,2015,41(5):99-110.(LI M J. The inheritance and innovation of ancient book collation paradigm in the digital environment[J]. Journal of Library Science in China,2015,41(5):99-110.)

[19] 胡小菁. 文献编目:从数字化到数据化[J]. 中国图书馆学报,2019,45(3):49-61.(HU X J. Cataloging from digitization to datafication[J]. Journal of Library Science in China,2019,45(3):49-61.)

[20] 姚伯岳,张丽娟,于义芳,等. 古籍元数据标准的设计及其系统实现[J]. 大学图书馆学报,2003(1):17-21.(YAO B Y,ZHANG L J,YU Y F,et al. On the design of rare book metadata standard and its system implementation[J]. Journal of Academic Libraries,2003(1):17-21.)

[21] MORETTI F. Graphs, maps, trees: abstract models for a literary history[M]. London: Verso, 2005.
[22] 王军. 从人文计算到可视化——数字人文的发展脉络梳理[J]. 文艺理论与批评, 2020(2): 18-23. (WANG J. From humanities computing to visualization: the evolution of digital humanities[J]. Theory and Criticism of Literature and Art, 2020(2): 18-23.)
[23] SALAH A A A. The evolution of classification systems: ontogeny of the UDC[J/OL]. CoRR, 2012[2021-02-28]. http://arxiv.org/abs/1204.3769.
[24] CHOI I. Visualizations of cross-cultural bibliographic classification: comparative studies of the Korean Decimal Classification and the Dewey Decimal Classification[J]. NASKO, 2017, 6(1): 39-55.
[25] SMIRAGLIA R P. Disciplinary, asynthetic, domain-dependent: NARCIS a national research classification in isolation[J]. Advances in Classification Research Online, 2017, 28(1): 7-10.
[26] CHOI I, PARK M S. Specificity and exhaustivity of bibliographic classifications—a cross-cultural comparison with text analytic approach[C]//iConference 2020 Proceedings. Bross, Sweden, 2020.
[27] SMIRAGLIA R P, SCHARNHORST A, SALAH A A A, et al. UDC in action[C]//Proceedings of the International UDC Seminar. Hague, Netherlands, 2013.
[28] SINCLAIR S, RUECKER S, RADZIKOWSKA M, et al. Information visualization for humanities scholars[M]// Literary studies in the digital age: an evolving anthology. New York: Modern Language Association of America, 2013.
[29] BURDICK A, WILLIS H. Digital learning, digital scholarship and design thinking[J]. Design Studies, 2011, 32(6): 546-556.
[30] SHNEIDERMAN B. The eyes have it: a task by data type taxonomy for information visualizations[M]//BEDERSON B, SHNEIDERMAN B. The craft of information visualization. San Francisco: Morgan Kaufmann Publishers Inc., 2003: 364-371.
[31] KEIM D A, MANSMANN F, SCHNEIDEWIND J, et al. Challenges in visual data analysis[C/OL]//Tenth International Conference on Information Visualisation (IV'06), 2006[2021-02-28]. https://dl.acm.org/doi/10.1109/IV.2006.31.
[32] CARD S K, MACKINLAY J, SHNEIDERMAN B. Readings in information visualization: using vision to think[M]. San Francisco: Morgan Kaufmann Publishers Inc., 1999: 1-34.
[33] FRIENDLY M. The golden age of statistical graphics[J]. Statistical Science, 2008, 23(4): 502-535.
[34] 杜泽逊. 文献学概要[M]. 北京: 中华书局, 2008. (DU Z X. Summary of documentation studies[M]. Beijing: Zhonghua Book Company, 2008.)
[35] HEINRICH J, WEISKOPF D. State of the art of parallel coordinates[C/OL]//Eurographics 2013, 2013[2021-02-28]. http://diglib.eg.org/bitstream/handle/10.2312/conf.EG2013.stars.095-116/095-116.pdf?sequence=1&isAllowed=y.
[36] BLANKE T, HEDGES M, DUNN S. Arts and humanities e-Science: current practices and future challenges[J]. Future Generation Computer Systems, 2009, 25(4): 474-480.

**李文琦** 北京大学信息管理系、北京大学数字人文研究中心博士研究生。北京 100871。
**王凤翔** 北京大学信息管理系、北京大学数字人文研究中心博士研究生。北京 100871。
**孙显斌** 中国科学院自然科学史研究所研究员。北京 100190。
**黄芷欣** 北京大学中文系博士研究生。北京 100871。
**李芃蓓** 中华书局文学编辑室助理编辑。北京 100073。

(收稿日期: 2022-02-11)