DOI：10.13998/j.cnki.issn1002-1248.21-0264

面向典籍内容分析的分类体系构建方法研究

艾毓茜，徐　健，何　琳*，戚　筠
（南京农业大学 信息管理学院，南京　210095）

摘　要：［目的／意义］随着面向典籍的数字人文研究的不断深入，对基于文本内容的细粒度分类要求不断提高，合理的分类已成为数字化典籍研究和有效利用的关键。［方法／过程］研究利用分面分类思想，以典籍文本数据及相关典籍词典为研究对象，结合概念语义信息，组织并描述典籍内容数据特征。［结果／结论］本文构建的分类体系突破典籍数量、体裁和种类的限制，从政治、经济、文化、社会和军事 5 个维度将典籍内容进行有序的组织与揭示，对典籍数字资源的深度开发和利用具有重要价值。

关键词：数字人文；典籍数字资源；分面分类法；信息组织
中图分类号：G350　　**文献标识码：**A　　**文章编号：**1002-1248（2021）09-0018-09

引用本文：艾毓茜, 徐健, 何琳, 等. 面向典籍内容分析的分类体系构建方法研究[J]. 农业图书情报学报, 2021, 33(9)：18-26.

A Construction Method of the Classification System Oriented to Content Analysis of Ancient Books

AI Yuxi, XU Jian, HE Lin*, QI Yun
（College of Information Management, Nanjing Agricultural University, Nanjing 210095）

Abstract: [Purpose/Significance] With the deepening of digital humanistic research on ancient books, the requirement for fine-grained classification based on text content is increasing continuously, and reasonable classification has become the key to the research and effective utilization of digital ancient books.[Method/Process] The research uses the concept of faceted classification, takes the text data of ancient books and related dictionary of ancient books as the research object, and combines conceptual semantic information to organize and describe the features of the

收稿日期：2021-04-12
基金项目：国家社科基金项目“基于典籍的中华传统文化知识表达体系自动构建方法研究”（18BTQ063）
作者简介：艾毓茜（ORCID：0000-0001-9141-9138），南京农业大学信息管理学院，研究方向为人文计算。徐健（ORCID：0000-0002-0230-5137），南京农业大学信息管理学院，研究方向为信息组织与数据挖掘。戚筠（ORCID：0000-0001-8157-4732），南京农业大学信息管理学院，研究方向为科学数据管理
*通信作者：何琳（ORCID：0000-0002-4207-3588），女，南京农业大学信息管理学院教授、博士生导师，研究方向为信息组织及文本挖掘。Email：helin@njau.edu.cn

content of ancient books. [Results/Conclusions] The classification system constructed in this paper breaks through the limitation of the number, genre and type of ancient books. The research selects five dimensions of politics, economy, culture, society and military to organize and reveal the contents of ancient books in an orderly manner, which is of great value to the in-depth development and utilization of digital resources of ancient books.

Keywords: digital humanities; digital resource of ancient book; faceted classification; information organization

1 引　言

典籍作为记录中华文明史的重要载体，承载着浩繁且丰富的传统文化知识，对于史学与古文化知识的研究具有重要意义。传统的典籍资源研究集中于古籍资源的保存、整合和利用等方面，通过开发古籍数据库系统，实现线上更新收录资源以便利用，如《汉语古籍电子文献知见录》[1]、奎章阁网站[2]等。但这类传统的典籍资源利用和开发方法，对于大规模古籍数据的深度开发和利用率较低，难以适应人文学科研究发展[3]。

数字人文技术的蓬勃发展，为古籍文本深度开发提供了新的技术与思路，利用文本挖掘等多种中文信息处理技术可以帮助充分揭示和组织古籍数字资源，使其成为立体的学术知识库，有效提高了古籍资源的开发利用效率[4]。如何借助数字人文技术对古籍资源进行深度挖掘与知识发现，以便成就学业、研治古文的人使用，具有重要的学术意义与价值[5]。

为实现基于典籍内容细粒度知识单元的管理、共享和重用的目的，需构建一个面向数字人文领域、以典籍内容分析为基础的系统、全面的分类体系，以实现更准确有效的信息检索。已有的分类体系研究大多面向典籍的外部形式，多以典籍的载体形式[6]、记叙手法[7]和语言结构[8]为依据进行分类，这类知识组织方法不能有效揭示典籍内在知识信息和语义关系，分类较粗糙。随着数字人文研究的不断深入，对基于文本内容的细粒度分类要求不断提高。在此背景下，本文意图构建一种较为通用的典籍内容分类体系，这一体系综合了分面分类理念和字词语义信息，从政治、经济、文化、社会和军事 5 个维度将典籍内容进行有序的组织与揭示。

2 相关研究

2.1 语义组织研究

知识单元是知识领域中知识控制与处理的基本单位，是一切知识管理活动的前提和基本对象[9]。在自然语言处理中，语义组织是对知识单元间的语义关系进行描述，并进行存储以便交流和传递，其目的是通过各种数字人文技术，将非结构化的文本数据资源转化为结构化数据，并将数据间的语义关系通过叙词表、元数据、本体等多种方式进行组织，以实现数据的关联化和智能化[10]，可以认为语义组织关系着信息服务和信息共享的质量和水平。面向数字人文领域的语义组织主要包括知识建模和知识抽取两个方面。

知识建模通过对知识单元的结构化、模型化表达，实现文本数据知识的语义化和共享化。传统知识建模以分类叙词表为主，早期多通过手工标引的方式，进行知识组织，如《历代进士登科数据库》[11]。随着数字人文技术的发展，利用分词、词性标注、命名实体识别、文本挖掘等自然语言处理技术，有效实现对大规模数据资源的语义组织，可以快速抽取出典籍数据资源中的人名、地名、官职等信息[12]，丰富了实体间关系的表示方式及更广泛的知识组织，为数字远读奠定了基础。

在数字人文领域，知识抽取主要用于识别大规模数据资源中潜藏的知识及其之间的语义关系，目前主要有基于规则匹配和基于机器学习两种方法。其中，基于规则匹配的方法通过人工对文本资源进行特征分

DOI：10.13998/j.cnki.issn1002-1248.21-0264

析，以相应的领域知识为基础构建正则表达式，从而实现基于规则的知识抽取，如 CBDB 项目[12]中领域专家以相应领域知识为背景，针对墓志铭等设计知识抽取正则表达式；丁君军等[13]针对学术文献中的概念属性描述，构建描述规则用以抽取学术概念属性。而基于机器学习的方法通过对少量语料数据进行标注，训练模型，以实现对大量文本的自动抽取，如意大利自然语言处理实验室设计的 LinguA、READ-IT、T2K 等工具，以实现文本标注、命名实体识别、可视化[14]。

2.2 典籍分类体系研究

典籍作为文化传承的重要载体，如何有效组织和利用典籍一直都是人文学者研究的重点之一。类书作为典籍的荟萃，将某一门类的古籍通过一定的方法加以组织以便寻检和征引，从魏晋南北朝的《皇览》到明清时期的《永乐大典》，对文献保存和学术研究起到了重要作用[15]。

20 世纪末，随着计算机技术的发展，中文古籍数字化逐渐成为国内典籍研究的重点。在典籍数字资源组织方面，王依民先生将传统文献学与数字技术相结合，提出“数字文献学”概念，研究涉及古籍文史资源的保存、整合、加工、传播和利用等方面[6]。此后，有学者提出“古籍电子文献学”，从古籍数字资源的分类与导航、古籍联合目录和古籍数字资源的评价研究 3 个方面，展开对古籍数字资源的目录学的研究[7]。为适应大规模典籍数据，学者们根据古籍数据资源分散、形式多元、数据格式多样等特点，建立多种估计数据库导航系统，如《汉语古籍电子文献知见录》[1]、奎章阁网站[2]等，在实现线上更新收录资源的同时，提高与用户的交互性。

随着典籍分类体系的深入研究，单一的使用《中图法》或《四库法》进行典籍分类组织，无法将典籍的表象主题与深层主题进行有效结合，研究者开始将分面分类法引入典籍分类研究中。罗艳秋等[16]在综合分析民族医药典籍内容特征的基础上，结合《中国中医古籍总目》，对民族医药古籍进行分类组织，共划分 11 个大类，并进一步细分三到四级小类以便使用。李娜等[17]以《方志物产》山西卷为研究对象，从物产、土产、食货、方产等方面对《方志物产》中的物产西信息进行分类组织，实现了物产类目信息的智能完善。而针对古籍数据库，张力元等[6]提出利用分面分类法，构建古籍数据库分面分类体系，包括主题、类型、建置主体、格式、权限和地区等 6 个维度，在粗粒度层面对现有古籍数据库资源进行了组织。

2.3 典籍数字资源内容分析研究

早期的典籍研究多以典籍词汇研究为主。古籍词汇研究始于汉代，学者在古籍的注疏中解释古代语词，如《尔雅》《说文解字》等，为后人研究奠定了基础[18]。目前多集中于词汇系统的发展、新词的产生与变化、词义演变以及构词法的发展等领域。社会制度、环境的演变使得词汇数量增多，典籍文本中词汇的变化反映着社会情况的变化，通过分析词汇发展脉络借以分析社会发展的情况[19,20]。

以词汇研究为基础，国内外学者面向典籍的文本内容展开了事件抽取、主题挖掘及相关知识组织研究。RYAN 等[21]对中国古代和中世纪的 500 多万字的语料库进行主题建模，从相交主题和不相交主题两个角度，对《论语》《孟子》和《荀子》的竞相关系进行了解释。彭炜明等[22]在实例挖掘的基础上，提出采用模式驱动的方式，构建《资治通鉴》历史领域本体，以实现对《资治通鉴》先秦史部分的深度开发。何琳等[4]利用词匹配算法抽取特征词语料，然后使用 LDA 主题模型对语料进行处理，并结合相关时间信息进行主题强度计算，从盟会、礼仪、战争、权力斗争和周礼治国等主题入手，对春秋时期社会发展态势进行了分析。

综上所述，本文借鉴分面分类思想，以语义组织中知识建模和知识抽取的方法和技术为支撑，提取典籍数据中的概念及其关系，从细粒度知识单元语义信息的层面对典籍内容进行组织和揭示。面向典籍内容分析的分类体系的构建可以突破典籍数量、体裁和种类的限制，有效地从典籍中抽取相关特征，为成就学业、研治古文的人删繁取要，进而推动对典籍内容的研究。

3　构建方法

3.1　分面分类法

典籍数字资源与文化传承及其相关活动密切相关，因此具有一定的领域独特性：①文化性，典籍资源产生于中华民族历史社会中的某一特定时期，一定程度上反映了当时环境下人类的人文、历史、艺术等情况，是国家和民族的文化积淀。②延续性，典籍资源记录了中华文明发展的历史进程，即使其所记录为数千年前之事，研究者们仍旧可以通过保存的典籍资源去发现历史奥秘。③分散性，典籍资源涉及的信息涵盖社会、经济、政治、军事、文化等多个领域，且分布广泛，很难在一部典籍中获得全部信息。④繁杂性，典籍数据资源的语义和形式都很复杂，且古汉语与现代汉语表达结构有很大差别。

针对典籍数据资源的以上特性，在构建面向典籍内容分析的分类体系时，需充分考虑典籍资源中数据的语义和形式特征，而分面分类法可以很好地根据不同的方面和范畴对数据进行有效划分，通过多个组合表达复杂主题[23]，因此本文利用分面分类法，考虑到分类体系的易用性，采用“分面–类目”结构，以实现对典籍内容多维度的组织与揭示。

3.2　数据来源

本研究所构建的典籍分类体系研究对象为典籍内容数据，而典籍的类型、编撰时间及其社会背景决定了典籍的内容。由于历史典籍一词多义现象严重、文本短、缺乏结构性，且在大量的古代用词，与现代常用词难以对应，因此本文广泛收集与历史典籍相关的主题词表和词典，如与《左传》相关的杨伯峻《春秋左传词典》等，这些词典是由专业人士编制的成熟的词典，一定程度上保障了信息准确性和有效性。在选词过程中，以词典中的词释义为主要依据，通过对词的释义进行解析，对词进行分类，并从相关历史典籍如《公羊传》《史记》等中进行抽词，从而保证自然语言环境下可以用典籍中的词语进行检索。

3.3　确定分类框架

为确定面向典籍内容分析的分类体系的具体分面，本研究结合《中图法》并参考相关古籍分类与内容分析研究文献，以深入知识单元的词义为主要分类依据，确定最能有效描述与划分典籍数据的类别维度为：政治、经济、文化、社会、军事。

概念体系的建立以一般叙词表的概念间逻辑关系为基础，采用分类法编制标签分类索引，来表示词间的等级关系和属性关系。通过对词典中的词数据进行初步标引，对每个大类下各小类进行简单划分，采用自下而上和自上而下相结合的方法，构建基于词典和史籍的分类体系的概念语义网络。

3.4　概念抽取

K-means 聚类算法自上世纪 50 年代被提出后，广泛应用于不同学科领域的聚类划分[24]。K-means 算法通过反复迭代，从初始 K 个类别开始计算，分别将数据划分至已知类别，并重新计算类别中心，最终使得各类别总距离平方和趋于最小值[25]。K-means 算法具有简单、高效等优势，且类别个数 K 值可通过人工指定，因此本文利用 K-means 算法，本研究根据分类框架设计二级类目，对杨伯峻的《春秋左传词典》以词释义为文本相似度计算对象，并引入《汉语大词典》对释义进行扩展，对词头进行分类，以实现面向典籍内容分析的分类体系二级类目的划分。其中相似度计算分为两部分：①分词后利用 TF-IDF 计算词向量间的相似度，相似度超过阈值（0.3），即认为两个词属于同类词。②若 A 词的词头出现于 B 词的释义中，认为 A、B 两词为同义词。

算法中 K 值设置为 6，迭代次数为 10 000，即分类结果共输出 6 类。观察输出结果，为其中 5 类赋予最接近的类名：政治、经济、文化、社会、军事，第六类为手工分类的补充数据。观察第一次聚类结果后，对赋予类名的 5 类词进行简单筛选，将不属于当前类的词剔除至作为手工分类补充数据，分别对 5 个类别进行二次聚类。对第二次聚类结果进行简单筛选后，

DOI：10.13998/j.cnki.issn1002-1248.21-0264

参考相关研究文献，设计二级类目。

3.5 语义关系组织

面向典籍内容分析的分类体系主要包括概念和概念间语义关系两部分。在本文构建的分类体系中，可以通过词释义对概念范围进行规范，用于语义关系的构建和组织。

3.5.1 等同关系

在词典编纂时，编者需要对词做必要的解释以便使用者了解其含义，在此过程中，多利用已知的同义概念即同义词对新概念进行综合性描述。因此在古代社会画像标签体系的构建中，可以利用词典中词定义，获取同义词以完善词间等同关系。主要通过以下 3 种途径。

（1）如果存在两个词 *A* 词和 *B* 词，*A* 词的词头出现于 *B* 词的释义中，且 *B* 词的词头出现于 *A* 词的释义中，即这两个词可以形成词头–词释义的映射，那么认为 *A*、*B* 两词为同义词。

（2）通过观察语料，发现在《春秋左传词典》中，这种利用同义词作术语诠释时，通常会运用特定的指示词，如“同”“见”“即”“又称”“或称”“参”“亦作”“亦称”“犹言”“借为”等。利用模式匹配的方法，根据上述语言标志寻找词典中的同义词。如表 1 所示，“甸”——“甸服”“幣帛”——“幣鑞”分别为一组同义词。

表 1 同义词示例

Table 1 Examples of synonyms

词头	词释义
甸	①甸服，蓋封地在周天子幾內千里之地中者；②官名，蓋卽周禮之甸師
甸服	同「甸」
幣帛	泛指禮品、貢獻品
幣財	義同“幣帛”，泛指禮品、貢獻品

（3）对词释义分词后利用 TF-IDF 计算词向量间的相似度，若存在两个词，其词向量间相似度超过阈值，即认为这两个词是同义词。

3.5.2 相关关系

基于文本获得相关关系时，通常通过计算两个词向量在多维空间中的距离来进行分析。Word2Vec 作为计算词间距离的重要方法，也被称为“Word Embedding”，可以将字词转化为向量的形式并用词向量的方式表征词的语义信息。通过将单词从原先所属的空间嵌入到一个多维空间里，使得语义上相似的单词在该空间内呈现较近的距离，该过程实质上即是一个映射[26]。

在现代语言环境下，语言表述具有一定的结构性，Word2Vec 可以很好地处理结构化文本以发现文本内容中的同义词，但相对于古文这种一词多义现象严重、文本短、缺少结构化的文本，Word2Vec 可以更多的用于发现相关词，以补充词间相关关系。通过对相关典籍语料进行分词后，去除特殊字符及停用词，利用 Word2Vec 训练模型，计算词间相似度后，抽取词间相似度高于阈值的词，认为抽取出来的词组具有一定的相关度。

4 分类体系框架

本研究构建的面向典籍内容分析的分类体系如图 1 所示。分类体系共设置 5 个分面，分别表示典籍内容数据的 5 个维度：政治、经济、文化、社会和军事。研究者可根据分面和类目实现对典籍内容的快速检索。

4.1 政治分面

历史研究中，通常以史籍为重要研究依据，而史籍记叙以国家大事为主，如《春秋》《史记》等，因此政治分面极大程度上反映了典籍内容的社会背景信息。本文将典籍数据的政治分面归纳为国家外交、律法及政权更迭等类目。其中外交为国家或证权对外交流情况，具体包括朝见、盟会、盟约、聘问、议和、断交、贿赂、人质等方面；律法为国家或证权对内管理情况，具体包括基本法、刑法、法典以及诉讼等方面；朝代更迭则反映了国家或政权变迁情况，具体包括治国政务、新皇即位、政令发布、政变叛乱、逃亡和国家迁移等方面。

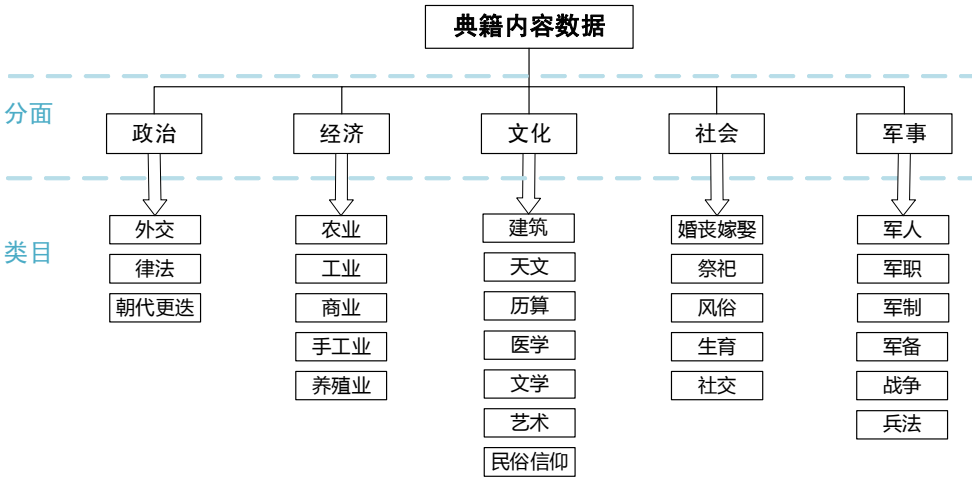

图 1　面向典籍内容分析的分类体系框架

Fig.1　Classification system framework for content analysis of ancient books

4.2　经济分面

经济分面揭示了典籍记叙内容的社会经济情况，本文将经济分面归纳为农业、工业、商业、手工业和养殖业等类目。其中农业数据包括农作物、非农作物和农用器具；工业包括工业过程使用的材料、工艺、器具以及生产的工业产品等；商业包括市场流通过程中使用的货币及对应产业数据；手工业则涉及具体的手工材料、工艺以及手工产品等；养殖业则为生产养殖涉及的牲畜、禽鸟、水产等信息；建筑业专指建筑材料及工艺。

4.3　文化分面

文化作为典籍研究的核心基础，对政治、经济有潜在的长期作用，本文将文化分面归纳为建筑、天文、历算、医学、文学、艺术及民俗信仰等类目。其中建筑专指古建筑类型，包括宗庙、宫殿、门、亭台等，具体建筑名称分别在对应的小类中进行描述；天文和历算多用作占卜、祭祀等，具体包括星象、天文现象、天像、节气和时间等方面；医学专指医学典籍数据，具体包括医药、病理等；文学以文学典籍、典籍载体和文学理论为主；艺术主要为舞蹈和乐曲，还包括棋、画、狩猎等休闲活动；民俗信仰则进一步分为信仰类、活动类和饮食习惯等，其中信仰包括宗教、禁忌和伦理道德等方面，活动类以婚、丧、祭祀、节日和宴席等活动形式为主。

4.4　社会分面

社会分面特指典籍中所记叙的社会习俗等信息，根据反映的社会活动形式对典籍数据进行划分。具体包括婚丧嫁娶、祭祀、风俗、生育以及社交等方面，利用典籍中这些社会活动中涉及的风俗习惯、礼仪、器具等数据进行表征。

4.5　军事分面

中华民族历经 23 个朝代、近千位君王，历史变迁中军事始终占据的重要地位，因此军事分面是历史研究中重要的组成部分。典籍数据中的军事分面具体包括军人、军职、军制、军备、战争、兵法等类目，其中军人、军职等类目从实体维度进行组织，而战争则是从时间维度进行组织，具体包括战前军事储备力量、战中使用兵法策略、战后交战双方胜负和领土归属等方面。

5　应用前景分析

本文构建的面向典籍内容分析的分类体系框架不仅涉及政治、经济等社会科学领域，还涵盖了文学、

DOI：10.13998/j.cnki.issn1002-1248.21-0264

宗教等人文学科知识。此分类体系可应用于典籍数字资源的深度开发利用，以文本内容为基础，从典籍的分类组织、知识导航和分析利用等方面为研究者提供便利。

5.1 基于分类体系的典籍资源分类组织

现有的典籍资源组织系统如古籍全文数据库、书目数据库和索引数据库，大多从从典籍的外部特征与主题角度对典籍数据进行组织和描述，缺少对于典籍内容特征及内在知识的组织。分类体系是学科知识组织与利用的框架，以分类体系为基础构建的分类表可系统地将知识资源加以分类组织，再通过浏览的方式逐层遍历，以选择需要的信息或资源。采用面向典籍内容分析的分类体系对典籍数字资源进行分类标引，将文本内容与其内在知识、语义相结合，对典籍资源进行知识层面的组织，可以帮助深度整合典籍资源，实现基于知识内容的典籍数据组织和基于语义的典籍信息检索。

5.2 基于分类体系的典籍资源知识导航

数字人文技术的发展，为典籍文本智能标注、语义分析、知识挖掘和数字化地图建设等智能导航提供了技术支持[27]。面向典籍内容分析的分类体系以规范数据为数据基础，从细粒度知识语义角度出发，对典籍内容进行重新组织，并利用规范数据对典籍中的实体信息提供参考，可以为读者提供典籍知识导航，降低阅读难度，帮助读者理解和利用典籍数字资源。

5.3 基于分类体系的典籍资源分析利用

典籍资源涉及时间跨度长、学科范围广，传统的文献细读方式效率较低，不适用于大规模典籍数据的开发利用。借助数字人文理论与技术，利用面向典籍内容分析的分类体系，可以有效挖掘典籍文本中潜藏的知识和规律，并进行清晰、直观的分析和展示。

笔者将分类体系应用于古籍文本内容分析，基于用户画像技术和数字远读技术，以本文构建的分类体系为基础，利用多种文本挖掘技术对典籍文本进行多维度特征抽取，通过构建和分析古代社会画像，全景化呈现社会发展状况，帮助研究者快速获得古代社会概貌[28]。

6 结　语

典籍数字化资源的出现，对于中华文化的传承与研究具有重要意义。随着典籍数字资源的不断深入开发和利用，传统的知识组织方式多以典籍外部载体形式特征为主，不能有效揭示典籍内在知识信息和语义关系，在一定程度上限制的学者对典籍资源的开发利用深度，同时在研究过程中浪费了大量的时间和精力。

本研究试图从细粒度知识单元语义信息的层面对典籍内容进行组织和揭示，提出基于典籍内容分析的分类体系，从政治、经济、文化、社会和军事 5 个维度将典籍内容进行重新组织与揭示，以期帮助研究者快速分析典籍内容，提高典籍数字资源的利用效率。但本文提出的分类体系框架具体分面与类目尚不完善，在分类实践中需考虑到具体分类目的、分面组配方式和分类深度等问题，需要更多的典籍数据及人文学者的意见进行细化和修订。

参考文献：

[1] 张三夕, 毛建军. 汉语古籍电子文献知见录[M]. 广州: 世界图书馆出版公司, 2015.
ZHANG S X, MAO J J. Knowledge of Chinese ancient electronic documents[M]. Guangzhou: World library publishing company, 2015.

[2] 唐宸. 奎章阁[EB/OL]. [2019-10-22]. https://www.kuizhangge.cn/.
Tang C. Kuizhangge[EB/OL]. [2019-10-22]. https://www.kuizhangge.cn/.

[3] 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.
OUYANG J. Visual analysis and exploration of ancient texts for digital humanities research[J]. Journal of library science in China, 2016, 42(2): 66-80.

[4] 何琳, 乔粤, 刘雪琪. 春秋时期社会发展的主题挖掘与演变分析——以《左传》为例[J]. 图书情报工作, 2020, 64(7): 30-38.

HE L, QIAO Y, LIU X Q. Topic mining and evolution analysis of social development in spring and autumn period – A case of studying Zuozhuan[J]. Library and information service, 2020, 64(7): 30–38.

[5] 魏晓萍. 数字人文背景下数字化古籍的深度开发利用[J]. 农业图书情报学刊, 2018, 30(9): 106–110.
WEI X P. Deep development and utilization of digital ancient books under the background of digital humanities[J]. Agricultural library and information, 2018, 30(9): 106–110.

[6] 张力元, 王军. 古籍数据库分面分类体系设计[J/OL]. 图书馆建设: 1–9[2021–04–10]. http://kns.cnki.net/kcms/detail/23.1331.G2.20200820.1048.002.html.
ZHANG L Y, WANG J. Design of faceted classification system of ancient book databases[J/OL]. Library development: 1–9[2021–04–10]. http://kns.cnki.net/kcms/detail /23.1331.G2.20200820.1048.002.html.

[7] 毛建军. 论古籍电子文献学研究范畴的确立[J]. 图书馆理论与实践, 2010(9): 46–48, 88.
MAO J J. On the establishment of the research category of electronic philology of ancient books[J]. Library theory and practice, 2010(9): 46–48, 88.

[8] 朱玲, 崔蒙, 杨峰. 中医古籍语言系统分类体系的构建[J]. 中华医学图书情报杂志, 2012, 21(6): 15–18, 28.
ZHU L, CUI M, YANG F. The construction of the language system classification system of ancient Chinese medicine books[J]. Chinese journal of medical library and information science, 2012, 21(6): 15–18, 28.

[9] 秦春秀, 刘杰, 马晓悦. 知识单元间的语义关系研究进展[J]. 情报理论与实践, 2017, 40(6): 128–133.
QIN C X, LIU J, MA X Y. Research progress of semantic relationship among knowledge units[J]. Information studies: Theory & application, 2017, 40(6): 128–133.

[10] 李章超, 何琳. 文化遗产语义组织研究进展[J]. 图书情报工作, 2020, 64(7): 4–12.
LI Z C, HE L. Case study on semantic organization of cultural heritage[J]. Library and information service, 2020, 64(7): 4–12.

[11] 龚延明. 重构宋代四万进士档案——浙大宋学中心龚延明、祖慧《宋代登科总录》(14 册)介绍[EB/OL]. [2020–01–30]. http://rwsk.zju.edu.cn/2015/1112/c2039a170378/page.htm.
GONG Y M. Reconstructing the archives of 40000 Jinshi in the Song dynasty – Introduction of Gong Yanming and Zu Hui's general records of Dengke in the Song dynasty (14 volumes) from the Song study center of Zhejiang university [EB/OL]. [2020–01–30]. http://rwsk.zju.edu.cn/2015/1112/c2039a170378/page.htm.

[12] 傅君勤, FULLERMA. 中国历代人物传记资料库用户指南 (中文版)[EB/OL]. [2019–10–25]. http://172.16.20.58/cache/2/03/projects.iq.harvard.edu/6ed82dd42b570535d90551ae3c305d66/cbdb_users_guide_ch_170126.pdf.
FU J M, FULLERMA. User's guide for biography database of Chinese historical figures (Chinese version)[EB/OL]. [2019–10–25]. http://172.16.20.58/cache/2/03/projects.iq.harvard.edu/6ed82dd42b570535d90551ae3c305d66/cbdb_users_guide_ch_170126.pdf.

[13] 丁君军, 郑彦宁, 化柏林. 基于规则的学术概念属性抽取[J]. 情报理论与实践, 2011, 34(12): 10–14, 33.
DING J J, ZHENG Y N, HUA B L. Extraction of academic concept attribute based on rules[J]. Information studies: Theory & application, 2011, 34(12): 10–14, 33.

[14] T2K(Text–To–Knowledge)[EB/OL]. [2019–10–25]. http://www.italianlp.it/demo/t2k–text–to–knowledge/.

[15] 刘全波, 何强林. 2014 年类书研究综述[J]. 古籍研究, 2017(2): 300–319.
LIU Q B, HE Q L. A review of the 2014 category book research[J]. Ancient books research, 2017(2): 300–319.

[16] 罗艳秋, 徐士奎, 郑进. 少数民族医药古籍文献分类体系构建研究(下)——民族医药古籍文献的分类体系研究[J]. 中医学报, 2014, 29(12): 1851–1854.
LUO Y Q, XU S K, ZHENG J. Classification system research of minority medical – Study on classification system of minority medicine ancient books (two)[J]. Acta Chinese medicine, 2014, 29(12): 1851–1854.

[17] 李娜, 包平. 基于《方志物产》的物产分类体系智能化研究——以《方志物产》山西分卷为例[J]. 中国农史, 2016, 35(4): 31–38, 143.
LI N, BAO P. Study on intellectualized processing of product classification system based on local chronicle: produce – Taking local chronicle: Produce of Shanxi for example[J]. Agricultural history of

DOI：10.13998/j.cnki.issn1002-1248.21-0264

China, 2016, 35(4): 31-38, 143.

[18] 车淑娅.《韩非子》词汇研究[D]. 杭州: 浙江大学, 2004.
CHE S Y. Research on the vocabulary of Han Feizi[D]. Hangzhou: Zhejiang university, 2004.

[19] 孙丽丽. 春秋时期词汇研究[D]. 济南: 山东大学, 2012.
SUN L L. Vocabulary research in the spring and autumn period[D]. Jinan: Shandong university, 2012.

[20] 胡明. 基于《汉语大词典》的战国—秦新词研究[D]. 济南: 山东大学, 2016.
HU M. Research on warring states - Qin new words based on Chinese dictionary[D]. Jinan: Shandong university, 2016.

[21] RYAN N, EDWARD S, KRISTOFFER N, et al. Modeling the contested relationship between analects, Mencius, and Xunzi: Preliminary evidence from a machine-learning approach[J]. The journal of Asian studies, 2018, 77(1): 19-57.

[22] 彭炜明, 宋继华.《资治通鉴》历史领域本体构建及其应用研究[J]. 中文信息学报, 2010, 24(2): 33-38.
PENG W M, SONG J H. Ontology construction and application research in the historical domain of Zi Zhi Tong Jian[J]. Journal of Chinese information processing, 2010, 24(2): 33-38.

[23] 曾熙, 谭旭, 王晓光. 文化遗产大数据二维分类框架研究[J]. 图书情报知识, 2020(1): 84-93.
ZENG X, TAN X, WANG X G. Two-dimensional classification framework for cultural heritage big data [J]. Documentation, information & knowledge, 2020(1): 84-93.

[24] ANIL K J. Data clustering:50 years beyond K-Means[J]. Pattern recognition letters, 2010, 31(8): 651-666.

[25] 王千, 王成, 冯振元, 等. K-means 聚类算法研究综述[J]. 电子设计工程, 2012, 20(7): 21-24.
WANG Q, WANG C, FENG Z Y, et al. Review of k-means clustering algorithm[J]. Electronic design engineering, 2012, 20(7): 21-24.

[26] 郭思成, 李纲, 周华阳. 基于 Word2Vec 的医学知识组织系统互操作研究——以词表间语义映射为例[J]. 情报理论与实践, 2019, 42(9): 160-165, 176.
GUO S C, LI G, ZHOU H Y. Interoperation for medical knowledge organization systems based on Word2Vec: Taking semantic mapping between thesaurus as an example[J]. Information studies: Theory & application, 2019, 42(9): 160-165, 176.

[27] 朱成林, 袁曦临. 中国古籍的数字化导读研究[J]. 图书馆建设, 2014(11): 50-55.
ZHU C L, YUAN X L. Study on the digital reading guidance of the Chinese ancient book[J]. Library development, 2014(11): 50-55.

[28] 艾毓茜. 基于文本分析方法的古代社会画像构建及其应用研究[D]. 南京: 南京农业大学, 2021.
AI Y X. Research on the construction and application of ancient social portraits based on text analysis methods[D]. Nanjing: Nanjing agricultural university, 2021.