古籍信息处理回顾与展望*

□刘忠宝　赵文娟

摘要　随着大数据时代的到来，古籍信息处理迎来前所未有的发展良机。此文从技术方法及其演进角度，对古籍数据挖掘、古籍信息系统构建、古籍信息服务等方面进行回顾和总结，并对未来的研究趋势进行展望。研究表明，越来越多的研究人员开始关注该领域并产生不少研究成果，但仍然面临一系列未解难题，如古籍数据挖掘能力不强、古籍信息系统构建效率低下、古籍信息服务水平尚有差距。未来研究应从古籍数字资源共享体系、高性能古籍信息处理模型、古籍信息服务评价等方面展开。此次梳理和展望有助于研究人员全面了解古籍信息处理现状，方便古籍信息资源的研究与利用；有助于推动古籍信息处理多学科融合与国际化合作。

关键词　古籍信息处理　古籍数据挖掘　古籍信息系统构建　古籍信息服务
分类号　G256
DOI　10.16603/j.issn1002－1027.2021.06.010

1　引言

古籍是中华优秀传统文化的重要载体，也是具有中华民族特色的文化宝藏。随着大数据时代的到来，数字化古籍数量持续增长，如何利用现代信息技术从古籍中挖掘出有价值的信息或知识，是中华文化得以延续与弘扬的重要保障。古籍信息处理正是在上述背景下提出的。古籍信息处理是利用现代信息技术对古籍文本的音、形、义进行加工和处理，并基于此实现古籍文本的数据挖掘和知识发现[1]。古籍信息处理涉及历史学、社会学、语言学、文献学、信息学等学科，研究内容涵盖古籍数字化、断句、自动标点、自动分词、词性标注、语义理解、知识组织、信息服务等方面。近年来，随着信息技术的发展，特别是自然语言处理技术的发展，古籍信息处理迎来了前所未有的发展良机，越来越多的研究人员开始关注该领域的研究。因此，有必要对古籍信息处理研究成果进行详细梳理，以发现现有研究存在的主要问题，明确下一步的研究方向，为研究人员从事古籍信息处理研究提供有益的参考。

笔者从技术方法及其演进角度，围绕古籍数据挖掘、古籍信息系统构建、古籍信息服务三方面展开研究。古籍数据挖掘主要包括古籍断句与标点、古籍分词与词性标注、古籍语义理解与知识发现、古籍知识组织与利用。古籍断句是对古籍中连在一起的句子进行切分，标点则是在断句的基础上为古籍中的句子加上标点符号；古籍分词是利用自动分词技术在词之间添加分隔符，词性标注则是为词语打上词性标签；古籍语义理解与知识发现主要围绕命名实体识别、语义消歧与对齐、知识挖掘等方面展开研究；古籍知识组织是对古籍知识进行整理、加工、控制等一系列组织化的方法。古籍信息系统构建主要包括古籍图像库、古籍版本库、古籍知识库、古籍元数据库。古籍信息系统构建方法主要有两类：一类是借助于人工构建，另一类是利用现代信息技术自动构建。由于第一类构建方法费时费力，效率低下，因此，目前研究人员更为关注第二类构建方法，这亦是文章探讨的重点。古籍信息服务是以现代信息技术和网络技术为基础，通过对古籍数字资源的搜索、整理、组织、分析，形成面向用户需求的个性化、专业化的古籍信息或解决方案。古籍信息服务主要包括古籍校勘、古籍索引、古籍翻译、古籍检索、古籍编撰。本研究除了对古籍信息处理具有重要的理论价值和现实意义，还有助于研究人员全面了解古籍信息处理现状，方便古籍信息资源的研究与利用；有助

* 国家社会科学基金一般项目“大数据环境下面向图书馆资源的跨媒体知识服务研究”（编号：19BTQ012）的研究成果之一。
通讯作者：刘忠宝，ORCID：0000-0002-0038-2462，邮箱：zbliu@blcu.edu.cn。

于推动古籍信息处理多学科融合与国际化合作。

2 基于规则的方法

2.1 古籍断句与标点

通过对已断句古籍的分析，得到断句特征模式，并基于此给出断句规则，以实现古籍断句。陈天莹等提出基于前后 N-gram 模型的古籍断句方法，该方法在较小规模语料集情况下，通过分析古籍文本的上下文信息，能够正确预测切分位置[2]。黄建年等以农业古籍为研究对象，利用模式识别中的正则表达式给出断句规则，在此基础上，构建古籍文本断句与标点的模式库，利用该模式对古籍文本进行断句和标点[3]。该方法断句速度快，准确率高，但存在两大不足：一是断句规则库构建费事费力；二是该方法只对某一特定类型古籍有效，对其他类型古籍适用性较差。

2.2 古籍分词与词性标注

在人工构建分词词典的基础上，利用匹配算法对古籍文本进行自动分词和词性标注。黄建年等综合采用切分标志、分词词典、N-gram 模型三种方法对农业古籍进行分词[4]。徐润华等将《左传》作为研究对象，引入《汉词语大典》作为知识来源，利用基于支持向量机的半监督方法进行语义标注[5]。黄水清等基于《汉学引得丛刊》中的《春秋经传注疏引书引得》制定词汇表，利用条件随机场(Conditional Random Field，CRF)模型，结合统计和人工方法确定特征模板[6]。王姗姗等基于《广韵》字表和《毛诗引得》领域词表，引入条件随机场模型，对多领域知识下的《诗经》自动分词问题进行研究[7]。该方法思路清晰，方便可行，但也存在一些不足：一是词典构建需要人工参与；二是受到词典限制，常常面临“未登录词”问题，这严重地影响了实际效果。

2.3 古籍语义理解与知识发现

在命名实体识别方面，利用中文命名实体的内部结构和上下文的边界特征等来手工建立命名实体识别的规则。潘正高重点关注确定实体边界的外部规则和描述实体内部结构的内部规则，将古籍文本的规则特征和统计特征结合起来，提出规则和统计相结合的中文命名实体识别方法[8]。这种方法使用的规则需要语言学和领域专家共同参与构建，时间代价较大，效率不高；这些规则依赖于设计者的主观直觉，适用性和可迁移性较差。

在语义消歧与对齐方面，肖晶等选取作者、机构、关键词等特征属性，结合模糊匹配和精确匹配，提出一种适用于国家科技图书文献中心馆藏资源的人名消歧规则库，并给出基于规则的人名消歧算法[9]。这种方法设计简单，操作方便，但规则库构建过于依赖人工，效率有限，且扩展性不高。

在知识挖掘方面，首先构建大量的知识挖掘规则，然后将规则与文本字符串进行匹配。李文林等在建立知识挖掘规则的基础上，对明清两代中医治疗疫病的药症关系进行关联分析，初步揭示了明清医家疫病诊疗的学术思想及治疫经验[10]。由于不同古籍的词语分布规则存在差异，因此将某一古籍的知识挖掘规则应用到另一古籍时通常会出现问题。

2.4 古籍知识组织与利用

古籍知识组织是对古籍知识进行整理、加工、控制等一系列组织化的方法。古籍知识组织能够降低存储知识的古籍文本过度分散的缺陷，通过对古籍知识的聚类、重组、关联分析，有助于得到古籍知识组织与利用的模式。以分类法、叙词表等为代表的传统知识组织手段，用等级结构或聚类揭示概念间的关系，人工组织更新缓慢，跟不上知识的增长速度。许雯等以“知识元”为核心、“知识分类”为基础，着重探讨中医古籍文献数字化中叙词表的构建问题[11]。鉴于《中国图书馆分类法》所设民族古籍类目过于分散和笼统的不足，李敏指出应从增设相关类目、类目注释、交替类目等方面进行改造[12]。基于规则的古籍知识组织在特定古籍情况下所得结果较为准确，针对性强，但对规则设定者的经验提出了较高要求，结果容易受个人主观性影响。

2.5 古籍图像库构建

史丽君在分析当前国内古籍图像库存在问题的基础上，提出古籍图像库建设应遵循成熟性、标准性、统筹性等基本原则，并结合首都图书馆工作经验，给出加强古籍图像库功能建设的对策和措施，包括加强图像色彩与内容管理、合理配置发布平台功能、完善古籍数字化工作机制等。该研究立足于首都图书馆实践，提出对古籍图像库构建的建议和思考，但缺乏理论支持和量化研究结果[13]。

2.6 古籍元数据库

夏翠娟等借鉴“循证实践”和“循证社会学”的思想，利用文献调研、数据建模、实证研究等方法，在分

析古今目录编排体例和古籍元数据标准规范的基础上,设计了一个可以融合不同来源、不同格式古籍目录、元数据、古籍全文和古籍知识的古籍数据模型[14]。近年来,古籍数字化不断发展,现有常见的古籍数据库达到数百种,这些资源缺乏有效的整合和组织,导致研究人员很难拓宽古籍资源获取途径,一些机构重复数字化相同资源,造成了财力、物力的浪费。为了解决上述问题,张力元等针对国内外常见的三百余种古籍数据库,借鉴都柏林元数据集,为每条古籍数据标注了元数据,并基于此从主题、类型、格式、地区、建制主体和权限六个维度给出古籍数据库分面分类体系,以此分类和描述古籍数据库[15]。

2.7 古籍校勘

古籍校勘是利用信息技术自动识别并标记出不同版本古籍文字之间的差异,并帮助专家利用各种校勘辅助工具进行勘误。常娥等依据校勘学有关理论,针对错文、脱文、衍文等自动校勘对象,设计了基于窗口匹配的自动校勘算法,探讨古代官名、人名和地名表的自动校勘工具的设计问题[16]。肖磊等对古籍版本异文自动发现方法进行研究,版本异文是同一古籍不同版本之间本应相同的字句出现的差异,该研究首先利用配对算法得到句珠之间的相似度;根据相似度发现最有可能的句珠配对,在异文句珠中去掉"同文"并输出"异文"[17]。上述研究成果均属于基于规则的方法,这些方法的性能依赖于规则库的构建,由于目前传世的古籍数量庞大,因此,如何高效地自动构建规则库有待于深入研究。

3 基于机器学习的方法

3.1 古籍断句与标点

引入机器学习模型,将古文断句和标点视为分类问题或序列标注问题。王川等针对古籍文本缺少标点符号的问题,引入层叠式条件随机场模型用于古籍文本断句和句读标记[18]。张开旭等在引入互信息和 t-测试差两类统计量作为输入特征的基础上,利用条件随机场模型对古籍文本断句标点方法进行研究[19]。古籍标点的准确率明显低于古籍断句的准确率。其主要原因是标点符号的多样性致使古籍标点难度较大,故准确率较低。这种方法自动化水平和断句标点效果明显优于基于规则的方法,但该方法依然受限于两大问题:一是针对特定类型古籍,需要事先给定特征模板;二是特征模板的适用性有限,无法适应于不同年代和不同体裁的古籍。

3.2 古籍分词与词性标注

通过对古籍文本中的字词信息进行统计分析,进而实现古籍的分词和词性标注。该方法不依赖于分词词典,克服了基于词典方法对词典过分依赖的不足。古籍词性标注和自动分词工作一般依次进行,这种做法便于理解但存在错误扩散问题。鉴于此,石民等以《左传》为研究对象,引入条件随机场模型,设计了自动分词和词性标注一体化模型[20]。为了解决古籍中术语、典故等造成的理解困难问题,姜欣等以《茶经》为研究对象,基于统计学习方法和树剪枝理论,提出快速树剪枝算法,进而完成古籍文本的快速切分;同时还针对《孟子》语料,先后采用基于条件随机场模型的自动分词方法和利用注疏文献的自动分词方法进行分词[21]。留金腾等在分析上古汉语词汇特点的基础上,将现代汉语作为基础语料,采用条件随机场模型和领域自适应方法进行自动分词和词性标注[22]。钱智勇等利用隐马尔科夫模型进行自动分词和词性标注,通过比较分词后的标注词性概率,取最大概率作为自动分词和词性标注的结果[23]。王国龙等在设计基于键值对模型的词性标记集基础上,利用基于词语联系的隐马尔可夫模型进行词性标注[24]。周澍绮基于机器学习模型和传统文献研究方法,引入文本工程通用框架平台 GATE 构建具有语义特征的《楚辞》词表和语义规则,结合语义抽取和数据库技术,较为高效地标注了《楚辞》语料[25]。王东波等利用条件随机场模型,结合字词结构、词长、拼音等特征信息构建组合特征模板,在对先秦古籍文本分词的基础上,设计了词性标注实验[26]。这种方法利用古籍文本中词语的概率分布实现古籍分词和词性标注,突破了词典的限制,但需要大规模的训练语料,实际效果与训练语料的质量紧密联系。

3.3 古籍语义理解与知识发现

在命名实体识别方面,汤亚芬结合古汉语人名的分布、长度、左右一元词汇与词性等内部和外部特征,利用条件随机场模型,训练先秦人名自动识别模型[27]。祁瑞华等从词汇、句子、语篇三个维度对典籍英译诗歌作品的特征进行分析,提出加权朴素信念不完整数据分类算法,并将该算法应用于作者身份识别[28]。黄水清等在分析古汉语地名统计特征

的基础上,构建了包括词汇、词性、词汇长度、虚词词性、左边界词、右边界词等特征的特征模板库[29]。王东波等在分析人名、地名、时间实体的内部数量和外部特征的基础上,利用条件随机场模型,构建了面向先秦古籍的历史事件基本实体构件自动识别的特征模板[30]。李娜将数字化的《方志物产》作为研究对象,在全文人工标注的基础上,构建基于条件随机场的别名抽取模型[31]。袁悦等基于已标注的《左传》《国语》语料,在南京师范大学先秦词性标注集的基础上,融合北京大学、中科院计算所词性标注集,引入条件随机场模型以及特征模板,比较上述三类词性标注集在实体抽取方面的差异性[32]。古籍命名实体识别的效果远低于现代书籍,其主要原因是古籍文本的语义理解能力有待于进一步提升,语义理解难点主要体现在三方面:首先,古籍中存在大量的通假字、生僻字、一词多义等情形;其次,古籍文本的句式复杂,多为复合长句,并采用大量的修辞手法;最后,部分实体表征缺乏统一标准。

在语义消歧与对齐方面,于丽丽等在分析现有词义消歧技术和方法的基础上,引入条件随机场模型,融合词的上下文语义和语言学特征,实现了《春秋左传》高频词词义消歧[33]。常娥等在古汉语义项词语知识库支持下,通过构建多义词向量空间模型,计算多义词待消歧上下文向量与该词各义项之间的关系,对农业古籍的消歧平均正确率达到79.5%[34]。丁长林等在分析中医古籍叙述性术语特点的基础上,将语义标注问题转化为有监督学习的短句序列标注或分类问题[35]。车超等为了解决古籍术语对齐问题,提出基于子词的最大熵模型[36]。上述研究成果均属于有监督的语义消歧与对齐方法,有效地解决了数据稀疏的问题,但需要事先建立大规模标注语料库,这在一定程度上限制了该类方法的推广应用。

在知识挖掘方面,马创新等针对古籍注疏存在的问题,通过引入知识表示方法组织古籍及其注疏的知识,探讨注疏知识网络的基础架构以及知识的组织方式和应用价值[37]。李娜等采用条件随机场模型对馆藏方志古籍的古代地名进行识别[38]。王东波等在 TF-IDF 提取类别特征词的基础上,先后利用支持向量机、条件随机场、深度学习模型对先秦古籍进行问句自动分类研究[39]。何琳等基于自然语言处理技术,探讨了先秦古籍本体构建方法[40]。潘俊基于中国历代人物资料库构建历史人物关系网络,并提出一种融合人物影响力的网络表示学习方法[41]。上述研究对古籍知识挖掘与利用进行了有效的探索和尝试,并取得了阶段性成果,然而,知识挖掘过程完全是数据驱动的,没有借助领域知识对知识挖掘过程进行指导,这在一定程度影响了古籍知识挖掘的效果。

3.4 古籍知识组织与利用

基于机器学习的知识组织可以实现知识的动态更新和管理,有助于人员从繁琐的工作中解脱,集中精力解决知识库设计和优化等更高层次的问题,降低成本,提高效率。柳长华基于信息论中的本体论与认识论,在对中医古文献的结构特征分析的基础上,先后提出基于知识元的知识表示方法和知识解析方法[42]。徐春波从知识组织与文献组织关系、中医古籍的知识组织、中医药古代文献知识库等方面,对基于知识单元的中医古籍知识组织方法进行了探讨[43]。李兵等在系统分析中医古籍中本草知识特点的基础上,以概念关系作为本草知识关联的依据,提出基于知识解析、概念类型分类和概念关系关联的本草古籍知识组织方法[44]。李娜等对适用于《方志物产》知识组织的技术和方法进行了系统探讨[45]。常颖聪等在总结古籍知识组织和关联数据的基础上,试图将关联数据应用于古籍知识组织,进而提出基于关联数据的古籍知识组织模式[46]。王国玺等探讨了医案古籍的知识组织方法[47]。基于机器学习的知识组织相当于对语料的归纳,实用性更强,结合机器学习可以快速学习古籍语料的规律,得出较为通用的结果,但由于实体之间具体关系识别欠缺、语料库往往不能系统涵盖各类语言现象等原因,知识组织结果的准确率有待提升。

3.5 古籍版本库构建

邓仲华等在分析当前古籍数据库优缺点的基础上,提出建设古籍版本主题库的必要性,利用本体库构建技术,通过设计古籍版本知识的类、属性和实例以及本体之间的关系,基于 Protégé 工具构建了古籍版本本体库,为后续古籍版本主题库建设奠定基础[48]。柳建钰等对计算机辅助古籍版本校勘库建设进行研究,该研究遵循校勘库建设规范,充分整合各类古籍资源,全面梳理相关资料,以期形成一个标准化、开放化、共享的计算机辅助古籍版本校勘库[49]。上述研究均是在中小规模语料基础上展开

的,而实际上古籍版本库构建涉及到的古籍数字资源规模庞大,现有的技术和方法能否适用需要进一步探讨。

3.6 古籍知识库构建

贾凤旭以《周易》及其注疏为研究对象,探讨利用计算语言学方法和计算机技术处理《周易》及其注疏,采用 XML 方式标注文献的知识结构,以知识结构化方法构建《周易》及其注疏知识库[50]。赵洪雅在整合古籍知识资源服务评价指标的基础上,利用质性量化研究方法,从用户视角构建古籍知识资源服务评价模型,引入探索性因子分析法对模型进行校验[51]。现有这些古籍知识服务大多属于理论探讨和实践总结,缺乏基于调研与实证的定量研究。

3.7 古籍索引

古籍索引是一种用于揭示古籍内容的特定形式,它将古籍中有关事物的名称、词语、字句、人名、地名、主题等分别摘录标引,通过注明出处、页码以及行数,以一定的排列方式编辑而成,供人们查询古籍内容。王雅戈等以两种版本的《道德经》索引编撰为研究对象,对索引之星、微软 Word 和自编索引软件等辅助工具进行比较研究[52]。韩琴在古籍索引数字化时代背景下,结合我国古籍索引工作的具体实践,对计算机技术与古籍索引编制相结合的古籍索引数字化实践进行了全面回顾,该研究有助于提升馆藏资源的知识服务质量[53]。黄建年利用 VFP 和 Word 等工具对多文本古籍索引编制问题进行研究[54]。目前,在古籍数字化领域,古籍索引数据的实践较为丰富,但相关研究相对滞后。肖禹在引入古籍索引概念的基础上,从资源揭示、文本碎片化、数据挖掘、创建新数据等方面对古籍索引数据应用实践进行研究[55]。上述研究成果利用现代信息技术自动构建古籍索引,一定程度上克服了人工构建索引效率低下的不足。

3.8 古籍翻译

郭锐等综合考虑句子长度、汉字字形、标点符号等因素,提出古籍汉语句子自动互译模型,引入遗传算法和动态规划模型,实现了古今汉语句子对齐功能[56]。王爽等基于语言学和机器翻译理论,在分析现有古文机器翻译研究成果的基础上,提出一种基于实例的古文机器翻译系统[57]。韩芳等在构建古汉语词典的基础上,依据黎锦熙提出的句本位句法规则构建知识库,引入词义消歧算法对古汉语进行机器翻译研究[58]。目前,直接可用的句对齐古今平行语料规模较小,直接影响到统计学习模型的训练水平,进而影响到古籍机器翻译效果。

3.9 古籍检索

何琳等以农业古籍为研究对象,在构建领域本体的基础上,探讨基于领域本体的语义检索机制,利用领域本体对概念的形式化描述,将目前面向字面匹配的检索提升至面向语义的检索层面[59]。如何从大量古籍语料中快速准确地找到与输入最相似的语句,是目前古今汉语机器翻译以及古今汉语平行语料库高级应用的基础性问题。传统逐句匹配的做法效率不高。鉴于此,杨志芹在分析现有古籍检索系统不足的基础上,基于信息抽取技术,构建语义智能检索系统[60]。郭伟玲等对古籍数字化的检索问题进行了深入研究,提出解决古籍数字化检索问题的基本路径[61]。白淑霞等以蒙古文古籍《甘珠尔经》图像为研究对象,提出一种与视觉语言模型相结合的隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)主题模型,通过引入查询似然模型实现蒙古文古籍关键词检索[62]。总的来说,古籍信息检索取得了一定进展,并涌现出一系列研究成果,这对于古籍研究与利用具有重要价值。然而,上述成果面临两方面的挑战:一是搜索结果并未考虑用户的个性化需求;二是所提技术和方法无法适用于海量古籍检索。

3.10 自动编撰

古籍自动编撰是利用计算机信息技术对古籍进行数据挖掘,从中摘录与某一主题密切相关的信息资源,并编撰成册。常娥等将农业古籍作为研究对象,探讨了农业古籍自动编撰算法,着重解决了分割农业古籍章节、提取字句关键词、计算机紧凑度和深度以及确定分割点等关键技术问题[63]。常娥面向农业古籍,设计并实现了农史专题资料自动编纂系统,并对其性能进行了测试和分析[64]。上述研究仅是对计算机辅助自动编撰的探索和尝试,其中不少环节仍需人工干预,远未达到智能化的程度,而且自动编撰效果有限,能否引入领域知识并结合最新的深度学习模型来进行古籍编撰值得深入思考。

4 基于深度学习的方法

4.1 古籍断句与标点

深度学习模型具有强大的特征学习能力,引入该模型能够提取更为丰富的古籍特征,进一步提高

古籍的断句和标点效率。王(Wang)等引入循环神经网络模型用于古籍文本断句研究，在《阅微草堂笔记》《宾退录》《朝野佥载》《南部新书》等古籍语料上具有优良的断句效果[65]。王博立等将古籍文本断句问题看作是基于字的序列标注问题，并据此提出基于门控制单元(Gated Recurrent Unit，GRU)的断句方法。该方法将句号、问号、感叹号、逗号、分号、冒号等六种标点作为断句标记，通过GRU神经网络能够自动地学习古籍文本上下文的语义特征，避免了人工给定特征的随机性和盲目性[66]。韩(Han)等将字根特征引入到BiLSTM(Bi-directional Long-Short Term Memory)和CRF混合模型用以完成古籍文本断句任务[67]。传统的古籍信息处理将古籍断句与标点以及分词与词性标注分开处理，容易造成错误的多层扩散，鉴于此，程宁等将古籍的断句、分词、词性标注等进行融合研究，并引入BiLSTM-CRF混合模型形成一体化的标签体系[68]。

4.2 古籍分词与词性标注

深度学习模型强大的特征提取能力在古籍分词和词性标注方面亦有较好的效果。郭利敏等引入卷积神经网络模型，将古籍汉字识别问题转化为分类问题[69]。针对已有自动分词方法依赖于大量人工标注的问题，俞敬松等融合非参数贝叶斯模型和BERT模型建立《左传》自动分词模型[70]。古籍数字化过程中常常对间隔、交错、粘连的汉字无法有效切分，鉴于此，倪劼受流水模式启发，基于古籍汉字特点，利用投影法与图像形态学对汉字进行列切分[71]。张琪等采用涵盖“经史子集”的25部先秦典籍作为训练语料，在未加入人工特征的前提下，基于BERT模型构建了先秦典籍分词词性一体化标注模型[72]。王莉军等在缺少中医领域开放数据集和大量中医词表的情况下，提出了使用BiLSTM-CRF模型完成中医领域文言文文献分词的任务[73]。魏一首次尝试使用无监督方法，将非参数贝叶斯模型与预训练模型结合，获得了较好的古文分词效果[74]。深度学习模型需要大规模的训练语料，而现阶段可以直接利用的古籍语料不多，上述矛盾限制了基于深度学习方法的推广应用。

4.3 古籍语义理解与知识发现

在命名实体识别方面，主要是将深度学习和统计模型结合使用，通过深度学习得到每个词的新向量表示，然后使用CRF模型输出对每个词的标注结果。高甦等利用BiLSTM-CRF混合模型对中医古籍的实体进行识别[75]。王菁薇尝试构建ALBERT-BiLSTM-CRF模型，提取《伤寒论》中疾病、证候、症状、处方、药物等实体[76]。刘江峰等尝试利用一系列BERT预训练模型，以《左传》《史记》《汉书》《后汉书》《三国志》等为实验语料，对人名、地名、时间词等三种历史事件的主要构成实体进行识别[77]。杜悦等分别利用Bi-LSTM、Bi-LSTM-Attention、Bi-LSTM-CRF、Bi-LSTM-CRF-Attention、Bi-RNN和Bi-RNN-CRF、BERT等深度学习模型，从中抽取构成历史事件的相应实体并进行效果对比[78]。上述研究表明，在英文和现代汉语语料集上表现优良的某些深度学习模型，不一定适用于古籍实体识别，其主要原因是古籍文本的表达习惯与现代汉语有较大差别，深度学习模型在古籍命名实体识别中的迁移学习能力有待进一步加强。

在语义消歧与对齐方面，梁继文等将典籍汉英句子自动对齐问题视为候选句对分类问题，利用LSTM-CRF模型对先秦典籍汉英句子进行对齐[79]。张春祥等以歧义词为中心，选取其中的词形、词性和语义类作为消歧特征，使用卷积神经网络来确定歧义词的语义类别[80]。虽然基于深度学习的方法消歧效率较高，但由于古籍规模较大，模型训练的工程很大，无法在短时间内得到语义消歧与对齐结果。能否将深度学习的方法与基于图的方法结合起来进行集体消歧，值得进一步研究。

在古籍知识挖掘方面，欧阳剑以大规模古籍文本为研究对象，采用大数据视域下数字人文的研究方法，在对古籍进行整理、标注、分词等处理后，引入可视化技术对古籍文本进行挖掘，构建了一个面向语言学、历史文献学、历史地理学等学科的古籍实时统计分析平台[81]。周莉娜等面向唐诗构建了领域知识服务驱动的唐诗本体模型，利用知识抽取、知识融合、知识推理等技术和方法自动构建唐诗知识图谱，以实现大规模唐诗数据的语义化表征[82]。笔者针对《史记》语料集，在BERT模型和LSTM-CRF模型的基础上，提出事件及其组成元素抽取方法，并基于此构建《史记》事理图谱，以揭示历史事件的演化规律，全面刻画历史人物的行为活动[83]。上述成果普遍存在可用的标注语料规模较小、训练得到的模型性能有限、获得的部分知识与实际情形存在偏差等问题。

4.4 古籍知识组织与利用

深度学习背景下的古籍知识组织与利用主要包括术语自动抽取、词表自动构建与丰富、自动标引等研究内容。术语是知识组织系统中的核心元素,术语自动抽取是指从特定领域文本抽取核心概念词语。张卫等针对古诗及鉴赏文本缺少学习语料的现状,提出一种基于"冷启动"的字序列自动标注方法,将汉字语言知识引入 BERT 模型,实现了大规模情感术语的自动抽取[84]。词表自动构建与丰富,通过挖掘词与词之间的关联,实现词语聚类。王晓光等围绕敦煌壁画叙词表建设,根据自顶向下与自底向上相结合的研究思路,深入探讨敦煌壁画叙词表的构建方法[85]。自动标引可将作者语言自动转换为标引语言,方便检索系统语言匹配。陈博等基于文本挖掘技术深入《英雄格萨尔》文献内容层挖掘主题词,并利用可视化工具直观呈现所获信息,在此基础上尝试构建可视化主题自动标引系统[86]。深度学习模型用于古籍标引不仅需要提前训练,而且需要一定规模的训练语料,如何获得大规模的标注语料是一个值得关注的问题。

5 研究趋势与展望

随着理论的完善和技术的成熟以及国家对传统文化的重视,古籍信息处理迎来了前所未有的发展良机。因此,笔者围绕古籍数据挖掘、古籍信息系统构建、古籍信息服务等问题,对未来研究趋势进行预测和展望。

5.1 古籍数据挖掘

随着大数据时代的到来,古籍数据挖掘面临两大挑战:一是古籍数字资源规模日益庞大,内涵知识呈现结构复杂、关联多样、价值密度低等特点;二是大多数据挖掘技术与方法是针对现代汉语或英语提出的,对于古籍文本的处理效果并不理想。因此,有必要利用历代注疏文献,借助最新的信息技术手段,创新古籍数字资源的标注方式,为利用大数据模型进行古籍数据挖掘提供必要的训练语料。此外,有必要整合现有研发力量,针对古籍数字资源的数据挖掘方法展开重点攻关,提出一系列行之有效的古籍数据挖掘方法,尽早建立古籍数据挖掘的方法体系。古籍断句与标点研究突出大规模语料环境下深度学习模型的设计,着重解决古籍文本可用特征较少的问题,引入迁移学习方法和注意力机制,探索深度学习模型针对不同古籍文本的适用性,建立一体化断句与标点模型。古籍分词的一些理论问题有待于深入研究,如什么是词、词和词素如何区分、词和词组有何差异、如何利用计算机辅助识别词等。解决上述问题的关键是探究词的定义,参照现代汉语分词方法,反向推演古籍文本分词的差异性。在词性标注方面,引入语义角色、句法结构、依存分析等多语言信息,借鉴集成学习思想,融合古籍分词、聚类、分类及模式匹配等多种技术手段来改善词性标注效果。在古籍语义理解和知识组织方面,进一步扩大古籍语料规模,改善模型的学习能力,通过构建面向古籍文本的顶层语义描述框架;推动历史学、社会学、信息学等领域的学者开展协同研究,着力解决古籍知识的可解释性问题。古籍知识组织与利用重点研究知识组织的扩展和共享、多源数据的歧义和噪声、用户参与知识组织的规范控制等问题。

5.2 古籍信息系统构建

古籍信息系统构建应遵循成熟性、标准化、统筹性等原则。成熟性原则要求古籍信息系统的立项和建设应建立在稳定的资金支持和切实可行的工作方案基础上;标准化原则要求古籍信息系统构建应遵循国家标准和行业标准,以便各类古籍信息系统互联互通;统筹性原则要求理顺古籍信息系统构建流程,及时解决建设过程中的常见问题并完善工作机制。确保古籍信息系统运行稳定、响应迅捷、使用方便,为用户使用古籍信息资源提供基础保障。特别是针对当前古籍信息资源呈现出数据量大、数据类型复杂、存储平台类型多样等特点,引入最新的信息技术与方法,破解影响古籍信息系统构建效率的关键问题,创新古籍信息系统构建的机制、体制,从系统性、全局性、完整性角度研究面向海量、异构、动态数据的古籍信息系统构建方法。加强古籍信息资源的组织、分类和整合研究,深入挖掘与揭示古籍信息资源的潜在价值,重视检索与交互服务平台建设,突出检索平台整合古籍资源,提供多层次、多角度的检索服务,交互服务平台以用户为中心,提供个性化咨询与交互服务。加强对全网古籍信息资源的整合,尽快建立统一的古籍信息系统平台,彻底解决现有系统存在的"信息孤岛"问题。进一步降低古籍信息系统的使用门槛,提高古籍资源的利用率,充分发挥古籍资源的学术价值和应用价值。

5.3 古籍信息服务

目前,古籍信息服务呈现出服务载体数字化、服务对象扩大化、服务形式个性化、服务内容多样化等特点。在今后的研究中,古籍信息服务尚需在理论、资源、数据、技术等方面持续发力,催生古籍信息服务新业态。研究的重点包括:完善古籍信息服务的理论体系,夯实古籍信息服务的资源基础,利用数据指导古籍信息服务,借助技术构建古籍信息服务生态。在古籍索引方面,针对索引数据的整合、标准化以及应用等问题展开研究,重点探讨基于索引学的古籍数字化方法以及基于古籍索引数据的索引学方法,这些研究都将产生一系列创新性成果。古籍翻译借鉴语义分析和知识图谱最新成果,统计古今汉语的词对齐结果,根据词对信息,修正词对翻译结果,或引入同义词林,细化语法规则库,以提高古籍翻译质量。在整合书目库、版本库、全文库和知识库的基础上,引入本体论思想,探索面向古籍数字资源的语义检索机制,着重解决语义检索、可视化检索、语义网发布等问题,实现由单一检索变为多元检索、由静态检索变为动态检索、由定向检索变为关联检索,为最终实现真正的智能检索奠定理论和技术基础。

6 总结

现有研究往往是对古籍信息处理进行阶段性的总结,关注的是古籍数字化过程中的主要任务,大多数成果成文较早,缺乏最新研究成果的支持。鉴于此,本研究根据古籍信息处理的生命周期,针对古籍信息处理过程中涉及的主要研究领域与方向,对古籍信息处理当前的研究状况进行总结,并分析未来的发展趋势。研究表明:随着信息技术的发展,特别是自然语言处理技术的发展,越来越多的研究人员开始关注该领域并产生了不少研究成果。然而,在现代汉语中表现优良的信息处理技术在古籍信息处理中效果欠佳,主要原因是古籍的表达习惯与现代汉语有较大差别,可利用的标注语料规模较小,现有模型对古籍文本的特征提取能力不足,知识组织面临一系列未解难题,古籍信息系统构建效率低下,古籍信息服务的个性化、专业化尚有差距。因此,笔者认为未来研究应从整合古籍信息资源、建立古籍数字资源共享体系、进一步扩大古籍语料规模以及寻找适用于古籍信息处理的高性能模型、古籍信息服务系统性能评价的主客体评价方法与工具、评价指标体系等方面展开研究。期望本研究有助于研究人员全面了解古籍信息处理现状,方便古籍信息资源的研究与利用;有助于推动古籍信息处理多学科融合和国际化合作。

参考文献

1 陈小荷.先秦文献信息处理[M].北京:世界图书出版公司,2013.
2 陈天莹,陈蓉,潘璐璐,等.基于前后文 n-gram 模型的古汉语句子切分[J].计算机工程,2007,33(3):192－196.
3 黄建年,侯汉清.农业古籍断句标点模式研究[J].中文信息学报,2008,22(4):31－38.
4 黄建年,侯汉清.中国古籍文本分词的一次试验[J].情报学报,2011,30(6):618－625.
5 徐润华,陈小荷.一种利用注疏的《左传》分词新方法[J].中文信息学报,2012,26(2):13－17,45.
6 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.
7 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.
8 潘正高.基于规则和统计相结合的中文命名实体识别研究[J].情报科学,2012,35(5):708－712.
9 肖晶,梁冰,张晓丹,等.一种面向篇级数据的作者名消歧规则和算法[J].现代图书情报技术,2012(5):55－59.
10 李文林,屠强,彭丽坤,等.基于关联规则分析明清古籍中疫病文献的药－症关系[J].时珍国医国药,2010,21(4):957－959.
11 许雯,柳长华,顾漫.试述中医古籍文献数字化中叙词表的构建[J].国际中医中药杂志,2015,37(1):1－3.
12 李敏.《中国图书馆分类法》组织民族古籍的可行性、局限及其改造[J].图书馆建设,2009(7):16－18.
13 史丽君.古籍图像数据库建设常见问题及对策研究——以首都图书馆馆藏古籍珍善本图像数据库建设为例[J].图书馆工作与研究,2016(9):62－66.
14 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.
15 张力元,王军.古籍数据库分面分类体系设计研究[J].图书馆建设,2021(3): 56－61.
16 常娥,侯汉清,曹玲.古籍自动校勘的研究和实现[J].中文信息学报,2007,21(2):83－88.
17 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.
18 王川,张小红,韩采华.古汉语句子切分与句读标记方法研究[J].河南大学学报(自然科学版),2009,39(5):525－529.
19 张开旭,夏云庆,宇航.基于条件随机场的古汉语自动断句与标点方法[J].清华大学学报(自然科学版),2009,49(10):1733－1736.
20 石民,李斌,陈小荷.基于 CRF 的先秦汉语分词标注一体化研究[J].中文信息学报,2010,24(2):39－45.
21 姜欣,姜怡,方淼,等.基于树剪枝的典籍文本快速切分方法研究——以《茶经》的翻译为例[J].中文信息学报,2010,24(6):10－13,42.

22 留金腾，宋彦，夏飞.上古汉语分词及词性标注语料库的构建——以《淮南子》为范例[J].中文信息学报，2013，27(6)：6－15.
23 钱智勇，周建忠，童国平，等.基于HMM的楚辞自动分词标注研究[J].图书情报工作，2014，58(4)：105－110.
24 王国龙，杜建强，郝竹林，等.中医诊断古文的词性标注与特征重组[J].计算机工程与设计，2015，36(3)：835－840.
25 周澍绮.基于GATE的楚辞语义标注研究[J].图书馆理论与实践，2015(11)：58－62，101.
26 王东波，黄水清，何琳.基于多特征知识的先秦典籍词性自动标注研究[J].图书情报工作，2017，61(12)：64－70.
27 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(7)：63－68.
28 祁瑞华，霍跃红，郭旭，等.典籍英译作者身份识别研究[J].现代图书情报技术，2015(1)：31－37.
29 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.
30 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(1)：65－77.
31 李娜.基于条件随机场的方志古籍别名自动抽取模型构建[J].中文信息学报，2018，32(11)：41－48，61.
32 袁悦，王东波，黄水清，等.不同词性标记集在典籍实体抽取上的差异性探究[J].数据分析与知识发现，2019，3(3)：57－65.
33 于丽丽，丁德鑫，曲维光，等.基于条件随机场的古汉语词义消歧研究[J].微电子学与计算机，2009，26(10)：45－48.
34 常娥，张长秀，侯汉清，等.基于向量空间模型的古汉语词义自动消歧研究[J].图书情报工作，2013，57(2)：114－118.
35 丁长林，白宇，蔡东风.基于有监督学习的医古文叙述性术语语义标注[J].中文信息学报，2015，29(2)：49－57.
36 车超，郑晓军.基于子词的历史典籍术语对齐方法[J].中文信息学报，2016，30(3)：46－51.
37 马创新，陈小荷，曲维光.经典古籍注疏文献的知识网络研究与设计[J].图书情报工作，2013，57(9)：124－128.
38 李娜，包平.面向数字人文的馆藏方志古籍地名自动识别模型构建[J].图书馆，2018(5)：67－73.
39 王东波，高瑞卿，沈思，等.基于深度学习的先秦典籍问句自动分类研究[J].情报学报，2018，37(11)：1114－1122.
40 何琳，陈雅玲，孙珂迪.面向先秦典籍的知识本体构建技术研究[J].图书情报工作，2020，64(7)：13－19.
41 潘俊.面向数字人文的人物分布式语义表示研究——基于CBDB数据库和古籍文献[J].图书馆杂志，2020(8)：94－102.
42 柳长华.基于知识元的中医古籍计算机知识表示方法[G]//《传统医药与人类健康》编委会.第三届国际传统医药大会论文集.北京：中医古籍出版社，2004：313－314.
43 徐春波.基于知识单元的中医古籍文献知识组织[G]//《中医药发展与人类健康》编委会.中医药发展与人类健康——庆祝中国中医研究院成立50周年论文集.北京：中医古籍出版社，2005：451－453.
44 李兵，张华敏，符永驰，等.面向知识服务的本草古籍知识组织方法研究[G]//世界中联信息专业委员会.世界中联信息专业委员会第四届学术年会暨大数据与中医药国际信息化发展论坛论文集.兰州：世界中医药杂志社，2014：8－11.
45 李娜，白振田，包平.基于《方志物产》的古籍知识组织路径探析[J].古今农业，2016(1)：105－113.
46 常颖聪，路程，翟军平.基于关联数据的古文知识组织应用研究[J].图书馆理论与实践，2019(2)：55－59.
47 王国玺，李兵，张华敏，等.基于知识组织的医案古籍知识库的构建与思考[J].西部中医药，2019，32(9)：49－52.
48 邓仲华，黄鑫，陆颖隽，等.论中文古籍版本本体库的构建[J].图书情报知识，2014(4)：80－87，93.
49 柳建钰，周晓文.计算机辅助古籍版本校勘资源库建设浅议[J].图书馆理论与实践，2017(3)：54－58.
50 贾凤旭.基于知识类聚的古籍知识库构建方法[J].图书馆学刊，2015(5)：45－48.
51 赵洪雅.古籍数字资源服务效能评价指标体系[J].图书馆论坛，2020，40(7)：150－160.
52 王雅戈，杜慧平.机编古籍索引探讨——以《道德经》语词索引自动编纂为例[J].图书馆论坛，2008，28(5)：34－37.
53 韩琴.试论古籍索引与古籍索引电子化[J].情报科学，2010，28(7)：1113－1116.
54 黄建年.基于VFP＋Word的多文本古籍索引编制实验[J].现代图书情报技术，2011(10)：85－89.
55 肖禹.古籍索引数据应用研究[J].新世纪图书馆，2017(5)：45－48.
56 郭锐，宋继华，廖敏.基于自动句对齐的相似古文句子检索[J].中文信息学报，2008，22(2)：87－91，105.
57 王爽，熊德兰，王晓霞基于实例的古文机器翻译设计与实现[J].许昌学院学报，2009，28(5)：88－91.
58 韩芳，杨天心，宋继华.基于句本位句法体系的古汉语机器翻译研究[J].中文信息学报，2015，29(2)：103－110，117.
59 何琳，曹玲.农业古籍本体的构建及其检索机制研究[J].现代图书情报技术，2006(12)：37－39，53.
60 杨志芹.基于信息抽取的古籍知识检索系统研究[J].情报科学，2009，27(8)：1219－1221，1226.
61 郭伟玲，戴艳清.论古籍数字化的检索问题[J].图书馆理论与实践，2011(10)：13－16.
62 白淑霞，鲍玉来.LDA单词图像表示的蒙古文古籍图像关键词检索方法[J].现代情报，2017，37(7)：51－54，88.
63 常娥，侯汉清.农业古籍自动编纂的设计和研究[J].南京农业大学学报(社会科学版)，2007，7(1)：99－104
64 常娥.农史专题资料自动编纂系统的构建与测试[J].图书馆学研究，2009(6)：10－14.
65 Wang B，Shi X，Tan Z，et al. A sentence segmentation method for ancient Chinese tests based on NNLM [G]// Dong M H，Lin J X，Tang X R. Proceedings of the 17th Chinese Lexical Semantics Workshop. Singapore：Springer，2016：387－396.
66 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.
67 Han X，Wang H，Zhang S，et al. Sentence segmentation for classical Chinese based on LSTM with radical embedding [J]. The Journal of China Universities of Posts and Telecommunications，2019，26(2)：1－8.
68 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.
69 郭利敏，葛亮，刘悦如.卷积神经网络在古籍汉字识别中的应用实践[J].图书馆论坛，2019，39(10)：142－148.

70 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.
71 倪劼.基于流水模式的古籍文献汉字切分算法[J].图书馆论坛，2021，41(9)：141－149.
72 张琪，江川，纪有书，等.面向多领域先秦典籍的分词词性一体化自动标注模型构建[J].数据分析与知识发现，2021，5(3)：2－11.
73 王莉军，周越，桂婕，等.基于 BiLSTM－CRF 的中医文言文文献分词模型研究[J].计算机应用研究，2020，37(11)：3359－3367.
74 魏一.古汉语自动句读与分词研究[D].北京：北京大学，2020.
75 高甦，金佩，张德政.基于深度学习的中医典籍命名实体识别研究[J].情报工程，2019，5(1)：113－123.
76 王菁薇，肖莉，骆嘉伟，等.基于《伤寒论》的命名实体识别研究[J].计算机与数字工程，2021，49(8)：1584－1587.
77 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[J/OL].图书馆论坛，2021，9(4)：1－14[2021－11－23].http://kns.cnki.net/kcms/detail/44.1306.G2.20210817.0904.002.html.
78 杜悦，王东波，江川，等.数字人文下的典籍深度学习实体自动识别模型构建及应用研究[J].图书情报工作，2021，65(3)：100－108.
79 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.
80 张春祥，赵凌云，高雪瑶.基于卷积神经网络的词义消歧[J].北京邮电大学学报，2019，42(3)：114－119.
81 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.
82 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.
83 刘忠宝，党建飞，张志剑.《史记》历史事件自动抽取与事理图谱构建研究[J].图书情报工作，2020，64(11)：116－124.
84 张卫，王昊，邓三鸿，等.面向数字人文的古诗文本情感术语抽取与应用研究[J].中国图书馆学报，2021，47(4)：113－131.
85 王晓光，侯西龙，程航航，等.敦煌壁画叙词表构建与关联数据发布[J].中国图书馆学报，2020，46(4)：69－84.
86 陈博，陈建龙.基于文本挖掘和可视化技术的主题自动标引方法——以《英雄格萨尔》为例[J].现代情报，2019，39(8)：45－51.

作者单位：刘忠宝，北京语言大学语言智能研究院，北京，100083
赵文娟，北京语言大学图书馆，北京，100083

收稿日期：2021 年 7 月 24 日
修回日期：2021 年 9 月 29 日
（责任编辑：支娟）

Review and Prospect of Research on Ancient Chinese Text Information Processing

Liu Zhongbao　Zhao Wenjuan

Abstract：With the advent of the era of big data，it appears an unprecedented opportunity to the development of ancient book information processing. This paper aims to review and prospect the research on ancient book information processing in the aspects of data mining，information system construction and information service of ancient book. It found that the ancient book information processing has attracted more and more researchers' interest and made great achievements. However，there still exists some unsolved problems，such as the poor performance of data mining of ancient book，the low efficiency of information system construction of ancient book and the unsatisfactory service level of ancient book. This paper points out the future research direction，including constructing the sharing system of ancient book，proposing a series of high-performance models，and conducting the research on the evaluation method. The paper is beneficial for researchers to comprehensively understand the advance of ancient book information processing in China，and it is also convenient for them to research and use the ancient book information resources. Meanwhile，it contributes to promote interdisciplinary research and international cooperation in ancient book information processing.

Keywords：Ancient Book Information Processing；Data Mining of Ancient Book；Information System Construction of Ancient Book；Information Service of Ancient Book