古文信息处理研究的现状及趋势*

黄水清[1,2]　王东波[1,2]

[1] 南京农业大学信息科学技术学院　南京 210095　[2] 南京农业大学领域知识关联研究中心　南京 210095

摘要：[目的/意义] 随着古文数字化、智能处理和相关人文计算研究的迅速发展，对这一领域的整体研究状况进行梳理，不仅有助于从以往的研究当中总结相应的规律，而且在一定程度上有益于后续探究的展开。[方法/过程] 厘定古文信息处理的概念，分析古文信息处理的研究现状，给出古文信息处理研究的整体概貌。同时，在统计分析的基础上，对古文数字化、智能处理和人文计算这 3 个方面的研究内容进行总结、回顾和研究趋势的展望。[结果/结论] 在古文信息处理研究中，古文数字化所取得的成就最大，古文智能处理在词汇级的探究上取得了一定的成效，而对于人文计算来说，与古文相关的研究则才刚刚起步。

关键词： 古文数字化　数字人文　信息智能处理　人文计算　古文信息处理

分类号： G255.1

DOI：10.13266/j.issn.0252-3116.2017.12.005

1　引言

古文信息处理是借助信息技术手段对古代汉语文本的音、形、义进行处理和加工[1]，并可在此基础上实现对古代汉语文本的深度挖掘与知识发现。古文信息处理涵盖了数字化、语音的自动处理、自动标点、词语切分、词性标注、专名标注、语义标注、事件抽取、句法分析、古今汉语的机器翻译、智能信息检索、文本挖掘、知识发现等多方面的研究内容。随着数字化古文献数量的快速增长和人文计算的迅速发展，古文信息处理的研究也日益朝着智能化与深语义化的方向发展。同时，如何更好地从古文献中挖掘出覆盖面全、层次多样的知识并进行有效的呈现是时代和社会的呼声，也是构建中国特色哲学社会科学的需要。在上述这一背景下，对古文信息处理的研究整体状况进行系统的梳理和分析，一方面有助于研究者了解以往研究的整体状况，另一方面也有益于研究者把握未来研究的趋势。

关于古文信息处理的研究现状，早期的研究者曾在不同的时期从不同的侧面进行过总结。1995 年，姚松[2]从典籍数字化的支撑环境与典籍数字化结果的应用两个方面对典籍整理的情况进行了较早的总结和分析。2000 年，王桂平[3]在总结书目数字化和善本数字化所取得成就的基础上，对古文数字化的发展方向进行了预测。李国新[4]则于 2002 年在总结古文数字化 4 个特征的基础上，从汉字字符集、古文输入、数字化的原则和智能化探究 4 个方面对古文数字化的研究内容进行了归纳。同年，潘德利[5]从典籍的规范化、标准化、网络化和系统化方面对古文数字化的研究方向进行了展望，李弘毅、厉莉[6-7]对古文数字化的研究状况进行了总结。2003 年，段泽勇和李弘毅[8]总结了古文数字化的 3 个阶段，并结合已有的数字化成果给出了古文数字化的主要流程。王冠中[9]于 2005 年对历年的古文数字化成果进行了总结和特征分析。最新的成果是 2014 年的，林竹鸣和朱翠萍[10]从统筹规划、版本一致性和字符集扩大等多个方面总结了古文数字化的发展现状，并结合新技术提出了未来的发展方向。

从上文可以看出，之前的研究者对古文信息处理的阶段性总结主要是针对古文的数字化并结合数字化过程中的各项特征进行的，且大多数成果成文较早。因此，对于古文信息处理全过程的研究状况，尚待吸收近年来的新成果进行更全面、更细致的总结与分析。本文将针对古文信息处理过程中所涉及的大的研究领域与方向，对古文信息处理当前的研究状况进行总结，

* 本文系国家社会科学基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”（项目编号：15ZDB127）和南京农业大学人文社会科学基金项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”（项目编号：SKPT2016001）研究成果之一。

作者简介：黄水清（ORCID：0000-0002-1646-9300），教授，博士生导师，E-mail：sqhuang@njau.edu.cn；王东波（ORCID：0000-0002-9894-9550），副教授，硕士生导师。

收稿日期：2017-02-13　修回日期：2017-05-08　本文起止页码：43-49　本文责任编辑：刘远颖

并分析未来的发展趋势。

2 古文信息处理概念的厘定

从概念的内涵和外延上看,古文信息处理是一个交叉的研究领域,涉及了数学、计算机科学、语言学和图书情报学等多个学科的理论、方法、知识和技术。从处理对象上看,本定义所指的古汉语没有时间上的限制性,即无论是五四运动之前的古文本还是今人以古代汉语撰写的文本,只要其音、形、义是以古代汉语的形式呈现出来的,均是古文信息处理的处理和加工的对象。从处理任务上看,古文信息处理囊括了古文的音素、音节、音调、字、词素、词汇、实体、短语、句子、段落、篇章和文集等不同语言单位上的标注、组织、挖掘和分析的任务。

古文信息处理的定义,强调了信息处理的对象是以古代汉语书写的文本,关注的是文本的内容,完全忽略文本的载体等外在特征。古文信息处理这一概念的命名与定义方式与"中文信息处理"一脉相承,体现了与中文信息处理的整体与分支关系。古文信息处理与现代汉语信息处理则同为中文信息处理下的分支,相互之间既有联系又有区别,所采用的研究方法、技术和理论基本上是一致的,但所处理的对象有严格区别,前者是古文,后者是现代汉语。

还有两个概念与古文信息处理意义相近,使用频度也比较高,即古籍信息处理、古文献信息处理。按《汉语大词典》的解释,"古籍"就是"古代典籍。泛指古书。"而"文献"则是"有关典章制度的文字资料和多闻熟悉掌故的人""后专指有历史价值或参考价值的图书资料"。这里所定义的文献,其实更接近古文献。显然,古文献的内涵比古籍大得多,包括了实物文献甚至包括掌握了知识的人。"古籍"专指古书,"古文献"则指一切历史性资料,"古籍"只是"古文献"的一个组成部分,或者说是其最主要的组成部分。

从语义上看,古籍包含了载体的概念在里面,即除了文本所呈现的内容外,古书的形态、装帧、印刷时代等都是古籍的组成部分。比如古籍整理就包括了对中国古代书籍进行校勘、刻印、版本甚至纸张及印刷技术认定等方面的工作。从时代上看,古籍一般指民国之前的书籍。

业界使用"古籍信息处理"或"古文献信息处理"这两个概念时,一般情况下其实并不涉及实物文献,也不涉及书籍的载体等外在特征,而只关注书籍的文本及文本呈现的语义。在这个意义上,使用"古文信息处理"比用"古籍信息处理"、"古文献信息处理"更贴切。而且,古文信息处理可以把今人创作的古汉语文本(包括格律词)也纳入研究范围内。比如《毛泽东诗词》中的旧体词和用古文撰写的《南京大屠杀死难者国家公祭鼎铭文》都不能称之为古籍,但它们是古文,它们不是严格意义上的古籍信息处理的对象,却是古文信息处理的对象。

古文中有一类的价值特别重大,就是被称为典籍的古代汉语文本。典籍专指重要的古代文本。比如,先秦诸子可称为典籍,而曾国藩编纂的《清代名人手札》虽然对研究清末的政治、军事和文化具有重要价值,可以称为古籍,却不适合称为典籍。

综上所述,古文信息处理可以比古籍信息处理、古文献信息处理等概念从内涵与外延上更精准地表达以古代汉语文本为对象的信息处理技术、内容、过程和范围。对古文信息处理的概念进行厘定,微观层面上有利于在整个领域朝着大数据、深语义和细颗粒度等方向发展的形势下聚焦古文信息处理的研究热点,宏观层面上有利于实现大数据环境下针对古文的知识挖掘与知识发现及向用户的传送与呈现,进而促进相关学科的发展,传承古代文化、讲好中国故事。

根据古文信息处理研究的整体状况,结合当下的研究趋势,本文对古文信息处理研究当中的数字化、智能处理和人文计算 3 个方面的研究内容做进一步的分析,并判断未来的发展趋势。

3 古文数字化的研究现状

古文数字化是目前古文信息处理成绩最为显著的领域,其核心内容就是基于现代信息技术对古文的文本进行加工和处理,以数字化的形式把古文存储在磁盘、光盘等介质上。古文数字化是对古文内容的再现和加工,是后续针对古文进行字频统计、自动分词、词性标注、分类、聚类、关联分析和语义知识挖掘等一切深度探究的前提和基础。经过近 40 年的发展,无论是在数字化的文献文本数量与类型方面,还是在相应的技术手段方面,均积累了大量的研究成果。

已经数字化的文本方面,美、英、日取得的成果较早。基于版本学、目录学和古汉语等专家学者编纂的索引,结合信息技术手段,1978 年,美国的 P. J. Ivanhoe 完成了针对《朱熹中庸章句索引》《王阳明大学问索引》《戴震孟子字义疏证索引》《王阳明传习录索引》《朱熹大学章句索引》的数字化,开启了世界性的古文数字化时代。此后,美国图书馆研究学会制定了开发"中文善本书目数据库"的计划,美国国会图书馆则在"美国记忆导航计划"项目中完成了中文古籍图书资

源的影像化。哈佛燕京图书馆在对 1800 种线装古籍进行编目的基础上开发了“线装古籍计算机检索系统”，同时通过与麦基尔大学的合作对明清时期的女性诗歌和著作进行了数字化并设计了检索工具。大英图书馆实施的“数字化图书馆计划”对《中国古代地图》和《急就篇》完成了数字化。日本京都大学主持编制了“全国汉籍书目数据库”。此外，东京大学针对比较珍贵的汉籍通过拍摄黑白或彩色的照片构建了“霞亭文库”“富士川文库”和“综合图书馆所藏古典籍”等 3 个数据库，在东亚的古籍数字化中具有重要的意义。上述数字化的探究为我国后续进行大规模的古文献数字化提供了可资借鉴的经验，尤其对我国古文献数字化流程的规范起到了直接而重要的推动作用。

受国外在古文数字化的技术、方法、策略和流程上的影响，中国大陆和台湾地区结合自身所拥有的丰富的典籍资源，在古文数字化资源建设方面取得了长足的发展。台湾地区自 20 世纪 80 年代开始编纂了大量的古文书目数据库，并且设计了一定量的书目索引数据库，其中较为有代表性的成果为覆盖全台湾的家谱联合目录。这些成果促使台湾地区迈出了古文数字化的第一步。在数字化书目和索引的基础上，台湾地区实施了针对特定领域的古文全文数字化的计划，并且取得了丰硕的成果。“古籍善本数据库”和“家族谱牒文献资料库”是其中的代表性成果，该成果是由台北故宫博物院研制。对特定领域的古文全文进行数字化的这一策略不仅有助于推进数字化向纵深方向发展，而且为实现针对古文献的全文本知识挖掘奠定了坚实的基础。

中国大陆地区的古文数字化工作始于 20 世纪 80 年代。但是，囿于古汉字字符集的规模，古文数字化的发展一直比较缓慢。随着“中易汉神 e”“龙语瀚堂四字节处理系统”等产品的问世，古汉语的生僻词输入问题在汉字的四字节存储体系下得以彻底解决。四字节存储体系对推动整个数字化的进程起到了加速器的作用。随着汉字字符集规模的不断扩展，古文数字化也得到了系统而全面的发展。1987 年，“史记全文检索系统”和“中国年历日历谱微机检索数据库”分别由哈尔滨工业大学和北京师范大学建成。这两项成果开创了中国对古文献全文进行字检索的先河。中国社会科学院计算机室的栾明贵在 20 世纪 90 年代初开发了“论语数据库”“永乐大典索引”“全唐诗索引”，在这一成果的基础上又启动了“中国古典数字工程”，并基于该工程于 2016 年建构了“子曰数据库”。栾明贵不仅完成了对古籍的数字化，而且基于数字化的成果进行了深入的挖掘和再加工。上海图书馆于 1996 年基本建成“中国古籍善本查阅系统”。1998 年，国家图书馆启动了“中国数字图书馆工程”，该工程在古文数字化方面涵盖了“数字方志资源库”“石刻拓片资源库”“甲骨文献资源库”“馆藏各类文献书目数据库”“永乐大典资源库”等 6 个子项目。百衲本《二十四史》电子版光盘、中国地方志宋代人物资料管理系统、续资治通鉴长编全文检索系统、全唐诗电子检索系统也由商务印书馆完成了开发。书同文公司与迪志公司合作开发的《文渊阁四库全书》和两种《古今图书集成》全文检索系统成为当时古文数字化的标志性成果。此外，由广西金海湾电子音像出版社和广西师范大学出版社完成的《古今图书集成》全文检索系统也是较为重要的数字化研究成果。国学公司开发的《中国历代基本典籍系列》则是重要的典籍文献全文数据资源。上述成果虽然完成了对古籍的数字化和检索，但由于没有对古籍文本进行自动分词的标注，所开发的检索系统都是基于字展开的，对查全率和查准率均有较大的影响。

古文数字化无论是在规模上还是质量上目前均取得了巨大的成就，但是，由于古文数字化更多的是实践性、工程性的工作，一定程度上导致了古文数字化的理论研究成果相对较少，也比较薄弱。在古文数字化的内涵与外延及整体架构上，栾贵明[11]给出了古文数字化的理论阐释。虽然该阐释相对比较宏观，但涉及到了古文数字化理论的主要方面。张普[12]针对古文数字化的进程，给出了“一个计算机与古籍整理相结合的新局面正在形成，更大规模的更加完善的古籍资料库和数据库正在筹划”的架构性判定。毛建军[13－16]认为古文数字化属于古籍整理的范畴，代表着古籍整理的未来方向，并给出了古文数字化的定义，同时在其 2008 年的博士论文中，从古典文献学和古籍整理学角度系统提出了古文数字化的基本理论框架，提出了古文电子索引、古文书目数据库以及古文全文数据库的概念，对古文数据库的规范提出了评价标准。在关于古籍数字化理论体系的架构上，毛建军的研究相对较充分和完善，对推动古籍数字化理论体系的构建起到了重要的作用。李弘毅[17]把古文数字化划分为准备阶段、自动化实施的过渡阶段、自动化发展的高级阶段。陈力[18－19]不仅从研究需求的角度指出了古文数字化存在的问题，而且认为古文数字化是数字图书馆重要的组成部分。罗凤珠、郁默、史睿、裴丽等[20－23]分别针对典籍特征、古文数字化与人文学术关系、国外数

字化的进度和中医学典籍的处理等多个层面进行了细致的研究。上述研究者从不同的角度对数字化的构成部分、归属和与不同学科的关系进行了论述，对古文数字化的理论构建具有一定的促进作用，但相对比较单一和分散。

4 古文智能处理的研究现状

随着古文数字化所涵盖的文本数量越来越多，如何从这些文本当中进一步提取、挖掘出更加有意义的知识成为了古文信息处理研究的重要任务之一。由于古汉语在篇章、段落、句子和字词上的特殊性，对数字化后的古文进行自动断句、词汇处理（分词、词性标记、命名实体识别）、语义和句法标注等智能处理成为了必不可少的环节。

古文断句是根据古代汉语句子的组合原则，结合现代汉语当中的句读集合，通过自动和智能化的策略完成对古代汉语自动添加句读的功能，进而实现对古代汉语句子的断句。目前关于古代汉语文本的自动断句主要有基于规则和基于机器学习两种策略。基于规则的古代汉语断句是通过人工总结古代汉语句子分布的情况，制定古汉语句子构成规则，并通过计算机自动实现对古汉语文本的断句。黄建年[24]探讨了如何利用计算机技术对农业古籍进行断句标点，构建了农业古籍断句标点的原型系统。该研究是基于规则进行断句的代表性成果。基于规则的方法由于规则的覆盖性较差、可迁移性较弱的原因，目前在古汉语断句中使用得越来越少，而通过人工标注的断句语料，结合隐马尔可夫模型、最大熵模型、条件随机场模型等机器学习模型构建古文本自动断句模型，已逐步成为了古语断句的主流方法。张开旭等[25]采用条件随机场模型基于《史记》和《论语》语料构建的古汉语自动断句模型，调和平均值接近 80%，是一项比较有代表性的基于机器学习策略的古汉语断句方法研究。虽然该研究取得了相对较好的结果，但如何更进一步地获取古文断句的特征知识并把该特征知识有效地融入到机器学习模型当中，是后续研究均需关注的问题。

古文的词汇处理涉及自动分词、词性标注和命名实体识别 3 个层面。自动分词是通过相应的技术，把由连续汉字构成的古汉语文本按照词的标准变成由连续词构成的文本。邱冰和皇甫娟[26]采用融合字的互信息和词汇频率知识的方法在先秦语料上进行了基于规则的分词实验。虽然分词效果整体上不是太理想，但在具体分词过程中有效地使用了统计的知识。古文自动分词的主流技术是基于机器学习的自动分词方法。梁社会和陈小荷[27]对《孟子》文本分别使用条件随机场模型和注疏文献方法进行了分词研究，发现两种方法在《孟子》中均取得了理想的分词效果。留金腾等[28]在人工校对的基础上，对《淮南子》完成了分词标注。王嘉灵[29]对《汉书》进行了详尽的分词研究，发现地名表、人名表加上注疏词表能达到最好的分词效果。古文的词性标注，是在分词的基础上，通过制定相应的词性标记集合，在统计和机器学习的基础上完成对词汇的名词、动词、形容词、副词、代词等词性的自动标注。石民等[30]实现了一种基于条件随机场模型的先秦汉语分词及词性标注一体化的系统。朱晓和金力[31]以《明史》为语料对象，验证了条件随机场在无边图模型、完全图模型以及嵌套图模型 3 种模型上的性能。张颖杰等[32]提出了一种新颖的先秦汉语词义标注方法，利用 SVM 对《左传》进行了半指导的词性标注实验，为缺乏训练语料的古文词义标注提供了一个可行的研究思路。古文的命名实体识别是指通过规则和机器学习的方法从古汉语文本中自动识别出人名、地名、机构名、官职、时间和谥号等具有独立语义内涵的词汇。无论是对历史事件演化历程的揭示还是对基于复杂网络人物关系的深层挖掘，古文的命名实体识别都具有重要的意义。通过对人名的分布状况的分析，汤亚芬[33]采用条件随机场模型完成了面向先秦语料的人名实体模型的构建。同样基于条件随机场模型，黄水清等[34]通过构建多特征模板完成了对典籍当中地名的自动识别。相较于基于规则和统计的方法，上述分词和词性标注的研究充分证明了机器学习的方法具有十分突出的优势，而且说明了构建可供机器学习模型训练的精加工语料的重要性。

古文的语义和句法标注是在词汇标注的基础上，对古文进行更深层的词汇语义知识和句法组合规则标注。语义标注对汉语，尤其是古代汉语来说，是一项非常困难的任务，因此目前这一方面的研究相对较少。比较有代表性的是于丽丽等[35]比较了条件随机场与最大熵和朴素贝叶斯统计模型，发现通过条件随机场进行古汉语词义消歧的实验效果最好。陈小荷[1]在对古汉语语义标注的内涵和外延进行界定的前提下，结合《春秋左氏传》的语料，验证了基于规则、基于统计和基于机器学习的 3 种策略的整体性能。古文的句法标注是针对古汉语句式的基本特征，基于某一种自动句法分析器，完成对古汉语文本当中词与词、词与短语、短语与短语之间主谓、动宾、动补、介宾、定中和状中等句法关系的标注。由于句法标注性能整体上相对

较差，目前这一方面的研究也较少，较有代表性的是冯秋香[36]基于数据库语义学及整理出的古汉语论元句法和并列句法特征，实现了对这两类句法结构的自动生成。从上述研究可以看出，目前针对古文的语义和句法标注还停留在相对较简单的层面。如何更进一步地对古文进行语义和句法标注，不仅需要更先进的技术，而且更需要针对古文标注的理论知识。

5 古文人文计算的研究现状

人文计算也称为数字人文。根据国内外相关的研究，人文计算是面向人文社会科学与计算之间的交叉领域开展研究，通过智能检索、文本挖掘、可视化等各种信息技术和手段达到研究目的。人文计算根据载体的不同又可以分为文本人文计算、语音人文计算、图像人文计算和视频人文计算，而在文本人文计算当中与古代汉语相关的人文计算则可称为古文人文计算。除了学者个人纷纷投身人文计算研究外，众多的与人文计算相关的研究机构也相继设立。其中，日本立命馆大学京都数字文艺研究中心、台湾大学数位人文研究中心、武汉大学数字人文研究中心在古文人文计算的理念、实践和方法上均有相应的涉猎。

在先秦典籍的古文人文计算方面，汪定明和李清源[37]从历时性与共时性的维度构建了老子汉英平行语料库，并比较了汉语典籍与对应英语在语义、语用和语体上的差异。许超等[38]利用 Pajek 软件，对《左传》中提取的人物和事件建立了社会网络，并以此为基础，对该时期的社会网络关系进行了定性定量探索，为探究古代社会关系提供了新思路。在对《楚辞》的特点进行分析的基础上，钱智勇等[39－40]论述了《楚辞》知识库及网站设计的实现步骤、技术难点及解决思路，提出将语义网技术应用于辞赋知识组织的技术构想。在唐诗宋词的古文人文计算方面，北京大学计算语言研究所的胡俊峰，俞士汶[41]开发了“唐宋诗计算机辅助研究系统”，不仅对唐宋诗的特点进行了深入的分析，而且可以自动生成相应的诗句。北京大学的李铎[42－43]则开发了全宋诗分析系统。余振山等[44]通过隐码的策略，成功开发了指定词牌后的词自动生成系统。在明清小说的古文人文计算方面，陈炳藻[45]利用计算机对《红楼梦》前 80 回和后 40 回的用字进行了测定，并从数理统计学的观点出发，推断出前 80 回与后 40 回的作者均为曹雪芹一人的结论。在把 120 回本的《红楼梦》看成一个样本的情况下，李贤平[46]通过聚类的算法，得出了与陈炳藻不一样的结论。基于自行构建的《红楼梦》汉英平行语料库，刘泽权等[47]系统地分析了汉英动词在《红楼梦》当中的分布特征以及所涵盖的文化特点。在二十四史的古文人文计算方面，董慧等[48－50]的研究团队从语义系统的角度，对中华史籍进行了多角度分析。通过“向下挖掘、向上组织”的国史知识语义揭示与组织方法，王颖等[51]在对隐藏于国史资源文本条目中的国史知识对象和相关事实进行语义挖掘和揭示的基础之上，通过国史知识对象的关联，构建国史知识网络，并基于时间、类属、层级及统计等关系，对国史知识内容进行更高层次的多维组织展示。从上述研究可以看出，虽然有一定的研究者对历时的典籍进行了人文计算的探究，但绝大部分的研究还是停留在某一部或者几部典籍的探究上，并且针对典籍人文计算的探究相对还停留在词汇、术语和实体这个层面上，缺乏系统而全面的针对句子、段落、篇章和典籍文本的研究。

6 趋势及展望

随着理论的发展、技术的成熟以及时代对传统文化、中国故事寄予的重托，古文信息处理将迎来发展的契机和增长的高潮。

6.1 古文数字化方面

基于已完成的古文数字化工程，对古文数字化过程中的相应经验进行总结，并结合古文献学、版本学、目录学、计算机科学和信息学，将形成古文数字化的基本方法和操作规范。其针对古文数字化过程中生僻繁体字无法被识别或录入的难题，在已有 GB2312、BIG5 和 GBK 字符集的基础上，将探讨并产生扩大字符集规模及解决字符集存储问题的新方法。对于由 OCR 识别而造成的古代汉语文本数字化方面的错误，在充分统计和分析所造成错误的类型和分布规律的基础上，结合统计和机器学习的策略，可构建错误恢复模型，提升古文 OCR 自动识别的精准率。对已经影印或扫描的古籍应加大转换为文本文件的力度，为后续的古文智能处理和人文计算奠定坚实的基础。从顶层设计的角度考虑，对分散在不同机构和单位的已数字化的古文进行统一的整合，构建历时而海量的古文数字化资源库，对于探究、厘清和把握中华文化的历史渊源、发展脉络和基本走向具有极为重要的作用。

6.2 古文智能处理方面

通过充分挖掘已有的语义化的古文词汇、术语、地名、人名、官职、谥号等资源，建立语义关联和整合资源，构建具有一定覆盖度的古文语义知识资源库，为后续的古文智能处理奠定坚实的基础，并为面向典籍文

本展开人文计算提供直接而有针对性的领域化知识。为实现真实语境下大规模古代汉语文本智能化处理任务,制定相应的标注规范并由受过专门训练的相关人员完成对一定量的古文的人工标注,形成所谓的"金本位"的古文语料库,并基于该语料库构建性能优越的各种机器标注模型,在将来的古文智能化处理过程中势必成为一种趋势。另外,全面而系统地把机器学习和深度学习的技术融入到古文智能化处理中,构建针对古文的词汇级、实体级、句子级、段落级和篇章级等各个层级的机器学习模型,也是古文智能处理研究的必然趋势。围绕某一历史时间段或者某一专题的古文语料,结合跨语言检索、本体和语义网的技术和方法展开多层面、立体和全方位的古文智能处理,将成为未来研究的趋势之一。

6.3 古文的人文计算方面

结合人文计算的理念,将通过智能处理技术从古文中获取的词汇、实体、句子、段落和篇章的知识提升到人文计算的层面,实现真正意义上的古文人文计算,是未来的热点问题。基于通过人文计算所获取的知识,结合虚拟现实、增强现实、地理信息系统和信息可视化等技术,构建针对人文计算结果的多维度和多视角的呈现方式,是未来研究的趋势之一。古文人文计算的研究不仅要充分发挥机器学习的优势,而且需要针对古文的人文性、社会性和历史性等选取特性并制定人文计算相应的特征和模板,进而提高古文人文计算的整体性能。结合人文计算的各种成果和在计算过程中所涉及的策略、知识和技术,针对计算过程中的特点与特征,探究古文人文计算的方法论和理论体系,是整个人文计算不可回避也无法回避的研究课题之一。

7 结语

本文在厘定古文信息处理的概念及回顾相关综述文献内容的基础上,通过检索得到与古文信息处理相关的研究文献,基于文献计量学的方法,统计和分析了古文信息处理的整体研究状况。然后,依次对古文信息处理的 3 个主要领域——古文数字化、古文智能处理、古文人文计算的研究现状进行了总结。随着数据驱动研究方法的兴起及深度计算的日益普及,古文信息处理研究在分词、词性标注、浅层句法标注、深层句法标注和语义标注等方面会积累越来越多的深层次标注资源和高效性能标注模型,同时在事件抽取、人文关系网络构建、篇章结构识别、文体特征挖掘和语体风格计算等人文计算的研究对象方面也会取得越来越多的成果。

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] 姚松. 计算机用于古籍整理研究的现状与展望[J]. 中国典籍与文化,1995(2):121-127.
[3] 王桂平. 我国古籍数字化的现状及展望[J]. 图书情报知识,2000(4):50-51,54.
[4] 李国新. 中国古籍资源数字化的进展与任务[J]. 大学图书馆学报,2002,20(1):21-26.
[5] 潘德利. 中国古籍数字化进程和展望[J]. 图书情报工作,2002,46(7):117-120.
[6] 李弘毅. 浅论古籍数字化的发展阶段[J]. 上海高校图书情报学刊,2002(2):24-27.
[7] 厉莉. 古籍数字化的现状及对策[J]. 图书馆研究,2002,32(1):57-58.
[8] 段泽勇,李弘毅. 古籍数字化的回顾与展望[J]. 图书馆理论与实践,2004(2):37-39.
[9] 王冠中. 中文古籍数字化成果与展望[D]. 长春:东北师范大学,2005.
[10] 林竹鸣,朱翠萍. 古籍数字化的历史、现状及问题探析[J]. 淮北师范大学学报(哲学社会科学版),2014(6):192-194.
[11] 栾贵明. 电脑中文的突破性进展——迎接第一个国际汉字标准确定[J]. 汉字文化,1992(2):38-41.
[12] 张普. 计算机在中国古籍整理研究领域中的应用(综述)[J]. 语文研究,1989(4):40-45.
[13] 毛建军. 古籍数字化的概念与内涵[J]. 图书馆理论与实践,2007(4):82-84.
[14] 毛建军. 古籍数字化研究的回顾与思考[J]. 国家图书馆学刊,2007,16(3):62-65.
[15] 毛建军. 国外中文古籍数字化资源概述[J]. 数字图书馆论坛,2006(12):30-34.
[16] 毛建军. 欧美地区中文古籍数字化概述[J]. 数字与缩微影像,2008(1):36-38.
[17] 李弘毅. 浅论古籍数字化的发展阶段[J]. 上海高校图书情报学刊,2002(2):24-27.
[18] 陈力. 中文古籍数字化的再思考[J]. 国家图书馆学刊,2006(2):42-49.
[19] 陈力. 中文古籍数字化方法之检讨[J]. 国家图书馆学刊,2005(3):11-16.
[20] 罗凤珠. 以"互动观念"建立"红楼梦网路资料中心"对红学发展之影响[J]. 红楼梦学刊,1997(S1):537-546.
[21] 郁默. 台湾中央研究院汉籍全文资料库[J]. 中国典籍与文化,1998(3):110-115.
[22] 史睿. 论中国古籍的数字化与人文学术研究[J]. 北京图书馆馆刊,1999(2):28-35.
[23] 裴丽,褚长海. 中医古籍文献资源数字化建设探讨[J]. 图书馆学研究,2001(6):67-68.
[24] 黄建年. 农业古籍的计算机断句标点与分词标引研究[D]. 南京:南京农业大学,2009.
[25] 张开旭,夏云庆,宇航. 基于条件随机场的古文自动断句与标点

方法[J]. 清华大学学报(自然科学版) 2009(10):163 -166.
[26] 邱冰,皇甫娟. 基于中文信息处理的古代汉语分词研究[J]. 微计算机信息,2008(24):100 -102.
[27] 梁社会,陈小荷. 先秦文献《孟子》自动分词方法研究[J]. 南京师范大学文学院学报,2013(3):175 -182.
[28] 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(6):6 -15,81.
[29] 王嘉灵. 以《汉书》为例的中古汉语自动分词[D]. 南京:南京师范大学,2014.
[30] 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(2):39 -45.
[31] 朱晓,金力. 条件随机场图模型在《明史》词性标注研究中的应用效果探索[J]. 复旦学报(自然科学版),2014(3):297 -304.
[32] 张颖杰,李斌,陈家骏,等. 基于词典信息的先秦汉语全文词义标注方法研究[J]. 中文信息学报,2012(3):65 -71,103.
[33] 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.
[34] 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.
[35] 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(10):45 -48.
[36] 冯秋香, 汪榕培. 数据库语义学在古汉语自动分析上的应用[J]. 大连理工大学学报, 2012(6):902 -907.
[37] 汪定明, 李清源.《老子》汉英翻译平行语料库建设[J]. 上海翻译, 2013(4):43 -47.
[38] 许超,陈小荷.《左传》中的春秋社会网络分析[J]. 南京师范大学文学院学报,2014(1):179 -184.
[39] 钱智勇,周建忠,贾捷. 楚辞知识库构建与网站实现研究[J]. 图书馆理论与实践,2010(10):70 -73.
[40] 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.
[41] 胡俊峰,俞士汶. 唐宋诗之计算机辅助深层研究[J]. 北京大学学报(自然科学版), 2001, 37(5):727 -733.
[42] 李铎,王毅. 关于古代文献信息化工程与古典文学研究之间互动关系的对话[J]. 文学遗产, 2005(1):126 -137.
[43] 李铎. 从检索到分析——计算机知识服务的时代[J]. 文学遗产, 2009(1):135 -137.
[44] 余振山, 黄刘生, 陈志立,等. 用宋词实现高嵌入率文本信息隐藏[J]. 中文信息学报, 2009, 23(4):55 -62.
[45] 陈炳藻. 从词汇上的统计论《红楼梦》的作者问题[C]//首届国际《红楼梦》研讨会. 麦迪逊:威斯康辛大学,1980.
[46] 李贤平.《红楼梦》成书新说[J]. 复旦大学学报(社会科学版), 1987(5):3 -16.
[47] 刘泽权, 刘超朋, 朱虹.《红楼梦》四个英译本的译者风格初探——基于语料库的统计与分析[J]. 中国翻译, 2011(1):60 -64.
[48] 董慧,徐雷,王菲,等. 语义分析系统研究(Ⅱ)——史籍推理机制[J]. 情报学报,2014,33(2):195 -203.
[49] 董慧,徐雷,王菲,等. 语义分析系统研究(Ⅲ)——中华史籍语义分析系统实现[J]. 情报学报,2014,33(2):204 -214.
[50] 董慧,徐雷,王菲,等. 语义分析系统研究(I)——史籍语义分析流程[J]. 情报学报,2014,33(2):183 -194.
[51] 王颖, 张智雄, 孙辉,等. 国史知识的语义揭示与组织方法研究[J]. 中国图书馆学报, 2015(4):55 -64.

作者贡献说明：

黄水清：提出相关概念及整体研究思路，修订完稿；

王东波：进行统计分析及初稿撰写。

Review and Trend of Researches on Ancient Chinese Character Information Processing

Huang Shuiqing[1,2]　Wang Dongbo[1,2]

[1] College of Information Science and Technology, Nanjing Agricultural University, Nanjing, 210095

[2] Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] With the rapid development of ancient Chinese character digitization, intelligent processing and humanities computing, a review about these researches could not only help to summarize the corresponding rules but also be beneficial to the subsequent researches. [**Method/process**] This paper defines the concept of ancient Chinese character information processing, and analyzes the overall situation of ancient Chinese character information processing based on preliminary statistics under the definition of ancient Chinese character information processing. Meanwhile, the researches on ancient Chinese character digitization, intelligent processing and humanities computing are summarized, reviewed and forecast on the later research trend. [**Result/conclusion**] On the whole researches of ancient Chinese character information processing, the achievements of ancient Chinese character digitization is the largest, and the researches of intelligent processing of ancient Chinese character in vocabulary level gain some effects. But the study of humanities computing relevant with ancient Chinese character is only just beginning.

Keywords: ancient Chinese character digitization　digital humanities　intelligent processing　humanities computing　ancient Chinese character information processing