数字人文视角下语料库在清代黄河问题研究中的应用

潘威,徐娟

(云南大学 历史与档案学院,云南 昆明　650091)

摘　要:清代黄河研究是历史地理学中的一项重要研究内容,历史地理信息化的发展为清代黄河研究积累了丰富的数字化史料,也为相关史料的整理、分析和知识挖掘等带了新的挑战。语料库作为能够满足多语种、历时性、大规模数据处理和分析的重要工具,将其引入到清代黄河研究中能够助力研究的深入发展。从分词、词性标注等方面入手,阐述了语料库在清代黄河研究中的必要性,提出了一套可行的清代黄河标注语料库建设的技术规范,并以部分清代"河道钱粮册"语料为例进行实证研究,研究发现,从"河道钱粮册"标注语料中基于标注的时间、地点、机构、数词、量词等词性抽取出历年朝廷黄河河银的来源、去向、数量、时间的关联信息,来源包括地点和白银项目。这些指标数据可以支持实现快速的知识整理和深层次的知识发现,能够反映清代河工银制度在不同时期的财政体制变化及其对黄河河工的影响。

关键词:数字人文;语料库应用;分析与词性标注;清代;黄河;"河道钱粮册"

中图分类号:K061　　**文献标志码**:A　　**文章编号**:1671-6248(2024)03-0125-14

收稿日期:2024-01-17

基金项目:教育部哲学社会科学研究重大课题攻关项目(22JZD039)

作者简介:潘威(1981–),男,上海宝山人,教授,博士研究生导师,历史学博士。

Application of corpus in the study of the Yellow River in the Qing Dynasty from the perspective of digital humanities

PAN Wei, XU Juan

(School of History and Archives, Yunnan University, Kunming 650091, Yunnan, China)

Abstract: The study of the Yellow River during the Qing Dynasty is a significant area of research in historical geography. The advancement of informatization in historical geography has provided a wealth of digital historical materials for studying the Yellow River in this period, while also introducing new challenges in organizing, analyzing, and extracting knowledge from these materials. Introducing corpus analysis into the study of the Yellow River in the Qing Dynasty can facilitate more in-depth research by enabling multilingual, diachronic, and large-scale data processing and analysis. This study begins with word segmentation and part-of-speech tagging to explain the necessity of using a corpus in researching the Yellow River in the Qing Dynasty. It proposes a set of feasible technical specifications for constructing an annotated corpus of the Yellow River in the Qing Dynasty. An empirical study is conducted using a portion of the "Financial Records of River Canals" corpus from the Qing Dynasty. The study extracts data on the source, destination, quantity, and timing of the subsidies allocated by the government for the Yellow River over the years from the annotated corpus based on parts of speech such as annotated dates, locations, organizations, numerals, and quantifiers. The sources include locations and silver items. This indicator data supports rapid knowledge organization and in-depth knowledge discovery, reflecting changes in the fiscal system of the river construction subsidy system during different periods of the Qing Dynasty and its impact on the Yellow River construction works.

Key words: digital humanities; corpus application; analysis and part-of-speech tagging; Qing Dynasty; Yellow River; "Financial Records of River Canals"

黄河是中华文明的摇篮,中华民族治理黄河的历史也是一部治国史,深入挖掘黄河文化蕴含的时代价值,延续历史文脉是实现中华民族伟大复兴的中国梦的重要力量[1]。清代作为黄河环境变迁及黄河治理发生重大变化的时期,对该时期的河工史料进行系统整理、充分挖掘各类史料中的信息对于深入开展黄河问题研究具有重要意义。

近年来,随着历史地理信息化和数字人文技术的发展,清代黄河问题的史料整理取得了重大进展,历史地理学的研究者们建设了一系列支撑各类专题研究的清代黄河数字

资源库[2],极大地推动了清代黄河研究信息化发展的进程。其中具有代表性的清代黄河数字资源库是由河南大学经济学院与云南大学历史地理研究所共同建设的“数字历史黄河”专题信息平台,该平台整合了包括正史河渠志、正史中的黄河水灾记录、治河类书、清代河务档案在内的数量庞大、种类多样的黄河历史文献资源。基于该平台学者们开展了一系列数字人文研究,例如潘威等以“数字历史黄河”地名库为例,将大数据技术引入历史地名数据库,提出了一种基于时空框架和时态地理信息系统的思路和方法[3]。但这一过程中也出现了一些新的问题,一方面随着数字化史料规模的不断扩大,史料的整理、阅读和检索难度急剧增加,迫切需要自然语言处理技术辅助研究者在更短的时间内掌握史料的基本内容,实现更加细粒度、多元化的检索,提高史料的阅读效率;另一方面,目前的研究主要停留在基于史料的归纳和演绎层面,在史料内容的整理、分析以及深层次知识挖掘方面有待提升。语料库作为支撑大规模、多来源数据的存储、检索、深层次语义分析和知识挖掘的重要工具,将其引入到清代黄河研究中将会在史料辅助阅读、文本分析与知识挖掘等方面带来一些新的突破。

清代“河道钱粮册”是反映河银拨款、河务稽查、河道疏浚及河患治理等事务的奏折类史料,具有极强的真实性和作为史料的可靠性,是研究清代黄河问题的重要史料[4]。鉴于此,本文在详细阐述语料库技术在清代黄河问题研究中的必要性的基础之上,以部分清代“河道钱粮册”语料为例,尝试探索清代黄河标注语料库的构建过程,提出一套可行的技术规范,并基于构建的黄河标注语料库从检索、计量分析、知识挖掘等方面探讨语料库在清代黄河问题研究中的应用前景,以期为相关研究提供参考和借鉴。

一、“河道钱粮册”语料库技术

语料库是存放一定规模的原始数据材料并利用现代化信息技术进行语料信息搜集、整理和处理分析的数字化资源库。语料库不仅是按照一定的标准而形成的某一领域的数字资源库,更是一种研究方法和新的研究思维,构建语料库是为了更方便研究某类问题并达到研究目的而服务的。语料库的构建涉及到多个学科领域,它融合了语言学、计算机科学、统计学等学科的知识,其优势在于可以支持从词汇和语义层面实现大规模数据中史料信息的精确检索及知识挖掘与发现。20 世纪 90 年代以来,得益于机器学习、自然语言处理技术和全文检索技术的飞速发展,第三代语料库技术已由最初的中文信息处理、语言学等领域相关的总体概念向词向量、深度学习、知识图谱、数据智能分析等技术发展[5],随着大量文献的文本化,语料库技术不仅向科学化、多元化转变,也更趋向于精准化、智能化。近年来不同学科的研究者们针对各学科的特性构建了一系列解决不同学术问题的语料库,语料库的出现和发展为各学科领域的研究问题提供了新的研究思路和数据支撑[6-11]。

分词、词性标注、命名实体识别是语料库建设过程中最核心的内容,它会直接影响到

语料库分析、处理和使用环节的精准度[10]。近年来,随着古籍数字化进程的不断加快,学者们开始意识到古文语料库建设的重要性。针对古代汉语具有的特殊语法结构和语言特点,国内语言学界、图书情报以及人工智能领域的学者们提出了一系列古文自动分词、词性标注、命名实体识别方面的方法,有力地推动了古文的检索、计量分析和知识挖掘等领域的发展。北京大学计算语言学研究所建立了唐宋诗语料库,收录了 641 万字的全唐诗及部分宋诗名家的古诗语料,为古诗词检索、古诗词语义计算以及古诗词的自动生成等问题的研究带来了新的突破[12]。中国台湾省率先构建了包含 48 部上古文献、69 部中古文献、20 部近代文献的较大规模的古汉语词性标注语料库[13],为上古至近代部分文献的语义分析和知识发现工作奠定了基础。古文分词方面,学者们通过对各类古文文献进行实验测试,验证了 N 元语法("N-gram language model"或者"N-gram statistical language model",简称 N-gram)、隐马尔科夫(hidden markov model,简称 HMM)、条件随机场(conditional random field,简称 CRF)、BERT(bidirectional encoder representations from transformers)等模型在古文自动分词和词性标注方面具有良好的效果[14-19]。目前针对古文的词性标注主要分为先分词再词性标注和分词标注一体化两种方式,通常情况下前者的效果要比后者要好[20]。词性标注一般采用序列标注算法实现,其中条件随机场(CRF)以其支持多特征融合的特点,在古文的词性标注方面具有良好的标注效果[21]。命名实体识别是指从文本中标记出人名、地名、时间、机构名等专有名词,命名实体的识别对于信息检索和信息抽取起着至关重要的作用。目前针对古文的命名实体自动识别的主流方法有条件随机场(CRF)及深度学习方法[22-26],部分学者积极探索了基于条件随机场和最大熵模型、序列化标注、深度学习等方法的古籍实体识别方法[27-28],为关联关系抽取和知识图谱构建提供了有力支持。清代黄河问题研究的相关史料具有古文的部分特点,以上研究对于实现清代黄河标注语料库的分词、词性标注和命名实体识别提供了研究方法的借鉴。

随着学术研究的深入,一些清史研究专家开始对清代奏折进行系统的整理,这些工作的开展有力地推动了清代河务奏报档案的整理与研究[3],对于清代"河道钱粮册"相关档案的摘录与整理工作逐渐开启,已陆续整理出大量"河道钱粮册"档案资料,"河道钱粮册"主要是清代奏折类史料,目前对于奏折类档案的研究大多停留在寻章摘句、史料的归纳和演绎之中,缺乏对史料内容的有效管理,在文本的处理、分析和解读方面存在诸多不足,借助语料库技术可以协助实现对"河道钱粮册"内容的高效管理,为研究者开展学术研究提供便利。清代"河道钱粮册"既具有一般古文的语法结构和语法特点,同时又具备独特的语言特点以及奏折类档案的特征,这种特征主要体现在文本中包含了大量与清代河务相关的专业术语,主要特点包括:(1)相同、相关意义联合式复合词较多,例如"项款""役夫""拣选""堵筑""捐输""地丁"等。(2)修饰人或事、动作行为及性质或状态的偏正式复合词较多,例如"解役"

“岁修”“捏报”“开列”等。(3)涉及到一些河段、河工、河务管理机构等与河务相关的专业术语及奏折类的规范用语,如“中牟大工”“谨奏”“跪奏”等。目前暂未发现专门针对清代“河道钱粮册”文本的分词、词性标注、命名实体识别的相关研究。

二、语料库在清代黄河问题研究中的必要性

数字人文、GIS 等技术的应用极大地推动了历史地理信息化的发展,为清代黄河问题研究注入了新的活力,但在史料的整理、分析与深层次的知识挖掘等方面仍存在诸多不足,许多有见地的研究成果因缺乏细粒度的量化指标、数据标准不统一、数据分析不足等问题,使得研究成果难以在更广泛的领域凸显其价值。针对这一问题,在对清代黄河问题相关史料进行收集、整理、考订及数字化的基础之上,借助语料库技术对大规模史料进行整理和标注,并借助自然语言处理的分析手段,不断提高清代黄河问题研究的深度和广度是一条切实可行的科学研究道路。

(一)协助清代黄河史料的管理

历史地理研究高度依赖史料,在历史信息化的推动下清代黄河问题研究的相关史料具有良好的信息化基础,积累了大量数字化的史料数据,清代的黄河史料数据具有的多语种、大规模、历时性等特点为史料的管理带来了很大的难度[3]。此外,传统清代黄河史料数据的管理更多侧重于对史料宏观特征的描述与揭示,很少深入到史料的内容层次,导致能够为研究者提供新的知识增益非常有限。20 世纪 90 年代以来,第三代语料库技术在设计、编码和数据的管理方面均取得了较大进展,可以很好满足清代黄河史料基于内容层次的管理需求[29]。目前历史地理学者针对清代黄河的研究主要围绕黄河水患、河道变迁、河政体制等几个方面展开[30],各个方向的研究相对独立,研究数据碎片化严重,存在严重的“数据孤岛”现象。学者们往往倾向于选择与自己研究方向最相关的部分史料开展大量微观的、个案的、区域性的研究,而忽略了其他各社会因素之间相互作用的影响,缺乏整体性、结构性的分析,导致无法看到特定历史时期黄河问题的整体面貌,从而可能造成研究结论不准确。通过构建清代黄河历时性标注语料库,利用语料库的编码、分词和标注等技术可以实现不同来源、不同类型、不同结构的各类清代黄河史料数据在同一个语言标注框架下的综合管理,破除“数据孤岛”现象,为清代黄河研究的学者们提供覆盖面更全、层次更加多样的史料信息,使清代黄河问题研究从选择性分析向整体性还原转变。

(二)助力实现史料的智能化检索

随着数据规模的不断扩大,清代黄河研究的史料资源也逐步进入大数据时代,如何快速、准确地从海量的史料文本中检索到研究所需要的信息,并为历史地理研究者呈现史料中隐藏的知识线索成为了历史地理研究的重要任务之一。简单将史料进行数字化并利用数据库进行存储,虽然在一定程度上可以方便历史地理史料的查阅和检索,但由于

组织粒度较粗、组织形式单一，很难满足多元、智能的检索需求[31]。由于各类清代黄河研究的专题数据库往往未对史料进行分词标注，使得清代黄河史料的检索大都停留在低层次的基于字符匹配的全文检索层面，检索的效率和准确率相对较低。通过借助语料库的衍生技术进行史料的分词、词性标注和命名实体识别等，可以实现基于词、词性以及人名、地名、时间等实体的跨文本或者跨库检索，打通库的界限，建立各类史料之间的联系。例如在对人名进行识别之后，通过检索某一人物名，便可以精准获取到与该人物相关的所有史料数据，进而大大提高史料检索的效率和准确率，减少人工阅读史料的时间。此外，分词、词性标注、命名实体识别等还是实现语义检索、主题检索、可视化检索的重要前提[32]，进而为最终实现清代黄河史料的智能化检索奠定基础。

（三）推动史料的深层次分析和知识挖掘

深入研究清代黄河问题既要吸收传统史学的定性研究方法的长处，也需要引入计量分析、语义分析、知识挖掘等定量分析方法，将微观分析与结构思维有机结合起来，力图实现研究方法的多元化以及史料分析的智能化。语料库的分词、词性标注、命名实体识别等是实现智能化分析和处理必不可少的环节，经过标注的语料不再是篇章的堆砌，而是成为可供研究者进行知识发现的有价值的数字资源，可以促使清代黄河研究史料的应用方式从“读”向“分析”转变，有效提高史料的复用性和实用价值。基于标注的语料库可以实现词汇级、实体级、句子级、段落级、篇章级等各个层级的史料分析和研究，字词的分解有利于为清代黄河问题研究相关史料中的各类属性提供描述性信息和量化指标，据此可以实现基于词汇的信息挖掘和计量统计，进而推动清代黄河问题研究由传统定性分析向定量分析延伸，同时有助于通过计量手段为研究者提供结构化思维。通过对清代黄河史料中的人名、地名、官职名、机构名、工程名、时间等实体进行标注和编码，以某些特定词性的词作为触发词可以实现关联关系的抽取，在此基础之上，结合自然语言处理技术和可视化技术能够开展深层次的知识挖掘，建立起各实体之间的语义联系，构建领域的知识链条，形成相应的知识网络，实现知识的外显化，更好地为历史地理学者的学术研究服务。

综上所述，语料库在清代黄河问题研究中的应用从微观层面上有利于整个研究领域在朝着大数据、深语义和细颗粒度的方向发展的形势下，聚焦史料文本信息处理的研究热点，从宏观层面上有利于实现大数据背景下针对清代黄河问题研究史料的深度分析、挖掘和知识发现，进而促进整个研究领域的发展。

三、清代黄河标注语料库的构建

清代黄河标注语料库的构建主要涉及到语料的采集选取、分词、词性标注三大块的内容，其中词性标注包含了对人名、地名、机构名、工程名等命名实体的标注，具体的技术路线和流程如图 1 所示。

本文是一个过渡性的研究，通过本文的

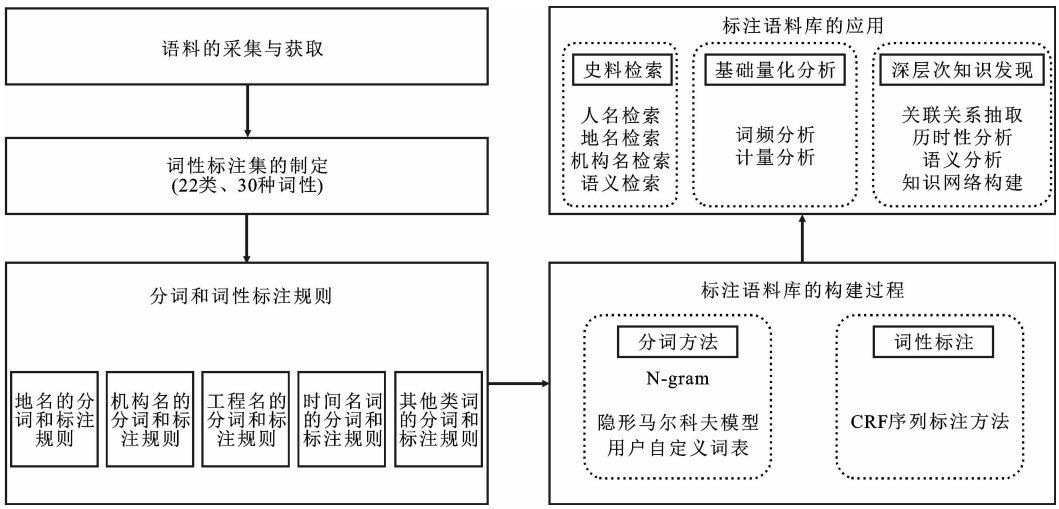

图 1 清代黄河标注语料库构建流程和技术路线

研究可以为后续实现全自动大规模的清代黄河史料的标注奠定基础,也可以为针对河流研究的语料库建设提供一套可参考的技术规范。

(一)语料库的数据来源及构建方法

本文主要以潘威等[30]整理的部分清代“河道钱粮册”档案为初始语料来展示标注语料库的构建过程及其在黄河问题研究中的应用价值。这部分数据以奏折类史料为主,具有重要的史料价值。本文主要采用自动标注和人工校对相结合的方法,构建集分词和词性标注于一体的清代河工财务档案标注语料库,以期为古文分词、命名实体识别、关系抽取和大规模清代黄河标注语料库的构建等一系列与清代黄河相关的自然语言处理研究工作奠定基础。

(二)清代黄河标注语料库的分词和词性标注规则

分词与词性标注是语料库建设的核心工作,由于清代“河道钱粮册”语料具有古汉语的特点,在分词时应注意切分单元的颗粒度,再加上语料中又涉及一些与河工相关的专业术语及专有名词,也为分词和词性标注工作增加了难度。考虑到以上因素,本文主要采用机器自动标注与人工校对的方式对语料进行分词及词性标注,在保证标注质量的同时兼顾标注速度。机器自动标注主要采用 Python 编程、“甲言”结合自定义词表的方法进行分词和词性标注,该方法无需提前构建训练集,在分词和词性标注的精度和速度方面都能够较好地满足研究需求。

1. 清代黄河标注语料库的词性标注集

人名、地名、机构名、官职名、工程名、时间等实体的标注在满足多元检索、文本分析和知识挖掘等方面具有重要意义,因此本文在“甲言”分词标准的基础之上将这几类名词进行了专门的标注,最终将词性标注集分为 22 类、30 种词性,见表 1。

2. 清代黄河标注语料库的分词和词性标注规则

(1)地名的分词和标注规则

标注的地名类型主要分为以下几种:

表 1　清代黄河标注语料库词性标注集

序号	词性	标记符	序号	词性	标记符	序号	词性	标记符
1	普通名词	n	3	形容词	a	13	缩略词	j
	时间名词	nt	4	数词	m	14	前接成分	h
	方位名词	nd	5	量词	q	15	后接成分	k
	处所名词	nl	6	副词	d	16	语素字	g
	人名	nh	7	代词	r	17	非语素字	x
	地名	ns	8	介词	p	18	标点符号	wp
	机构名	ni	9	连词	c	19	非汉字字符	ws
	官职名	ng	10	助词	u	20	名词修饰语	b
	工程名	ne	11	叹词	e	21	描述性词语	z
2	动词	v	12	拟声词	o	22	习用词	i

①行政区划:行政区划是地名中重要的内容,清代的行政区划主要包括了“省”“府”“州”“厅”“县”等,由于行政区划通常为“专名 + 通名”的结构形式,且名称使用比较规范,在地名标注时作为一个整体标注,不予切分,如“东昌府/ns”。②山川自然地理名称:由专名或普通名词加地形地貌的名词(如江、河、湖、海、山、岛、平原、峡谷等)构成的地名。③黄河治理相关的聚落名称:主要包含了河流沿线的一些村落名、集镇、街路巷、关隘等相对较小的标识性地点名称,这类地名是河务事件的重要载体,需要重点标注出来。④河段名称:主要是指基层河务机构或组织所管辖的河段名称。⑤对于由缩略词构成的特指某些地点的名词作为一个整体标注出来,如“常三关/ns”。

(2)机构名的分词和标注规则

标注的机构名主要包括以下两类:①一般性的中央行政机构名:包括宗人府、内阁、六部、都察院等机构名。②清代黄河管理机构及基层水利组织名:包括了“河 - 道 - 厅 - 汛 - 堡”五级管理机构下所包含的所有机构或组织名,均作为一个整体的机构名标注,如“祥符上汛十九堡/ng”。

(3)工程名的分词和标注规则

工程名主要是对大工和具有具体名称的另案工程进行标注,如“郑州大工/ne”“马营坝大工/ne”等,对于一般性的岁修仅作为普通名词进行标注。

(4)时间名词的标注规则

时间名词的类型主要分为以下几种(表2):①日历型时间:如“顺治五年三月三十日”,这类时间信息描述具体完整,将其作为一个整体标注为时间名词(nt)。②表示季节、节气、频率的时间类型:如“春”“夏”“秋”“冬”“霜降”“每年”“每岁”“历年”等统一标注为时间名词(nt)。③参照型时间:如“本年”“来年”“该年”“明岁”等统一标注为时间名词(nt)。④带有时间修饰成分“年底”“左右”等的时间类型,如“明岁二月底”等作为一个时间整体进行标注。⑤混合型时间:“频率 + 季节”“频率 + 节气”“参照时间 + 月份”等结合型的时间类型,如“每年春夏”“每年冬”“每年霜降”“来年二月”等,为保证其文本意义的完整性,将其作为一个整体标注,不予切分。

(5)其他类型词的分词和标注规则

一般情况下数词与量词的组合需要进行

表 2　时间名词标注示例

序号	时间类型	语料标注示例
1	日历型时间	乾隆五十七年二月十七日/nt 奉旨/v 户部/ni 驳/v 穆和兰/nh 酌/v 改/v 徵收/v 河工/n 帮价/n 章程/n
2	季节、节气、频率型时间	现/nt 已/d 霜降/nt 安澜/a,/wp 应/v 即/c 赶/v 购/v 明年/nt 岁料/n 以备/v 岁抢修/n 之/u 用/n
3	参照型时间	历年/nt 核奏/v 清单/n 后/nd 即/d 于/p 司库/n 找拨/v 还款/v,/wp 仍/d 循例/v 存贮/v 道库/n 以备/v 垫发/v 来年/nt 要工/n 之/u 用/n。
4	带有修饰性成分的时间	截止/v 明岁二月底/nt 上/nd 续收/v 银两/n 一并/d 尽数/n 拨用/v,/wp 以/c 济/v 工需/n
5	混合型时间	即/c 由/p 河工/n 之/u 银/n 由/p 布政司/ni 每年春夏/nt 二/m 季/n 照/v 额/n 于/p 司库/n 正项银两/n

切分,如“沁河长水二十二次”切分为“沁河/ns 长水/v 二十二/m 次/q”。当数词与所修饰的名词构成特指含义时,不进行切分,如“三省”“六部”皆有特指,不予切分。涉及到具体工程所用银两数时,为了便于研究,将银两数作为一个整体标注,不予切分。“余”“来”“多”等在数词之后表示约数或者余数时,与数字合并为一个分词单位,如“十七万余/m”。

由两个意义相近或相反的语素构成的特定含义的词作为一个整体切分,如“项款/n”“赏罚/n”“堵筑/v”“羡余/a”等。

如果官职名前连着机构名或地名的视为一个整体标注,不予切分,如“工部尚书/ng”“江南河道总督/ng”等。

否定副词与其他词连用,构成的常用词无需切分,如“不得不”“不实”“不免”等。介词或连词在句中出现且构成特定词时,不予切分,如“考成之法/n”。

由“名词 + 名词”或者“动词 + 动词”构成的具有特定意义的偏正式复合词,作为一个整体标注,如“捏报/v”“解役/n”等。语义上以朝廷政事、官员选拔等为主构成的特定词作为整体标注,如“捐监/v”“委署/v”等。

清代奏折中常用的代替标点断句的用语或者奏折中的专用词,作为一个整体标注,不予切分,如“等因”“等情”“钦此”“具题”“谨奏”等。

3. 清代黄河标注语料库的构建过程

如上文所述,国内外学者通过实验验证了 N-gram、隐马尔科夫模型、条件随机场方法在古文分词、词性标注以及人名、地名、时间名词的识别方面具有良好的效果。因此,文章主要采用以上几种方法进行语料库的构建,具体实现过程如下:

第一步:将语料数据导入 Python 中,导入“甲言”古文自然语言处理工具包,利用 N-gram和隐马尔科夫模型进行无监督的自动分词。通过分词结果发现,对于清代黄河管理相关的机构名、官职名的切分效果不是特别理想,主要源于系统词典中缺乏相关的词汇。

第二步:为了提高分词的准确度,我们通过查阅相关文献[33]自建了清代黄河治理相关的机构名、官职名的“自定义用户词表”,将该词表作为分词词表载入 Python 语言进行分词操作,得到准确度较高的分词结果,再对结果进行人工校对,把切分不准确的词按照正确形式添加到“自定义用户词表”,不断提高分词的准确度,最终获得分词语料。

第三步:在分词的基础之上,利用条件随机场(CRF)的序列标注方法,依据词性标注规则进行词性标注,进一步对标注结果进行人工校对,最终建成清代河工财务标注语料库。

四、清代黄河标注语料库的应用

本文试图在语料库技术的支持下,推动清代黄河问题研究,基于上文构建的清代“河道钱粮册”标注语料库分别从史料检索、量化分析及深层次知识发现 3 个角度,探索了语料库在清代河工财政问题研究中的

应用。

（一）基于标注语料库的检索和计量分析

1. 标注语料库在史料检索方面的价值

清代黄河标注语料库由于完成了分词、词性标注、人名等专有名词的标注工作，使得其除了能够实现传统数据库基于字符串匹配的全文检索功能之外，还可以支持实现基于词、词性、人名、地名、机构名、时间等检索方式，将检索单位从单字层面拓展到词汇层面，进而可以更好地满足研究者们多元的检索需求，同时也可以避免字符匹配的检索方式造成的冗余、缺漏、误配问题，提高检索的准确度。此外，由于标注语料库实现了分词和词性标注，就可以支持通过词的相似度计算，进而实现语义检索，为真正意义上的智能检索奠定基础。

2. 基于标注语料库的基础量化分析

基于构建的标注语料库可以轻松地实现词性、人名、地名等方面的信息挖掘和量化分析，可以为研究者提供研究线索，也可以辅助研究者尤其是初始研究者们快速掌握重点信息。在史料研究中，区分高低词频对于挖掘史料中的重点内容具有现实必要性[34]，基于标注语料库可以非常方便地区分高低词频，能够帮助研究者快速地捕获到高词频内容，深挖背后的内涵，从而提高研究效率。以人名为例，我们通过检索词性“nh”即可获得语料库中所有的人物信息，共计出现人名 222 次，涉及人物 118 人。通过对标注语料中的“人名－官职－时间”关联关系进行提取，获取了高频人物的信息，其中出现频次最高的前 7 位人物如表 3 所示。

对于标注语料库中实体或词的历时性分析是其重要价值之一，通过对语料中人名、地名、机构名等实体在不同历史时期所出现频次进行分析，可以探索各类实体随时间的变化趋势，以人物为例，基于上文提取的“人名－频次－时间”关联数据，利用 ECharts 绘制了人物出现频次的历时性分布如图 2 所示（频次≥4），可以直观地发现不同的历史时期与清代河工财务关系密切的人物。通过对

表 3　高频次人名表

人名	频次	官职	人名出现的时间
黎世序	14	江南河道总督	嘉庆十八年六月二十八日、嘉庆十八年六月二十八日、嘉庆十九年二月初三日、嘉庆二十三年二月十八日、嘉庆二十三年六月十三日、嘉庆二十四年六月初十日、嘉庆二十四年七月二十日、道光元年九月初三日、道光二年五月二十七日
严烺	10	河东河道总督	道光二年十月十八日、道光三年四月二十日、道光四年六月二十四日、道光四年八月二十二日、道光四年九月三十日、道光五年五月初九日、道光六年三月十八日、道光七年二月十三日、道光七年六月十八日、道光十六年六月二十七日
方受畴	9	河南巡抚	嘉庆二十一年二月初四日、嘉庆二十四年七月二十二日、道光二年十月十八日、道光四年六月二十四日、道光四年九月三十日、道光七年六月十八日、道光十六年六月二十七日
孙玉庭	7	两江总督	嘉庆二十三年二月十八日、嘉庆二十三年五月二十一日、嘉庆二十三年六月十三日、道光元年九月初三日、道光二年五月二十七日
程祖洛	5	河南抚臣	道光二年十月十八日、道光四年六月二十四日、道光四年八月二十二日、道光四年九月三十日、道光七年六月十八日
谭廷襄	5	河东河道总督	光绪元年四月初十日、光绪六年四月二十三日、光绪十六年六月初一日、光绪二十一年九月十三日
许振祎	5	河东河道总督	光绪十六年六月初一日、光绪十九年十一月十八日、光绪十九年十一月二十二日、光绪二十年十月初六日、光绪二十一年九月十三日

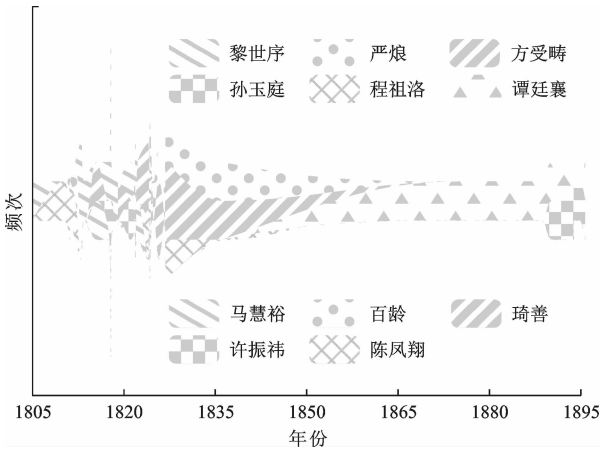

图 2 人物出现频次的历时性分布(出现频次≥4)

注:该图运用 ECharts 绘制,ECharts 是一款基于 JavaScript 的数据可视化图表库,通过编写代码提供直观、生动、可交互、可个性化定制的数据可视化图表。该图表类型为“主题河流图”,主要用来表示事件或主题等在一段时间内的变化,其中不同标识的条带状河流分支编码了不同的事件或主题,河流分支的宽度编码了原数据集中的数据值,原数据集中的时间属性,映射到单个时间轴上。

人名频次进行分析可以发现,像“方受畴”“谭廷襄”等对河道财政影响较大的官员不仅提及频次高,甚至在其去世多年之后仍被多次提及,而像“靳辅”“杨方兴”等清代非常重要的治河能臣却鲜少被提及,这从侧面反映了清代河务管理更多地关注财政,而对治河能臣的重视度不够。观察图 1 还可以发现嘉庆十一年(1806)至道光八年(1828)之间,呈现出人物数量多、频次高的特点,直接原因是这期间上奏请款的奏折数据较多。究其深层原因可能与这一时期“河患”日趋加重有关:一方面体现在黄河河患频发,据统计仅嘉庆八年至道光四年间(1803—1824 年)就发生了 10 次黄河大工①,再加上“岁修”、另案工程等,导致河道耗费巨大。另一方面体现在黄河河工带来的日趋沉重的财政负担,使得这一时期的河银制度发生了重大变化,原有的河工财政体系无法满足河工用银的需求,定额河银制度日趋崩溃。

(二)基于清代黄河标注语料库的深层次知识发现

基于标注语料库的细粒度词性划分,可以支持实现快速的知识整理和深层次的知识发现。以“河道钱粮册”标注语料为例,为快速了解清代朝廷拨款对黄河河工的影响,可以从语料中基于标注的时间、地点、机构、数词、量词等词性抽取出历年朝廷黄河河银的来源、去向、数量、时间的关联信息(表 4),其中来源包括地点和白银项目,这些指标数据的变化反映了清代河工银制度在不同时期的财政体制变化,有助于窥察财政制度的变化对黄河河工的影响。

表 4 河工数据示例

时间	来源	去向	数量
乾隆五年	两淮盐课(盐课银)	直隶山东河南	四十七万六千两
	江安浙江三省(地丁银)	直隶山东河南	二十二万余两
乾隆二十二年	淮安、泸墅、扬州(地丁银)	渠汊港等	二十余万两
乾隆五十五年	江宁藩库借支	砀山县王平庄	十八万七千六百八十余两
	司库耗羡银	六合、桃源、萧县、通州、泰兴、海州等	二千一百七十一两

通过对河银去向进行分析发现,康雍乾嘉时期河银去向管理比较模糊,河银的具体去向体现并不明确,只有一些笼统的地名或者工程名,并未形成明确的河银去向明细,也没有要求对超出额定的河工开支进行核算,对河银的管理缺乏严格的审核机制,直至道

① 该统计数据来源于殷继龙《清代黄河大工研究》,吉林大学 2022 年博士学位论文,第 226-237 页。

光十五年（1835）道光皇帝才下令要求对每年的超额银两进行独立核算，这一财务管理上的重大漏洞也是清代雍正、乾隆时期的河库道设计存在的缺陷之一。清初至乾隆期间河务相关的各项制度日趋完善，但为何没有形成规范的审计制度，这一问题有待深入探究。

为了深入分析白银项目来源的历时性变化，依据提取的“来源（白银项目）－时间”的关联数据绘制了白银项目来源的历时性变化（图3）。通过分析可以发现，长期以来地丁银一直是河工银的主要来源，但地丁银受河患的影响比较大，河工危机不解除难以收到地丁银，地丁银无法按时收缴反过来又会影响河工，从而形成恶性循环。为了打破这一格局，从数据来看最晚自乾隆初年盐课银也成为河银的来源之一，随着堤防加长、险工增多以及物价上涨等因素的影响河工开销剧增，原有的河银供给难以维持，乾隆中后期以加价为目的的帮价银开始成为了河工正项。乾隆末年虽然有帮价银的支持仍然无法满足河工用银需求，嘉庆中后期开始采用“生息”的做法试图弥补河工用银上的供需矛盾。至道光年间，河银的供给开始越来越依赖国家财政体系之外的名目，其中最重要的来源之一为捐纳。

本文仅以构建的“河道钱粮册”标注语料库为例，简单探讨了语料库在清代河工财政问题研究中的应用，随着语料数据的不断丰富、语料库规模的不断扩展将可以实现更多维度的应用，更好地辅助清代黄河问题研究。

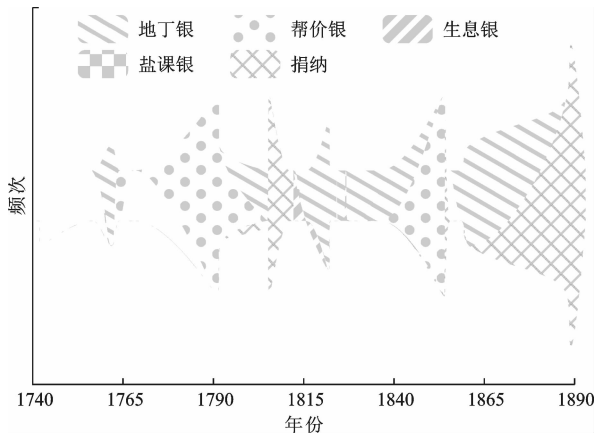

图3　河银白银项目来源出现频次的历时性分布

注：该图的绘制方法、类型等，同图2一致。

五、结语

语料库的分词、词性标注和命名实体识别等技术为清代黄河研究史料的组织、整理、检索和分析提供了重要基础，对今后清代黄河研究史料文本的处理和分析工作具有重要的意义。本文的贡献在于分析了语料库技术在清代黄河研究中的必要性，并探索出一套理论上可行的清代黄河标注语料库构建的技术规范和建设流程，并基于部分“河道钱粮册”语料进行了实证分析，为后续语料库在黄河问题研究中的应用提供了一个可行的思路。在未来的研究中，一方面可以将提出的语料库标注规范和流程应用于清代其他河流研究史料的语料库建设中，进而实现清代河流研究史料数据的标准化、规范化；另一方面，基于统一的标准规范可以实现不同河流数据库之间的跨库检索，便于实现不同流域河流治理之间的比较分析，同时也可以探索不同史料对同一事件记载的区别，加强研究的准确性、全面性。随着清代黄河档案整理

和信息化工作的快速推进和清代黄河研究数据基础设施建设的日趋完善,语料库在清代黄河研究中将会拥有越来越广阔的应用空间,将有可能为历史时期的河流研究带来全新的研究局面。

参考文献:

[1] 习近平. 在黄河流域生态保护和高质量发展座谈会上的讲话[EB/OL]. (2019-10-15)[2024-01-15]. http://www.xinhuanet.com/politics/leaders/2019-10/15/c_1125107042.htm.

[2] 潘威,夏翠娟,张光伟,等. 历史地理信息化与图情研究融合的必要性与可行性——以“数字历史黄河”为中心的考察[J]. 图书情报知识,2021(3):37,50.

[3] 潘威,白江涛,夏翠娟,等. 基于 TGIS 的专项历史地名库设计与搭建——以“数字历史黄河”地名库为例[J]. 数字人文研究,2022(1):13-24.

[4] 《清代河务档案》编写组. 清代河务档案[M]. 桂林:广西师范大学出版社,2022.

[5] 黄水清,王东波. 国内语料库研究综述[J]. 信息资源管理学报,2021(3):4-17.

[6] 林玉萍,龙红,李彪,等. 基于医学影像和病历文本的甲状腺多模态语料库构建与应用[J]. 西北大学学报(自然科学版),2021(2):198-206.

[7] 曾凡斌,陈荷. 基于谷歌图书语料库大数据的百年传播学发展研究[J]. 现代传播(中国传媒大学学报),2018(3):135-145.

[8] 宋鹏飞. 大气污染专题语料库构建与语料空间化方法研究[D]. 青岛:山东科技大学,2020.

[9] LIN B,YIP P,On the construction and application of a platform-based corpus in tourism translation teaching[J]. International journal of translation, interpretation, and applied linguistics,2020(2):30-41.

[10] 马海群,张涛. 文献信息视阈下面向智慧服务的语料库构建研究[J]. 情报理论与实践,2019(6):124-130.

[11] 付璐,李思,李明正,等. 以清代医籍为例探讨中医古籍分词规范标准[J]. 中华中医药杂志,2018(10):4700-4705.

[12] 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(5):727-733.

[13] 柯永红,江琛. 古代汉语词性标注语料库建设述评[J]. 语料库语言学,2021(1):97-111.

[14] 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.

[15] 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(2):183-193.

[16] 王晓玉,李斌. 基于 CRFs 和词典信息的中古汉语自动分词[J]. 数据分析与知识发现,2017(5):62-70.

[17] FU X,YUAN T,LI X,et al. Research on the method and system of word segmentation and postagging for ancient Chinese medicine literature[J],Bioinformatics and biomedicine,2019(1):2493-2498.

[18] 徐玉慧. 中文 N-gram 分词模型改进[D]. 天津:天津财经大学,2018.

[19] 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(6):44-54.

[20] 林立涛,王东波. 古籍文本挖掘技术综述[J]. 科技情报研究,2023(1):78-91.

[21] 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.

[22] 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(12):135-140.

[23] 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.

[24] 崔竞烽,郑德俊,王东波,等. 基于深度学习模型的菊花古典诗词命名实体识别[J]. 情报理论与实践,2020(11):150-155.

[25] 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(3):100-108.

[26] 余馨玲,常娥. 基于 DA-BERT-CRF 模型的古诗词地名自动识别研究——以金陵古诗词为例[J]. 图书馆杂志,2023(10):87-94.

[27] 刘浏. 古汉语典籍中的实体知识挖掘研究[D]. 南京:南京大学,2018.

[28] 王一钒,李博,史话,等. 古汉语实体关系联合抽取的标注方法[J]. 数据分析与知识发现,2021(9):63-74.

[29] 潘威,岳佳雲. 关于数字人文进入清代河流研究的若干想法[J]. 史学月刊,2023(1):116-121.

[30] 潘威,张丽洁,张通. 清代黄河河工银制度史研究[M]. 北京:中国社会科学出版社,2020.

[31] 邓三鸿,胡昊天,王昊,等. 古文自动处理研究现状与新时代发展趋势展望[J]. 科技情报研究,2021(1):1-20.

[32] 刘石. 文献学的数字化转向[J]. 文学遗产,2022(6):10-13.

[33] 贾国静. 水之政治:清代黄河治理的制度史考察[M]. 北京:中国社会科学出版社,2019.

[34] 林伟杰,杨阳,文玉锋,等. 古籍知识组织中的知识计算:理论特性与基础指标[J]. 图书与情报,2022(5):24-30.

(责任编辑:杨海挺)