DOI:10.13530/j.cnki.jlis.160011

面向数字人文研究的大规模古籍文本可视化分析与挖掘

欧阳剑

摘　要　传统的古籍开发与应用模式已难以适应人文学科研究的需要,人文学科研究者期待一个技术逻辑和人文逻辑相耦合的数字人文研究范式的出现。本文从古籍文献深层次开发与利用出发,利用新的信息技术与面向数字人文研究跨学科方法,以大规模中国古籍文本为研究对象,采用大数据研究理念,对古籍进行整理、标注、自动分词等处理,以词频分析统计为研究核心,采用数据降噪、基于窗口时间单位的统计分析计算、滑动窗口预测等分析与挖掘方法,采用大数据实时分析技术,实现了实时、在线、立体、可视化、定量分析字词的历史词频分布规律,创建了一个以语言学、历史文献学、历史地理学等人文学科研究为主的古籍实时统计分析平台,可辅助研究者在大量的古籍文献中发现新的模式、现象、趋势等,实现古籍开发与应用模式创新的初步尝试。图 11。参考文献 36。

关键词　数字人文　文本可视化　数据挖掘　古籍文献

分类号　G250

Visual Analysis and Exploration of Ancient Texts for Digital Humanities Research

OUYANG Jian

ABSTRACT

Digital humanity, a new research pattern, brings consequently a new way of research for traditional humanity and social sciences for traditional development and utilization mode of the ancient literature resources that no longer fit the requirements of humanity researches. This paper aims at the deep development and utilization of ancient literature resources by using new information technology and method of digital humanity with the ancient Chinese literatures as to construct a new platform for real-time textual statistic analysis of linguistics, studies of historical literature and historical geography etc.

This study adopts a big data concept, and applies sorting and labelling to Chinese ancient texts for the construction of a corpus of more than 40 000 kinds of ancient texts. This study also adopts means of dictionary superposition of piecewise and Bigram model to carry out word segmentation of Chinese ancient texts and also with the application of Grubbs method for data denoising and the maximum elimination of problematic data. With word frequency statistical analysis as the research focus base on ancient corpus, we use time window unit analytical computing to analyze the word frequency, apply the idea of memory real-

通信作者:欧阳剑,Email:oyjjj@163.com,ORCID:0000-0001-5867-2852(Correspondence should be addressed to OUYANG Jian, Email: oyjjj@163.com, ORCID: 0000-0001-5867-2852)

time computing to solve the bottleneck problem of reading big data. The results of the statistics and analysis are displayed by the micro-level scatter plot and the macro-level curve graph based on the time axis as the main line. With the author of the ancient books as the main line, we use the geographic information system (GIS) technology to integrate and display digital ancient books, and with the retrieval of the ancient literature as a clue to show the geographical distribution of the authors. This study improves the efficiency of real-time inquiry and realizes the visualization of the scatter diagram and curve graph of the word frequency according to the years. A statistical and analytical platform of ancient literatures and documents in linguistics, history and historical geography will be established based on the new methods and pattern.

The study not only extends the research paradigm and method of the humanities, but also enriches the research tools of the humanities research. This research broadens the dimension of the utilization and development of ancient literature and texts, and expands the scope of humanities materials. The platform has a vast application prospect in linguistics, history and historical geography.

This research is a new attempt in the deep development and utilization of ancient texts and documents by means of digital humanity within the scope of big data. First of all, this study builds a large-scale ancient text corpus of more than 40 000 kinds of ancient books; secondly, this study uses statistical methods and superposition of word segmentation method to implement word segmentation in ancient texts; finally, with the help of big data technique, this study improves the efficiency of real-time inquiry and realizes the visualization of the scatter diagram and curve graph of the word frequency according to the years, which provides a direct visual display of the result of the analysis.

Due to the insufficient vocabulary database, the accuracy of word segmentation needs to be improved; in addition, in order to improve the quality of the corpus, the information of edition of ancient books and the authors also requires verification. The extraction of the entity in corpus of ancient books, such as persons, historical events, places, titles and names needs to be developed further. 11 figs. 36 refs.

KEY WORDS

Digital humanities. Text visualization. Data mining. Ancient literature.

0 前言

数字人文(又称人文计算,Humanities Computing)是一个将现代计算机和网络技术深入应用于传统的人文研究的新型跨学科研究领域,数字人文给传统的人文社科研究提供了新的研究方法和研究范式[1]。文本内容挖掘及可视化分析已成为数字人文研究的重要手段。文本挖掘技术是数据挖掘技术的拓展与延伸,与传统结构化的数值数据挖掘不同,文本挖掘是以非结构化的文本数据为挖掘对象,文本挖掘能够实现从海量的非结构性文本中发现新的模式、规则、趋势等[2],为用户非结构化的文本挖掘与分析研究带来便利。新信息环境下,庞大的信息量使人们处理和理解信息的难度大增,传统的文本分析技术难以满足人们现阶段信息浏览及筛选的需要,而可视化技术可以将文本中复杂的或难以表达的内容以视觉符号的形式表达出来,为人们提供一种理解海量复杂文本的内容、结构和内在规律等信息的有效手段[3],使人类视觉认知、关联、推理的能力得到充分发挥,可视化技术已成为帮助人们理解非结构化文本和发现其中所隐含知识的新方法与有效途径。文本内容挖掘及可视化分析目前在很多人文学科领域得到了广泛应用。

1 国内外面向数字人文的文本挖掘与可视化分析研究

可视化分析与文本挖掘的结合为探究文本中所蕴含的新知识和复杂的结构模式分析提供了一种有效的方法,可视化分析及文本内容挖掘分析在人文学科的数字人文研究中越来越受到国外学者的重视。2011 年,谷歌与哈佛大学的研究人员共同开发的一套数据库,可以对 1800 年至 2000 年出版的近 520 万本书籍的单词和短语的使用频率进行统计,从而帮助了解文化和语言的变迁过程,并为人文学科的研究提供新方法[4]。2008—2013 年由德国联邦教育及研究部(BMBF)资助并由德国莱比锡大学古代历史系承担的数字人文项目 eAQUA,是在古典文学研究领域的一次尝试,项目通过将计算机科学与古代研究知识相结合,从古代研究的需求及要求出发,通过文本挖掘技术从古典文献资料中抽取特定领域知识[5],并通过开放获取的方式向研究者免费提供这些知识。2010 年,受美国国家人文基金(NEH)资助,北德克萨斯大学与斯坦福大学合作开展了一个历史领域文本挖掘与可视化应用项目,以记载美国历史的旧报纸数据库中的 23 万页数字化新闻文本为样本,抽取在历史学研究中出现的相应人名、地名等特定信息,并把这些特定信息以可视化形式呈现,辅助研究人员发现历史事件随时间和空间的演变过程及变化规律[6]。2014 年 8 月,Schich 等通过获取公元前 600 年到公元 2012 年间150 000名不同领域的历史杰出人物的出生和死亡地点数据,描绘了这些著名人物的迁徙模式,通过网络和复杂性理论的工具,识别特征模式,确定文化和历史的关联,使用大规模可视化和定量工具从宏观的角度绘制了三千年欧洲和北美的文化史图,获得文化中心发展的历史趋势[7]。Cho 等人开发的罗马历史可视分析系统(VAiRome),是一个集时空分析与文本分析为一体的可视分析系统,运用文本分析技术和多种直观的可视化视图,向学者展示了罗马的历史,揭示了其中重要的时间、地点、事件以及它们之间的关系[8],为学者分析研究罗马历史提供了极大便利。

可见,文本挖掘与可视化方法在人文社会科学研究中的应用虽然才刚刚起步,但已显示出广阔的应用前景,为人文社会科学研究带来了新范式与方法,文本挖掘与可视化方法也为古籍深层次利用与开发带来了新的模式。

2 大规模古籍文本可视化分析与挖掘背景及思路

2.1 古籍文本开发利用现状

目前,我国古籍的数字化已经比较成熟,而且也具有一定的规模,以《文渊阁四库全书》《四部丛刊》《中国基本古籍库》《国学宝典》《中国数字方志库》等为代表的古籍数字化文本大量出现,为人们利用古籍提供了极大的便利。但现有的古籍数据库普遍只有检索功能,而统计、分析功能较少,目前的古籍利用普遍以检索浏览为主,只能按原始资料的结构进行浏览,不能将原始的资料信息根据自己的研究需要进行自动重组,也不能深度挖掘潜藏的信息,利用方式还处于比较原始的纸质替代状态,且只有时间的序列性展示,缺少空间的序列性展示[9-11]。国内对古籍文本的利用与开发研究基本上还是采用传统的研究方法与模式,缺乏新的人文研究范式及方法,导致当前规模庞大的古籍文本数据与较低的古籍深度利用率之间的矛盾比较突出。

随着 E-考据及数字人文等新理念的出现,传统古籍利用与开发模式的局限性越来越明显,更多的学者已开始认识到古籍数字化带来的不仅仅是庞大的古籍存储,“数字化”为技术与人文的合流构筑了新平台,可通过技术逻辑和人文逻辑相耦合的“数字人文”的研究,构建可持续完善和丰富的数据集和分析工具,充分利用新的信息技术与跨学科方法对古籍进行深

层次的分析与挖掘。

2.2 大规模古籍文本可视化分析与挖掘思路

埃雷兹·艾登等在《可视化未来 数据透视下的人文大趋势》专著中以“谷歌图书”项目为背景，通过500多万本电子书讲述了大数据在研究历史文化、人类语言、社会名望、群体记忆等方面的重要作用，大数据对社会科学的变革意义凸显，体现了科学与人文之美[12]。大数据时代的各种思潮和视角在不断涌现，大数据作为一种全新的数字化研究资料，与传统资料相比，其样本量具有庞大、丰富及时间跨度大等特点，为社会科学经典理论的验证和拓展提供了更大的研究空间[13]。传统人文学科的实证研究强调在理论的前提下建立假设，大数据时代重在发现知识与现象，在没有理论假设的前提下，从海量的数据中发现隐藏在数据中的模式、知识和趋势，从而帮助人们揭示事物现象与发展规律[14]，大规模的古籍文本扩大了人文学科资料的范围，提供了人文学科研究新的研究空间，拓宽了古籍利用的研究领域。

数字人文的出现促进了人文学科与技术的融合，为古籍数字化的深度开发与利用提供了新的理念与独特的创造性思维，将古籍文本集成化、可视化，使古籍利用融资料查询、计量分析、知识发现等功能为一体，使得数字化古籍深度开发与利用成为可能。古籍资源涵盖面广，包含历史、地理、文化以及社会等诸多方面，因此，数字人文研究中的许多前沿实践都能为古籍文本深度开发利用所借鉴，运用人文计算、可视化分析及文本挖掘等方法，对大规模古籍文本进行可视化分析与挖掘，为语言学、历史文献学、历史地理学等人文学科研究探索新的研究范式与方法，在此基础上，为研究者提供一套方法较为科学、客观的分析工具与平台，挖掘古籍在传统人文学科研究中的新空间和新的增长点。

大规模古籍文本可视化分析与挖掘以古籍文本为基础，采用大数据的研究理念，通过大数据实时分析技术，以词频分析为手段，采用数据降噪、基于窗口时间单位的统计分析计算、滑动窗口预测等分析与挖掘算法，定量分析字词的历史词频分布规律，对古籍文本中的人物、历史事件、地名、官职、称谓等实体进行抽取及关系的建立，呈现每个时期古籍文献共时性的空间分布与变化，从时空二维立体地展示语言、文化、历史等的发展变化。

3 大规模古籍可视化分析与挖掘研究

3.1 数据来源及处理

数据已成为数字人文的基础和核心，古籍文本的收集、整理是本研究的基础，本研究古籍文本主要有网络数据采集与专业数据库文本获取两种。目前，网络上分布着大量公开的古籍文本。为了收集到尽可能多的古籍文本，本研究通过设计一些爬虫软件对特定的网站或专业数据库进行数据采集，对于离线版的专业数据库则通过其他技术手段进行提取。目前，本研究收集、整理了41 563种古籍（总计48.35亿字），时间跨度上从上古到民国，文献分布比例为：秦及以前0.69%，汉1.85%，三国、晋2.71%，南北朝、隋、初唐7.59%，唐中至五代十国1.50%，北宋27.80%，南宋1.16%，元13.22%，明17.43%，清20.41%，民国2.11%，未进行作者及文献朝代核对的3.53%。从种类上来看包括经、史、子、集等文献，其中经部占6.37%，史部占25.43%，子部占31.32%，集部占28.46%，未进行分类的文献占8.42%，形成了一个比较综合、全面的古籍语料库。

数据的规范及一致性是分析及统计准确性的重要前提。由于采集的数据格式包含PDF、WORD、HTML等多种形式，为了研究的需要，通过数据抽取的方式提取其中的文本，并且把UTF-8、Unicode、UTF-16BE、GBK等编码转换成统一的Unicode码；同时，采用厦门大学、教育部语言文字应用研究所、北京师范大学联合开发的“汉字简繁文本智能转换系统”进行简繁异体

字转换,形成统一的简体字[15]。

词频分析是文本挖掘中的一种重要研究方式,也是文本可视化的一种重要模式,Google 实验室推出的 Books Ngram Viewer 就是以词频分析研究为基础[16]。词汇是古代汉语研究中的重要内容,本研究对古籍文本内容进行可视化分析与挖掘主要是通过词频来进行,因此,分词是古籍文本分析与挖掘的前提。古代汉语的词汇有一个长期的从单音节词向双音词及多音节词演化的过程[17]。古代汉语在词汇和语法等诸多方面与现代汉语不同,尽管现代中文信息处理技术已经在很多领域取得了快速发展,然而这些研究成果主要针对现代汉语,现代汉语已有的分词研究成果并不能完全照搬到古代汉语分词领域中。与现代汉语分词相比,古籍分词所需要的词库与训练语料及语法规则基本空白,因此,词库与训练语料构建是分词的关键与基础,词库建设更是核心。

王力[18]、史存直[19]、潘允中[20]等对古代汉语词汇的发展过程及特点做了比较深入的研究,基本反映了古代汉语词汇发展的总体面貌与特点,从中可知对古籍文本的处理必须遵循古代汉语词汇的发展特点。在缺乏一定规模的断代词典的情况下,大规模地对不同朝代的古籍文本进行分词,准确性难以保证。采用分朝代、分词汇表的方式切分才符合古代汉语词汇的发展规律,即切分不同朝代的古籍文本语料时采用相应朝代的词汇表,这是本研究进行古籍文本语料切分的主要思想,这样可以最大程度保证古籍分词的准确率。

古代汉语的发展是一个渐变的过程,无泾渭分明的分水岭,就汉语史分期而言,学者所持立场不同。日本著名汉学家太田辰夫把汉语史的发展分为上古、中古、近古、近代和现代五期[21],王力从汉语语法、语音变化角度出发,在《汉语史稿》中提出了古代汉语分上古、中古、近代和现代四阶段说[22]。具体到各个阶段的上下限,学者们众说纷纭,目前没有确切的意见[23-26]。为了系统处理方便并结合相关专家建议,笔者根据古代汉语词汇发展的特点、断代词典及词汇专书的研究现状,将古籍文本切分为四个时间段,如图 1 所示。考虑到词汇使用的连续性,词库构建采用分段叠加的方式进行。所谓分段叠加是指后一个词库在前一个词库的基础上通过添加当前朝代的新词汇的方式进行累加,分词时按古籍文本的年代分别调用相应词库,比如:当待分文献属于战国时,则采用词库 1 来分词,当文献属于元朝时则采用词库 3 来分词。

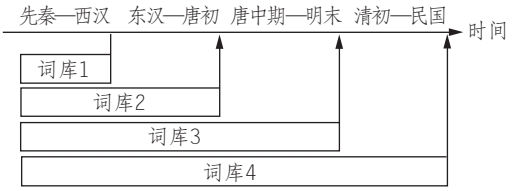

图 1 词库分段叠加的构建方法

目前,对古籍分词的研究还主要集中在就某个特定时间段的分词[27-28],在缺乏一定规模的断代词典的情况下去进行分词难以保证准确性。古代汉语词汇史研究一直受到古代汉语研究者的关注,多年来已出版大量古代汉语词汇词典及专书,包含大量的古籍词汇。因此,古代汉语词典及专书是古籍分词词库建设的重要来源,不但数量多,而且古代汉语词典及专书的词汇质量有保障,通过整理已有的 20 余种断代词典及专书语言词典,可快速建成相应的四个断代基础词典库。为了便于计算机自动提取词汇条目,所选词典及专书从内容编排上来说,其结构基本上属于半结构化数据,且大部分只需提取词汇索引部分即可,极大加快了词汇表的构建速度,对那些散落在其他古代汉语词汇词典与专书中的词汇后续将逐步采取人工方式加入。除从文献中抽取词汇外,利用字词间互信息、共现度、凝固度相结合的方法从分段的语料中提取词汇,从提取效果来看,分段文献文本数量越多,提取效果越明显。

实体名词的识别一直是中文信息处理的难题,而且古籍中的实体名词数量及类型众多,因此,除了不同时期的核心词库,我们还建有适用

于所有古籍的文献题名词表、历代人物词表(主要来源于中国历代人物传记资料库)、称谓词表、官职词表、年号词表、地名词表及古代虚词表,实体名词词表的建立提高了分词的准确率。

目前,针对古籍分词主要采用两种分词方法:一是基于字符串匹配,一是基于统计及机器学习的分词。在古籍分词中条件随机场模型(CRF)取得了比较好的效果[29-30],但在实施中由于标注人工训练模板需要花费大量人力与时间,所建训练模板规模对分词有直接影响,所以本研究自动从《汉语大词典》(去掉现代汉语部分例句)等词典中提取例句中的词汇二元语法(Bigram)特征。因此,对古籍分词以 Bigram 模型为主,CRF 模型为辅,CRF 模型对古籍分词中的新词提取效果更佳。采用以上分词方法,以《左传》为例,准确率达到 87.3%。随着词汇及特征模型的完善,分词准确率会得到一定的提高。

3.2 系统平台建设与开发

3.2.1 实时统计分析

为了满足用户任意字符检索及统计的需要,在统计分析时需要对所有全文数据进行查询。采用传统建索引的处理方法对 4 万余种古籍的文本进行统计分析,不但耗时,而且由于分词原因容易出现漏检、漏统计。传统的处理方法所遇到的瓶颈主要在读取数据与快速统计分析上,而大数据分析处理技术使快速的实时统计分析成为可能。

首先,大数据处理的内存实时计算思想可以很好地解决读取数据的瓶颈问题。在系统平台开发的过程中,通过设计一个内存数据存储器的方式,在系统启动时把所有数据加载到内存上,每次进行统计分析时直接从内存数据存储器中提取数据,这样就避免了频繁地读取硬盘所带来的长时间等待问题。出于实时统计分析的需要,内存数据存储器需要支持高并发、高吞吐量的线程安全等特性,而 ConcurrentHashMap 是 Java 5 中支持高并发、高吞吐量的线程安全 HashMap,具有很好的读取并发性能,由于大多数情况下读取 HashMap 时都没有用到锁定,因此读取操作几乎是完全的并发操作。

在实时分析方面,MapReduce 的大数据处理思想是目前处理大数据的一种常用方法,主要用于大规模数据集的并行运算。词频统计是 MapReduce 经典的功能演示案列,最能体现其思想,它能够统计一系列文本文件中每个单词出现的次数,但不能统计每个词在不同文本中出现的频次,因此需要对 MapReduce 功能进行改进,通过对文本分组进行并行统计,然后对结果进行汇总即可解决此问题,本研究处理的整个流程如图 2 所示。通过以上处理,4 万余种古籍的文本在 16G 内存、CPU 为 Intel Core i7 的 PC 机器上实现了 1 秒内实时统计分析。

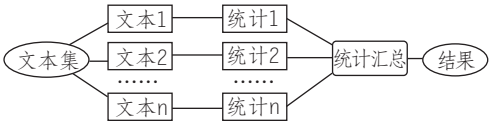

图 2 文本实时分析框架

3.2.2 数据可视化分析及挖掘

分析与挖掘与实时分析统计紧密关联,实时统计是分析与挖掘的基础。以词条及词频时间分布为视角的古籍利用分宏观层次与微观层次,宏观层次用于宏观趋势、规律的呈现,微观层次则用于具体信息考证及例证。

(1)以时间轴为主线的微观散点图分析

词语首见考证的可视化的目的在于帮助研究者快速从图像上判定词条的首见年代,是一种表示型的可视化,需要简单明了地表述词语的首见年代,而散点图(Scatter Chart)则是比较适合其可视化的表现形式。以系统实时统计分析结果为基础,通过散点图获得各个时期词语的词频信息(见图 3),从而反映汉语词汇系统的量化发展过程。词语首见年代能比较直观地反映汉语词汇的具体产生时间,因此,对历史词汇学的研究具有极其重要的价值。以词条首见年代为视角,以可视化的形式展现字词的年代及词

频分布,力求量化地呈现不同时期的词汇分布趋势,便于学者从微观视角进行词义考释,考察词义演变以及词语、词义产生年代的信息分布及发现规律。散点图分析支持用户交互,可选择一段时间区间内缩放。

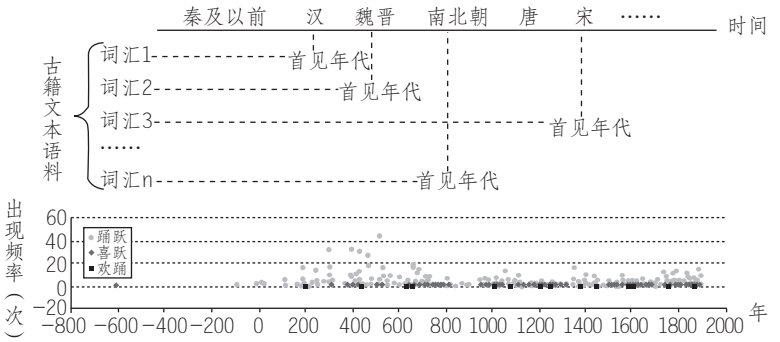

图 3　词汇词频时间轴的微观分布

(2)以时间轴为主线的宏观曲线分析

以时间轴为主线的微观研究注重于具体的某本古籍的词频及内容研究,虽然提供了散点的密度分布,但无法从定量上进行更准确的宏观描述,无法直观地描述事物的整体发展规律、趋势及模式。宏观研究是在微观研究的基础之上,对大量的微观数据进行定量计算、统计归纳、聚合,进行大量的微观数据的整体展示,反映的是宏观层次的表现。古籍的词频在宏观层面的构成蕴含着极为丰富的社会、经济、文化、历史等信息,反映当时社会历史、经济文化等发展规律与变迁。古籍的词频在宏观层面的描述以前面的词频信息为基础,采用数据降噪、基于窗口时间单位的统计分析计算、滑动窗口预测等分析与挖掘算法定量分析字词的历史词频分布规律。

本研究的总体思想是采用大数据的理念,在大量的样本数据中少量的噪点数据可以忽略,但为了更好的效果,去噪点数据是必要的。由于存在原始古籍文本分词准确性的限制,为了避免一些佛经之类古籍中特定俗语或固定用语词频偏高的影响,在分析过程中需要对词频进行数据降噪,过滤掉那些高出正常值的数据,在过滤的过程中采用单位时间窗口滑动技术(如图 4),即以朝代为单位,在每个朝代中采用格拉布斯法(Grubbs)来做异常值判断。根据格拉布斯法的定义可知,在待分析处理的数据集中,如果某些个别数据的值与待分析的数据平均值相差太大,那么就可以认为这些数据是存在问题的,即他们是“可疑值”,根据实际的需要,需要去掉那些高频率的“可疑值”,使其不参与平均值的计算。因为样本过少将会影响分析的结果,为了提高分析的可靠性,在数据分析的过程中只考虑那些单位时间窗口内样本数据大于 10 的数据。

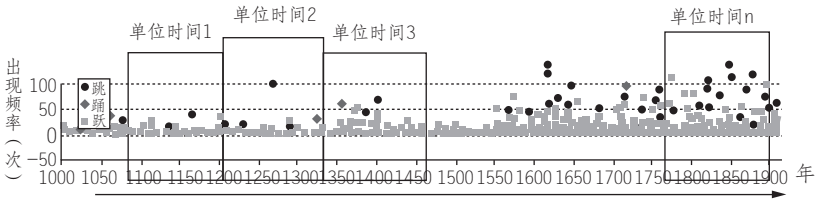

图 4　单位时间窗口滑动分析方法示意

词频分析是文本分析及内容挖掘的重要方法之一，通过词频能挖掘出一些语言现象、历史事实、风俗面貌、社会文化及地理等信息，单位词频越高说明语言表现得越明显，历史事实越突出，风俗越盛行等，从而可以更清晰地发现一些规则、趋势。古汉语的词汇演变过程具有连续性及持久性的特点，即词汇的变化过程在一定的时间范围内（或某个朝代内）是稳定的，因此，与数据去噪处理相似，采用单位时间窗口滑动技术对词频分布进行分析，其中单位时间内的词频计算与 Google Books Ngram Viewer 的算法相似，具体如下：

$$单位时间内词频=\frac{单位时间该词总数}{单位时间总词数}$$

将数据库中每一个数据项作为单个图元（Primitive）元素表示是数据可视化技术的基本思想，即通过大量的数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示并加以展现，从而可以通过图形的方式从不同的维度观察数据，以便对数据进行更深入的观察和分析。本系统采用 D3 JS、Highcharts JS 等可视化工具，因此，系统既提供微观数据的浏览与查证，也提供宏观的数据可视化分析，微观数据可视化采用时间轴的方式，显示词条在所有文献中的分布，方便研究者溯源及比较，而宏观分析结果则通过曲线时间轴来展示，通过大量的数据的统计计算，以曲线的变化来反映字词频率在不同历史时期的变化，系统提供多个字词的比较分析功能，辅助研究者的对比研究（见图 5）。

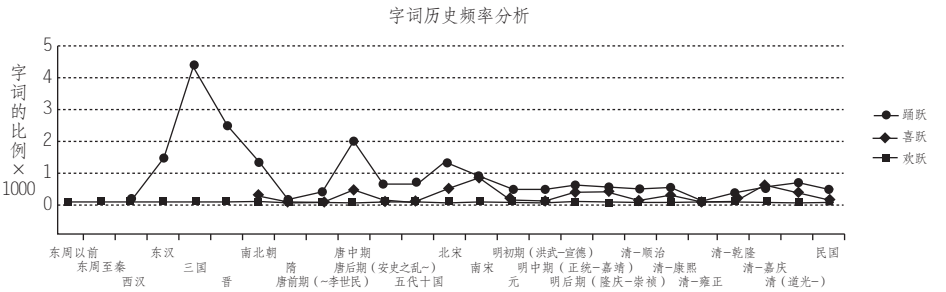

图 5　以时间轴为主线的宏观曲线分析

3.2.3　空间信息可视化分析

空间信息可视化分析主要是通过地理信息系统（Geographic Information System，GIS）来处理、组织和显示古籍分析结果，将古籍的空间数据和属性数据结合在一起，将地图独特的视觉化效果、地理分析功能和层信息的划分方式在时间和空间两方面直观展示，本研究中的主要地理信息来自于哈佛大学的中国历代人物传记资料库（CBDB）[31]与复旦大学的中国历史地理信息系统（CHGIS）[32]，中国历代人物传记资料库（CBDB）中有丰富的历代人物信息，包括官职、籍贯（包括 GIS 信息）、生卒年、人物关系等，缺失的部分地理数据则采用人工方式通过百度地图采集。

空间信息可视化以古籍作者为主线，利用地理信息系统（GIS）技术，将我国庞大的、静态的、分散的数字化古籍进行大规模的集成和地图展示，以古籍文献的检索为线索在地图上呈现相关作者的地理分布，方便研究人员分析语言现象、历史事实、风俗面貌、社会文化及地理等的分布，可帮助研究者在大量的古籍中归纳一些语言、文化及历史事件等在地理上的分布特点，有助于学者探究语言、文化、历史和地理环境的互动，以深度开发中国古籍中的多元文化价值，适应学术研究和教学的深度需求。

在数字人文研究中，古籍是语言学方言研究及文学地理研究的重要研究对象，作者作品中所反映的方言现象具有极强的地域性，而文

学地理则是探讨文学和人文地理空间的关系[33]，文学地理作为一种新兴的跨学科研究，“场景还原”和“版图复原”是文学立体图景研究的重要内容[34]。通过古籍作者与中国历代人物传记资料库(CBDB)中的人物信息关联，中国历代人物传记资料库(CBDB)与复旦大学的中国历史地理信息系统(CHGIS)中的地理信息关联，即可通过中国历代人物传记资料库(CBDB)中的人物籍贯GIS信息构建一幅以古籍检索为主线，跨时间维度与空间维度的中国古籍立体图景(见图6)，并提供基于统计的时空聚合分析(见图7)，实现时间和空间上的交互式缩放分析，历史、文化地理学者利用空间信息可视化可发现历史过程及文化发展中的隐含信息。

图6　古籍时间维度与空间维度立体图

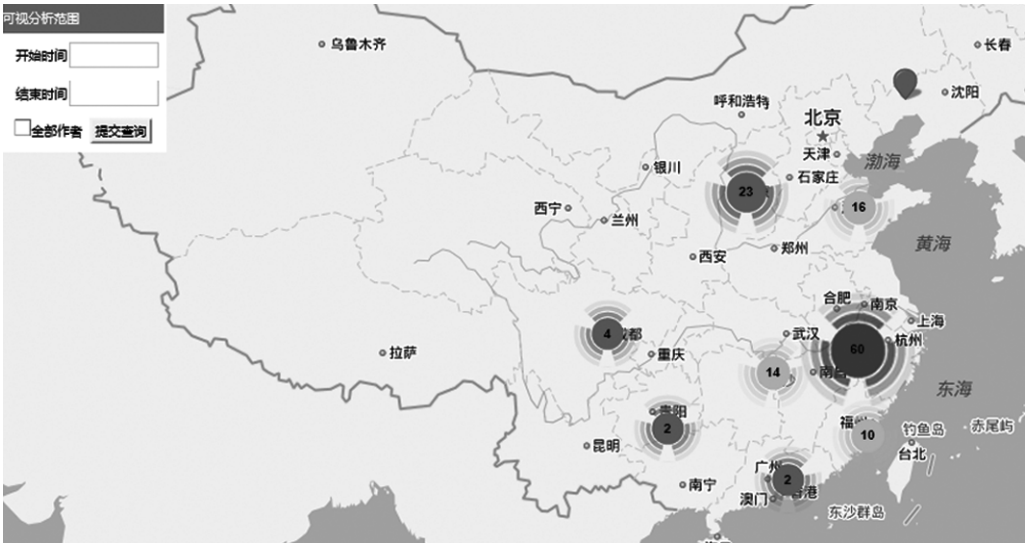

图7　古籍作者空间聚合分析

3.3 平台使用场景及实际使用效果

本平台以古籍的深度开发与利用为出发点，探索古籍开发与应用新模式，引入大规模定量计算分析方法，为古籍在语言学、历史文献学、历史地理学等人文学科研究中的应用探索新的研究范式与方法，平台可用于多种人文学科研究领域的定量分析与规律发现。

3.3.1 语言学词语首见年代考证

在语言学词汇与语法研究中，字词及语法现象的首见年代查找与发现极其重要，但是传统研究方式却费时费力，资料总量的快速增加给研究者带来了巨大的挑战，庞大的资料超过学者一般的阅读、分析和理解所能处理的范畴，本系统给研究者提供了语言学研究新的研究方法，有助于他们发现利用历史资料。

本文在研究过程中形成了一个庞大的古籍语料库，目前也是国内外收藏种类较全、数量较多的古籍文本语料库，系统能以可视化的形式从微观与宏观角度展现字词的年代分布及词频分布，可辅助研究者进行词义考释，考察词义演变，词语、词义产生年代的信息分布，发现规律。浙江大学汉语史研究中心真大成博士 2014 年 5 月在上海师范大学语言所做题为“文献考误与汉语词汇史研究”的学术报告，报告中认为，东晋时代“赶”字尚未产生，至晚唐五代方才出现。通过本系统的可视化分析可快速发现“赶”这个字首次大量涌现的年代是在唐中期（见图 8），但这是一个估算的大概首见年代，如果需详细查证字词具体首见文献，则可通过微观散点图快速查看文献全文并进行字词定位分析与判断（见图 9），最早“赶”字出现在窥基的《金刚般若经赞述》里：“如赶远质以临台睇影颜而府己也”，比真大成博士所得结论有所提前，反映出大数据给研究带来的巨大优势，对于词条首见考证来说无疑加快了研究的步伐。

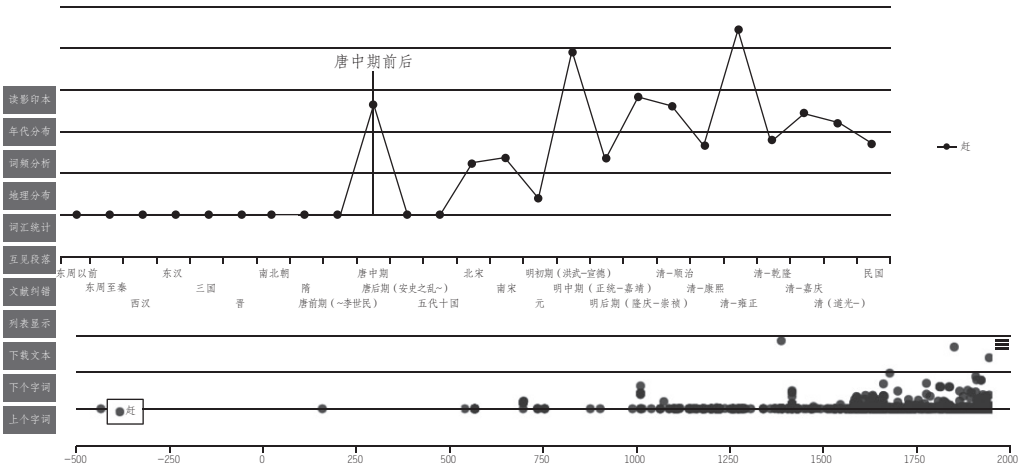

图 8 “赶”字产生考证

3.3.2 史学研究中的定量分析

在历史分析方面，计量史学是历史学研究的一种重要方法，运用自然科学中数学与统计方法对历史资料进行定量分析，使史学趋于精确[35]，计量史学在经济史、政治史、社会史、人口史等领域研究中取得了很多的研究成果，发挥了重大作用。

大规模古籍文本具有时空跨度大、材料面广的特点，很大程度上可以避免资料选择时的疏漏与偏废，弥补史学家惯用的“选精”与“集粹”研究方法之缺陷。古籍文本经过可视化定量分析后，非常容易有一些“不期而遇”的发现

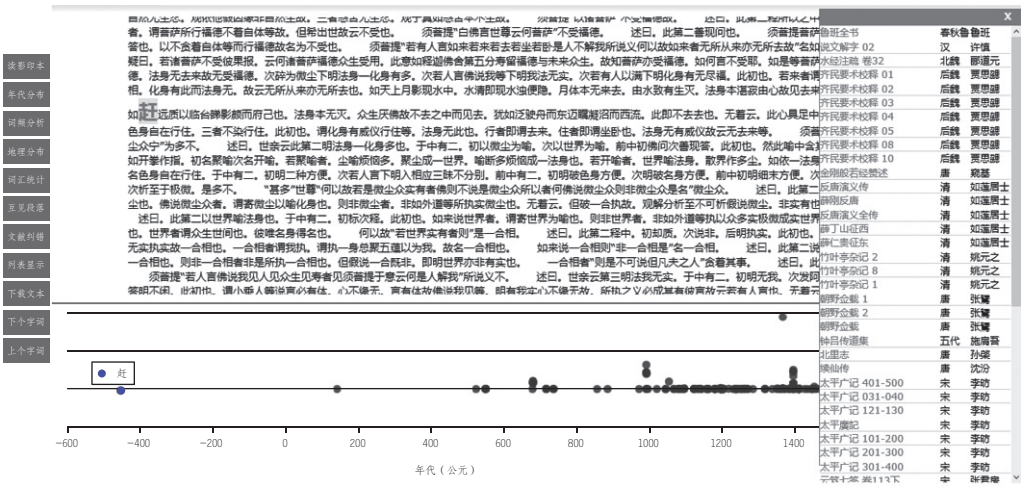

图 9　文献全文中的“赶”字浏览

及未知现象，而且很多发现还与我们以往的预设和已知有很大不同，以大规模数据为基础的量化研究还能较好地纠正研究的主观性，实现研究从常见的理论或问题驱动转向数据或经验驱动。基于大规模古籍文本，系统对一些历史事件与现象进行定量分析，能很快完成传统研究方法所不能完成的工作，使结论更具说服力。

中国传统观念一直受“重学轻术”这一思想的影响[36]。学、术在我国古代分别具有不同含义，按《汉语大词典》《康熙字典》等的解释，“学”基本是指钻研知识、获得知识、掌握知识，而“术”则多指技艺、方法。可见，学术在古代国人的知识体系中是一分为二的，前者为白领的学问，后者为蓝领的技术，学是儒道之经，术是“奇技淫巧”。从图 10 中也基本上可见“学”字的频率在中国古代文献中要远高于同期“术”字的频率，这在一定程度上反映了古代重学轻术的传统对仕人思想的影响。

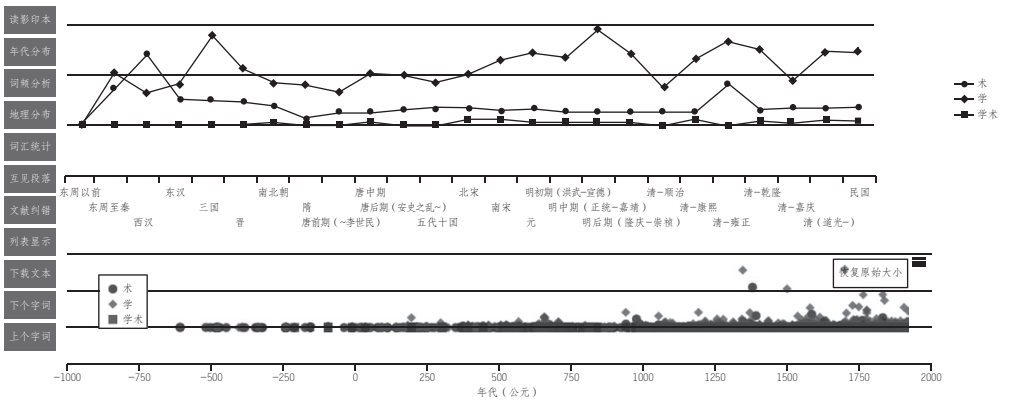

图 10　历史文献中“学”“术”“学术”的变化

宋朝时期的战史研究是宋辽、宋金关系中的一个突出问题，因此，主战与求和成为宋朝两派争议的焦点，但现阶段很难从文献角度去量化当时战与和的舆情，从已有的古籍中学者只能定性得出宋朝时议和为主，通过该系统对宋朝时期的文献中“战”“和”的分析可以定量得出议和为主的结论（见图 11）。

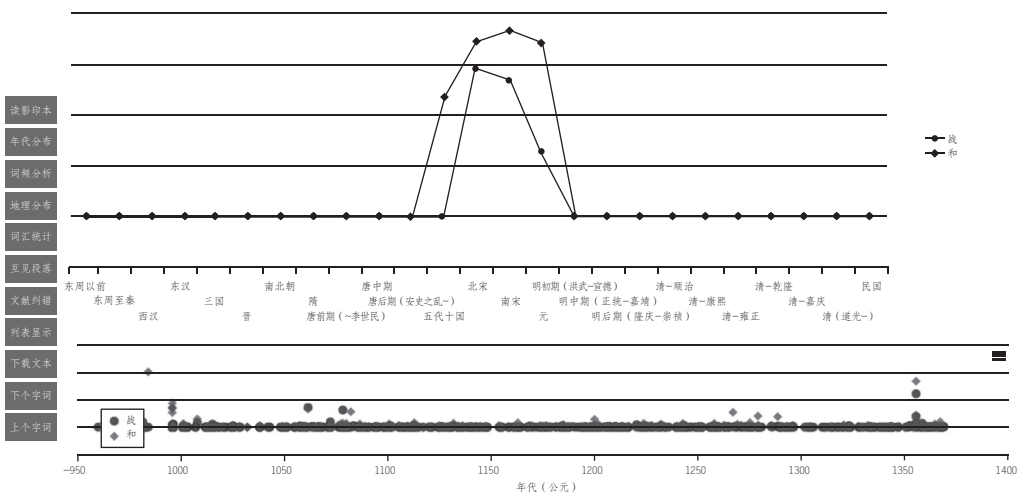

图 11　宋朝时期的文献中“战”“和”话题分析

4　面临的主要问题及解决方案

目前面临的问题及未来研究重点主要在三个方面。

首先，由于采用断代词典进行分词，而现有的断代词典收录得并不完整，因此准确性有待提高，未来会在现有断代词典的基础上不断加入新词，并实现系统对新词的自动发现，随着断代词典的不断扩充，分析的准确性将会有更大的提高。

其次，年代的考证问题，由于统计分析的时间单位计算主要是以作者的卒年为依据（不确定卒年时，系统分析统计是按不同朝代分段进行的，即只要在能确定作者所属朝代的情况下，卒年跨度可以设在该朝代内），因此作者年代的考证结果直接影响分析结果，而年代的考证难度比较大，同时，古籍文本的版本考证也是一个难题，存在诸多争议，这涉及史书语料时代性这一难度较大的问题，相关专业人士的核查将是分析准确性的保障。

再次，对古籍文本语料中的人物、历史事件、地名、官职、称谓等实体，需要在已有知识库的基础上进行抽取及关系的建立，这是文本空间信息可视化分析的核心与基础。

5　结语

数字人文的出现为传统的人文学科研究带来了新的活力，为数字化古籍的深度开发与利用带来了新的视角。本研究采用大数据视域下人文学科的数字人文研究方法及研究范式，研究古籍文本的开发与利用，以大规模中国古籍文本为研究对象，通过对古籍进行整理、标注、自动分词等处理，并采用新的可视化分析方法对古籍文本进行挖掘，创建了一个可辅助研究者进行语言学、历史文献学、历史地理学等人文学科研究为主的古籍实时统计分析平台。通过应用场景的分析发现，该平台的建设不仅拓展了人文学科的研究范式与方法，也丰富了人文学科研究工具，更是对古籍深层次开发与利用的一次尝试。虽然系统还面临诸多问题，但希望能抛砖引玉，以推进数字化古籍的开发与利用朝着更深、更广的领域发展，充分挖掘古籍的价值。

参考文献

[1] 王晓光. 数字人文:概念、现状与思考[EB/OL]. [2015-01-26]. http://meeting.lib.szu.edu.cn/conference/zh-hans/infomation?v=07000003. (Wang Xiaoguang. Digital humanities: concept, current situation and thinking [EB/OL]. [2015-01-26]. http://meeting.lib.szu.edu.cn/conference/zh-hans/infomation?v=07000003.)

[2] 陆宇杰,许鑫,郭金龙. 文本挖掘在人文社会科学研究中的典型应用述评[J]. 图书情报工作,2012(8):18-25. (Lu Yujie, Xu Xin, Guo Jinlong. Review of text mining application in Humanity and Social Science[J]. Library and Information Service,2012(8):18-25.)

[3] 唐家渝,刘知远,孙茂松. 文本可视化研究综述[J]. 计算机辅助设计与图形学学报,2013(3):273-285. (Tang Jiayu, Liu Zhiyuan, Sun Maosong. A survey of text visualization[J]. Journal of Computer-Aided Design & Computer Graphics,2013(3):273-285.)

[4] Michel J B, Yuan K S, Aiden A P, et al. Quantitative analysis of culture using millions of digitized books[J]. Science, 2011,331(6014): 176-182.

[5] Universitat Leipzig. eAQUA[EB/OL]. [2015-10-09]. http://www.eaqua.net/index.php.

[6] Torget A J, Mihalcea R, Christensen J, et al. Mapping texts: combining text-mining and geo-visualization to unlock the research potential of historical newspapers[EB/OL]. [2015-09-03]. http://mappingtexts.stanford.edu/whitepaper/MappingTexts_WhitePaper.pdf.

[7] Schich M, Song C M, Ahn Y Y, et al. A network framework of cultural history[J]. Science,2014,345(6196): 558-562 .

[8] Cho I, Dou W, Wang D X, et al. VAiRoma: a visual analytics system for making sense of places, times, and events in Roman history[J]. IEEE Transactions on Visualization & Computer Graphics, 2016,22(1):210-219.

[9] 王兆鹏. 三大功能:对未来数字化古籍的期待[N]. 中国社会科学院院报,2007-09-18(3). (Wang Zhaopeng. The three major functions: digitization of ancient books of expectations for the future[N]. Chinese Social Sciences Today,2007-09-18(3).)

[10] 杨琳. 大陆古籍数字化的现状及存在的问题[C]//第一届中国古籍数字化国际学术研讨会论文集,2007:13. (Yang Lin. Present situation and existing problems of Chinese ancient books digitization in the Chinese mainland[C]//The First International Symposium on Chinese Ancient Books Digitization,2007:13.)

[11] 王兆鹏. 利用 GIS 技术提升中国古代文学研究的数字化水平[C]//第三届中国古籍数字化国际学术研讨会论文集,2011:5. (Wang Zhaopeng. Using GIS technology to promote the digital level of Chinese ancient literature research[C]//The Third International Symposium on Chinese Ancient Books Digitization,2011:5.)

[12] 埃雷兹·艾登,让-巴蒂斯特·米歇尔. 可视化未来 数据透视下的人文大趋势[M]. 王彤彤,沈华伟,程学旗,译. 杭州:浙江人民出版社,2015. (Aiden E, Michel J B. Uncharted: big data as a lens on human culture [M]. Wang Tongtong, Shen Huawei, Cheng Xueqi, trans. Hangzhou: Zhejiang People's Publishing House,2015.)

[13] 陈云松,黄超. 大数据推动社会科学研究深挖潜力[N]. 中国社会科学报,2015-01-12(B01). (Chen Yunsong, Huang Chao. Big data dig potential for Social Science research[N]. Chinese Social Sciences Today, 2015-01-12(B01).)

[14] 沈浩,黄晓兰. 大数据助力社会科学研究:挑战与创新[J]. 现代传播(中国传媒大学学报),2013(8):13-18. (Shen Hao, Huang Xiaolan. Big data potential for Social Science research: challenges and innovation[J]. Modern Communication(Journal of Communication University of China),2013(8):13-18.)

[15] 汉字简繁文本智能转换系统[EB/OL]. [2015-10-06]. http://jf.cloudtranslation.cc. (Intelligent conversion

system between simplified Chinese characters and traditional Chinese characters[EB/OL]. [2015-10-06].http://jf.cloudtranslation.cc.)

[16] Google Ngram Viewer[EB/OL]. [2015-10-06].https://books.google.com/ngrams.

[17] 董秀芳. 词汇化:汉语双音词的衍生和发展(修订本)[M].北京:商务印书馆,2011.(Dong Xiufang. Lexicalization: the origin and evolution of Chinese disyllabic words (revised edition)[M].Beijing: The Commercial Press,2011.)

[18] 王力. 汉语词汇[M].北京:商务印书馆,1993:1-18.(Wang Li. Chinese vocabulary[M]. Beijing: The Commercial Press,1993:1-18.)

[19] 史存直. 汉语词汇史纲要[M].上海:华东师范大学出版社,1989:79-96.(Shi Cunzhi. Outline of the history of Chinese vocabulary[M]. Shanghai: East China Normal University Press,1989:79-96.)

[20] 潘允中. 汉语词汇史概要[M].上海:上海古籍出版社,1989:1-15.(Pan Yunzhong. Summary the history of Chinese vocabulary[M]. Shanghai: Shanghai Ancient Books Publishing House,1989:1-15.)

[21] 太田辰夫. 汉语史通考[M].重庆:重庆出版社,1991:2-3,63.(Oota Tatsuo. History of Chinese language[M]. Chongqing: Chongqing Publishing House,1991:2-3,63.)

[22] 王力. 汉语史稿[M].北京:中华书局,2004:43-44.(Wang Li. Chinese language history draft[M].Beijing: Zhonghua Book Company,2004:43-44.)

[23] 吕叔湘. 近代汉语指代词[M].上海:学林出版社,1985:1.(Lü Shuxiang. Modern Chinese pronoun[M]. Shanghai: Xuelin Publishing House,1985:1.)

[24] 吕叔湘. 吕叔湘文集:第4卷 语文散论[M].北京:商务印书馆, 2004:466.(Lü Shuxiang. Lü Shuxiang collected works:volume 4 language theory[M]. Beijing: The Commercial Press,2004:466.)

[25] 方一新. 20世纪近代汉语研究概说[J].浙江大学汉语史研究中心简报,2001(1):14-27.(Fang Yixin. The research of modern Chinese in twentieth century[J]. The Briefing News of Research Center for History of Chinese Language,2001(1):14-27.)

[26] 方一新. 从中古词汇的特点看汉语史的分期[J].汉语史学报, 2004,4(1):178-184.(Fang Yixin. From the characteristics of the ancient words look at the stage of Chinese history[J].Journal of Chinese Language History, 2004,4(1):178-184.)

[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.(Liang Shehui, Chen Xiaohe. Methodological study of automatic word segmentation in the pre-Qin document *Mencius*[J].Journal of School of Chinese Language and Culture Nanjing Normal University, 2013(3):175-182.)

[28] 段磊,韩芳,宋继华. 古汉语双字词自动获取方法的比较与分析[J]. 中文信息学报,2012,26(4):34-42.(Duan Lei, Han Fang, Song Jihua. A comparative study on the automatic extraction of two-character word from ancient Chinese[J].Journal of Chinese Information Processing, 2012,26(4):34-42.)

[29] 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.(Shi Min, Li Bin, Chen Xiaohe.CRF based research on a unified approach to word segmentation and POS tagging for pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010,24(2):39-45.)

[30] 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.(Huang Shuiqing, Wang Dongbo, He Lin. Exploring of word segmentation for fore-Qin literature based on the domain glossary of *Sinological Index Series*[J].Library and Information Service,2015,59(11):127-133.)

[31] 中国历代人物传记资料库 (CBDB)[EB/OL].[2015-10-13].http://isites.harvard.edu/icb/icb.do?keyword

=k35201.(The China biographical database[EB/OL].[2015-10-13].http://isites.harvard.edu/icb/icb.do?keyword=k35201.)

[32] 中国历史地理信息系统(CHGIS)[EB/OL].[2015-10-12].http://yugong.fudan.edu.cn/views/chgis_index.php?list=Y&tpid=700.(China historical GIS[EB/OL].[2015-10-12].http://yugong.fudan.edu.cn/views/chgis_index.php?list=Y&tpid=700.)

[33] 杨义.中国文学与人文地理[N].人民日报,2010-04-02(24).(Yang Yi. Chinese literature and human geography[N]. People's Daily,2010-04-02(24).)

[34] 曾大兴.构建文学地理学"立体图景"[N].中国社会科学报,2011-11-08(11).(Zeng Daxing. Constructing the literary geography "three dimensional picture"[N]. Chinese Social Sciences Today, 2011-11-08(11).)

[35] 罗德里克·弗拉德.计量史学方法导论[M].王小宽,译.上海:上海译文出版社,1997.(Floud B. An introduction to quantitative methods for historians[M].Wang Xiaokuan, trans. Shanghai: Shanghai Translation Publishing House,1997.)

[36] 罗志田.走向国学与史学的"赛先生"——五四前后中国人心目中的"科学"一例[J].近代史研究,2000(3):59-94,2.(Luo Zhitian. "Mr. Science's" turn towards national studies and history: an example of "science" as seen by Chinese during the May Fourth Period[J]. Modern Chinese History Studies,2000(3):59-94,2.)

欧阳剑 上海师范大学语言研究所计算语言学博士研究生,广西民族大学图书馆研究馆员。上海 200234。

(收稿日期:2015-10-16;修回日期:2015-11-19)