历史典籍的结构化探索——《史记·列传》数字人文知识库的构建与可视化研究

郑童哲恒[1]，李斌[1]，冯敏萱[1]，常博林[1]，王东波[2]
1. 南京师范大学文学院，江苏 南京 210097；
2. 南京农业大学信息管理学院，江苏 南京 210095

摘要
中国古代典籍文献浩如烟海，蕴藏了大量的历史人文知识。以电子化和全文检索为主要方法的古籍数字化开发应用模式已经成为语言文学、历史、哲学等学科的重要基础资源和工具。随着人工智能与大数据技术的发展，数字人文的研究范式不断演进，将传统典籍的文本转换为高度结构化的新型数字人文数据库是一项新的探索，将文本中词汇、人物、地理实体等要素有机组织起来，对于历史现象可视化、历史规律量化具有重大意义。以《史记·列传》为对象，进行古汉语自动分词及词性标注、人工校对以及实体信息人工标注，形成多层次、高质量的数字人文知识库，实现包含古籍词汇、人物、地点等要素的定量分析与可视化检索，挖掘出《史记·列传》人物和地点分布情况、人物关系、人地关系等信息。得出：《史记·列传》共出现人物1 787位、地点1 173个；相比《史记·本纪》和《史记·世家》，《史记·列传》特有人物共1 092位，特有地点共556个。本文研究内容为古籍数字人文知识库的构建提供了新的思路与框架。

关键词
数字人文；《史记·列传》；知识服务；大数据；古汉语信息处理

中图分类号：G250 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022067

Explore the structuration of historical books: the construction and quantitative analysis of digital humanities database of the Biographies of the Shiji

ZHENG Tongzheheng[1], LI Bin[1], FENG Minxuan[1], CHANG Bolin[1], WANG Dongbo[2]
1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097, China
2. College of Information Management, Nanjing Agricultural University, Nanjing 210095, China

Abstract
Ancient Chinese classical books are vast and contain a lot of historical and humanistic knowledge. The development and application mode of the digitization of ancient books based on digitization and full-text retrieval has become

注：为保证本文研究的真实性和可信性，文中保留涉及历史典籍《史记·列传》原文内容的繁体字。

an important basic resource and tool for language and literature, history, philosophy and other disciplines. With the development of artificial intelligence and big data technology, the research paradigm of digital humanities is constantly evolving. It is a new exploration to convert the text of traditional books into a highly structured new digital humanities database. Organizing elements such as words, characters, and geographical entities in the text organically is of great significance for the visualization of historical knowledge and the quantification of historical information. The Biographies of the Shiji was selected as the object. The automatic word segmentation and part-of-speech tagging, manual proofreading and manual annotation of entity information were performed to construct a multi-level and high-quality structured digital humanities knowledge base, realize quantitative analysis and visual retrieval of elements, such as words, characters and locations of ancient books, and excavate information such as distribution of characters and locations, relationship between characters and relationship between people and locations. It was concluded that there are 1 787 persons and 1 173 locations in the Biographies of the Shiji, and compared with Benji and Shijia of the Shiji, there are 1 092 unique persons and 556 unique locations of the Biographies of the Shiji. New ideas and frameworks for the construction of digital humanities knowledge base of ancient books were provided.

Key words
digital humanities, the Biographies of the Shiji, knowledge service, big data, ancient Chinese information processing

0 引言

古籍是我国传统文化的重要载体，是民族精神的集中体现。我国古籍浩如烟海，在绵延千年的发展过程中历久弥新，蕴藏了大量的历史人文知识，是研究传统文化和挖掘历史信息的珍贵材料。在众多古籍之中，《史记》意义重大，它是中国历史上第一部纪传体通史，记载了从传说中的黄帝时代至汉武帝时期共3 000多年的历史，对后世文学和史学发展具有重要指导意义。《史记》共130篇，其中列传有70篇，共24万余字，占《史记》全文篇幅的一半左右，记载了众多历史人物的言行事迹，具有很高的研究价值。

古文信息处理是指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]。数字人文（digital humanities）也被称为人文计算（humanities computing），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[2]。近年来，随着古文信息处理技术、人工智能与大数据技术的持续发展，数字人文研究范式在古籍研究中的应用范围不断扩大、应用方式不断演进[3]。古籍数字人文研究为解决古代典籍因卷帙浩繁、晦涩难懂而不易开发利用等问题提供了新思路，为深入挖掘古籍文本信息、全面检索古籍文本内容、直观展示古籍文本内涵提供了新方法。

本文继承南京师范大学开发的《左传》[4]、《史记·本纪》（以下简称为《本纪》）[5]和《史记·世家》（以下简称为《世家》）3个数字人文知识库，创新性、发展性地以《史记·列传》（以下简称为《列传》）为研究语料，首先进行自动分词和词性标注并进行人工校对，再进一步人工标注人物和地点等实体信息，得到《列传》高质量标注文本。在此基础上构建《列传》数字人文知识库和检索平台，并据此完成词汇、人物、地点3个方面的信息挖掘与计量统计，力图运用大数据技术推动历史典籍的结构化探索，进而为历史文献学、历史地理学、语言学等学科的研究提供服务。

1 相关研究

古籍数字化开发分为表层和深层两个层次[6]。表层古籍数字化包括古籍的录入、数字化存储、网络传播等，深层古籍数字化则包括古籍的信息标注、内容加工和知识检索。表层古籍数字化研究与实践始于20 世纪 70 年代末[7]，在其发展初期涌现出以文本录入为基础实现全文检索的古籍语料库。如中国社会科学院开发的《全唐诗》速检系统，提供字、诗句、标题检索[8]；爱如生公司开发的中国基本古籍库，提供分类、条目、全文检索[9]。由于没有对古籍文本进行深加工，上述表层古籍数字化成果的功能较为单一，查全率和查准率亦不够理想。

随着人们对古籍数字化的认识不断发展，数字化古籍文本的知识加工不断完善，迈向更深的“知识域”，进入深层古籍数字化阶段。深层古籍数字化旨在对古籍内容进行标注并构建知识网络，进而推动古籍文本可视化、文本信息挖掘等工作。对古籍文本进行词语切分和词性标注，是突破基于“字”的全文检索、构建词汇级别古籍数据库的必要条件。古代汉语标注语料库目前较为稀少，主要有：台湾的上古、中古汉语标记语料库；南京师范大学先秦、中古[10]汉语标注语料库；留金腾等人[11]以《淮南子》为文本构建的上古汉语分词及词性标注语料库。针对目前古汉语标注语料库数量少、深度不足的问题，本文对古籍文本进行了更深层次的数字化加工。

21世纪初兴起的数字人文研究以古籍数字化为基础条件，对古籍内容进行数据统计、信息和知识挖掘等处理[12]。基于知识本体（ontology）的古籍知识库建设取得进展。唐振贵等人[13]在时间轴上由粗至细系统梳理了中国古代时间谱系，构建了涵盖时间系统等五大主要模块的中国古代时间本体。中国历代人物传记资料库（China biographical database，CBDB）通过创建关系型数据库，记录了史料中保存下来的历史人物的职业、亲属关系、社会关系等数据[14]。古籍专书数据库亦取得成果。钱智勇等人[15]论述了楚辞知识库和网站设计的实现步骤、技术难点及解决思路，力求实现辞赋知识的多维度关联与智能检索。在南京师范大学先秦语料库的基础上，许超等人[16]提取《左传》中的人物、事件，使用社会网络分析软件Pajek建立春秋时期的社会网络，并对其进行定性、定量探索性研究。李斌等人[4]在词语切分、词性、人物ID信息标注的基础上进一步标注时间、地点坐标信息，构建深度标注的《左传》知识库，实现了一系列基于词语、实体和时间地理信息的统计与可视化。相同的思路也被应用于南京师范大学《史记·本纪》和《史记·世家》数字人文知识库的构建当中。

《史记》在汉籍当中至关重要，因此相关数字化研究很受重视。1987年，哈尔滨工业大学建成《史记》全文检索系统，这是中国对古文献全文进行字检索的开创性成果。《鼎秀古籍》等古籍典藏数据库将《史记》收录在内，提供全文检索功能，完成了《史记》的表层数字化工作。随着《史记》数字化走向深层阶段，《瀚堂典藏》数据库收录《史记》，并运用人工智能分词技术，实现了古籍文本基于词的检索。2014年中华书局推出收录《史记》在内的《中华经典古籍库》，提供专名查询（包括人名、事件、地名、纪年、职官机构）、联机字典、纪年换算等检索功能[17]。

近年来，《史记》专书数字人文研究亦有发展。张琪等人[18]探究基于深度学习方法的古籍分词词性一体化标注技术，并

将其应用于《史记》，统计出《史记》中人名、地名、动词、时间词4种词类的高频词。刘忠宝等人[19]提出面向《史记》的历史事件及其组成元素抽取方法，并基于此构建《史记》事理图谱。南京师范大学开发的《史记·本纪》数字人文知识库，提供词汇、人物、地点与地理信息系统（geographical information system，GIS）信息检索功能。

综上可知，《史记》专书深层数字化和数字人文研究已有一定成果，词汇级别的、提供实体信息查询的《史记》数字人文知识库正在逐步建设当中。本文有效结合词汇、实体信息、GIS技术等方面，完成《史记》中《列传》部分的内容标注与知识挖掘，为建成完整的《史记》数字人文知识库补充大量语料，也为后续进行综合性、多层次的《史记》全文文本知识挖掘、计量分析与可视化检索提供可能。

2 《史记·列传》数字人文知识库的建设

知识库是存储、组织和处理知识以及提供知识服务的重要知识集合[20]。数字人文视域下的古籍知识库建设是在古籍文本录入的基础之上，对生文本进行词性、句法、语义等不同层面的标注，提取时间、地点、人物、事件等不同类型的实体，通过大数据技术重组古籍文献知识，并支持可视化分析。为建设《史记·列传》数字人文知识库，首先对《列传》进行自动分词和人工词性标注，再为每个人物、地点指定唯一的ID编号，进一步完善命名实体信息。人物方面补充人物别名、性别、国别，地点方面补充今地名和GIS坐标，由此实现了《列传》词类标注基础上的历史时间、地点、人物信息全面标注，得到6张数据表：文本表、文本标注表、人物表、地点表、人物同现表、人地同现表。进而以6张一维线性序列表为基础，构建多维《列传》知识网络，打通人物库与GIS库，使《史记·列传》数字人文知识库成为基于词和实体的、结构化、一体化的知识集合。

2.1 数据来源与分词和词性标注

《史记·列传》数据库的原始数据来自《史记》（点校修订本）[21]的《列传》部分。首先使用南京师范大学开发的古汉语分词与词性标注规范和自动分析工具[22]，对《列传》全文24万余字进行自动分词和词性标注，词性标记共分为32类：形容词（a）、连词（c）、副词（d）、方位词（f）、词缀（i）、兼词（j）、数词（m）、普通名词（n）、书名（nb）、国名（ng）、年号（nh）、民族（nn）、官职（no）、人名（nr）、地名（ns）、专名（nx）、介词（p）、量词（q）、代词（r）、拟声词（s）、时间词（t）、助词（u）、动词（v）、使动用法（vs）、为动用法（vw）、意动用法（vy）、标点（w）、其他语素和字（x）、语气词（y）、形容词作状语（za）、名词作状语（zn）、动词作状语（zv）。再根据《二十四史全译》[23]等工具书，对自动分词和词性标注结果进行人工校对。在人工校对的基础之上，对《列传》全文进行二次实体信息人工标注（标注内容包括人物信息和地点信息等），由此形成了《列传》高质量、多层次的标注文本。多层次标注样例见**表1**。

2.2 实体信息标注

2.2.1 人物信息标注

《列传》中人物和名称往往不是一一

表1 多层次标注样例

标注层级	样例
原始文本	管仲夷吾者，颍上人也。
自动分词与词性标注	管仲/nr 夷吾/nr 者/u ，/w 颍/v 上人/n 也/y 。/w
人工校对	管仲/nr 夷吾/nr 者/u，/w 颍上/ns 人/n 也/y 。/w
实体信息人工标注	管仲/nr[人物ID：1964] 夷吾/nr[人物ID：1964] 者/u ，/w颍上/ns[地点ID:1052] 人/n 也/y 。/w

对应的，异名同指（一人对应多个名称）、同名异指（一个名称对应多人）的情况时有出现。人物与名称的参差对应使后续计量分析的准确性受到很大影响，因此本文采取为每个人物标注唯一人物ID编号的方法，选取其最具代表性和概括性的、为人们所熟知的称呼为“正名”，其余归为“别名”，同一人物的不同名称都指向同一个ID。如果某人物在《史记》的《本纪》和《世家》部分出现过，则沿用其先前被匹配的人物ID，如果是在《列传》中出现的新人物，则为其标注新的ID。除人物ID、正名、别名之外，《史记·列传》数据库中收录的人物信息还包括每个人物的性别、国别、备注，人名表示例见**表2**。

2.2.2 地点信息标注

《史记·列传》知识库收录的地点信息包括文中每个地点的地点ID、地名、今地名、类别（一般地名、诸侯国名、河流、山名等）、百度地图GIS坐标，地名表示例见**表3**。同样，如果某地点在《史记》的《本纪》和《世家》部分出现过，则沿用其先前被匹配的地点ID；如果是在《列传》中出现的新地点，则为其标注新的ID。笔者参考《史记地名考》[24]等文献以考证文中古地名的今地点，在此基础上利用百度地图应用程序接口（application program interface，API）解析今地点，获得对应的GIS坐标数据。

2.3 数据库架构

在经过二次校对的分词和词性标注、人物信息标注、地点信息标注的基础之上，完成了《列传》文本的历史时间、地点、人物信息的全面标注，形成文本表、文本标

表2 人名表示例

人物ID	正名	别名	性别	国别	备注
1964	管仲	管 管敬仲 管氏 管夷吾 管仲 夷吾	男	齐	政治家
7542	趙堯	堯 江邑侯	男	汉	西汉官吏

表3 地名表示例

地点ID	地名	今地点	类别	来源	百度地图GIS坐标
6507	昭關	安徽含山县北	吴越地名	钱穆《史记地名考》	118.10 31.92
6575	三梁	河北永年县	魏地名	钱穆《史记地名考》	114.55 36.75

注表、人物表、地点表、人物同现表、人地同现表，构建了《史记·列传》数字人文知识库，知识库结构如图1所示。

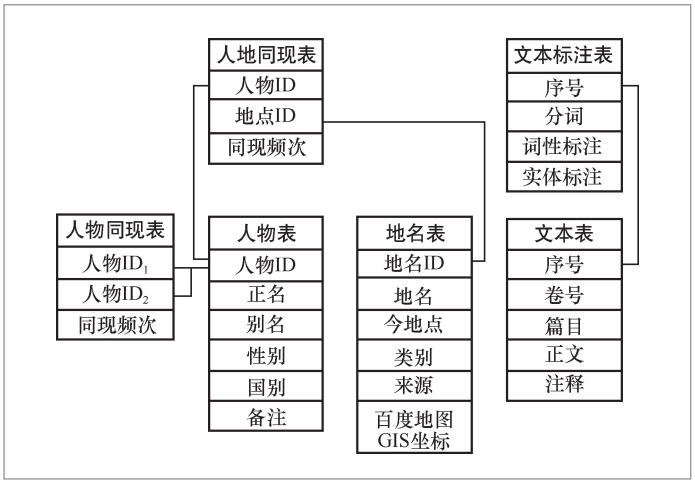

图1 《史记·列传》数字人文知识库结构

3 《史记·列传》数字人文知识库与地图平台

3.1 检索框架

本文构建的《史记·列传》检索平台包含全文检索、人物检索、地名检索三大功能，全文检索包括“文本”“词频词性”检索功能，而人物和地名实体查询需要依托实体ID，其中人物检索包括“人物基本信息”“原文追踪”和“人物关系”检索功能，地名检索包括“地点基本信息”和“人地同现”检索功能。检索平台结构如图2所示。

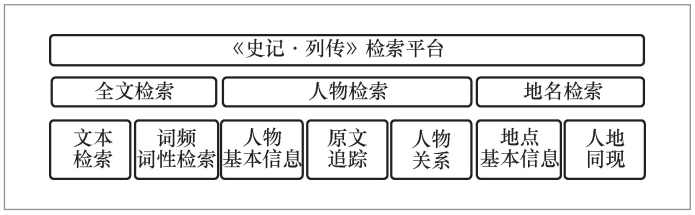

图2 检索平台结构

3.2 全文检索

在全文检索方面，本检索平台除提供基础的文本字符匹配检索之外，还提供词频词性检索。词频词性检索可以基于词，如检索“者”，可得“者”在《列传》中以助词（u）词性出现2 714次，以代词（r）词性出现1 812次，以名词（n）词性出现86次。从不同词性的应用比例来看，在《列传》中“者”主要以助词和代词形式出现，尤以助词为主，这可以为《史记》的词汇研究提供支撑材料。词频词性检索也可以基于词性，如检索名词（n），可得《列传》中的名词按频次由多到少排列分别为“人、王、兵、臣、國……”，从高频名词可以看出，这是一段群雄交锋、英雄辈出、战争四起的历史岁月。词频词性检索示例见表4和表5。

表4 词频词性检索示例（词：者）

词/词性	频次/次
u	2 714
r	1 812
n	86

表5 词频词性检索示例（词性：名词）

词/词性	频次/次
人	4 120
王	2 100
兵	1 942
臣	1 792
國	1 350

3.3 人物检索

相较于传统的人物检索，本平台的人物检索功能更加全面、准确、直观。人物检

索页面能够为用户提供所查询人物的基本信息（人物ID、正名、别名、性别、国别）、上下文信息（出现次数、原文追踪）以及人物关系（交往人物、交往频次）。以检索“公孫敖”为例，首先在人物检索页面输入“公孫敖”，继而呈现“公孫敖”的人物基本信息，可知其人物ID为7731。以人物ID为线索，进一步检索可得“公孫敖”在《列传》中以各种称谓出现的24个文段。“公孫敖”人物检索示例见**表6**，原文追踪示例见**表7**。

3.4 地点检索

地点检索页面供用户检索《列传》中所有地点的基本信息（地点ID、地名、今地点、类别），并使用百度地图API，添加地图控件，将《列传》中出现的地名还原为精确的地图坐标，并做出相应标记，使用户能够从地图上直观感受《列传》地名的具体位置。

3.5 人物地图——人地同现轨迹图

人物游历轨迹是历史研究中的重要问题之一，但用传统方法进行研究往往需要进行大量考证，且文字描写不够直观。为了用更加简洁且直观的方式来展现《列传》中人物的游历轨迹，运用近似计算和可视化方法，根据人物和地点在文本中的同现信息（在用逗号或句号分隔的一个句子中同时出现）生成人地同现轨迹图，并在检索平台网站上提供地图信息查询功能。

以“李廣”为例，平台检索“李廣”的高频同现地点见**表8**。由此可以推断出“李廣”的游历轨迹，生成人地同现图，为“李廣”事迹研究提供可视化线索。

4 数据分析与可视化

《史记·列传》数字人文知识库及检索平台进行了历史典籍的结构化探索，在数据的丰富性和检索的层次性上远超传统全文检索数据库。本节将在此基础上，对《列传》进行词汇、人物、地点、实体同现等层面的知识挖掘与计量分析。

4.1 《列传》词汇基本面貌

不同于以往基于字的古籍数据库，本文构建的《史记·列传》数据库以经过大

表 6 “公孫敖”人物检索示例

人物ID	正名	别名	性别	国别	备注
7731	公孫敖	敖 合騎侯	男	汉	无

表 7 “公孫敖”人物原文追踪示例

序号	原文
1	將軍李廣爲匈奴所得，復失之；公孫敖大亡卒：皆當斬，贖爲庶人。
2	而是時公孫敖新失侯，爲中將軍從大將軍，大將軍亦欲使敖與俱當單于，故徙前將軍廣。
3	公孫敖出代郡，爲胡所敗七千餘人。

表 8 平台检索“李廣”的高频同现地点

地名	同现频次	今地点	百度地图GIS坐标
匈奴	23	内蒙古一带	111.04 41.99
右北平	6	热河平水县	117.97 40.96
上郡	3	今陕西延安、榆林一带	109.49 36.58
雁門	3	山西右玉县南	112.48 39.81
隴西	2	甘肃省陇西县	104.64 35.01

量切分和标注工作得到的《列传》分词标注文本为基础，实现了基于词的检索，能够从词汇层面对《列传》全文进行穷尽式的统计，将《列传》全文的计量分析从单字层面拓展到词汇层面。据统计，《列传》共有216 942个词（247 540个字），其中单字词有189 683个，双字词有23 175个，三字及以上词语有4 084个，全文以单字词为主，平均每词1.1个字。

运用《史记·列传》数据库可以进行以往基于字的数据库无法完成的多字词统计，这是没有分词的数据库无法实现的工作。《列传》高频多字词（前10位）见表9。构词方面，《列传》中的多字词以双字词为主；词性方面，《列传》中的多字词以名词为主，其他词性较少出现；词义方面，高频多字词均与国家、政治体系、军事、民族等相关，符合《史记》记叙朝代兴替、帝王与人臣事迹的文本特点。《列传》高频多字词词云如图3所示。

除了对词汇长度进行统计，还可以从词性角度对各词性内部的词汇分布进行计算，得出各词类的高频词。如《列传》全文中副词共出现16 956次，其中最高频的前5个副词见表10，由此可知文中最常用的副词是“不”，频次高达4 453次，远远超过其他副词。

表 9 《列传》高频多字词（前 10 位）

多字词	频次/次	多字词	频次/次
天下	654	諸侯	227
匈奴	362	將軍	212
太子	299	單于	212
天子	278	大王	194
丞相	238	公子	189

图 3 《列传》高频多字词词云

4.2 《列传》实体信息统计

4.2.1 人物分布

不同于《本纪》和《世家》，《列传》

表10 《列传》高频副词（前5位）

副词	频次/次
不	4 453
皆	664
必	496
亦	424
未	417

主要记录人臣事迹，所涉人物必然相应地与前两部分有所不同。对文中记录的历史人物进行频次层面的梳理，有助于把握《列传》的重点人物和事件。据统计，《列传》出场人物共1 787位，其中未在《本纪》《世家》出现的《列传》特有人物共1 092位。

统计《列传》高频人物有助于把握《列传》的人物事件主基调，而高频人物往往有多个不同称谓，这给人物统计增加了难度。本文使用的为每个人物标注唯一人物ID的方法，不仅在很大程度上降低了“异名同指”和“同名异指”问题对人物统计造成的负面影响，还为《列传》人物研究提供了人物的不同称谓频次方面的研究材料。《列传》中按出场频次排序前10位的人物如**图4**所示，由内圈至外圈分别为人物ID、人物主名以及该人物的不同称谓占比。

4.2.2 地点分布

传统的古籍地点研究往往以某地在文本中出现的若干处例句为对象，研究方法以列举、归纳为主，研究结果也多停留在文字层面。而通过穷尽式的统计与可视化的检索，本文可收集《列传》任意地点的所有出处，并将其定位至百度地图，这为《列传》地点研究提供了更精细的语料、更高效的方法、更直观的结果。

据统计，《列传》共提及地点1 173个，按频次排序前10位的高频地点（不包括诸侯国）见**表11**，出现范围最广、次数最多的地点多为河流、古都城。

黄河作为频次最高的地点，在《列传》乃至《史记》全文中的地位一目了然，这印证了北方黄河流域是《史记》所记载历史的主要地理背景。**表11**中排名第二的邯郸为赵国国都，排名第八的咸阳为秦国国都（秦朝都城），再次为赵国和秦国的影响力提供了佐证。值得注意的是，《列传》中邯郸的频次高于咸阳，与《本纪》中情况相反，这正体现了秦国和赵国的不同历史地位：赵国为战国七雄之一，但后被秦军攻灭；而秦国兼并六国进而完成统一大业，建立了中国历史上首个统一封建王朝，因此在以王朝更替为主的《本纪》之中，秦国都城的出现频次自然比赵国都城高得多。这足以证明从《史记》地名的分布规律中可以窥见历史信息，为古籍研究提供材料。

《列传》中出现的1 173个地点中，有556个未在《本纪》和《世家》中出现过。为了更好地探索《列传》独特的历史地理信息，本文统计得出《列传》独有的高频地点前5位（不包括诸侯国），具体见**表12**。

《列传》独有高频地点前5位中包含“乌孙”“康居”两个西域地名，可见《列传》有许多前文较少涉及的与西域相关的历史事件描写，这值得相关学科的研究人员特别关注。

4.3 实体关系

传统古籍研究很难自动地、全面地挖掘人物、地点等实体间的关系，并以客观统一的标准对其进行衡量。本文在对《列传》进行全文实体标注的基础上，计算实体ID

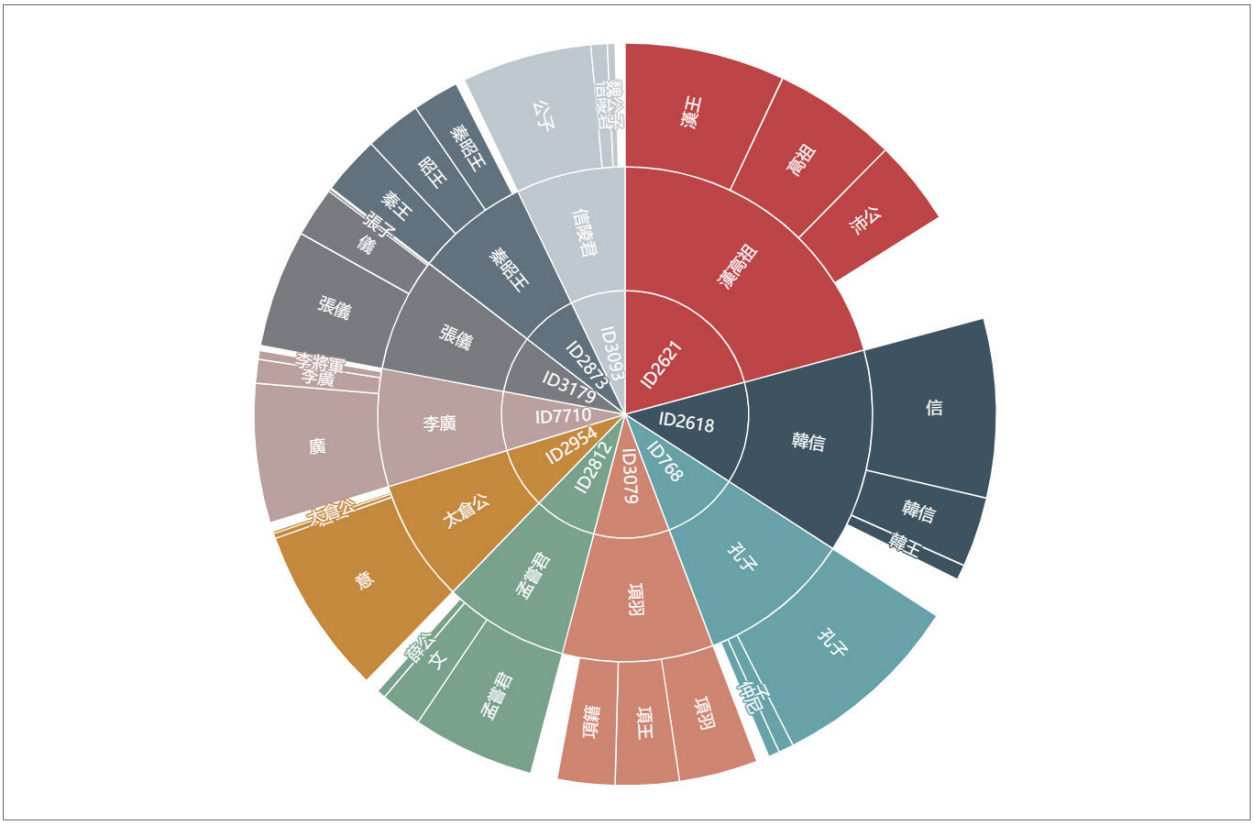

图 4 《列传》高频人物及称谓分布（前 10 位）

间的同现情况并进行统计，实体同现次数越多则相关度越高。据此本文进行了《列传》人物关系密度和广度、人物的同现地点数、地点的同现人物数的统计和汇总。

表 11 《列传》高频地点（前 10 位）

地点	频次/次	地点	频次/次
（黄）河	103	梁	48
邯鄲	68	淮南	43
大宛	68	咸陽	35
長安	62	雲中	33
滎陽	50	江	32

4.3.1 人物关系密度

两个人物之间的同现次数可以作为估算人物关系的指标，往往联系越紧密的两个人同现次数越多。本文在《列传》中选取**表13**所示的3对同现人物进行分析，高频同现人物对多与历史事件、血缘亲族、君臣关系等相关。汉高祖刘邦和项羽在《列传》中为最高频同现人物对，这正是“楚汉争霸”的缩影；汉文帝与汉景帝是父子关系，并共同造就“文景之治”；秦昭王和蔺相如同现多次，这源于“完璧归赵”和“渑池会盟”。由此可见，统计《列传》中的高频同现人物对可以为众多历史人物和历史事件的研究提供量化参考。

为了更好地展现《列传》众多人物间的关联以及交往密度，本文选取《列传》同现人物高频前120对，借助ECharts技术绘制人物关系网络（如**图5**所示）。**图5**中节点表

表 12 高频《列传》独有高频地点（前 5 位）

地名	频次/次	类别	今地点	百度地图GIS坐标
烏孫	25	西域国名	吉尔吉斯坦伊塞克湖东南伊什提克	77.15 42.25
定襄	20	西北边地名	杀虎口北，归绥南，绥远和林格尔县治	112.33 40.60
晝邑	18	齐地名	临淄县西北	118.02 36.98
雁門	17	郡名	山西右玉县南	112.48 39.81
康居	17	西域国名	约在今巴尔喀什湖和咸海之间，王都卑阗城	60.81 44.29

表 13 《列传》高频人物同现对

人物1	人物2	同现频次/次
漢高祖	項羽	42
漢文帝	漢景帝	16
秦昭王	藺相如	15

示人物，边表示交往关系，根据图中节点大小、关系网疏密，可以直观地把握人物交际网络。从整体上看，《列传》中的人物交际关系网主要以汉高祖、秦始皇、韩信、项羽、秦昭王等人物为核心。

4.3.2 人物关系广度

广度同样是衡量人物交往情况的参考依据。某一特定人物对的同现频次可以显示两人之间的关系疏密，而某一特定人物拥有的同现对数量，则可以显示该人物的交往范围。统计出某一特定人物共拥有多少对人物关系后，可以进一步细化查询该人物分别与哪些人物有过几次同现，在研究历史人物生平时便可比较完整地把握其人际关系。借助ECharts绘制的“李廣”在《列传》中的人物关系图如**图6**所示。中心节点为“李廣”，周围节点为与其有同现关系的人物，节点越大说明同现关系越多，也即关系越紧密、相关度越高。由**图6**可见，“李廣”在《列传》中共与29人有过同现，其中相关度最高的是“公孫敖”，“衛青”“李敢”“程不識”3人次之。

4.3.3 人地关系

人物–地点关系是古籍研究的重要问题之一，有助于探究历史人物生平经历、把握历史地点重要程度。但使用传统研究方法很难从量化的角度让人们对古人游历情况有直观的了解。本文在计算人物–地点同现关系的基础上估算《列传》人物游历地点，分别从人物角度计算人物的同现地点数量、从地点角度计算地点的同现人物数量，这可以作为推断某特定人物在《列传》中所记录的游历轨迹、某特定地点在《列传》中的重要程度的参考。

《列传》中同现地点数最多的前5个人物和同现人物数量最多的前5个地点见**表14**。可以看出所列人名和地名与前文统计得到的高频人物、高频地点、广交人物、密交人物多有重合。

5 结束语

古籍数字化不断向深层方向发展，将

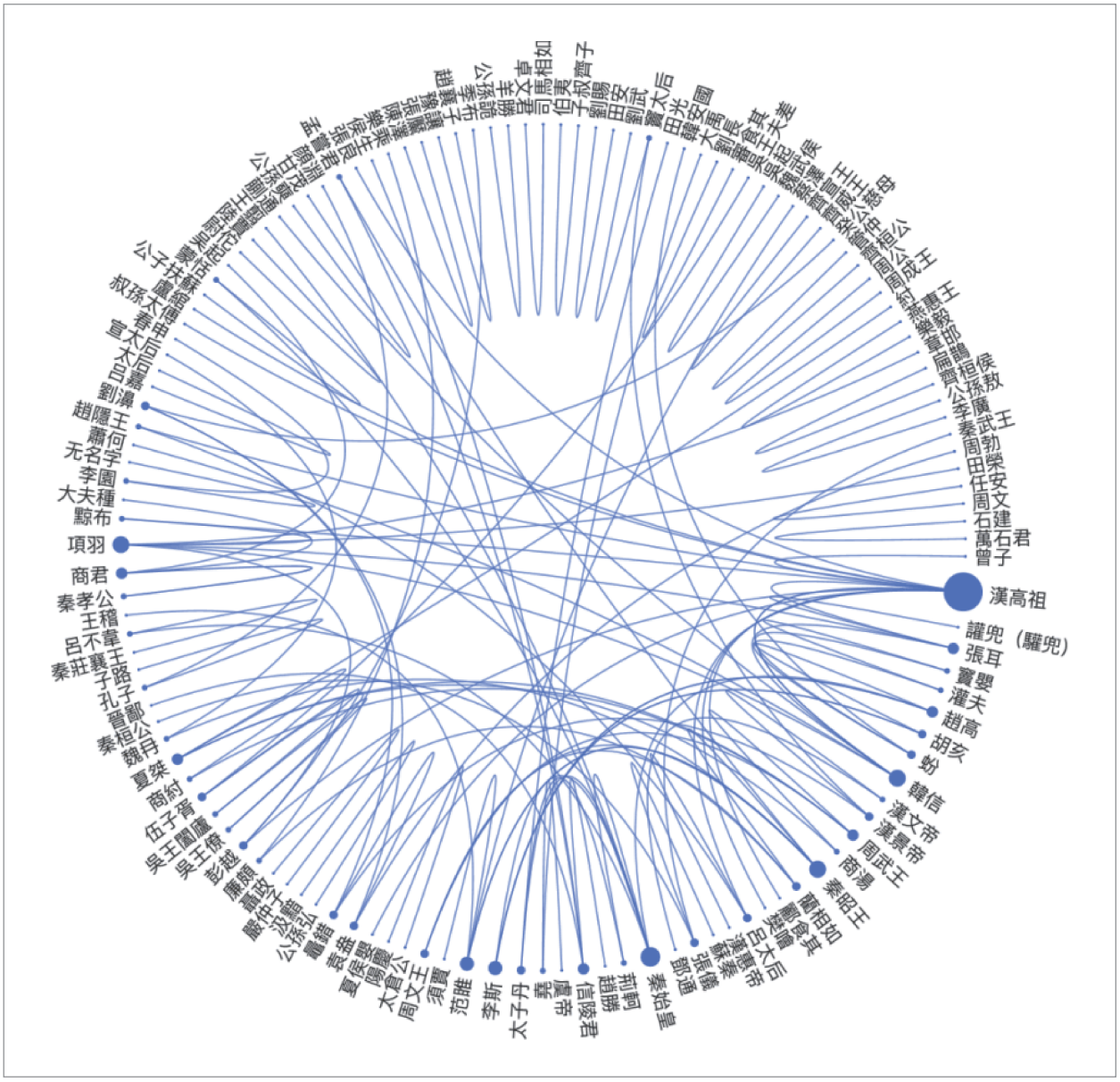

图 5 《列传》同现人物关系网络（前 120 对）

传统典籍的文本转换为高度结构化的新型数字人文知识库，将文本中词汇、人物、地理实体等要素有机组织起来，推动古籍文本可视化、文本信息挖掘等工作，对我国古籍的研究与传承意义重大，对语言学、历史文献学、历史地理学等学科具有积极的推动作用。本文为进行历史典籍的结构化探索、推动《史记》深层数字化工作，以《列传》为对象，将传统典籍的文本转换为高度结构化的新型数字人文知识库，主要完成了以下工作。

● 对《列传》进行词性、实体标注，完善《列传》人物表、地名表等6张数据表，在此基础上建成了基于词和实体的、结构化、一体化的《史记·列传》数据库。这对南京师范大学开发的《史记·本纪》《史

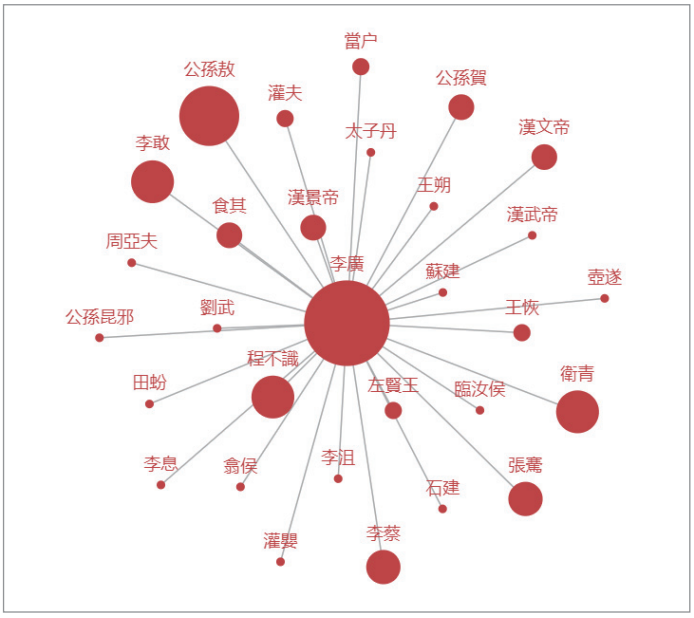

图 6 “李廣”在《列传》中的人物关系图

表 14 《列传》高频共现人物、地点（前 5 位）

人–地关系		地–人关系	
人物	同现地点数/个	地点	同现人物数/位
漢高祖	102	（黃）河	66
項羽	81	長安	62
韓信	73	邯鄲	50
灌嬰	51	代郡	38
孔子	42	梁山	36

记·世家》数字人文知识库起到了重要的承接作用，为《史记》整体数据库的构建做了丰富的内容补充。

● 基于数据库开发线上检索系统，检索功能包括全文检索以及传统数据库无法实现的基于深度标注的词频词性检索、人物检索、地点检索等，并结合百度地图实现人物关系、人地关系的可视化。

● 在数据库和检索平台的基础上，本文进行了一系列数据统计和可视化分析。首先描写《列传》多字词的基本面貌，计算得到《列传》平均每词1.1个字。其次统计《列传》人物、地点分布情况，列出了《列传》的高频人物和地点，得出《列传》共出现人物1 787位、地点1 173个。且较之《本纪》和《世家》，《列传》特有人物共1 092位，特有地点共556个，量化了《列传》与《本纪》《世家》的差异。最后，量化《列传》人物关系和人地关系，对人物–人物、人物–地点的交往密度和广度进行计量。

但受制于时间、人力等因素，本文研究仍存在不足之处有待在未来的工作中不断改进，具体如下。

● 完善标注规则，提高标注准确性。本文数据正在持续校对当中，后续将对细节问题进行补充和校正。在此过程中需要及时记录并整理所遇到的问题，相应地对标注规则进行细化。亦可通过开放在线标注校正系统，为邀请各界专家学者加入标注校对工作提供便利条件，最终形成系统性的标注规范，使《列传》标注文本具有更高的准确度。

● 后续将继续扩大数据规模，将《本纪》《世家》和《列传》三部分数据库进行整合，形成更加完整的《史记》数据库。

● 尝试运用多种数字化技术，对包括人物关系、人地关系在内的实体关系计算进行改进，使其突破限于近似估算的水平。

● 尝试设计交互可视化系统，使可视化效果更加多维、丰富。优化检索平台性能，使检索平台更好地为社会服务，起到科研和科普作用。还可以与其他学科和数据库联动，拓宽研究思路，得出更加多层次、宽领域的研究成果。

参考文献:

[1] 陈小荷，冯敏萱，徐润华，等. 先秦文献信息处理[M]. 北京：世界图书出版公司北京公司，

2013.
CHEN X H, FENG M X, XU R H, et al. Pre-Qin literature information processing[M]. Beijing: World Book Publishing Company Beijing, 2013.

[2] 黄水清, 王东波. 古文信息处理研究的现状及趋势[J]. 图书情报工作, 2017, 61(12): 43-49.
HUANG S Q, WANG D B. Review and trend of researches on ancient Chinese character information processing[J]. Library and Information Service, 2017, 61(12): 43-49.

[3] SCHREIBMAN S, SIEMENS R, UNSWORTH J. A companion to digital humanities[M]. Oxford: Blackwell, 2004.

[4] 李斌, 王璐, 陈小荷, 等. 数字人文视域下的古文献文本标注与可视化研究: 以《左传》知识库为例[J]. 大学图书馆学报, 2020, 38(5): 72-80, 90.
LI B, WANG L, CHEN X H, et al. Digital humanity based ancient text annotation and visualization—a case study on Zuozhuan knowledgebase[J]. Journal of Academic Libraries, 2020, 38(5): 72-80, 90.

[5] LI B, LI Y X, YANG Q, et al. From history book to digital humanities database: the Basic Annals of the Shiji[J]. Journal of Chinese History, 2020, 4(2): 528-536.

[6] 马创新, 曲维光, 陈小荷. 中文古籍数字化的开发层次和发展趋势[J]. 图书馆, 2014(2): 104-106.
MA C X, QU W G, CHEN X H. The exploitation hierarchy and development trend of digitization of Chinese ancient books[J]. Library, 2014(2): 104-106.

[7] 李明杰, 张纤柯, 陈梦石. 古籍数字化研究进展述评(2009—2019)[J]. 图书情报工作, 2020, 64(6): 130-137.
LI M J, ZHANG X K, CHEN M S. Review on the research progress of the digitization of ancient Chinese books (2009-2019)[J]. Library and Information Service, 2020, 64(6): 130-137.

[8] 毛建军. 古汉语电子语料库资源与类型概述[J]. 辞书研究, 2011(6): 83-93.
MAO J J. Overview of Ancient Chinese electronic corpus resources and types[J]. Lexicographical Studies, 2011(6): 83-93.

[9] 王大盈. 《中国基本古籍库》和《瀚堂典藏》两大古籍数据库比较研究[J]. 情报杂志, 2011, 30(S1): 157-158, 161.
WANG D Y. A comparative study of two databases: Chinese basic ancient books and HYTUNG BOOKS[J]. Journal of Intelligence, 2011, 30(S1): 157-158, 161.

[10] 董志翘. 为中古汉语研究夯实基础: “中古汉语研究型语料库”建设琐议[J]. 燕山大学学报(哲学社会科学版), 2011, 12(1): 1-6.
DONG Z Q. To lay a solid foundation for the study of medieval Chinese: on the construction of research corpus of medieval Chinese[J]. Journal of Yanshan University (Philosophy and Social Science Edition), 2011, 12(1): 1-6.

[11] 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, 81.
LAU K T, SONG Y, XIA F. The construction of a segmented and part-of-speech tagged archaic Chinese corpus: a case study on Huainanzi[J]. Journal of Chinese Information Processing, 2013, 27(6): 6-15, 81.

[12] 陈力. 数字人文视域下的古籍数字化与古典知识库建设问题[J]. 中国图书馆学报, 2022, 48(2): 36-46.
CHEN L. Digitalization of ancient books and construction of classical knowledge repository from the perspective of digital humanities[J]. Journal of Library Science in China, 2022, 48(2): 36-46.

[13] 唐振贵，罗锦坤. 中国古代时间本体：细化数字人文研究的时间轴向[J]. 图书馆杂志，2022，41(4)：87–95，37.
TANG Z G，LUO J K. Ancient Chinese time ontology：refining the time dimension of digital humanities research[J]. Library Journal，2022，41(4)：87–95，37.

[14] 包弼德，王宏苏，傅君劢，等. “中国历代人物传记资料库”(CBDB)的历史、方法与未来[J]. 数字人文研究，2021，1(1)：21–33.
PETER K B，WANG H S，MICHAEL A F，et al. The history，methods，and future of the China biographical database(CBDB) project[J]. Digital Humanities Research，2021，1(1)：21–33.

[15] 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.
QIAN Z Y，ZHOU J Z，JIA J. Research on knowledge base construction and website implementation of ChuCi[J]. Library Theory and Practice，2010(10)：70–73.

[16] 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.
XU C，CHEN X H. Social network analysis of spring and autumn period based on Zuo Zhuan[J]. Journal of School of Chinese Language and Culture Nanjing Normal University，2014(1)：179–184.

[17] 季培培. 常见10种古籍全文数据库的比较研究[J]. 图书馆学研究，2020(20)：71–80.
JI P P. A comparison study of ten normal ancient book textual databases[J]. Research on Library Science，2020(20)：71–80.

[18] 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.
ZHANG Q，JIANG C，JI Y S，et al. Unified model for word segmentation and POS tagging of multi–domain pre–Qin literature[J]. Data Analysis and Knowledge Discovery，2021，5(3)：2–11.

[19] 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.
LIU Z B，DANG J F，ZHANG Z J. Research on automatic extraction of historical events and construction of event graph based on historical records[J]. Library and Information Service，2020，64(11)：116–124.

[20] 张斌，魏扣，郝琦. 国内外知识库研究现状述评与比较[J]. 图书情报知识，2016(3)：15–25.
ZHANG B，WEI K，HAO Q. Review and comparison of research status of knowledge base at home and abroad[J]. Documentation，Information & Knowledge，2016(3)：15–25.

[21]《史记》修订组. 史记(点校修订本)[M]. 北京：中华书局，2013.
The Shiji Revision Group. The Shiji (revised version) [M]. Beijing：Zhonghua Book Company，2013.

[22] 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 M，LI B，CHEN X H. 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.

[23] 许嘉璐，安平秋. 二十四史全译·史记[M]. 上海：汉语大词典出版社，2004.
XU J L，AN P Q. A complete translation of Twenty–four History Books·The Shiji[M]. Shanghai：The Publishing House of the Chinese Dictionary，2004.

[24] 钱穆. 史记地名考[M]. 北京：商务印书馆，2004.
QIAN M. An investigation of place names of the Shiji[M]. Beijing：The Commercial Press，2004.

作者简介

郑童哲恒（1998– ），女，南京师范大学文学院硕士生，主要研究方向为计算语言学、数字人文。

李斌（1981– ），男，南京师范大学文学院副教授，主要研究方向为计算语言学、数字人文。

冯敏萱（1978– ），女，南京师范大学文学院副教授，主要研究方向为语言信息处理、语料库语言学、数字人文。

常博林（1999– ），男，南京师范大学文学院本科生，主要研究方向为数字人文、计算语言学、语料库语言学。

王东波（1981– ），男，南京农业大学信息管理学院教授、博士生导师，主要研究方向为信息智能处理、自然语言处理。

收稿日期： 2022-02-19

通信作者： 李斌，libin.njnu@gmail.com

基金项目： 江苏省社会科学基金项目（No.20JYB004）；国家社会科学基金资助项目（No.18BYY127，No.21&ZD331）

Foundation Items: The Social Science Fund of Jiangsu (No.20JYB004), The National Social Science Foundation of China (No.18BYY127, No.21&ZD331)