分类号：　k233　　学校代码：　10697
密　级：　公开　　学　　号：　201920529

数字人文视阈下的秦汉简牍文本挖掘研究
——以里耶秦简牍（一二卷）为例

学科名称：科学技术史

作　　者：朱琳　　指导老师：陈镱文 教授

西北大学学位评定委员会

二〇二二年

Research on Qin and Han Bamboo Slips Text Mining from the Perspective of Digital Humanities
——Take Li-ye Bamboo Slips (Volume I and II) as an Example

A thesis submitted to

Northwest University

in partial fulfillment of the requirements

for the degree of Master

in History of Science and Technology

By

Zhu Lin

Supervisor: Chen Yiwen Professor

June 2022

西北大学学位论文知识产权声明书

本人完全了解西北大学关于收集、保存、使用学位论文的规定。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。本人授权西北大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。同时授权中国科学技术信息研究所等机构将本学位论文收录到《中国学位论文全文数据库》或其它相关数据库。

保密论文待解密后适用本声明。

学位论文作者签名：朱琳 指导教师签名：陈鎏文

2022年 6 月 11 日 2022年 6 月 11 日

西北大学学位论文独创性声明

本人声明：所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，本论文不包含其他人已经发表或撰写过的研究成果，也不包含为获得西北大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：朱琳

2022年 6 月 11 日

摘要

简牍的出土，极大程度上改善了秦代史料奇缺的情况，为研究战国晚期至秦代的历史提供了丰富而可靠的资源，对当前的史学研究具有不可替代的史料价值。

简牍文献中文书占比较大，相较于静态的规则制度而言，实用文书中往往包含着更为丰富的信息，从不同的角度来对当时的社会面貌及规范进行了诠释。如何从简牍语料中挖掘出有效信息并且进行分析研究，逐渐被广泛应用的文本挖掘技术为我们提出了一个较好的解决方法。

本文着眼于利用文本挖掘手段研究简牍文本，将文本挖掘分析方法引入简牍研究领域，以里耶秦简一二卷为语料来源对其进行文本分析，探索如何提高简牍文本研究效率、改变简牍文本研究方式,从而实现对于简牍文本的快速把握及内容揭示，亦可促进秦代历史知识发现研究的发展。

本文的主要工作有：一是将简牍文本进行数字化并对其进行文本预处理相关工作。与简牍小组成员共同将里耶秦简一二卷的内容进行数字化处理，并通过两轮核对校验确保简文数据的准确性，本文结合简牍文本特征及后续实验的实际需要，对简牍文本进行了分词、去停用词及特征项权重计算等预处理工作。二是利用关键词提取及词频统计方法对里耶秦简一二卷进行关键词的共现分析。以排名前十的关键词为索引，分别提取与其相关的所有语料并对其进行词频统计，分析实验结果得到里耶一二卷中存在作徒簿文书、粮食管理类文书及债务文书。三是利用 TextRank 文本摘要模型进行摘要自动抽取。通过对简牍文本进行降维并抽取出其中的关键信息，从而以事件的角度来获取并把握简牍文本中的主题内容，实验共获得 6 条摘要，涉及到债务类文书、粮食管理类文书及律令类文书。四是利用主题模型计算简牍文本主题-词概率分布。文章采用 LDA 主题建模的结果之一，即主题-词概率分布体现里耶秦简一二卷文本的主题及具体内涵，从主题的角度来对简牍内容进行把握。引入困惑度作为模型评价指标，根据困惑度及人工反复实验选择以 12 个主题作为最优主题，且其中 4 个主题具有明显主题倾向。

本文研究认为，里耶秦简一二卷中主要以往来公文为表现形式，具体内容涉及文书邮传、债务、粮食、徒簿、案件、徭役等等内容，计算结果能够较好的概括其主要内容，并且与目前里耶校释第一卷中给出的主题内容相对应。本文对于简牍研究这一领域的文本挖掘研究具有一定的理论意义及参考价值，尤其是首次尝试将简牍和文本挖掘技术结

合起来，有效地提高了简牍资料的处理效率和信息获取效率，为简牍材料的研究提供了新思路，证实了可借助文本挖掘方法来理解、诠释中国丰富的文化遗产，具有一定的可行性及实践意义。

关键词：里耶秦简；文本挖掘；文本摘要；主题建模

ABSTRACT

The excavation of bamboo slips has greatly improved the shortage of historical materials in the Qin Dynasty, provided rich and reliable resources for the study of the history from the late Warring States period to the Qin Dynasty, and has irreplaceable historical value for the current historical research.

Bamboo slips and Chinese books account for a large proportion. Compared with the static rule system, practical documents often contain more abundant information and interpret the social appearance and norms at that time from different angles. How to mine effective information from bamboo slips corpus and analyze it, the text mining technology, which is gradually widely used, puts forward a better solution for us.

This paper focuses on the use of text mining means to study the bamboo slips, introduces the text mining analysis method into the research field of bamboo slips, analyzes the text with the first and second volumes of Li-ye Bamboo Slips as the corpus source, and explores how to improve the research efficiency of bamboo slips and change the research methods of bamboo slips, so as to realize the rapid grasp and content disclosure of bamboo slips, and also promote the development of the research of historical knowledge discovery of the Qin Dynasty.

The main work of this paper includes: first, the digitization and preprocessing of bamboo slips. Together with the members of the bamboo slips team, we digitized the contents of the first and second volumes of the Li-ye Bamboo Slips, and ensured the accuracy of the bamboo slips data through two rounds of verification. Combined with the characteristics of the bamboo slips and the actual needs of subsequent experiments, this paper carried out preprocessing steps such as word segmentation, de deactivation words and feature item weight calculation. The second is to analyze the co-occurrence of keywords in the first and second volumes of Li-ye Bamboo Slips by using the methods of keyword extraction and word frequency statistics. Taking the top ten keywords as the index, extract all relevant corpora and make word frequency statistics. The analysis results show that there are apprenticeship documents, food management documents and debt documents in riyer Volume I and II. The third is to extract the text summary of bamboo slips by using textrank text summary model. By compressing the text information of bamboo slips, the key information is extracted, so as to obtain and grasp the subject content of bamboo slips from the perspective of events. A total of

6 abstracts are obtained in the experiment, involving debt documents, food management documents and legal documents. Fourth, the topic model is used to calculate the topic word probability distribution of bamboo slips. From the perspective of probability distribution, one of the themes of the first volume is to grasp the theme of the second volume, that is, the theme of the second volume is to reflect the content of the text. The confusion degree is introduced as the evaluation index of the model. According to the confusion degree and manual repeated experiments, 12 topics are selected as the best topics, and 4 of them have obvious topic tendency.

This paper holds that the first and second volumes of Li-ye Bamboo Slips are mainly expressed in the form of official documents, and the specific contents involve documents, postal transmission, debts, grain, apprentice books, cases, corvee and so on. The calculation results can better summarize its main contents, and correspond to the subject contents given in the first volume of Liye's proofreading. This paper has certain theoretical significance and reference value for the text mining research in the field of bamboo slips research. In particular, it is the first attempt to combine bamboo slips and text mining technology, which effectively improves the processing efficiency and information acquisition efficiency of bamboo slips, provides new ideas for the research of bamboo slips, and proves that China's rich cultural heritage can be understood and interpreted with the help of text mining methods, It has certain feasibility and practical significance.

Keywords: Li-ye Bamboo Slips; Text mining; Text summary; Topic modeling

插图索引

表格索引

目录

第一章 绪论

1.1 研究背景

近年来出土的秦代简牍材料，其内容丰富，数量浩大，为研究战国晚期至秦代的历史提供了丰富而可靠的资源，其中也涉及秦朝时期的各类基层管理制度，如当时设立的法律制度、军事制度、户籍与土地管理制度等，对中国早期的社会制度研究具有重要的史料价值。

随着数字信息时代的技术更迭和信息化程度的不断提高为科学研究提供了更加多样的研究方法，从数字人文的视角来对历史文本进行深度的研究越来越受到学者们的重视。数字人文方法是将文本的研究与计算机技术相结合，借助信息处理技术为人文研究提供了极大地便利，为史学研究者提供了新的研究工具与范式。通过数字人文方法，可以对历史文本资料进行组织、挖掘、分析与应用。

相较于其他简牍，里耶秦简数量巨大且内容丰富，仅已出版的一二卷就有八万余字，而较早出土的睡虎地秦简与张家山汉简一共也仅有十万余字。除了文本数量优势外，更是少见的释读如此完整的简牍，已出版的《里耶秦简校释》一二卷对简文进行了深入的释读，为后续准确理解诠释实验结果提供了基础[1, 2]。除此之外，文本分词对后续的挖掘研究至关重要，但简牍中未登录字、残断问题突出，已有的分词研究不能直接适用于简牍文献，语料处理难度极大，而目前课题组同学已经针对里耶秦简一二卷的文本进行了分词，至此语料已具备进行文本挖掘研究的前期基础条件。

里耶古城是湖南通往川渝的必经之地，其地处多个乡镇的交界，更是当时的军事要塞，凭借其优越的地理位置，里耶古城的贸易往来十分活跃。里耶秦简出土于里耶镇里耶古城 1 号井，共 38000 余枚简，数据规模十分庞大，所载多为秦始皇二十五年至秦二世元年间的官署档案，涉及当时社会经济、政治、文化的各个层面。在秦代的正史记载中，关于其社会生活甚至整个秦代的行政制度内容都十分有限，因此里耶古城中出土的三万余枚简牍，通过文字形式生动形象的复活秦王朝的历史，从根本上改变了几千年来战国时期的秦汉学术研究面貌。相较于静态的规则制度而言，实用文书往往包含着更为丰富的信息，或者说从不同的角度来对当时的社会面貌及规范进行诠释，里耶秦简其作为实录性材料，对于研究秦代的历史文化、了解秦代社会全貌具有重要意义。但由于简牍特殊的保存环境，存在大量的散乱、残断简牍，不仅如此，其中更是主题庞杂，涉及

诸多人物、事件且文本晦涩难懂，给人工分析带来了很大困难，影响了简牍文献材料的开发与利用。

数字时代的到来使得利用信息处理技术对史料进行文本挖掘研究成为可能。里耶秦简牍文本目前已经具备利用文本挖掘技术对其进行分析研究的基础，通过对相关文本进行分析将不仅能够降低阅读难度，还可以有效提高主要信息获取与利用程度。利用文本挖掘技术不仅为简牍研究提供了一个新的研究范式，能够更好地梳理简牍文本内容，还能从大规模文本数据中挖掘新事实、产生新认识。

1.2 研究目的与意义

1.2.1 研究目的

此前学者们对于里耶秦简牍的研究已经积累了一定的成果，但是大多数研究都是基于传统史学研究方法，而并未涉及利用数字人文方法或信息处理技术发现简牍背后的隐藏知识。

本文的研究目的就是从数字人文的角度出发，运用文本挖掘的研究方法和范式将零散的简牍内容串联起来，为学者们研究简牍文本开辟一种新的方法，从而达到提高知识获取的效率，揭示内容关联的目的。将简牍研究与文本挖掘技术相结合，可以有针对性的提高简牍材料处理效率及有效性，以一种直接高效的方式来把握简牍的主要内容，为简牍材料的研究提供了新思路，具有一定的可行性及实践意义。同时对简牍中的内容要素进行分析，帮助学者更好地理解其主旨，进而借助里耶秦简牍中所记载的事件、主题、制度等对其背后所反映的秦代社会制度及发展产生更为深刻的认识。

1.2.2 研究意义

出土简牍是我国先秦时期文化、思想和智慧的载体，挖掘简牍文献中的信息对史学、人文学科及社会科学的研究有着十分重大的意义。简牍数据因其非结构化的特点难以被研究人员及计算机直接解读利用，但其中却隐含着许多有价值的信息。随着先秦时期的简牍材料不断面世，其文本内容研究分析必然需要提上日程，但从目前的情况来看，关于简牍材料的研究大多采用人文学科的传统研究方法，因此，开展数字人文视阈下的简牍文本挖掘研究，无论在理论上还是在实践上都有着重要的意义，它不仅可以丰富文献资源的开发与利用，而且对数字时代的简牍研究也有一定的促进作用。

本文提出利用文本挖掘方法来对简牍内容进行研究，对实现里耶秦简牍的文本活化及深入研究具有重要意义，实现了数字人文方法在简牍研究领域的初步应用，有利于高

效地从残断简牍中获取知识，使得对简牍内容的研究由点转向面，解决现有简牍研究大多聚焦于某个点的问题。不仅如此，利用数字人文方法进行简牍材料研究具有重要的研究意义和应用价值，相较于传统的人文研究方法，数字人文方法从深度和广度两方面扩展了人文研究模式，对后续史学研究的数字化提供参考和借鉴，同时也是数字化时代人文社科获得创新性成果的必然之路。

本文利用文本分析技术对《里耶秦简》文本进行主题挖掘研究，该方法打破了传统的人文学者研究模式，主要通过关键词提取、词频统计、文本摘要及主题模型等方法来对简牍材料中的社会制度、主题进行不同维度的提取和挖掘，并为后续的研究提供参考与借鉴。

1.3 国内外研究现状

近年来，各行各业的数据量都呈井喷式增长，使得如何从大规模的数据中获得有价值的信息成为当前的研究重点。而文本挖掘技术则随着计算机的飞速发展得到了空前的进步，逐渐成为了文本研究的主流方法论[3]。随着各种古籍数字化和文献数据库建设的项目层出不穷，计算机技术逐渐渗透到人文领域的各个方面,使得人文研究各个环节的效率都得到了有效提升[4]。大规模的历史文本研究作为与文本信息密切相关的领域之一，需要时刻跟进时代及科技的发展步伐，本文主要是分析在数字人文的大背景下，探究如何利用文本挖掘技术来对简牍文本进行分析，并且使用不同的挖掘方法进行尝试，因此，下面主要探索国内外针对历史文本的一些文本挖掘研究成果。

1.3.1 国外研究现状

文本挖掘作为多项技术的交叉领域起源于文本分类（text classification）、文本聚类（text clustering）和文本自动摘要（automatic text summarization）等单项技术，国外的文本挖掘研究相对较为成熟，20 世纪 50 年代时，文本分类和聚类就已经作为模式识别的应用技术崭露头角，其主要是针对图情领域的图书分类研究需求而提出的，1958 年 H.P.Luhn 提出了自动文摘的思想，为文本挖掘领域增添了新的内容。20 世纪 80 年代末至 90 年代初期，美国政府资助了一系列有关信息抽取（information extraction,IE）的研究项目，并组织了多次消息理解会议（Message Understanding Conference,MUC）使得信息抽取技术迅速成为领域热点。继信息抽取之后一系列文本处理技术相继产生并且得到了空前的发展，如情感分析（text sentiment analysis）、观点挖掘（opinion mining）、话题检测与跟踪（topic detection and tracking,TDT）等等技术。

近几十年来，国内外不断开展针对中国古籍的文本挖掘研究，并且研究对象主要集中于传世文献。加州大学开展的“东亚研究巨视显微镜”项目，主要利用主题建模等方法对《全唐诗》进行文本挖掘分析，通过分析得到文本中的潜在语义模式并建立了交互界面，使得我们可以利用其交互界面来对诗歌的主题进行分析[5]；R.Nichols 等人建立了一个 500 多万字的中国古代哲学语料库，并对其进行主题建模，利用文本中所蕴含的主题内容，对《论语》、《孟子》及《荀子》之间的关系进行解释验证[6]；西安交通大学与印第安纳大学合作构建了 Handian(汉典古籍)语料库，为了能够较好的解释和分析语料库中发现的主题，学者们针对该语料库专门开发了一个分析工具（InPhO Topic Ex-plorer）[7]；J.W.Chen 等人利用 Gephi 网络关系图、GIS 方法对《世说新语》中的人物、主题以及其中的地理空间信息进行挖掘与量化分析[8]。

不仅是主题挖掘模型，可视化工具也对人文社科研究影响巨大。英国艺术与人文研究委员会(AHRC)赞助的研究计划为我们探索古代世界的空间观念提供了一种新的途径，它通过抽取希腊历史学家希罗多德的历史著作中所提及的地方以及它们的地理特征，为我们提供了一个可视化的工具，其能够更好地挖掘出文字的拓扑结构[9]。北德克萨斯大学与斯坦福大学于 2010 年受到美国国家人文基金(NEH)的资助，将数字化后的美国旧报纸新闻文本作为原始语料，对原始语料进行信息抽取，主要抽取出其中的人名、地名等等实体信息，并利用可视化技术将其中的实体信息进行呈现，以此来直观研究随着时间空间的变化历史事件的演变过程及变化规律[10]。Cho 等人为了将罗马历史的时空与文本分析结合起来，研究开发了一个可视化分析系统（VAiRome），系统中不仅可以利用多种文本分析技术还有多种可视化视图呈现方式，通过可视化系统深刻剖析了罗马历史中的重要时间、地点及事件及其关系，能够为研究罗马历史提供极大便利[11]。2014 年 8 月，Schich 等人对公元前 600 多到二十一世纪间的十几万余名多个领域中的历史杰出人物进行研究，利用可视化工具对他们的出生及死亡地点进行分析，得到了三千年来欧洲及北美文化中心的发展转移趋势，绘制出了一幅文化史图[12]。

1.3.2 国内研究现状

在国内，将文本挖掘应用于传世文献领域的时间较晚，尚未形成一个完善的理论体系与测评标准，尽管如此，针对大规模的历史文本挖掘研究中还是取得了许多不错的成果。

典籍分类方面，王东波等人利用 SVM 模型对《论语》《老子》等 9 种先秦诸子典籍进行自动分类研究，实验结果表明模型的分类准确率达到了 99.21%，取得了很好的推

广和应用价值[13]。吴斌等人对中国唐诗进行情感分类，并将其方法应用于唐代和宋代各个时期的情感分析以及代表流派分析[14]。李晖等人从格律诗的情感方面入手，发掘出中国古代各个历史阶段的社会情感状况[15]。

典籍聚类方面，马创新等人对先秦各学派间的相关度进行了定量的考察，并对各学派的主题思想进行了统计和分析，从中选出了具有代表性的主题词，根据研究结果得到了诸学派中相关度最高及相关度最低的学派[16]。何琳等人通过聚类方法对《左传》文本进行了多个不同角度的定量研究，揭示了春秋时代各个诸侯国家和各个领域的发展状况[17]。除此之外，还对古籍文本进行了命名实体识别、社会网络构建及可视化等方面的研究。针对汉典古籍的主题建模项目，王小红等人不仅对其中人文知识的统计分析所带来的新的语义解读及文本阅读方式等问题进行了深入的剖析，还对计算机方法在人文领域中的应用难点进行了揭示[18]。申资卓等将与“八音”有关的诗词作为研究对象，提取出《全唐诗》、《全宋词》等历史文本中的相关诗句，利用 LDA 模型及 NMF 模型等方法，从整体到局部，通过多种视角来对其中的中国古典乐器进行分析研究[19]。

信息抽取方面，朱锁玲等针对物产和地名的自动识别问题，提出了一种全新的命名实体识别技术，在模型中融合了规则与统计方法，以地方志资料《方志物产》为语料，取得了较好的识别效果[20]。皇甫晶等对纪传体中的古汉语中文姓名进行识别，主要采用了规则方法对其进行研究[21]。李章超等人利用信息抽取技术对《左传》中的战争事件进行研究，抽取得到了春秋时期参战频率较高的国家如晋国、楚国等,并且实验同时得到了战争的具体信息如战争的主要进攻方及防守方等[22]。

可视化方面，欧阳剑等人创建了一个中国古籍实时统计分析平台，通过平台对数据进行实时分析，实现了实时、可视化分析字词的历史词频分布规律研究[23]。张建立等人尝试针对李白诗词中的创作地点、其中涉及到的及联想到的地名诗词中描述的地名及作者联想到的地名利用 GIS 技术进行地理制图，同时对其空间分布特点进行归纳分析，从而明确了诗词的创作的空间分布格局[24]。张玮等人根据文本的关联度与时空可视化分析的方法，通过对不同年代宋代词人的生平轨迹与生活背景进行对比分析，在此基础上得出不同时代和经历的宋词在文本主题上的相关性和独特性[25]。

近年来国内外的学者们大多是将文本挖掘技术应用于传世文献的研究中，而针对出土文献则更注重相关数据库的建设，例如香港中文大学中国古籍研究中心制作的“汉达文库”，台湾中研院历史语言研究所制作的“简帛金石资料库”等等[26,27]，均收录了大量简帛资料，但甚少涉及利用文本挖掘技术对简牍文本进行分析。

由上可知，文本挖掘技术的蓬勃发展已经为大规模的历史文本分析提供了坚实的基础，学者们也已经成功利用计算机技术对《全唐诗》、《全宋词》、《汉典》、《左传》等传世文献进行文本分析研究，《里耶秦简》作为实用文书，对研究秦代历史具有及其重要的研究价值，然而简牍自身的特性以及其中庞杂的主题、事件、人物交织，为人工分析文本带来了巨大的挑战。

且目前针对《里耶秦简》的研究，基本上都是人文学科领域内的专家通过人工阅读等等方法进行分析，且主要集中于对残断简牍的缀连复原、简文的释读以及对秦代基层的职官、文书、土地和赋税等制度进行研究，这种方法对于大规模的简牍文本研究非常不利。在文本挖掘技术发展迅猛的今天，通过文本挖掘技术对简牍文本进行主题分析研究，能够更高效便利地挖掘出简牍中隐含的知识，对于历史研究领域具有重要的意义。

本文从《里耶秦简》一二卷简文文本数据出发，利用文本分析技术对里耶秦简文本进行研究，该方法有别于传统的人文学者研究模式，主要通过关键词提取、词频统计、文本摘要及主题模型等方法对简牍材料中涉及的主要制度内容进行不同维度的提取和挖掘，并为后续的研究提供参考与借鉴。

1.4 研究框架

根据目前的史学研究可知，里耶秦简中记录的内容多为遷陵县廷的来往文书，具体往来单位包括其上级洞庭郡府及下属司空、仓官、田官诸署以及都鄉、啓陵、贰春三乡等，主要内容涉及郡县与官署设置、官吏的考核步骤、赋税、徭役、诉讼、作物、邮传等等[1]2。

本文的研究框架如图 1 所示，首先选取已出版的《里耶秦简校释》第一卷与第二卷作为原始语料与简牍小组成员共同将一二卷的内容进行数字化处理，并通过两轮核对校验确保简文数据的准确性，结合简牍文本特征及后续实验的实际需要，对里耶秦简文本进行了分词、去停用词及特征项权重计算等预处理步骤，将简牍语言转化为结构化文本数据。在此基础上，首先运用特征词与词频统计对里耶秦简一二卷进行关键词的共现分析，以排名前十的关键词为索引，分别提取与其相关的所有语料，对其进行词频统计分析；其次利用 TextRank 文本摘要模型对简牍文本进行自动摘要提取，得到完整的具有代表性的事件，以事件的角度来获取并把握简牍文本中的关键信息；最后利用 LDA 主题模型的建模结果之一即主题-词概率分布来得到里耶文本的主要主题及具体内涵，从主题的角度来对简牍的主要内容进行把握，并且引入困惑度作为模型评价指标，根据困

惑度及人工反复实验选择最优主题数，对里耶秦简牍材料中的主要内容及制度进行词、事件、主题等不同维度的挖掘。

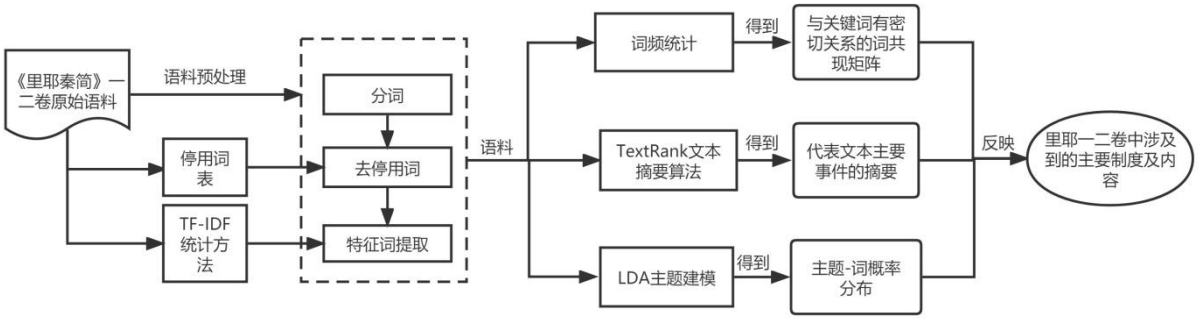

图 1　研究框架

1.5 论文组织结构

全文共分为五章撰写，以下为各章节的具体内容：

第一章：绪论。本章首先阐述了数字人文视阈下关于里耶秦简牍文本挖掘研究的研究背景，其次介绍了本研究的目的及意义、国内外的相关研究进展、本文的主要研究框架，最后对本文的组织结构进行阐述。

第二章：相关理论与方法。本章主要介绍了研究中涉及的主要理论知识，首先对数字人文及文本挖掘进行阐述，其次对本文研究中用到的模型即文本摘要模型和主题模型进行介绍，主要包括模型的概念、理论基础等等基础知识，为后续的文本挖掘实验提供理论基础。

第三章：数据收集与预处理。本章首先介绍了实验的原始数据来源及处理过程，其次对出土简牍文本的数据预处理难点进行阐述，最后进行分词、去停用词及计算特征项等等预处理步骤。

第四章：里耶秦简文本挖掘实验过程。本章首先通过词云图等方法对数据进行描述性分析，其次分别介绍词频统计、文本摘要及主题模型的实验流程、具体实验过程及实验结果分析。

第五章：总结与展望。对本文所做的关于里耶秦简与文本挖掘分析进行总结，指出本研究的不足之处，并对利用文本挖掘技术进行简牍研究的未来进行展望。

第二章 相关理论和方法

2.1 数字人文与文本挖掘研究

2.1.1 数字人文概述

数字人文(Digital Humanities，DH)，也被称为人文计算(Humanities Computing)，它将现代的计算机技术引入到传统的人文研究和教学领域当中，是一种新型的跨学科研究方法。不仅如此，数字人文以一个全新的角度审视数字化趋势与人文研究之间的联系以及其对人类社会的影响，其研究项目及团队不仅包含了传统的人文领域如哲学、历史学等学科的学者，同时也涵盖了计算机方面的专家，为人文科学学者对大规模数字资源进行组织及分析挖掘等方面提供了新的工具和平台。

数字人文的概念最早起源于 20 世纪中期，神父罗伯特·布萨（Roberto Busa）于 1949 年与 IBM 公司合作，首次利用计算机手段对拉丁文神学作品进行数字化编辑并建立数据库，这是学界公认的标志性事件[28]。上世纪 90 年代后，人文计算的应用版图不断扩大，随着计算机手段的蓬勃发展逐步从传统的语言学辐射至文学、历史学、艺术学、社会学等等相关领域，并且随着网络的广泛应用而逐步走向成熟[29]。可以说人文计算的产生和扩展，为各种信息资源的数据库的构建打下了坚实的基础，同时大大的提高了人们对文本数据进行分析、整合及利用的能力。随着互联网的兴起和运用，学者们不断扩大人文计算的研究范围和应用范围的，同时也对人文计算自身提出了更高的要求。至此，“数字人文”就产生了，其可以解释和包含更多复杂的人文计算和应用场景。进入二十一世纪，随着约翰娜·德鲁克在 ACH/ALLC 大会上做了题为 Reality Check: Projects and Prospects in Digital Humanities 的发言，于是“Digital Humanities”这一术语便成为了人文计算的升级概念，开始在英语领域中频频出现，并且逐渐向如超文本、图像、视频、音频和虚拟现实等多媒体领域应用发展[30]。同时，《数字人文指南》也在同一年问世，它是世界上第一本以“数字人文”命名的书籍，从那时起，“数字人文”一词悄然兴起，成为了一个全新的跨学科领域研究代名词。伴随着新兴互联网技术的飞速发展，数字人文方法在各个学科方向的应用和扩展越来越广，不仅如此还引发了一股更为广阔的实践浪潮。

总体来说，数字人文目前仍然处在一个不断发展和进步的过程，学者们根据各自不同的视角对数字人文方法有多种不同的认识，但是目前仍未有权威的定义。有学者认为

虽然目前给予数字人文概念一个精确的定义还是非常困难，但是我们可以从当前的实践中不断进行总结，认为数字人文是一种从源头上不断对人文知识的发现过程进行改进的方法，且其本质上是一种人文研究方式的创新与升级[31]。

综上，数字人文是一种将数字技术与人文社会科学的理念融合在一起的新型研究方法，并利用这种新方法对资源进行挖掘、整合、利用。其以数字文本资源为基础，人文研究需求为导向，以数据分析方法为手段，同时具有跨机构、跨学科、开放性、协作性等显著特点[32]。

2.1.2 文本挖掘概述

文本挖掘（Text Mining,TM）, 又被称为文本数据挖掘（Text Data Mining，TDM）或文本知识发现（Knowledge Discovery in Texts，KDT），是从海量的无结构的文本信息中去发掘潜在可能的数据模式、内在联系、规律、发展趋势等，得到散落在文本各处有价值的知识信息，并且利用这些知识更好地组织信息的过程[33]。文本挖掘并不是单指某种单一的技术或理论，而是融合了多个领域的知识技术如数据挖掘、机器学习、数据库等。它是一个通过融合多种计算机技术，抽取出文本中的重要信息进而挖掘文本知识的过程。

文本挖掘实质上属于数据挖掘的一部分，只是将研究的对象从结构化数据拓宽至非结构化数据，能够很好的帮助我们从大规模的无结构数据中发现新的知识、模式、规则及趋势，目前其已经广泛应用于多个领域。

通常一个完整的挖掘分析过程如图 2 所示，主要步骤包括预处理、挖掘分析、信息展示等，其中每一个步骤都包含了多种文本处理与挖掘技术，如预处理过程中的分词、特征表示、特征提取等，其后涉及到如文本结构分析、文本摘要抽取、文本分类及聚类、可视化技术等。通常一个完整的实验中需要融合多项文本挖掘算法而不是仅仅使用某种单一技术。

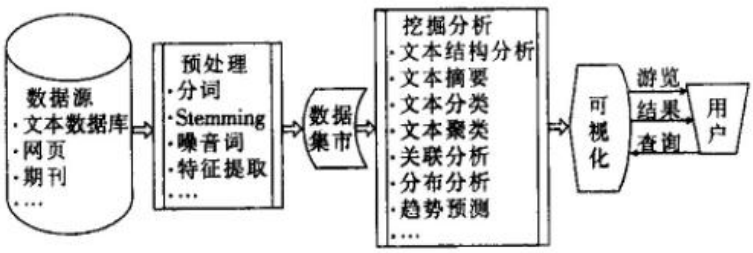

图 2　文本挖掘的一般流程

数据挖掘的起点是数据集，文本挖掘过程也不例外。文本挖掘的研究对象是各种形

式的文本数据集，所以文本挖掘的第一步便需要考虑如何将获取文本数据，其来源可以是各种文本数据库、网页或期刊等等。

在获取数据源之后需要对数据进行预处理，即将不规则的文本数据处理为结构化的数据。文本挖掘处理中文与英文文本的不同之处在于需要对中文文本进行分词。根据待挖掘文本数据集的不同在多种分词方法中选择一种将其准确切分为一列有意义的特征项，随后根据一定的规则对特征项进行筛选，从而得到待挖掘文本对象的特征项集合。

形成特征项集合后需要利用特定的数据模型，将文字特征集合转换成结构化的数据，便于计算机进行识别与分析。然而，由于分词后的文本特征项通常具有较大的维数，因此还需要对其进行特征项的维数进行缩减。在进行维数缩减后，我们需要明确文本挖掘的目的和需求，并根据需要对文本挖掘的方法和算法进行选择，从具体的应用目标中抽取出相应的知识模式，可能是分类、聚类等知识模式。

最后是对挖掘成果的有效性评价。如果实验的结果达到了预期的目的和工作需求，那么这种知识模型就会被保留下来，供类似研究使用，若未达到目标则回到之前的研究阶段，寻找失败的原因，然后针对性的进行研究分析并改进，直到得到符合预期和工作需要的知识模型。

文本挖掘的分类标准根据其应用的学科、技术、成熟度等等而不同，但是由于其研究过程中步骤繁杂，对其进行全面而又准确的分类十分困难，甚至在某些情况下根本无法明确其应用属于哪一类学科。郭金龙等人根据挖掘的层次对其各种应用进行分类，即将文本挖掘应用分为简单的初级挖掘和深层的高级挖掘，将是否产生了新的知识模式作为标准[34]。其中，初级挖掘是指较为传统的词频统计与关键词提取等等挖掘方法，高级文本挖掘中包括文本的分类、聚类、信息抽取、本体构建、可视化方向等。

2.2 文本摘要

2.2.1 文本摘要的定义

文本摘要（Text Summarization）是一种信息压缩技术，通过对文本信息进行压缩，来抽取其中的关键信息，在保留绝大部分信息的前提下压缩文本量[32]。通过摘要抽取可以在一定程度上降低需要深入理解的简牍文本数据量，同时提高关键信息的获取效率。自动摘要是指通过非人工方式自动生成文本摘要的过程，随着自动文摘技术的发展，人们逐渐摆脱了效率低下的人工摘要过程，具有很大的应用价值和研究意义。其根据不同的标准可以分为不同的类型。

按照数据有无标注，可以分为有监督（Supervised Text Summarization）和无监督（Unsupervised Text Summarization）两种。有监督模型需要对训练数据进行标注，即文本需要有“参考答案”，监督模型需要不断去学习训练文本与“参考答案”之间的关系。而无监督方法中，直接对文本进行检索得到其有效信息进而生成摘要即可，无需使用任何有标注的训练数据。相较于有监督模型，无监督模型的构建成本较低，且不需要通过大量人力物力去标注数据，但其相对来讲整体效果不如有监督模型精准。

根据文本的数量，能够分为单文档摘要抽取方法（Single Document Summarization）或多文档摘要抽取方法（Multi Document Summarization）。单文档摘要是指其抽取对象是单个文档；而多文档摘要的抽取是指从文章集合中抽取能够概括多个文章中心内容的摘要。相较于单文档摘要，多文档摘要的文本数据量更大且长度较长，更有可能出现多个中心点，因此对其进行自动抽取的难度更大。

按照摘要的生成方式又可以分为生成式方法（Abstractive Text Summarization）和抽取式方法（Extractive Text Summarization）。前者是通过对输入的文档的理解而形成的，模型根据理解文本内容来生成结果，为了接近摘要的本质，可能会生成原文中没有的词语。后者是通过对文本进行分析，根据原始文本提取出的关键信息形成摘要结果。采用这种方法所得到的摘要通常会保留原文章的显著信息，并且有着正确的语法。

2.2.2 自动文摘方法介绍

近年来随着自动文本摘要抽取技术不断丰富并且逐渐成熟，目前主要有基于统计的方法、基于机器学习的方法以及基于图模型的方法等等。

长期以来，在自然语言处理领域乃至整个计算机领域中，与统计学相关的技术得到了非常广泛的应用。同样，最早的自动摘要技术研究也是基于统计学的。与其他方法相比，基于统计学方法的原理和步骤都相对较简单，首先需要计算文本特征，主要通过数据中的词或句子来提取得到若干权重较高的句子，也是文本中最能反映文本特征的句子，最后通过这些句子组成最终的文本摘要。其过程如图 3 所示。

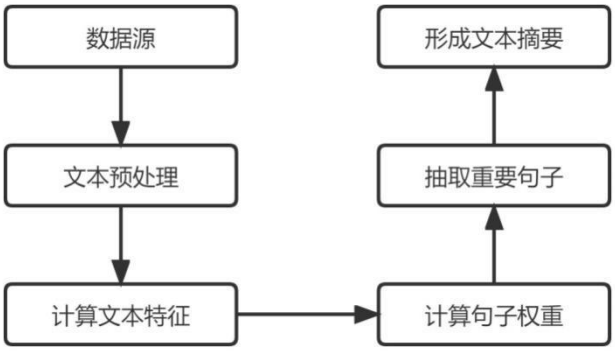

图 3　基于统计方法的自动文摘过程

基于统计的自动摘要方法虽然计算简单且使用十分便利，但这种方法主要取决于文本的统计特征，并且在文本的特征提取方面，大多数时候仅仅关注词语和句子的表面特征上，而很少涉及到深层次的文本语义信息。因此根据统计方法所提取得到的句子有可能会不连贯的且无法高度概括文本，故而其更适合于对固定格式文本的摘要进行抽取，在实际的应用中仍有一定的局限性。

随着近年来自然语言处理技术的广泛应用，机器学习技术逐渐成为了解决自然语言处理问题的重要方法，机器学习在机器翻译等应用领域上的发展为其在文本摘要自动抽取的研究打下了良好的基础，学者们逐渐将人工标注数据集的方法应用至自动文本摘要任务上，使得文本摘要的效果显著提高。为了更好的应用文本中的语义信息，机器学习方法需要将数据集进行人工标注，具体表现为对数据集中的所有文档进行人工标注，将“非摘要句”和“摘要句”分别进行标注；随后需要提取已标注好的句子的特征，机器学习方法实现自动文本摘要最为关键的就是在训练集中学习的过程，提取特征的准确程度将直接影响实验结果；最后选择合适的模型进行训练，实验结果表示为此句是摘要句的概率，再根据概率得分的排序抽取文档中句子组成文本摘要。目前在自动摘要方向使用较为普遍的机器学习模型有朴素贝叶斯模型、决策树模型及隐马尔可夫模型等等。

除以上两种方法之外，还有基于图排序的算法。其主要实现方法是将文档分为句子集合，以集合中的句子作为图的顶点，再以句子间的相关性大小作为图模型的边，建立得到一个以文档为基础的无向图，其后利用算法来计算图模型中边的权值及顶点权重，即计算句子相似度和句子权重，最后将计算结果按权重大小排名，将排名靠前的句子输出形成摘要。在实际的应用过程中，基于图模型的文本摘要抽取方法的具体实现流程主要分为文本预处理，图模型构建，图顶点权重计算，根据结果选取摘要句等步骤，在任一环节中都需要各类算法的高效支持，最终得到较为理想的结果。基于图排序的方法主

要基于构建图模型来准确得到文档中句子之间以及单词之间的关系，最终能够计算得到较为精确的句子权重，且其实现方法简单高效，能够较好的提升抽取式文本摘要的最终效果。

2.2.3 TextRank 文本摘要模型及其理论基础

在自动文摘领域几十年的发展过程中，相关的技术和方法层出不穷，其中 TextRank 算法作为自动文本摘要方向的经典算法，主要通过构建图模型的方式迭代计算文本单元的权重，得出文本集中相关概率分布。对于自动文本摘要的抽取都具有较高的应用价值。

TextRank 算法是基于 Google 的 PageRank 算法改进的，经过改进后是一种较为适合做文档处理的算法[35]。其通过对图模型的构建来凸显文本中的特征，首先将单个文档按照一定的规则划分为若干文本单元，然后以文本单元为图顶点，各顶点之间的联系即文本单元的相似度为边形成相关图结构，采取类似投票的方法来对图中的各顶点权重进行计算并排序，最后实现关键词及文本摘要等的选取[36]。TextRank 文本摘要算法方法具有如下优势：首先不需要大量的文本，仅需单个文本即可进行实验，其次是不需要预先训练，这一点与传统的条件随机场模型、决策树模型等有监督模型不同。

在基于 TextRank 算法的自动摘要抽取过程中，其构建的一般模型可表示为一个以句子为图中的文本单元顶点，句子相似度为各顶点之间的边权重，在具体实现上该模型由顶点集 V 以及边集 E 组成，图中不同顶点间的联系表示为 ω，对于某个顶点 V_i 来讲，$In(\mathrm{V}_i)$ 和 $Out(\mathrm{V}_i)$ 分别为指向该点集合和该点指向的点集合，顶点 V_i 权重的计算如式(2.1)所示：

$$WS\left(V_i\right)=\left(1-d\right)+d\times\sum_{V_i\in In(\mathrm{V}_i)}\frac{\omega_{ji}}{\sum V_k\in Out(V_j)\omega_{ji}}WS(\mathrm{V}_j) \tag{2.1}$$

其中 d 为一个节点指向另一个节点的概率，其值在 0 至 1 之间，通常为 0.85。V_i 为图模型中任意节点，V_j 为指向顶点 V_i 的一个集合，V_k 为经过顶点 V_j 连接出去的所有顶点集合，w_{ji} 为顶点 V_i 与顶点 V_j 之间的边的权重，$WS\left(V_i\right)$ 为顶点 V_i 的计算得出的权重结果。若要计算图中各个文本单元的权重，则需要先对相应的图顶点赋予一个初始权重，再根据公式迭代计算直至收敛。而 TextRank 摘要提取的具体步骤如下所示：

（1）文本预处理。包括分词、去除停用词、词性标注及词位标注等预处理步骤，将文本分词后的结果作为词的集合，此集合即为图模型中的顶点集合。

（2）计算边权重，即计算句子间的相似度。其计算方法有很多，包括但不限于余弦相似度、编辑距离、杰卡德系数、TF-IDF 计算等等计算方法，其计算公式如下所示：

$$\text{Similarity}(S_i, S_j) = \frac{|\{w_k \mid w_k \in S_i \,\&\, w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \tag{2.2}$$

（3）句子权重的计算。将计算结果及各数值代入权重计算公式即式（2.1）进行计算，得出文本的权重值。

（4）形成摘要。将计算得出的文本权重值进行排序，按照排序结果选择句子形成摘要。其中基于 TextRank 文本摘要算法构建的图模型如图 4 所示[37]。

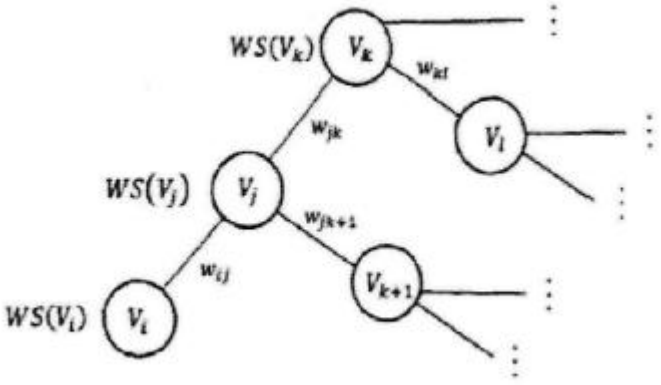

图 4　TextRank 图模型展示

TextRank 文本摘要算法其为图模型的代表，具体的实现基础是对文本的图模型进行构建，在进行文本摘要的抽取时将文本抽象为图模型，且无需对数据集进行人工标注，能够对句子结构进行综合考虑，其实现较为简单，也具有较强的有效性及普适性，这也是其应用广泛的原因。综上，本文利用 TextRank 文本摘要算法将里耶秦简的主要信息从庞杂的语料中抽取出来，形成抽取式摘要，以摘要计算结果来对简文的主要探究简文的主要内容。

2.3 主题建模

2.3.1 主题和主题模型

想要理解主题模型的相关知识，首先需要明确主题的概念。主题大多时候被人认为是主要内容，亦指在文艺作品、社会活动中体现出的中心价值思想。除此之外，主题还与艺术家本身的经验和价值判断紧密相关，主要的表现形式为将文本的主要内容表示为一系列的相关词项。而主题模型就是通过主题建模方法来探寻文本中的隐含主题[38]。

主题模型为无监督模型，它通过无监督的方式对文本中隐含的语义结构进行分析，其本质上是一种文本聚类和降维的处理方式，在模型建立过程中，首先会在主题与词汇

之间建立关联，其次再按照主题对文档进行分类。相较于更为传统的文献计量方法，主题模型能够对文本中的隐含主题进行发现、跟踪及预测等等，能够对文本间的语义关系进行深度挖掘，研究的文本包括新闻报导、文学作品及科技类的文献等等。

主题模型实质上就是一种通过概率生成模型来对文本中的隐含主题进行建模的方法，模型首先需要对语料库设立假定条件，基本思路是假定在文本中存在多个独立的隐含主题，然后利用该主题的概率分布，得到语料库各文档中的全部词语，由此将该文本视为一个特定的隐含主题的分布[39]。除了应用于主题挖掘方面，其在文本检索、文本分类、社会网络分析等多个方面亦有着广泛的应用。

主题模型的前身是潜在语义分析模型（ LSA, La-tent Semantic Analysis），该模型通过将奇异值进行分解从而将一个高维的文档向量映射至一个低维的潜在语义空间中，不仅能够降低维度同时也能解决文本当中普遍存在的同义、一词多义等等问题[40]。在潜在语义分析模型基础上，Hofmann 将概率统计方法引入了主题模型之中，提出了概率潜在语义分析模型(PLSA, Probabilistic Latent Semantic Analysis)，概率潜在语义分析不仅能够巧妙的避开复杂计算，更是第一个完整意义上的概率主题模型 [41]。但是在潜在语义分析模型中，缺少了混合权重的假设，导致模型中的参数取决于某些特定文档, 从而产生过拟合的现象。为了能够更好地表示文本生成过程，在 2003 年 Blei 等人提出了 LDA 模型，将其中的参数作为随机变量，从而将用来控制参数的超参数引入至模型中，最终实现彻底的“概率化”[42]。

在话题演化的基础上，Blei 等人提出了 DTM(Dynamic Topic Model)动态主题模型，该模型在 LDA 模型的基础上，假设其中的超参数以时间为序不断变化，并且后面时间的超参数取决于前一时间的超参数，从而反映出生成的主题随时间动态变化的特性[43]。随后 Alsumait 等人提出一种新的主题模型，即 OLDA(On-line LDA)模型，其能够对动态文本流实时进行主题建模，OLDA 模型能在文本流更新时利用现有的主题模型数据，逐步对现有模型进行增量式更新，且不需要重复访问之前存储的数据，该模型能够较好地实时获取文本中随时间变化的主题结构[44]。

2.3.2 LDA 主题模型定义及理论基础

LDA（Latent Dirichlet Allocation）是 D.M.Blei 于 2003 年在 PLSA（概率潜在语义分析）模型的基础上提出了一种更加泛化的文本主题模型，称作潜在狄利克雷分布，模型主要通过无监督方法来探寻文本中的隐含主题。

LDA 主题模型是一个典型的三层的贝叶斯结构，其中主要为词项、主题和文档三

层。模型对于语料中的任意文档，都具有一个固定的生成步骤：首先，在每篇文档中，选择一个主题；然后，在前述被选择到的主题中抽取相应的词汇分布；最后，将以上步骤进行多次重复，直到循环到每一个词。在这种情况下，文档到主题、主题到词均遵循多项式的分布。

其文本信息中每个词语的出现概率为：

$$P(\text{词项/文档})=\sum \text{主题}P(\text{词项/主题})*P(\text{主题/文档}) \tag{2.3}$$

将公式用矩阵表示如下图所示：

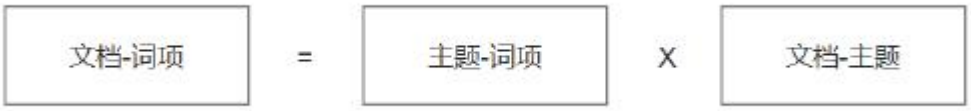

图 5　文档-主题-词矩阵表达

“文档-词”概率分布表示的是每个词项在某篇文档中的概率分布；

“主题-词”概率分布表示的是每个词项代表某个主题的概率分布；

“文档-主题”概率分布表示的是每个主题出现在某篇文档中的概率分布。

模型以词袋模型为前提，可以将文本中的有效文本信息变成模型建立所需要的数字信息。下图为 LDA 模型的图模型。

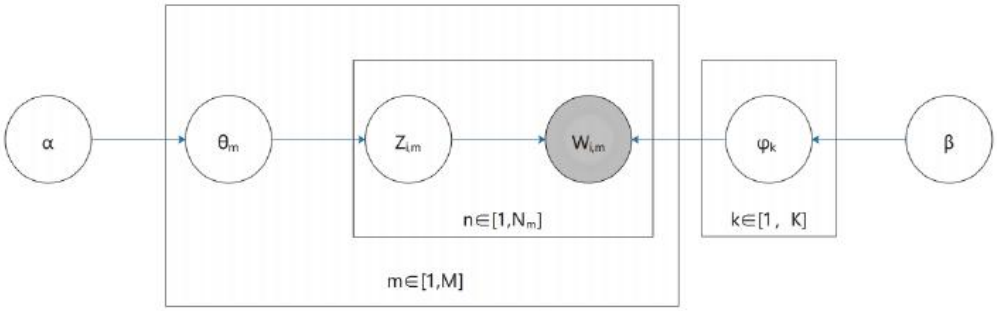

图 6　LDA 图模型

其中 M 为文档数量；K 为主题个数；V 为词袋长度；N_m 为第 m 篇文档中的词的数量；α、β 均为超参数，分别为每篇文章的主题分布及每个主题的词分布的先验分布 Dirichlet 分布的参数，通常为手动设定；θ_m 为第 m 篇文章中的主题的概率分布，其中 θ 为一个 $M*K$ 的矩阵；$Z_{i,m}$ 为第 m 篇文章第 i 个词被赋予的主题；$W_{i,m}$ 为第 m 篇文章第 i 个词；φ_k 表示第 k 个主题的词分布，其中 φ 为一个 $K*V$ 的矩阵。

LDA 模型根据词汇在每篇文档当中共同出现的频次来挖掘得到文本中所潜藏的语义结构，通过词袋假设模型来解决文本中存在的一些不确定性及噪声的干扰等等问题，

同时主题模型也被视为一种聚类算法或数据降维的手段，在文本挖掘的诸多任务上得到了广泛而成功的应用。

根据 LDA 模型的原理可知，模型生成过程中一般采用 Gibbs 及 EM 算法等进行参数估计，其中已知参数与未知参数的联合分布为：

$$p(w_m, z_m, \theta_m, \varphi|\alpha, \beta) = \prod_{n=1}^{N_m} p(w_{m,n}|\varphi_{z_{m,n}}) p(z_{m,n}|\theta_m) p(\theta_m|\alpha) p(\varphi|\beta) \tag{2.4}$$

本文利用里耶秦简特征词语料的主题建模结果之一，即主题-词概率分布来获取里耶秦简的主要主题及具体内涵，从主题的角度来对简牍内容进行分析。

2.3.3 LDA 主题模型最优主题数的确定

在 LDA 主题模型中，主题数目是模型结果的关键因素，若主题数目过少，可能会将一些不相关的主题合并在一起，若主题数量过多，则会使相关内容被分裂为独立的主题，从而产生大量的无关主题。在主题模型中，最优主题数目的确定我们通常使用困惑度指标，若困惑度的计算结果较小则可以证明模型得到的概率分布较为合适[45]。

困惑度通常是指模型的不确定程度，其主要指对于某篇文档 d 来说，不确定其为某个主题的程度，在其他的参数及条件不变时，过多的主题数目会带来困惑度数值过小，从而产生过拟合的问题。所以确定最优主题数目的重要参考标准是适当大小的困惑度数值。在 LDA 主题模型中，困惑度计算公式如下：

$$Perplexity(\mathrm{D}) = \exp\left\{ -\frac{\sum_{d-1}^{M} \log p(\mathrm{W}_d)}{\sum_{d-1}^{M} N_d} \right\} \tag{2.5}$$

其中，D 为测试集，M 为文档的具体数量，N_d 表示每篇文档 d 中的词语数量，W_d 为 d 中的词，$p(\mathrm{W}_d)$ 即为 W_d 的产生概率。

本文最优主题数目的评价采用困惑度与人工验证相结合的方法。由于简牍文本量较小，且仅根据困惑度指标最小来确定最优主题数目存在选取数目偏大的问题，这会致使我们抽取的结果之间辨识度不够等问题，因此本文采用人工进行辅助验证。

2.4 小结

本章主要介绍了后续研究中所涉及到的理论及模型，主要包括数字人文及文本挖掘的概念、文本摘要模型及主题模型等。首先对数字人文及文本挖掘的概念进行阐述，包

括定义、起源、发展及其应用。其次对本文中所涉及主要模型的基本理论进行阐述，分别介绍了文本摘要模型及主题模型的概念、TextRank 文本摘要模型的定义和理论基础以及 LDA 主题模型的定义、理论基础及最优主题数的确定方法。本章的理论知识等内容为后续的实证分析提供了坚实的基础。

第三章 数据收集与预处理

3.1 原始数据收集

本文实验数据来自于里耶秦简，其为2002年在湖南省龙山县里耶古城出土的秦代简牍，共约38000余枚，总计约20余万字。湖南省文物考古研究所计划共编著5卷《里耶秦简》，现已出版前两卷。其中第一卷包括一号井第5、6、8层出土的简牍，总计2627枚；第二卷录入第九层出土的3423枚简牍，两卷共计6050枚简牍。武汉大学历史学院作为湖南省文物考古研究所的合作方，几乎同时出版了《里耶秦简牍校释》一二卷，其中汇集了大量的字词校释及简牍缀连研究成果。

研究中使用的原始语料是来源于湖南省文物考古研究所出版的《里耶秦简》一二卷，首先将其进行数字化处理，同时为保证语料质量及完善程度，课题组参考陈伟团队推出的里耶秦简牍校释一二卷，对简文内容进行两次校对，并采用其中新的校释及改释结果。经过校对后的语料共八万余字。

3.2 简牍语料预处理的难点

由于简牍文本中存在大量未登录字及残断情况，为语料的预处理带来了极大的困难。简牍材料的挖掘研究较少的原因在于运用现代化计算机处理技术处理简牍文本时，简牍文本存在的一些自身属性大大增加了研究难度。以《里耶秦简》为例，一是由于简牍的保存环境使其存在较为严重的残断和破损情况；二是简文中存在大量的未登录字及无法辨识的字；三是对于简牍语料的语义还是存在诸多争议；四是简牍语料的分词步骤存在较大难度。

3.2.1 简牍残断

简牍语料与传世文献的载体不同,简牍记载于竹简之上,由于其保存的环境较为特殊,多存于井中,使得其出土后存在较为严重的残断及破损情况,极大的影响了简牍语料的数字化应用，本文采用的语料中，残断或字数不能确定时，用“……”表示，简牍残缺可确定有文字残去时，用字符“☒”表示。不仅如此，简牍内容的书写方式与现代汉语有较大的不同，采用了分栏、分行的书写方式，加剧了文本处理的难度。图7为湖南省文物考古研究所计划编著5卷《里耶秦简》中的第一卷部分内容，可以明显看出其中的残断及分行书写。残断的存在使得简牍语料内容表达不完整且多为短文本，极大的影

响了文本的数字化应用及利用计算机手段进行文本分析的难度。

图 7　里耶秦简部分原文

3.2.2 未登录字

未登录字，也称为集外字，是指在基于词典分词系统中未曾收录到的词语，同时也指在训练集中从未出现过的字[46]。

在过去的数十年里，古籍数字化的研究和实践已经取得了巨大的成就，并涌现出一大批具有一定影响力的古籍数字化工程。但是绝大部分的古籍数字化工程仅仅能够实现查询、浏览等功能，无法对文本进行更进一步的分析处理。限制其发展的最大原因还是古籍存在中大量的避讳字、异体字、少数民族文字、草体字等，对古籍的数字化与检索均有较大的制约[47]。目前，对于古书中的集外字，一般有替代、造字法和描写法三种[48]。替代法是把集外字转换成如符号、图形或者集内字等等可以录入的格式；描述方法是用一个字符串来描述集外字的字形；造字法是将集外字定义于字符集中的自定义区域中，同时根据集外字字形进行对应。

但目前还未有学者针对简牍中出现的集外字进行处理，本文原始语料中存在大量的未登录字，这些未登录字计算机无法识别，也无法对其进行分析，表 1 为小组整理得到的里耶秦简集外字表的部分内容。

表 1　《里耶秦简》部分集外字

集外字	简牍

宽	8-458，8-760，9-177，9-1164，9-1547，9-2041，
𥫱	8-461，
犴	8-474，8-565，8-575，8-763，8-765，8-800，8-1239，8-1286，8-1360，8-1580，9-813，9-948，9-99，9-1454,9-2334，
雔	8-487，8-1351， 9-1099，9-1139，9-1271，9-1451，9-1766，9-2067，9-2203,9-3295，

3.2.3 语义争议

秦代的郡县制度使得县、乡官吏的设置及文书制度较传世文献录载的更为复杂，目前对于简牍语料的语义还是存在诸多争议，这些争议为非历史学出身的作者带来了极大的解释性挑战。

例如里耶秦简中 5-1 中“守丞”，张春龙、龙京沙认为其为代理守丞，同时又指出由于其官员设置较为复杂，“守丞”之“守”也不一定是“试”或“代理”之意[49]。杨宗兵以简文的文例分析，认为“守”“丞”“守丞”这三词之间的含义没有太大区别，认为秦代县廷中的“守”“丞”“守丞”行使县令之权[50]。陈松长认为这里的“守”为掌管、主管之意，在秦简中，除了“郡守”是一种固定官名外，其余的如“司空守”“少内守”等“守”字应均表示为掌管或主管的含义[51]。

3.2.4 分词处理

在文本挖掘实验中，分词结果对后续的挖掘研究至关重要，直接决定了挖掘实验结果的准确性及可理解性。但是由于时代久远，简牍文献中存在较多的未释字、集外字，极大增加了语料预处理的难度。简牍材料一般残断问题突出，且各简牍之间在时代、题材等方面封闭性较强，语料预处理难度极大，所以目前已有的分词研究并不能直接适用于简牍文献。简牍文献的分词研究，需要从简文的实际特点出发，选择合适的方法、模型应用于简文的分词工作。

3.3 语料预处理

文本预处理是文本挖掘技术应用于简牍文本的前提。简牍文本为非结构化语言，所以在进行后续的文本挖掘实验前，必须要将其转化为计算机所能够识别的结构化文本数据，本文的预处理环节主要分为以下两个方向即文本分词及文本特征降维两方面，文本分词即对简牍文本进行切分，文本特征降维环节包括去停用词以及特征权重的计算。

3.3.1 分词

中文分词是利用中文分词算法对连续的文本按顺序进行切分，是对文本进行分析处理的基础环节。当前分词方法主要包括基于匹配的词典分词法、基于标注的机器学习算法和基于理解的深度学习算法[52]。词典分词是指将文本内容与事先建立的分词词典中的字或词条一一匹配对应，其算法的实现较为简单但无法去保证分词的准确率[53]。机器学习算法中决定分词效果好坏的是特征的选择，模型需要在经过标注的训练集中进行大量学习，通过学习使模型能够在一定程度上理解句子，并能够对句子进行语义分析。深度学习方法在一定程度上是对于机器学习方法的补充，因为其不需要人工进行特征选择，并且能够将长距离句子信息有效的保留下来。还有学者尝试将深度学习方法与前述两种方法相结合以获得更高的分词准确率[54]。

由于简牍语料的特殊性，本文研究中直接采用了西北大学出土文献小组的分词结果，下表为部分分词结果。

表 2　分词前后对比

原始文本	分词后
8-21 計以具付器計廿八年不來報敢言之 /□□□□□□☒I寫移令史可以律令從事敢【告】☒II	8 - 2 1 計 以 具 付 器計 廿八年 不來 報 敢 言之 / □ □ □ □ □ □ ☒ I 寫移 令史 可 以律令從事 敢【告 】 ☒ II
8-60a 十二月戊寅都府守胥敢言之遷陵丞☒I□公士樊道西里亭貲三甲為錢四千卅二☒II□□□□□謁告樊道□責有☒IIIb 六月乙亥水十一刻刻下二佐同以來/元手 ☒	8 - 6 0 a 十二月 戊寅 都府守 胥 敢言之 遷陵 丞 ☒ I □ 公士 樊道 西里 亭 貲 三 甲 為 錢 四千卅二 ☒ II □ □ □ □ □ 謁告 樊道 □ 責 有 ☒ III b 六月 乙亥 水十一刻 刻下二 佐 同 以來 / 元 手
8-71a 卅一年二月癸未朔丙戌遷陵丞昌敢言之遷☒I佐日備者士五梓潼長親欣補謁令☒IIb 二月丙戌水十一刻刻下八守府快行尉曹 ☒	8 - 7 1 a 卅一年 二月 癸未朔 丙戌 遷陵 丞 昌 敢言之 遷 ☒ I 佐 日 備 者 士五 梓 潼 長 欣補 謁令 ☒ II b 二月 丙戌 水十一刻 刻下八 守府 快 行 尉曹 ☒

3.3.2 去停用词

直接利用分词后的数据会使得数据的特征项集合维度较大维数过大会导致后续模型计算量过大且影响算法精确性导致计算结果不尽如人意。为了解决文本特征项的问题，我们通过去停用词及特征权重计算来对数据进行降维处理。主要是为了从分词后的数据中选择出一部分有效的特征，删除那些对于文本主题实验作用不大的特征项。

停用词（stop words）主要指功能词（functional words），其通常会在多种文档中频

繁出现但又极少附带实际的文本信息，如英文中的 that、is、which 等和汉语中的“的”、“呢”、“啊”等等助词、介词、连词。为了得到更好的文本分析结果，通常需要将这些会产生混淆的词语去除掉，这个过程即为去停用词，是文本数据预处理必不可少的一个环节。

针对现代汉语，有许多已经整理好的停用词次表例如百度停用词表、哈工大停用词表等，在进行现代汉语实验时直接进行下载调用即可。但是由于简牍语料具有先秦古汉语的特殊性，并且每部简牍材料都具有其自身的数据特性，目前并没有成熟的针对简牍材料的停用词典，需要针对其构建领域停用词典，根据里耶秦简语料的特点，我们构建了里耶秦简停用词表，其中主要为与语义表达无关的数字简号、符号表示及时间等词，尽量保留最多的有含义词项，表 3 为部分去停后结果。

表 3　部分去停用词结果

去停用词前	去停用词后
8-21 計 以 具 付 器計 廿八年 不來 報 敢言之 / □ □ □ □ □ □ ☒ I 寫移 令史 可以律令從事 敢【告 】 ☒ II	計 具 付 器計 不來 報 敢言 寫移 令史 可 以律令從事 敢告
8-60a 十二月 戊寅 都府守 胥 敢言之 遷陵丞 ☒ I □ 公士 樊道 西里 亭 貲 三 甲 為錢 四千卅二 ☒ II □ □ □ □ □ 謁告 樊道 □ 責 有 ☒ III b 六月 乙亥 水十一刻 刻下二佐 同 以來 / 元 手	戊寅 都府守 胥 敢言 遷陵 丞 公士 樊道 西里亭 貲 甲 為 錢 千 謁告 樊道 責 乙亥 水刻刻下 佐 同 以來 元 手
8-71a 卅一年 二月 癸未朔 丙戌 遷陵 丞 昌 敢言之 遷 ☒ I 佐 日 備 者 士五 梓 潼長 欣補 謁令 ☒ II b 二月 丙戌 水十一刻 刻下八 守府 快 行 尉曹 ☒	癸未朔 丙戌 遷陵 丞 昌 敢言 遷 佐 備 者 士五 梓 潼長 欣補 謁令 丙戌 水刻 刻下 守府 快 行 尉曹

3.3.3 特征词抽取

在对语料进行分词、去停用词任务后，与实验无关的符号、数字等无效信息都已经去掉。其余的一些特征项目，尽管与简牍的主题内容之间存在着一定的关联，但是其关联的紧密程度却并不尽相同。若特征项与主题之间的关联度大，那么其权重也大，若与主题关联度小，相应其权重也小。在文本挖掘的处理过程中通常是根据权值大小对特征项进行排序，从中选出权值排名位于前列的特征项，以便在后续的文本挖掘实验中使用。

特征项权重大小可以通过计算得到，目前主流的计算方法包括布尔（Bool）权重、特征频率（Term Frequency,TF）、倒文档频率（Inverse Document Frequency,IDF）、特征

频率-倒文档频率（TF-IDF）等等方法。在文本挖掘实验中，使用 TF-IDF 算法计算权重是最为常见的做法。

本文采用 TF-IDF 权重方法计算特征项，用来评价对于某份文档、某个文件集，某个字词对其的重要程度。公式中的*TF* 为词频，用来表示某个词语在整篇文档中出现的次数；*IDF* 为逆文档频率，表示某个特征词在总的文本集中出现的频率的倒数。TF-IDF 为 TF 和 IDF 的乘积。其计算公式为：

$$TF-IDF_{i,j}=TF_{i,j}*IDF_i \tag{3.1}$$

词频*TF* 计算较为简单，假设词*i* 在文档 *j* 中一共有 *N* 个词，那么词频为：

$$TF_{i,j}=\frac{n}{N} \tag{3.2}$$

假设语料库中有*M* 个文档，其中*m* 个文档中出现了词*i*，那么逆文档频率 IDF_i 为：

$$IDF_i=\log\frac{M}{m+1} \tag{3.3}$$

式中分母$m+1$则是为了防止除数为o。

对经过分词并去停用词处理的语料进行 TF-IDF 权重计算，经过比较，我们选取了排名前 20 的词汇作为文本的特征词，得到《里耶秦简》特征词语料。部分特征词如下：

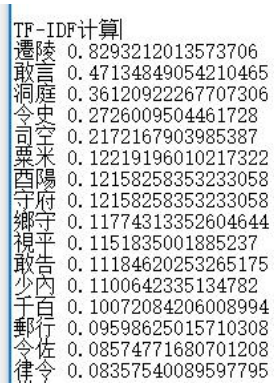

图 8　里耶秦简特征项

3.4 小结

本章主要介绍了简牍文本的预处理过程，包括简牍语料的收集、简牍语料预处理的难点、及预处理的主要步骤等。简牍文本预处理是文本挖掘工作的起点。首先介绍了本文所使用的简牍原始语料来源、数据收集及处理过程；其次介绍了简牍文本预处理工作的难点，主要包括目前先秦简牍中存在的普遍问题即残断、未登录字及语义争议等等问

题，残断多出现于竹简或木牍中，而未登录字不止出现于简牍中，还出现于铭文器物中，均对出土文献的分析利用产生较大影响；最后介绍了本文主要的文本数据预处理步骤，介绍了目前的分词主要技术及文本所采用的分词结果，同时构建了里耶秦简牍停用词表用于去除停用词，后利用 TF-IDF 权重计算方法计算文本特征项。

第四章 里耶秦简牍文本挖掘研究

随着简牍材料不断出土，文本数据量不断增多，如何从大量的散碎文本中提取有效信息成为已经成为待解决的问题之一。大数据环境下，文本挖掘技术为缓和简牍材料的人工研究效率低下与数据量不断增多这一矛盾具有重要意义。

4.1 文本描述性分析

描述性分析是利用统计分析对海量数据进行初步的整理、归纳，并从中发现其内在的规律，对实验语料进行描述性分析，可以在一定程度上把握文本的主题内容。

4.1.1 词频统计分析

词频统计通过统计文本材料中每个词出现的频率，以描绘词汇规律，发现隐藏在文本中的信息。其中最具影响力的当属对于《红楼梦》的一系列探讨。陈炳藻从数理统计学角度对《红楼梦》文本进行统计研究，通过计算机手段分别对其来前 80 回与后 40 回的用词规律及频率进行统计，结果得出这 120 回均出自曹雪芹一人之手。而李贤平教授则是通过聚类方法得出 120 回并非一个作者的结论，他将《红楼梦》整个 120 回作为 120 个样本，并且选取其中的虚词作为特征向量其进行分析，得到前 80 回与后 40 回分别聚为一类[55]。虽然在经过一系列探讨之后并没有给出作者是谁的定论，但它启发了相关学者对其他作家的作品进行一系列类似的研究。

利用词频统计除了可以对作者归属进行研究外，还可以用来分析领域热点、问题特征及作品的情感差异等。有学者对知网中与哈佛燕京学社引得编纂处和《汉学引得丛刊》有关的文献内容进行词频统计，通过词频统计结果得出该领域内近年来的研究热点与重点[56]；唐榕等通过对《全宋词》中高频词的统计分析，直观得到出宋词的文体特征[57]；李华鹏等通过对文本的词频和句子量进行分析，同时将其与传统的文学赏析方法相结合，对不同时间段及不同局势下的诸葛亮作品侧重点及情感、风格差异进行分析[58]。

我们对经过文本预处理的语料整体进行词频统计分析，通过高频词把握其中较为清晰的主要内容，表 4 为词频分析排名前 20 的词。

表 4　里耶一二卷词频结果

词	词频
人	773

遷陵	648
朔	577
手	575
敢言	429
洞庭	354
发	351
令	349
庭	344
守	308
丞	299
斗	295
者	271
佐	247
书	233
司空	225
陵	218
令史	213
石	208
仓	206

从表 4 可看出“遷陵”“洞庭”“庭”“陵”等词均为地名，“敢言”“书”“手”“发”等词均为文书用语或收发时所用词语，“守”“丞”“佐”“司空”“令史”等词均为秦代郡县的官职名，“石”“仓”等为粮食管理计量方面的用语。分析可得，通过词频统计可以看出文本的部分内容，但其所得过于笼统无法得到其细节内容。

4.1.2 词云图

里奇•戈登（Rich Gordon）是美国西北大学新闻学副教授、新媒体专业主任，“词云”的概念最早就是由他来提出的。词云（Word Cloud），又称文字云、标签云（Tag Cloud），词云图的核心是将文本数据通过一种可视化的呈现方式来对大量数据背后所蕴含的价值文本信息进行分析挖掘，通过这种呈现方式我们可以直观地得到关键词在文本中的词频分布，并以不同的颜色、字符大小来表达其相对重要程度。

经过前述文本预处理工作后，我们可以通过预处理后的数据来了解里耶秦简牍的内容重点，词云图如图 9 所示，具体分析如下：里耶秦简主要为秦代遷陵县的往来文书，主要内容涉及民生的方方面面，其中“敢言”“手”“書”等词涉及文书的记录，“發”“行”

等词涉及文书的邮行，“令史”“佐”“守”“尉”等均为秦代县廷官职名，“食”“石”“斗”“倉”“錢”等词均为往来文书中所涉及到的内容，包括粮食、钱帛债务等多项内容。通过词云图我们可以大致明了里耶秦简一二卷中的主要内容方向。相较于简单的词频统计，通过词云图可以更加直观的观察到文本中的重点词汇，并且其可以观察到更加细致的文本主要内容方向。

图 9　里耶秦简牍词云图

4.2 词频共现统计研究

4.2.1 实验流程

由于简牍文本晦涩难懂且其中存在着严重的残断情形，如果利用普通的文本分析方法难以达到预期的效果，所以本节利用 TF-IDF 特征词与词频统计相结合的方法，旨在将简牍文本数据转化为一个反映文本主要内容及其联系的词频共现矩阵，从词的角度来探究残断简牍之间的内容联系，通过分析共现词语间的联系来反映里耶秦简中的主要内容。其研究框架如下：

图 10　词频统计研究框架

该方法以文本关键词为节点，以关键词组合共同出现的次数为节点间的关系，通过

关键词的共现词情况来分析简牍语料的具体内容。基于词频统计对与关键词在同一语料中出现的字词进行统计，通过统计分析构建出与关键词相关联的词共现网络，这种共现网络分析相对于传统的内容分析来讲更加客观，通过词的关系勾勒出残断简牍内容间的共性，通过不同的层面去分析文本，最终呈现出简牍文本的主要内容。

4.2.2 实验结果分析

如上述研究框架所示，本实验采用 TF-IDF 算法计算得出的前十个特征词为关键词，分别提取出里耶秦简中与这十个关键词相关的所有语料，对这部分语料分别进行词频统计，得到与特征词具有高度相关性的前 20 个词，将文本数据转化为一个词频共现矩阵。部分结果如表 5 所示（具体见附录）：

表 5　词频共现统计表

遷陵	词频	敢言	词频	洞庭	词频	令史	词频	司空	词频	粟米	词频
遷陵	616	敢言	412	洞庭	270	令史	208	司空	212	斗	103
洞庭	174	朔	161	遷陵	151	視平	63	敢言	84	粟米	101
丞	160	上	107	署	84	斗	55	朔	73	石	72
朔	132	署	92	敢言	78	稟人	36	人	66	朔	34
守	107	謁	91	報	63	出稟	26	守	65	令史	27
手	87	手	85	朔	60	朔	37	署	60	少半	26
敢言	84	報	75	陽陵	58	感	37	受	48	出稟	13
行	56	令	75	縣	54	遷陵	34	報	45	升	23
令	52	遷陵	72	手	52	手	34	陽陵	40	佐	22
署	48	丞	70	發	45	史	33	縣	39	徑廩	20
發	46	守	67	丞	44	佐	33	謁	36	稟人	19
郵行	43	司空	54	守	43	石	29	爲	34	視平	19
敢告	39	今	53	司空	42	監	27	手	33	半	17
律令	37	陽陵	53	尉	42	粟米	26	丞	32	倉守	14

由词频计算结果可得“遷陵”是秦代洞庭郡下属县，敢言、 敢告、律令、從事等等词均为文书的固定用语，發、郵行为文书的传递用语，可得里耶简语料主要是以洞庭郡遷陵县的文书档案记录为主，具体往来官府涉及洞庭郡、酉陽县、下属啓陵、貳春、都鄉三乡等地。秦代国家通过发达的文书制度对社会的进行了有效控制，文书的上传下达保证了机构间的高效沟通，具体分析如下：

（1）作徒簿文书

秦代刑徒数量十分庞大，刑徒管理对维持一个国家的高效运转具有重要意义，正确

的认识刑徒问题，对理解秦代的国家统治了一个新的视角。秦代对于刑徒的管理，不仅通过法律上的各项规定，还通过严密的簿籍来对其日常活动进行严格的控制，作徒簿类文书即是一个窥探秦代刑徒管理相关问题的窗口。睡虎地秦简中的主要记载了各类法律规定，但其所反映的大多为制度规定，无法体现出其具体的运作细节，而里耶简中的作徒簿文书为机构档案文书，其中记载了秦代丰富的政治生活资料，生动细致的展现了其刑徒管理的具体细节。

作徒簿为里耶中自行命名的一种簿籍，有时也被称为徒簿，对遷陵县中“徒隶”的劳作情况进行详细记录，记载了诸多刑徒的工作分配情况[59]22。简牍 8-686+8-973 是一条完整的作徒簿文书，我们可以通过其探究文书的基本形式：

廿九年八月乙酉庫守悍作徒薄受司空城旦四人丈城旦一人舂五人受倉隸臣一人・凡十一人 AI城旦二人繕□□□ ☒AII城旦一人治輸□□ ☒AIII城旦人約車登 ☒AIV丈城旦一人約車缶 BI隸臣一人門負劇 BII舂三人級姇□娃 BIII廿廿年上之☒C（正）

八月乙酉庫守悍敢言之疏書作徒薄牒北上敢言之/逐手I

乙酉旦隸臣負解行廷☒II（背）8-686+8-973[1]203

我们可以看到除标题之外，整个徒簿文书向我们传达了这三部分的信息：首先为接受刑徒的具体明细及数量。其次是接受刑徒的具体分工。最后是部门官员针对刑徒管理对上级的汇报[59]23。而从标题可以看出，此简为“库”这一机构的作徒簿记录。同样作为接收作徒簿的机构还有直属机构（畜官、田官等）以及下属诸乡，可分别见里耶简牍 8-199+8-688、简牍 8-285、简牍 8-3034、简牍 8-142 等。

不仅是刑徒的接收方，刑徒派遣方也会对刑徒派出情况进行相应的记录。秦代的刑徒基本都是由司空管理，其不仅需要对刑徒的劳作情况进行监管，还需要负责修建维护其管辖范围内的土木、水利以及交通种种设施及对士卒的徭役情况及“居赀赎责者”的征发进行合理安排，如简牍 8-2008 所记载：

一人□【朝】AI一人有獄訊目 AII一捕鳥城 AIII一人治船疵 BI一人爲作務且 BII一人輸備弓具 BIII（正）

後九月丙寅司空□敢言☒(背)[1]416

除了司空这一机构外，里耶简中也存在着仓向外派遣劳力的记录，派遣劳力为隶臣妾，如简牍 8-663 中就存在非常详细的记载。

通过上述资料结合词频结果分析我们可以看出，其结果与人文学者研究基本一致。秦代编制作徒簿的机构有司空和仓，实验结果中司空和仓均与作徒簿共现。不仅如此，

朱圣明认为刑徒分派结果大多有明确目的地，大多分配至下属各乡或是下属各部门，在里耶中存在分配至遷陵县下属三乡即“启陵”“贰春”“都”（即“都乡”）的情况[60]57，而实验结果中此三乡均与作徒簿共现，分别共现 31 次、26 次和 20 次。

（2）债务文书

债务问题是社会经济史中的重要部分。观察实验结果可得，陽陵县与“敢言”、“洞庭”、“司空”这三个主题词分别共现了 53 次、58 次、40 次，经查陽陵在里耶文本中共出现了 61 次，其中 59 次出现于里耶简牍 9-1 至 9-12，大意为陽陵司空负责向洞庭郡写信追还士五、上造、公卒等所欠债务，请求此人所在地官府还债，洞庭郡负责调查此人目前在何縣署，吴方基认为此批文书为跨县追债的一种情况[61]。这部分文书完整的反映了秦代跨县追债的具体流程与细节，同时也侧面反映出司空的主要职责之一为追讨债务。

（3）粮食管理

从上述词频结果中可以看出，粮食管理记录在里耶简牍中占据了非常重要的地位，关键词“令史”“粟米”“鄉守”“視平”中均有涉及。

仓储是粮食管理中的重要一环，而“廩”的设置与管理，是秦代仓储稟食的核心环节，也是秦代稟食制度的关键。廩在里耶中较为常见，且大多数都有名称，里耶秦简一二卷中所记载的廩至少有四处，分别为径廩、乙廩、丙廩、西廩[1]42。

从里耶简所载看，稟给的绝大部分粮食种类都是粟，且其形态无一例外都是粟米[62]62，这一点由其排名第 6 的关键词“粟米”体现，其为里耶简前十个关键词中唯一一个作物名称。而与“粟米”关系最为紧密的“廩”为“径廩”，其与“粟米”共现 20 次，而“廩”在里耶语料中共出现过 30 次，由此可推测在秦遷陵县辖区内，“径廩”曾是最常使用的一个粮仓，这与谢坤在对里耶中的“廩”进行分析时所得出的结论一致[63]。

官方的粮食管理机构向特定人员发放粮食的行为称为稟食，其为粮食管理中的重要职能之一。稟食过程需要遵循一定的程序。我们目前可知稟食的流程应当至少包括“发放前的准备”“量谷发放”“文书归档整理”等主要环节[64,65]。而在实际发放的过程中，需多人共同参与完成。睡虎地《效律》中简 27-28 记载“某廩禾若干石，仓啬夫某、佐某、史某、稟人某”，此简内容为秦关于稟食记录的官方依据[66]。

从上述结果词频共现结果可得，“粟米”、“令史”、“视平”、“鄉守”等关键词均指向稟食的管理以及记录，与上述几个关键词共现的稟人、史、佐、倉、倉守、鄉守等均为稟食发放的管理者，里耶秦简所见稟食文书记录也可大致总结为“某廩粟米若干石，某年月仓/仓守/乡守/田官守某、佐/史某、稟人某出稟某，令史/令佐某视平，某手”，

如简牍 8-760、8-1557 等[62]58。由此记录总结中可看出，稟食的发放至少需要三人以上参与，并且参与人员各司其职、分工明确，仓、仓守、乡守、田官守等负责主理，佐、史负责记录监督，稟人负责具体劳作。

规律结合简文，我们发现可以通过“佐/史某”推测稟食发放中记录者的信息，如里耶 8-45“稻四 卅一年五月壬子朔壬戌倉是史感稟人□☒令史尚視平 ☒”[1] 39-40中缺乏记录者，而根据史感可推测记录者为感，利用此规律可对部分简牍中的信息进行复原，里耶秦简中 8-763、8-764 及 8-1550 等多条简牍均可作为验证。

“感”与“令史”、“视平”两个关键词分别共现 37 次和 36 次，经查，感在里耶简一二卷中共出现 81 次，其以“史感”出现 31 次，以“感手”出现 32 次，其多次参与稟食发放监督及记录且主要集中在卅一年，根据上述记录总结，我们可推测其为秦始皇三十一年间粮食发放的要管理者。监督者中令史逐视平 14 次、令史尚视平 12 次、令史扁视平 10 次，可推测这三人均为监督稟食发放的重要管理人员。

本部分实验通过词频统计的角度来对简牍文本进行定量研究，分析可得，里耶秦简中主要为洞庭郡遷陵县的文书档案，包括其与上属洞庭郡府，同级酉陽县，下属啓陵、貳春、都鄉三乡的往来文书，且其中多为债务、粮食及徒簿类文书。

相较于单纯使用词频统计或词云图，我们可以看出从词频共现角度对简牍文献进行分析能够得到出更加细致具体的结论，根据共现的频率也可以为我们把握文本的主旨提供一个更加直观的角度，因此不难发现，利用关键词共现对文本主题进行分析具有一定的可行性。

4.3 TextRank 文本摘要研究

4.3.1 TextRank 模型实现

单独通过 TF-IDF 算法来提取摘要时，那些能够准确概括表达文本内容的词被命名为关键词，通常认为越重要的句子关键词数量都较多，而越重要的句子被选为文本摘要的可能性就越大，但是此种算法仅仅考虑了词频而未考虑句子的位置信息等其他因素，因此在摘要抽取方面并不准确。TextRank 算法是一个基于图的模型算法。在文本摘要的自动提取过程中，该算法利用文档中的全部句子作为顶点，构造一个无向有权图，并以句子间的相似度作为边，最后得出摘要结果，但这种方法仅仅利用了语料中句子间的相似度作为重要程度的判断标准，而忽略了词频、位置信息等因素的影响，使得最后的抽取结果也呈现出片面性。在此基础上，本节在 TextRank 算法的基础上，将 TF-IDF 与

TextRank 的文本摘要算法结合起来，通过将影响句子重要性因素相融合，来提高自动文本摘要的计算结果。其整体实现流程如图 11 所示。

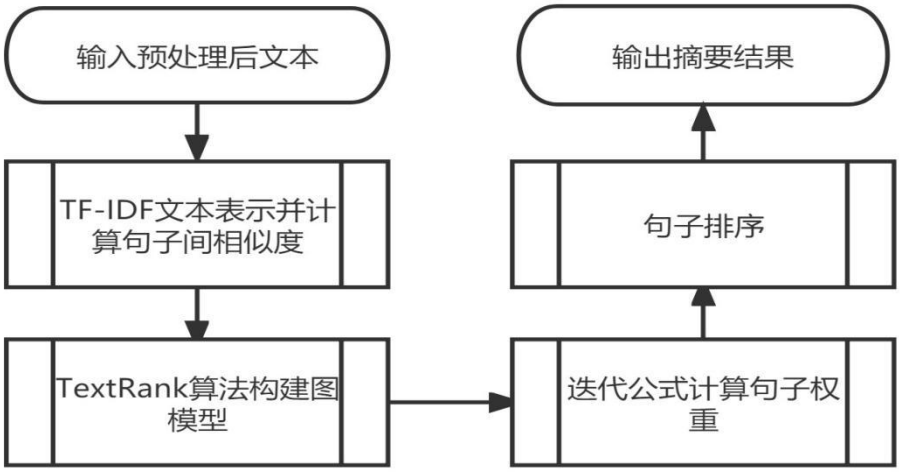

图 11　自动文本摘要算法流程图

首先利用经过预处理手段的语料获取句子的统计信息，从而得到文本的句子集合；再利用 TF-IDF 算法分别计算句子间相似度，在此基础上构建文本的图模型；之后利用图模型的迭代公式计算句子的权重，再根据权重值的大小进行排序，选取排名靠前的句子作为最终摘要，最后将排名靠前的最终摘要输出。

4.3.2 实验结果分析

利用 Python 对搭建实验环境，根据上述研究框架抽取文本摘要，从事件的角度对里耶内容进行分析，抽取结果及分析如下：

表 6　TextRank 文本摘要算法结果

摘要	内容
摘要 1	壬午朔 癸卯 左公田 丁 敢言 佐 州 里 煩 故 為 公田吏 徙屬 事 荅 不備 分 負 各 石 少半 斗 直錢 百 煩 冗佐 署 遷陵 今 上 責 校券 謁告 遷陵 令 官計 者 定 錢 百 受 旬陽 左公田 錢計 問 可 計 付 署 計 為報 敢言 辛亥 旬陽 丞 滂 敢告 遷陵 丞 主 寫 移 移券 可 為報 敢告主 兼 手　庚子 遷陵 守丞 敬 告 司空主 律令從事 言 懷 手 即 走 申 行 司空
摘要 2	戊戌朔 癸卯 尉守 竊 敢 洞庭 尉 遣 巫 居貸 公卒 安成 徐 署 遷陵 今 徐 壬寅 事 謁令 倉貣 食 移 尉 展 約 敢言 癸卯 遷陵 守丞 膻 告 倉主 律令從事 逐 手 即 徐 入
摘要 3	辛丑朔 丙午 司空 騰 敢言 陽陵 逆都 士 越人 貲錢 千百 越人 戍 洞庭郡 不智 何 縣 署 今 爲 錢校券 上 謁令 洞庭 尉 令 越 人 署 所 縣 責 受 陽陵 司空　司空 不

	名計 問 何 縣 官計 付 署 計 爲報 已 訾 其 家 家貧 弗能 入 乃 移 戍所 報 署 主責 發 敢言 戊申 陽陵 守丞 廚 敢言 寫上 謁報 報 署 金布 發 敢言 儋 手 癸巳朔朔 陽陵 遬 敢言 至今 未報 謁追 敢言 堪 手
摘要 4	甲子朔 癸酉 啓陵 鄉守 意 敢言 廷下 倉守 慶 書 言 令佐 贛 載 粟 啓陵鄉 今 已載 粟 石 為 付券上 謁令 倉守 敢言 甲子朔 乙亥 遷陵 守丞 𣪊 告 倉 主下券律令從事 壬 手 乙亥 旦 守府 印 行
摘要 5	戊申朔 癸亥 洞庭 叚守 昌 謂 遷陵 丞 遷陵 上 坐 反 適 辠 當 均 輸 郡中 者 人今 皆 輸 遷陵 其 聽書從事 它如律令 新武陵印 行事 丁酉 遷陵 守丞 敦 狐 告 司空主 律令從事 夫 手 走 郤 即 行
摘要 6	庚戌朔 丙子 司空守 樛 敢言 前 言 競陵 灊陰 狼 假 遷陵 公船 袤 丈 尺 名 曰 求故荆 積瓦 未 歸 船 狼 屬 司馬 昌官 謁告 昌官 令 狼 歸 船 報 曰 狼 律在 復獄已卒史 衰 義 所 今 寫 校券 牒 上 謁言 卒史 衰 義 所 問 狼 船 存 所 其 亡 為 責券移 遷陵 弗 屬 謁報 敢言 庚辰 遷陵 守丞 敦狐 卻 司空 自 叚 狼 船 何 故 弗辟 今 而 誧 曰 謁 問 覆獄 卒史 衰 義 衰 義 事已 不智 所 居 其 聽書從事 懷 手即令 走 行 司空

其中摘要1为因工作原因导致政府与官吏之间产生的记载其债务或赔偿关系的文书。大意为官吏“烦”在任时没有按照要求准备充足的粮食，导致其需要赔偿部分粮食，虽然其已迁至他地，但这项债务却一直未偿还，于是发文询问其赔偿的金额要由哪方县政府记录，其后将与此债务有关的校券移交至现居县[67]129。从摘要中显然可看出，秦朝的律法中明确规定了若债务人尚未清偿债务而迁出的问题，这一点在其他的秦简中也同样有迹可循，例如在《睡虎地秦简·金布律》中就记载了此类情况的权责分配情况： 有责（债）于公及货、赎者居它县，辄移居县责之。 百姓（假）公器及有责（债）未赏（偿），其日 以收责之，而弗收责，其人死亡……令其官啬夫及吏主者代赏（偿）之[66]38。

摘要2描写了一个因欠债而在遷陵县戍守的外郡戍卒。籍贯为巫县的徐为外郡人，由洞庭郡尉派遣至遷陵，与派遣至阳陵戍卒如出一辙，“謁令倉貣食”反映了其外郡戍卒的粮食也是由遷陵仓负责供应的，其很可能是因欠债而在遷陵县戍守的外郡戍卒。判断是否为外郡戍卒，如简8-1014：“☑□出貣居貲士五巫南就路五月乙亥以盡辛巳七日食。Ⅰ缺手Ⅱ”[1]262与简8-1083：“士五巫南就曰路娶貲錢二千六百☑Ⅰ卅一年四月丙戌洞庭縣官受巫☑Ⅱ”[1]275，“路”在两简中身份均为“士伍”、籍贯均为巫县南就里，无疑为同一人。路为外郡人，又有货錢，又获得仓“出貣”，极有可能是在洞庭戍守的外郡戍卒。

摘要3为里耶简牍9-8简文内容，是一种洞庭郡行政事务内部协调的文书。由上述

词频统计实验结果可知，此记录为跨县追债记录，完整的记录了展现了跨县追讨债务的流程和细节[68]。里耶简中存在着一批格式相同且内容类似的简牍，在这批文书的背后，其反映了秦代的经济司法制度，并且从中可以看出，秦代的法律中明确规定，若债务人出现意外情况导致债务的偿还出现问题时，应如何对债务人进行追索。

摘要 4 即为一条粮食的运输记录，记载了秦代遷陵县利用公船转运粮食的情况，是秦代粮食运输情况的缩影。从古至今，水路运输都是运输业中及其重要的一环，秦统一后，水上运输在国内粮食调配上仍占有重要地位，从遷陵县的各类档案记载即可窥见，例如里耶简牍 8-2191。不仅如此，从里耶简中与运输相关内容可得，遷陵县中的公船主要被用于运输粮食、刑徒，摆渡行人及货物等等。

摘要 5 为里耶简 9-23 内容，记录了坐罪者的具体服刑安排，包括由哪个官府来安排服刑地点、由谁来管理等[69]。其为一条地方官吏押解刑徒的记录，也是官吏徭使任务的一种，里耶秦简所记载的遷陵官吏徭使任务类型多种多样，主要包括押解护送各类人员、采购物资、上事郡府等等。押解及护送刑徒对于秦代的地方官吏来讲是较为常见的任务在里耶简牍中也较为常见，可见简 8-1023、9-712 等。

摘要 6 为一官府的公船管理记录，为政府与私人之间的债务往来问题，此摘要抽取出了一个完整的事件，“狼”向政府借用公船后却因船只已丢失无法归还，于是将事情经过详细记载于券上并上交给官府[67]128。观察摘要 6 我们可以发现，在出借船只的时候，至少需要对债务人的姓名、籍贯、出借财物名称、具体形制、出借财物的具体用途、出借时间等等关于债务人及出借财物的详细信息进行登记，为了对此类借入财物未还的情况进行记录，官府还专门设计了“责券”用来对财物借出情况进行登记存档。

根据抽取结果可以看出，里耶文本中涉及到大量的债务文书，主要包括居貲赎债。为无法足额缴纳貲罚以劳役代偿称为居貲；无力缴纳赎金而以劳役代偿称为居赎（摘要 2、3）；而居债则是因为欠了公家债务，其本身并未犯罪，秦简当中经常可见官府与百姓间的借贷往来记录，在经济借贷过程中，若因自身原因造成官府损失需要按律赔偿，此时相当于“有债于公”（摘要 1、6）[70]。

由词频实验可知粮食记录是其档案文书中重要的组成部分，摘要抽取结果中得到的粮食相关记录为运输管理记录（摘要 4），综合两项实验结果可知秦代对于粮食的运输、储存、管理、发放相关环节均十分重视。

律令类记录多见于其他秦简，如睡虎地秦简中的《秦律十八种》、《法律答问》，岳麓中的《秦律杂抄》、《秦令杂抄》等等，相比之下里耶简牍中的律令类内容则主要是通

过徭使任务管理（摘要 5）、仓储管理、邮驿津渡管理（摘要 6）等等管理类的具体记录来侧面体现，通过这些记录，勾勒出了秦代官府的日常运作及各类制度，鲜活的反映了秦代县级政府日常运作的面相。

综合上述分析可知，文本摘要的计算结果主要分为三类，分别为官府与官吏或私人间的债务问题、粮食管理运输、官吏徭役记录。根据内容来看，抽取结果均为里耶材料中较为完整且具有代表性的事件，能够较好的涵盖简牍材料的主要内容，且能与里耶校释第一卷中所给出的主要内容对应起来。

不难看出，相较于词频共现分析，文本摘要计算结果抽取出了完整的关键事件记录，更为细致生动的展现了秦代官府之间的沟通、管理的具体细节，并且其与里耶中的主要内容也可以较为贴切的对应起来。摘要结果可以正确的表达出文本的核心思想与主要内容，综上，利用 TextRank 文本摘要算法对简牍文本进行处理具有一定的可行性和有效性。

4.4 LDA 主题建模研究

4.4.1 实验环境及流程

本节选择 LDA 主题模型对里耶秦简牍进行建模实证分析，通过其建模结果之一主题-词概率分布来对里耶文本中所涉及到的主题进行分析。根据里耶秦简牍文本特点及模型特点，本部分实验设计流程如图所示.

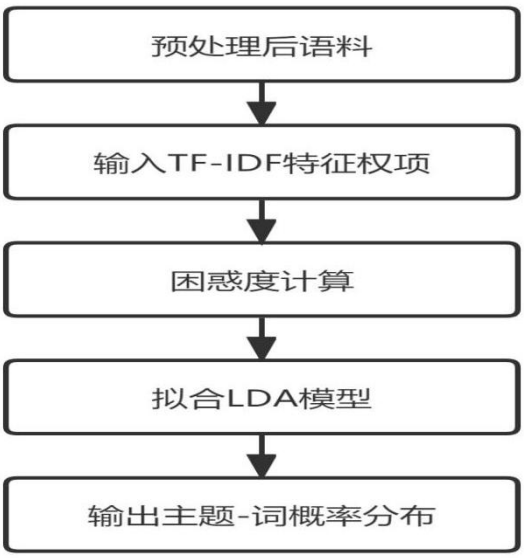

图 12　LDA 主题模型拟合流程图

根据以上流程图可知，本实验中需要利用经过分词及去停用词等预处理程序后的语料来确定最优主题数目，确定文本聚类实验中的最佳主题数目对于后续的分析来讲是极

其重要的一步，而确定最优主题数目的一种有效方法即为计算其困惑度。根据手肘法可知，当困惑度曲线位于最低的拐点处困惑度最小的位置就是最佳主题的位置，此时困惑度数值最小则主题数目最优。得到最优主题数目后，利用主题数目及前述计算得出的特征权项，对 LDA 模型进行拟合训练，最终得到里耶秦简的主题-词概率分布，以其概率分布来对里耶文本一二卷所涉及到的主题进行分析。

4.4.2 实验过程及结果分析

依据前述研究框架，首先需要对困惑度进行计算并绘制里耶文本困惑度与主题数目之间的折线图来确定最优主题数目。实验采取多次循环迭代来计算困惑度大小并分别绘图，如图 13 所示，在选取最优主题数目时，主题颗粒度过大会造成主题过于集中，而主题颗粒度过小则会造成主题过于分散，均对后续分析产生影响。经过观察，我们发现困惑度在主题数大于 15 时无法判断其拐点，而在主题数为 12 及 15 的位置处困惑度曲线位于最低的拐点处，在此基础上结合人工的反复验证判断，最终确定 12 个主题数目为里耶秦简一二卷内容中最为合适的主题数量。

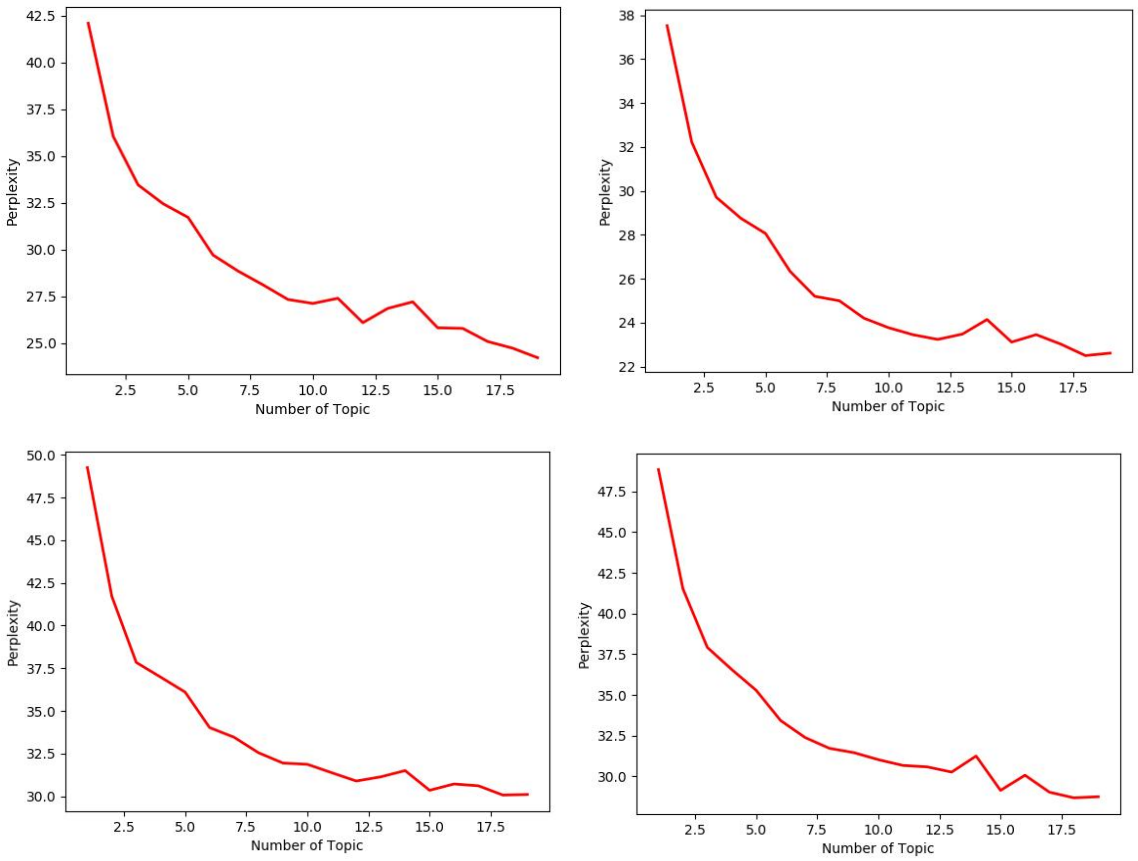

图 13　里耶主题-困惑度曲线

在确定了最佳主题数目后，开始进行 LDA 主题模型的拟合实验。实验以 12 个主题数为基础，设置最大迭代次数 max_iter=50, 利用 LDA 模型得到《里耶秦简》一二卷主

题-词概率分布，分别描述了每个主题概率分布排名靠前的 15 个关键词，根据前述所使用的特征词提取方法，使得分布中的词大多具有高度语义，代表了《里耶秦简》一二卷中的重要主题。结果如下表所示：

表 7　里耶主题-词概率分布

主题	关键词
主题 0	手 守 急 如 史 已 食 毋 庫 酉 心 少 平 丙午 爵
主题 1	令 署 書 言 者 下 其 敢言 縣 手 行 上 守府 當 報
主题 2	遷陵 發 洞庭 敢言 廷 丞 司空 朔 郵 官 行 鄉 金布 主吏 守丞
主题 3	過 可 尺 傳 酉陽 笥 此 寸 田 未 遷 中 在 買 丈
主题 4	廷 陵 尉 少內 元 留 及 壬 黔首 丙申 乙亥 水刻 追 丁巳 刻下
主题 5	爲 百 曰 所 皆 今 移 主 問 弗 敢言 各 見 具 計
主题 6	斗 石 佐 半 來 令史 出 正 入 升 視平 武 夫 敬 手
主题 7	人 倉 付 吏 捕 敢 其 徒 臣 取 盡 里 戶 稟 隸妾
主题 8	郵行 洞庭 遷陵 行 郵 欲 乙丑 癸酉 癸 賜 辛 大女子 毋 手 枳
主题 9	甲 貲 到 不 與 子 監 令佐 它 事 遣 旦 嘉 論 就
主题 10	遷陵 粟米 手 斗 士五 石 是 稟人 辰 等 倉守 令史 數 少半 出稟
主题 11	錢 受 得 不更 獄 千百 都鄉 兩 牒 當 斤 辛未 課 校 殿

根据表 7 的主题-词分布结果，我们对里耶主题进行分析，分析如下所示：

主题 2 中“遷陵”、“洞庭”均为地名；而“廷”、“丞”、“守丞”、“司空”等均为官职或机构名称，“廷”为县廷，杨宗兵从简文例中分析，“守”、“丞”及“守丞”互见无别；“發”为拆阅文书、“郵”、传达文书的机构，其均为传递、收发文书时所用；所以主题 2 可以定义为洞庭郡遷陵县的文书郵传。

在秦汉时期，公文是政府发布政令的主要载体，而文书制度则是维持秦汉王朝统一的重要方法。中央政府的文件是否能够迅速、准确地传递到各地，并被相关机构所采纳、实施，将直接影响到政令的执行、行政效率乃至整个官僚体系的稳定[71]。文书郵传制度为秦汉社会的最重要的基本制度之一，更为里耶秦简中的重要主题之一，里耶中的郵传文书内容，补充了学界对于秦代公文的认识，让大家更加清晰的认识了秦汉时期公文系统的下达方式。

主题 6 中“斗”“石”“半”“升”等词均为粮食计量时所用；“令史”“佐”为官职，秦简显示，当粮食及财物于仓、少内等机构进出时，需要安排官吏监督，称之为“视平”，或“视”“平”“监”。从简文例中可以看出，监督人主要为令史，但偶尔也可见到令佐

担任监督人的情况，所以主题词中的“佐”有可能为令佐；“敬”多为债务文书的记录者，涉及对欠债戍卒的粮食发放，所以主题 6 可定义为欠债者的粮食发放。

主题 9 中排名最前的主题词为“甲”“贲”，“贲”为有罪被罚款或罚物者，“贲甲”“贲盾”皆可以代偿贲款或物，“贲甲”为财产惩罚的一种；而官职“令佐”的职责也包括参与司法事物、监督财物的进出，简牍 9-119 中记录“令佐贛入贲錢七百七十六 元年八月庚午 令佐贛監”[2]70-71，所以主题 9 可定义为贲罚问题。

主题 10 中“粟米”为粮食；“斗”“石”“少半”为粮食的计量单位；“稟人”由徒隶担任，管理谷物的收藏出纳；“仓守”为仓的管理人员；在出粮时大多需要“令史”来监督，简牍 8-217 中记录“稻四 斗八升少半半升 卅一年八月壬寅倉是史感稟人堂出稟隸臣嬰自槐廇 令史悍平 六月食 感手”[1]116-117，其中出稟便是出粮的一种方式，多采用按月出粮，单独发放给个人；“手”为记录者；“數”为数量，可将主题 10 定义为出粮数量的管理及记录。

本节基于 LDA 主题模型进一步对里耶简牍文本进行文本分析，通过计算困惑度数值大小确定最优主题数，经模型拟合后获得了若干主题，凝练出了邮传、貲罚、粮食、债务者粮食发放等等主题类，由于简牍文本的特殊性和简文的多义性，其处理难度极大，其余主题中的主题词较为散乱，目前难以给出确切的定义，仍然需要进行进一步的调整文本研究，从主题的视角出发提供了一个分析简牍文本的新方法，并且具有一定的可行性及实践意义。

4.5 小结

本章为里耶秦简简牍文本的实证研究，首先对文本进行描述性分析，主要利用词频统计分析及词云图展现里耶秦简的主要内涵，结果中汇集了里耶简牍绝大部分具有高度语义的词；之后利用特征词及词频统计方法对文本进行共现分析，分析可得里耶中的三类文书即作徒簿文书、债务文书及粮食管理记录；其次利用 TextRank 摘要算法对文本进行摘要抽取，共得到六条摘要，经过整理分类可知，六条摘要均为官府与官吏或私人间的债务问题、粮食管理运输、徭役管理记录等三类内容；最后利用 LDA 主题模型对文本进行分析，通过困惑度及人工验证确定最优主题数为 12，分析得邮传、貲罚、粮食、债务者粮食发放等等主题。本章利用三种不同的文本挖掘方式，实现了从词、事件及主题三个角度对文本进行全面的分析，并且可以看出，通过文本挖掘方法得到的主题内容结果与里耶校释第一卷中给出的简牍主题均能对应，证实了这三种挖掘方法应用于简牍

材料的分析均具有一定的可行性和参考价值。

第五章 总结与展望

5.1 总结

文本挖掘方法在人文社科类研究中的应用，提供了一个行之有效的计算机辅助解释方法。本文基于近年来国内外相关研究，以里耶秦简一二卷为例，针对简牍文本进行特征词提取，构建里耶相关的词表、停用词表等，结合文本挖掘技术进行语料抽取与统计，运用词频统计、文本摘要及 LDA 主题模型等方法，分别从词、事件以及主题三个角度来对其中的相关可定义内容进行挖掘研究。研究显示一二卷中主要为洞庭郡遷陵县的文书档案记录为主，具体往来官府涉及洞庭郡、酉陽县、下属启陵、贰春、都乡三乡等地。其中涉及文书郵传、债务、粮食、徒簿、案件、徭使管理等等内容，与目前校释中给出的结论基本一致。

本文对于简牍研究这一领域的文本挖掘研究具有一定的理论意义及参考价值，尤其是首次尝试将简牍和文本挖掘技术结合起来，有效地提高了简牍资料的处理效率和信息获取效率，为简牍材料的研究提供了新思路，证实了可借助文本挖掘方法来理解、诠释中国丰富的文化遗产，具有一定的可行性及实践意义。本文通过一个全新的分析角度来看待秦史材料，为描绘秦代历史提供了一个新的研究视角，给国内相关研究的开展提供了一定的启示。本文主要工作如下：

（1）通过文献调研来确定目前文本挖掘手段目前在传世文献及简牍文献中的应用。发现目前文本挖掘方法在传世文献的历史分析与挖掘中应用广泛，包括简单的词语分析应用及复杂的知识发现应用，简单的文本分析主要为利用词频统计及关键词提取等手段，而复杂的知识发现应用主要包括利用文本的分类、聚类、信息抽取及可视化分析等手段对大规模的历史文本进行分析挖掘，例如对于《全唐诗》、《汉典》、《孟子》及《左传》等传世文献的分析，而在简牍文本分析中甚少运用，目前利用计算机手段对简牍文本进行处理主要集中于简牍文本的数字化以及各类相关数据库的建立。

（2）简牍文本数字化及预处理。与简牍小组成员共同将里耶秦简一二卷的内容进行数字化处理，并通过两轮核对校验确保简文数据的准确性，在此基础上对数据进行预处理，包括分词、去停用词及特征项权重计算等步骤，将非结构化的文本语言表示为计算机能够识别的结构化语言，为后续的文本挖掘实验打下了基础。

（3）利用关键词提取及词频统计方法对里耶秦简一二卷进行挖掘研究。以排名前

十的关键词为索引，分别提取与其相关的所有语料，对其进行词频统计，得到与关键词共现的词频排名前 20 的词，分析结果得到里耶一二卷中存在作徒簿文书、粮食管理类文书及债务文书。

（4）利用 TextRank 文本摘要模型进行摘要自动抽取。通过对简牍文本信息进行压缩抽取出其中的关键信息，可以在保留绝大部分关键信息的前提下压缩文本量。通过提取文本摘要可以在一定程度上降低需要理解的简牍文本数据量，从事件的角度来获取简牍文本中的关键信息，实验共获得 6 条摘要，涉及到债务类文书、粮食管理类文书及律令类文书。

（5）LDA 主题模型计算里耶一二卷主题-词概率分布。本文利用里耶秦简特征词语料的主题建模结果之一，即主题-词概率分布来获取里耶秦简的主要主题及具体内涵，从主题的角度来对简牍内容进行分析。引入困惑度作为模型评价指标，根据困惑度及人工反复实验选择以 12 个主题作为最优主题，且其中 4 个主题具有明显主题倾向。

5.2 展望

文本挖掘技术与大规模的出土文献文本相结合的研究，将大规模历史文本的研究置于一个新的研究场景中，随着近年来出土文献数量的不断增多，更多的历史文本将会以非结构化的形式呈现，基于文本挖掘技术的大规模历史研究将会不断发展，在新的历史研究场景中文本挖掘方法将会有越来越多的应用。本文对利用文本挖掘方法讨论研究简牍文本进行了一些有益的探索，但研究过程中还存在以下问题，今后还需要在下述多个方面进行深入的研究：

（1）随着后续里耶秦简牍三四五卷出齐，可进一步深化对里耶秦简牍文本挖掘研究。从目前得到的结果来看，利用文本挖掘方法分析简牍文本提供了一个新的研究视角，并且结果显示，与目前人文学者研究结果一致，待后续几卷出齐后，可进一步对里耶文本进行更加完整的分析。并且由于各简牍之间在时代、题材等方面封闭性较强且语料处理难度极大，故后续会逐步针对睡虎地秦简、张家山汉简等内容更为完整的简牍进行语料处理，依次实现文本挖掘研究。

（2）不断完善简牍的分词及预处理工作。数据预处理工作是运用文本挖掘方法进行分析的第一步，也是及其关键的步骤，直接决定了后续实验结果的精确程度，进一步提高分词准确率、尝试不同的特征项权重计算方法，更加有利于提升后续实验的准确程度。

（3）模型本身的不足。抽取型文本摘要算法虽然可以为我们提供一个较为完整的结果，但其在抽取时会倾向于抽取有着较为完整语义的内容，里耶简牍中存在较多的残断现象，在这方面会不可避免的存在偏颇。而在主题模型试验中仍有一些主题词较为散乱，需要进一步对实验进行调整，针对内容较为清晰的简文进行更加深入的分析。

（4）出土文献自身的局限性。文本在出土时即存在较为严重的残断情况及无法辨识的未登录字，并且有相当数量的简文未被明确释读，其中某些字词可能会含有关键性信息，例如在分析粮食监管者时，令史犴监督 13 次，但由于其为未登录字无法识别，这些文本自身的特征为我们的分析带来了极大的挑战。所以未来针对简牍文本的分析，除了需要着眼于内容较为清晰的简文，更需要针对残断及未登录字情况进行具体的研究分析，更加充分地利用目前出土的有限简牍数据。

参考文献

[1] 陈伟.里耶秦简牍校释:第一卷[M].武汉.武汉大学出版社,2012.

[2] 陈伟.里耶秦简牍校释:第二卷[M].武汉.武汉大学出版社,2012.

[3] 袁军鹏,朱东华,李毅,李连宏,黄进.文本挖掘技术研究进展[J]. 计算机应用研究,2006(02):1-4.

[4] 郭金龙,许鑫.数字人文中的文本挖掘研究[J].大学图书馆学报,2012,30(03):11-18.

[5] CHEN J W.East Asian studies macroscope [EB/OL]..http://macroscope.cdh.ucla.edu, 2022-02-20.

[6] NICHOLS R,SLINGERLAND E,NIELBO K,et al.Modeling the contested relationship between Analects,Mencius, and Xunzi:preliminary evidence from a machine-learning approach[J].The journal of Asian studies,2018,77(1):19-57.

[7] ALLEN C,LUO H L,MURDOCK J,et al.Topic modeling the hàn dian ancient classics [J/OL]..https://arxiv.xilesou.top/ftp/arxiv/papers/1702/1702.00860.pdf. [2022-02-17]

[8] Chen and Borovsky and Kawano. The Shishuo xinyu as Data Visualization[J]. Early Medieval China, 2014, 2014(20) : 23-59.

[9] Hestia[The Open University.Hestia[EB/OL].http://Hestia.open.ac.uk/,2022-02-03.

[10] Torget A J,Mihalcea R,Christensen J,et al.Mapping texts: combining text-mining and geo-visualization to un-lock the research potential of historical newspapers [EB/OL].http://mappingtexts.stanford.du/whitepaper/MappingTexts-WhitePaper.pdf,2021-12-30.

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

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

[13] 王东波,何琳,黄水清.基于支持向量机的先秦诸子典籍自动分类研究[J]. 图书情报工作,2017,61 (12):71-76.

[14] 吴斌,吉佳,孟琳,石川,赵惠东,李仪清.基于迁移学习的唐诗宋词情感分析[J]. 电子学报,2016,44 (11):2780-2787.

[15] 李晖,张天垣,金纾羽.古代中国格律诗中的社会情感挖掘[J/OL].计算机工程与应用:1-9.http://kns.cnki.net/kcms/detail/11.2127.TP.20200528.1413.006.html, 2021-12-22.

[16] 马创新,梁社会,陈小荷.先秦诸家学派的相关系数与特征词研究[J]. 中文信息学报,2019,33(12):129-134.

[17] 何琳,乔粤,刘雪琪.春秋时期社会发展的主题挖掘与演变分析——以《左传》为例[J].图书情报工作,2020,64(07):30-38.

[18] 王小红,科林·艾伦,浦江淮,李颖娜.人文知识发现的计算机实现——对“汉典古籍”主题建模的实证分析[J].自然辩证法通讯,2018,40(04):50-58.DOI:10.15994/j.1000-0763.2018.04.008.

[19] 申资卓,杨莹,邵艳秋.基于主题模型的古典乐器诗词文本挖掘[J].中文信息学报,2019,33(03):79-86.

[20] 朱锁玲,包平.方志类古籍地名识别及系统构建[J].中国图书馆学报,2011,37(03):118-124.

[21] 皇甫晶,王凌云.基于规则的纪传体古代汉语文献姓名识别[J].图书情报工作,2013,57(03):120-124.

[22] 李章超,李忠凯,何琳.《左传》战争事件抽取技术研究[J]. 图书情报工作,2020,64(07):20-29.

[23] 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(02):66-80.

[24] 张建立,李仁杰,傅学庆,张军海.古诗词文本的空间信息解析与可视化分析[J].地球信息科学学报,2014,16(06):890-897.

[25] 张玮,谭思危,刘凯,石磊,陈思明,陈为.宋词研究的新视角:文本关联与时空可视分析[J]. 计算机辅助设计与图形学学报,2019,31(10):1687-1697.

[26] 李静.浅析简帛文献的数据库建设[J].江汉考古,2017(05):128-130.

[27] 毛建军.甲骨文献全文数据库的建设与思考[J].图书馆学研究,2010(23):37-38+36.

[28] 刘炜,叶鹰.数字人文的技术体系与理论结构探讨[J].中国图书馆学报,2017,43(05):32-41

[29] 杨文.数字人文视阈下的社会记忆构建研究[J].情报资料工作,2019,40(05):38-45.

[30] 邓君,宋先智,钟楚依.我国数字人文领域研究热点及前沿探析[J].现代情报,2019,39(10):154-164.

[31] 朱本军,聂华.跨界与融合:全球视野下的数字人文——首届北京大学“数字人文论坛”会议综述[J].大学图书馆学报,2016,34(05):16-21.DOI:10.16603/j.issn1002-1027.2016.05.003.

[32] 刘炜,叶鹰.数字人文的技术体系与理论结构探讨[J].中国图书馆学报,2017,43(05):32-41.DOI:10.13530/j.cnki.jlis.170020.

[33] 李尚昊,朝乐门.文本挖掘在中文信息分析中的应用研究述评[J].情报科学,2016,34(08):153-159.

[34] 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,2012,56(08):18-25.

[35] Sato S.,Okurmura M.Advances in Automatic Text Summarization[M].Massachusetts:MIT Press,1999.

[36] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[C]. 2004 conference on empirical methods in natural language processing.2004:404-411.

[37] 黄波,刘传才.基于加权 TextRank 的中文自动文本摘要[J].计算机应用研究,2020,37(02):407-410.DOI:10.19734/j.issn.1001-3695.2018.07.0528.

[38] 曹洋.基于 TextRank 算法的单文档自动文摘研究[D].南京大学,2016.

[39] 徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011,34(08):1423-1436.

[40] 祖弦,谢飞.LDA 主题模型研究综述[J].合肥师范学院学报,2015,33(06):55-58+61.

[41] Deerwester S, Dumais S, Furnas G, et al.Indexing by latent semantic analysis.Journal of the American Society for Information Science, 1990, 41 (6) :391-407.

[42] David M. Blei and Andrew Y. Ng and Michael I. Jordan. Latent Dirichlet Allocation[J]. Journal of

machine learning research, 2003, 3(4/5): 993-1022.

[43] Blei D,Lafferty J.Dynamic topic models.In:Proc.of the ACM SIGKDD,2006:424-433.

[44] AlSumait L, Barbara D, Domeniconi C.On-line LDA:Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking[J].In:Proc.of the IEEEInternational Conference on Data Mining, 2008:3-12.

[45] 关鹏,王曰芬.科技情报分析中 LDA 主题模型最优主题数确定方法研究[J].现代图书情报技术,2016 (09):42-50.

[46] 邢富坤.中文分词中未登录词分布规律及处理方法研究[J].解放军外国语学院学报,2013,36(05):27-32.

[47] 陈力.中文古籍数字化的再思考[J].国家图书馆学刊,2006(02):42-49.

[48] 肖禹.古籍数字化中的集外字处理问题研究[J].图书馆研究,2013,43(05):27-30.

[49] 张春龙,龙京沙.湘西里耶秦代简牍选释[J].中国历史文物,2003(01):8-25+89-96.

[50] 杨宗兵.里耶秦简县“守”、“丞”、“守丞”同义说[J].北方论丛,2004(06):11-14.

[51] 陈松长.《湘西里耶秦代简牍选释》校读(八则)[J].简牍学研究,2004(00):21-26.

[52] 唐琳,郭崇慧,陈静锋.中文分词技术研究综述[J].数据分析与知识发现,2020,4(Z1):1-17.

[53] 孙茂松,左正平,黄昌宁.汉语自动分词词典机制的实验研究[J].中文信息学报,2000(01):1-6.

[54] Liu J , Wu F, Wu C, et al.Neural Chinese word segmentation with dictionary[J].Neurocomputing,2019, 338(APR.21):46-54.

[55] 周朴雄.基于神经网络集成的 WEB 文档分类研究[J].图书情报工作,2008(07):110-112.

[56] 马学良,刘玲玲.词频统计与主题分析视角下的《汉学引得丛刊》研究述评[J].河北科技图苑,2018,31(01):66-72+96.

[57] 唐榕.基于词频统计的宋词文体特征分析[J].重庆第二师范学院学报,2013,26(04):54-57+175.

[58] 李华鹏,薛峥,郭彩霞.不同局势下诸葛亮部分作品差异的统计分析[J].山西大同大学学报(自然科学版),2020,36(06):29-32.

[59] 沈刚.《里耶秦简》(壹)所见作徒管理问题探讨[J].史学月刊,2015(02):22-29.

[60] 朱圣明.里耶秦简所见秦代遷陵县公船相关问题研究[J].古代文明,2014,8(02):47-59+113.DOI:10.16758/j.cnki.1004-9371.2014.02.015.

[61] 吴方基.里耶秦简“校券”与秦代跨县债务处理[J].中国社会经济史研究,2017,(04):26-35.

[62] 刘鹏.秦代地方稟食的几个问题[J].中国农史,2018,37(01):57-68.

[63] 谢坤.出土简牍所见秦代仓、廪的设置与管理[J].中国农史,2019,38(03):46-57.

[64] 谢坤.简牍所见秦代地方稟食原则再探[J].中国农史,2020,39(05):65-72.

[65] 谢坤.蔡华祥.出土简牍所见秦代仓储管理制度[J].南通大学学报(社会科学版),2021,37(05):133-140.

[66] 睡虎地秦墓竹简整理小组编.睡虎地秦墓竹简[M].北京.文物出版社, 1990.

[67] 张燕蕊.简牍所见秦汉时期债务偿还问题刍议[J].史学月刊,2018(06):128-132.

[68] 朱红林.里耶秦简债务文书研究[J].古代文明,2012,6(03):44-50+113.

[69] 王勇.里耶秦简所见秦代地方官吏的徭使[J].社会科学,2019(05):154-162.DOI:10.13644/j.cnki.cn31-1112.2019.05.015.

[70] 刘鹏.秦简牍所见居赀赎债问题再探[J].北京社会科学,2021(08):44-56.

[71] 于洪涛.从里耶简看秦代紧急公文种类与递送方式——兼谈秦汉《行书律》相关问题[J].档案学通讯,2018(06):83-87.DOI:10.16113/j.cnki.daxtx.2018.06.018.

附录

表 5　词频共现统计表

遷陵	词频	**敢言**	词频	**洞庭**	词频	**令史**	词频	**司空**	词频	**粟米**	词频	**酉陽**	词频	**守府**	词频	**鄉守**	词频	**視平**	词频
遷陵	616	敢言	412	洞庭	270	令史	208	司空	212	斗	103	酉陽	84	守府	91	鄉守	90	視平	87
洞庭	174	朔	161	遷陵	151	視平	63	敢言	84	粟米	101	遷陵	29	敢言	30	朔	47	令史	61
丞	160	上	107	署	84	斗	55	朔	73	石	72	丞	21	朔	29	人	35	稟人	41
朔	132	署	92	敢言	78	稟人	36	人	66	朔	34	人	19	遷陵	26	敢言	33	出稟	35
守	107	謁	91	報	63	出稟	26	守	65	令史	27	敢告	17	行	23	啓陵	31	感	36
手	87	手	85	朔	60	朔	37	署	60	少半	26	守	13	手	23	貳春	26	手	34
敢言	84	報	75	陽陵	58	感	37	受	48	出稟	13	令	13	丞	21	都	20	史	32
行	56	令	75	縣	54	遷陵	34	報	45	升	23	洞庭	11	人	21	佐	20	斗	32
令	52	遷陵	72	手	52	手	34	陽陵	40	佐	22	朔	10	守	19	受	18	佐	28
署	48	丞	70	發	45	史	33	縣	39	徑廣	20	主	10	令	18	稟	18	朔	25
發	46	守	67	丞	44	佐	33	謁	36	稟人	19	過	10	上	17	倉	13	石	20
郵行	43	司空	54	守	43	石	29	爲	34	視平	19	報	9	來	17	司空	12	粟米	18
敢告	39	今	53	司空	42	監	27	手	33	半	17	手	9	書	16	石	12	尚	14
律令	37	陽陵	53	尉	42	粟米	26	丞	32	倉守	14	它	9	泰守府	15	斗	12	升	14
報	37	發	51	謁	39	丞	22	上	31	感	13	司空	8	曰	15	令史	12	壬	14
謁	36	言	49	郵行	34	敢言	21	遷陵	29	泰半	12	上	8	洞庭	15	粟米	11	倉	13
尉	36	曰	47	令	32	守	20	洞庭	29	守	12	令史	8	發	13	上	10	援	12
今	34	洞庭	44	上	31	倉	19	付	28	史	11	書	8	者	12	作徒簿	10	倉守	12
告	30	縣	44	叚	28	尚	19	發	28	士五	9	署	8	言	12	令	10	尚	12
從事	30	者	40	行	26	升	19	今	27	田官	8	發	8	快	12	手	10	扁	11

攻读硕士学位期间取得的科研成果

[1] 数字人文视阈下的秦汉简牍文本挖掘研究——以里耶秦简牍（一二卷）为例[J].渭南师范学院学报,2022,v.37;No.314(06):88-93.

致谢

时光荏苒，硕士生涯转眼间已经接近尾声，在西北大学学习的这三年使我受益良多。在学习上逐渐真正理解和投入到科研中，在生活中得以快速成长，在思想上迅速成熟。

首先要感谢我的导师陈镱文教授。在这三年的研究生生涯中，陈老师一直引领着我不断前进，在我迷茫的时候给予我建议，使我逐渐明确了自己的课题和方向。在我的论文写作过程中，陈老师时时督促着我，给予我细致耐心的指导，使得我能够顺利撰写完成毕业论文。除了在学术上给予我悉心指导外，在生活中也一直帮助和照顾着我，每一次与老师的交流、讨论都让我收获良多。

其次要特别感谢曲安京教授，曲老师学识渊博，为人谦和，经常对我们的科研工作进行积极的引导，并帮助我们打开视野，鼓励我们寻找真问题，经常与同学们深入交流。在曲老师的引领和教导下，我逐渐明白了应该如何去进行科研，同时也学会了许多做人做事的道理。除此之外还要感谢高研院所有的老师，感谢唐泉老师、袁敏老师、王昌老师、赵继伟老师、高洋老师、李威老师、胡鹏老师、陈明老师在课堂上传授我各类专业知识，为我进行学术研究打下了坚实的基础，还在我的实验和研究方面提供了许多建议，为我的论文撰写提供帮助和支持。

感谢各位师兄师姐，感谢各位同门对我学习上的帮助和支持，这三年时光与大家相伴学习是我的荣幸。

最后要感谢我的父母，在我这三年的求学过程中给予我精神及生活上的极大支持与关心，使我能够专心完成学业、健康成长。

三年的时间虽短，但使我受益终生。