【秦地文化研究】

数字人文视域下秦汉简牍文本挖掘研究

——以里耶秦简牍(一、二卷)为例

朱 琳,冯慧敏,刘 铭,张鹏雷,唐 杰,陈镱文
(西北大学 科学史高等研究院,西安 710127)

摘 要:近年来出土的秦简牍材料,为研究战国晚期至秦代的历史提供了丰富而可靠的资源。文章以目前已经校读出版的里耶秦简一、二卷为语料来源,通过特征词提取、词频统计、文本摘要及主题模型等文本挖掘技术,从词频、事件、主题三个维度对简牍内容进行挖掘与分析。实验表明,一、二卷中主要为官府文书,内容涉及文书邮传、债务、粮食、徒簿、案件、刑徒管理等内容,能够较好地反映里耶一、二卷的大体内容,并且与校释第一卷中给出的主题内容大体一致。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。

关键词:里耶秦简;文本挖掘;文本摘要;主题建模
中图分类号:K233 文献标志码:A 文章编号:1009-5128(2022)06-0086-08
收稿日期:2021-12-13
基金项目:陕西省重点研发计划项目:数字化文化资源平台的智能分析与利用研究(2019ZDLGY17-03)

作者简介:朱琳,女,陕西咸阳人,西北大学科学史高等研究院硕士研究生,主要从事数字人文研究。

DOI:10.15924/j.cnki.1009-5128.2022.06.009

近年来出土的秦代简牍材料,为研究战国晚期至秦代的历史提供了鲜活而可靠的资源。其中里耶秦简内涵丰富,作为研究秦历史文化的重要实录性材料,对于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,其文本晦涩难懂,涉及诸多人物、事件,给人工分析带来了很大困难,严重影响了简牍文献材料的开发与利用。

近年文本挖掘技术的发展为大规模的历史文化资料定量分析提供了坚实的基础,学者已经成功将文本挖掘技术应用于《全唐诗》[1]《汉典》[2]《孟子》[3]及《左传》[4]等传世文献的分析。而针对出土文献则更注重建立相关数据库[5-6],甚少涉及利用文本挖掘技术对简牍文本进行分析。

相较于其他简牍,里耶秦简数量巨大且内容丰富,仅一、二卷就有八万余字,除了文本数量优势外,释读得如此完整的简牍更是少见,已出版的《里耶秦简校释》一、二卷[7-8]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。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。虽然文本挖掘实验是一种探索性、创新性工作,但不难看出,里耶秦简一、二卷的主题挖掘结果与校释一卷中给出的主题内容徭役、邮传、禀食、案件诉讼等等均能对应,证实了通过文本挖掘方法来把握其主题内容具有一定的可行性。

文章利用文本分析技术对里耶秦简一、二卷文

本进行研究,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。

二、研究方法

(一) 研究框架

根据目前的史学研究可知,里耶秦简中大多是迁陵县廷与其上级洞庭郡府和下属司空、仓官、田官诸署以及都乡、启陵、贰春三乡的往来文书和各种簿籍,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。[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,对其进行数据预处理,包括分词、去停用词以及运用 TF-IDF 算法对语料进行特征词提取,得到其特征词语料。在此基础上,运用词频统计、文本摘要及 LDA 主题建模等多种挖掘方式对其中的主要内容及制度进行词、事件、主题等不同维度的挖掘。

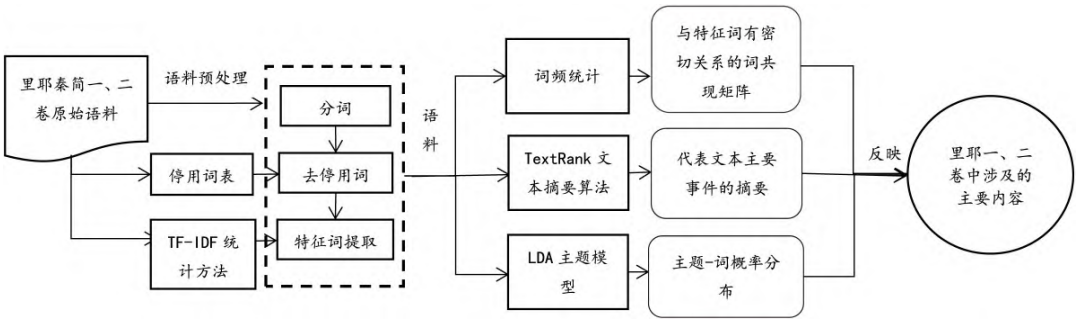

图 1　里耶秦简一、二卷文本挖掘研究框架

(二) 关键技术

1.语料预处理

语料预处理包括对简牍文本进行分词、去停用词和特征词的提取,由于简牍文本中存在大量未登录字及残断情况,为语料预处理带来了极大困难。根据文本特点,本文的语料预处理方案如下: 第一,分词。简牍语料分词工作相对于现代汉语及传世文献来说存在更大的处理难度。实验采用本课题组得到的分词结果。第二,去除停用词。为了得到更为有效的结果,需要去除其中与主题无关的词以及符号,本文采用的停用词表为针对里耶秦简一、二卷人工制作的停用词表。第三,基于 TF-IDF 的特征词提取。在经过去停后,还需要对语料进行关键词提取以提高模型数据输入的有效性,TF-IDF (词频—逆文档频率) 算法是信息检索中最常用的一种文本关键信息表示法,用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。其计算公式为

$$TF\text{-}IDF = TF \times IDF。 \quad (1)$$

其中 TF 为词频,IDF 为逆文档频率。对经过(1)(2) 步处理的语料进行 TF-IDF 计算,经比较,我们选取了排名前 20 的词汇作为文本的特征词,得到一、二卷的特征词语料。

2.TextRank 文本摘要算法

目前,文本摘要算法主要分为抽取式方法和生成式方法,前者是从原始文档中提取关键文本单元来组成摘要,文本单元包括但不限于单词、短语、句子等。抽取式方法产生的结果会留有正确的语法及原文中的显著信息,后者则会根据对原始文本的理解来形成摘要,模型试图去理解文本的内容,为了接近摘要的本质,可能会生成原文中没有的内容。根据前述的简牍文本特点,本文利用抽取型的 TextRank 算法来处理语料。TextRank 模型是一种基于图模型的关键词排序算法。[3] 首先其不需要很大的文本量,只有单篇文档亦可以进行摘要抽取; 其次无需事先对文档进行训练,这与机器学习中的隐马尔可夫模型、条件随机场等模型不同。

综上,本文利用 TextRank 文本摘要算法将里耶秦简的主要事件从庞杂的语料中抽取出来,以抽取结果来探究简文中的主要内容。

3.LDA 主题模型

LDA(Latent Dirichlet Allocation) 是 D. M. Blei 于 2003 年在 PLSA(概率潜在语义分析) 模型的基础上提出的一种更加泛化的文本主题模型,称作潜

在狄利克雷分布，通过无监督的学习方法发现语料库中隐含的主题信息。[10] 本文利用LDA模型的主题建模结果之一，即主题—词概率分布来获取一、二卷中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。

在LDA主题建模中，主题数目对于结果的分析至关重要，若主题太少可能会将不相关的主题融合在一起，若主题数量太多又会导致相关内容分裂为单独的主题，造成主题冗余或不相关主题的积累。困惑度是用来评估语言模型优劣的指标，较小的困惑度意味着模型对新文本有较好的预测能力。[11] 在LDA主题模型中，困惑度计算公式如下：

$$Perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(W_d)}{\sum_{d=1}^{M}N_d}\right\}。\quad (2)$$

其中，D表示语料库的测试集，共M篇文档，N_d表示每篇文档d中的单词数，W_d表示文档d中的词，$p(W_d)$即文档中W_d产生的概率。

本文最优主题数目的评价采用困惑度指标最小与人工验证相结合的方法。根据困惑度最小可以确定最佳主题数目，但是仅根据困惑度选取的主题数目往往偏大，从而导致抽取的主题之间辨识度不高，因此本文采用上述两种方法结合验证。

三、实验结果与分析

本部分根据前述研究框架对语料进行处理，运用其中的实现方法分别对其进行处理，并根据实验结果对里耶秦简一、二卷中的主要内容进行分析。

（一）词频统计分析

本实验采用TF-IDF算法计算得出前十个特征词为关键词，提取出与其相关的所有语料，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。词频统计结果见表1。

表1 里耶一、二卷文本关键词共现结果

关键词(词频)	相关词(词频)									
迁陵(616)	洞庭(174)	丞(160)	朔(132)	守(107)	手(87)	敢言(84)	行(56)	令(52)	署(48)	发(46)
	邮行(42)	敢告(39)	律令(37)	报(37)	谒(36)	尉(36)	今(34)	告(30)	从事(30)	
敢言(412)	朔(161)	上(107)	署(92)	谒(91)	手(85)	报(75)	令(75)	迁陵(72)	丞(70)	守(67)
	司空(54)	今(53)	阳陵(53)	发(51)	言(49)	曰(47)	洞庭(44)	县(44)	者(40)	
洞庭(270)	迁陵(151)	署(94)	敢言(78)	报(63)	朔(60)	阳陵(58)	县(54)	手(52)	发(45)	丞(44)
	守(43)	司空(42)	尉(42)	谒(39)	邮行(34)	令(32)	上(31)	段(28)	行(26)	
令史(208)	视平(63)	斗(55)	禀人(36)	出禀(26)	朔(37)	感(37)	迁陵(34)	手(34)	史(33)	佐(33)
	石(29)	监(27)	粟米(26)	丞(22)	敢言(21)	守(20)	仓(19)	尚(19)	升(19)	
司空(212)	敢言(84)	朔(73)	人(66)	守(65)	署(60)	受(48)	报(45)	阳陵(40)	县(39)	谒(36)
	为(34)	手(33)	丞(32)	上(31)	迁陵(29)	洞庭(29)	付(28)	发(28)	今(27)	
粟米(101)	斗(103)	石(72)	朔(34)	令史(27)	少半(26)	出禀(13)	升(13)	佐(22)	径廪(20)	禀人(19)
	视平(19)	半(17)	仓守(14)	感(13)	泰半(12)	守(12)	史(11)	士五(9)	田官(8)	
酉阳(84)	迁陵(29)	丞(21)	人(19)	敢告(17)	守(13)	令(13)	洞庭(11)	朔(10)	主(10)	过(10)
	报(9)	手(9)	它(9)	司空(8)	上(8)	令史(8)	书(8)	署(8)	发(8)	
守府(91)	敢言(30)	朔(29)	迁陵(26)	行(23)	手(23)	丞(21)	人(21)	守(19)	令(18)	上(17)
	来(17)	书(16)	泰守府(15)	曰(15)	洞庭(15)	发(13)	者(12)	言(12)	快(12)	
乡守(90)	朔(47)	人(35)	敢言(33)	启陵(31)	贰春(26)	都(20)	佐(20)	受(18)	稟(18)	仓(13)
	司空(12)	石(12)	斗(12)	令史(12)	粟米(11)	上(10)	作徒簿(10)	令(10)	手(10)	
视平(87)	令史(61)	禀人(41)	出禀(35)	感(36)	手(34)	史(32)	斗(32)	佐(28)	朔(25)	石(20)
	粟米(18)	尚(14)	升(14)	壬(14)	仓(13)	援(12)	仓守(12)	尚(12)	扁(11)	

迁陵是秦代洞庭郡下属县，敢言、敢告、律令、从事等词均为文书的固定用语，发、邮行为文书的

传递用语，可得里耶简语料主要是以洞庭郡迁陵县的文书档案记录为主，具体往来官府涉及其上级洞庭郡、同属洞庭郡府酉阳县及迁陵县下属启陵、贰春、都乡三乡等地。秦代国家通过发达的文书制度对社会进行了有效控制，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。

通过词频结果分析可以看出，其结果与现有研究基本一致。徒簿类文书为里耶中的一类重要文书，记录了秦代迁陵县“徒隶”劳作的相关信息，记载了诸多刑徒的工作分配情况。[12] 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，在里耶中存在分配至迁陵县下属三乡即“启陵”“贰春”“都”（即“都乡”）[13] 的情况，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，阳陵县与“敢言”“洞庭”“司空”这三个主题词分别共现了 53 次、58 次、40 次，经查阳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，大意为阳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县署，吴方基认为此批文书为跨县追债的一种情况。[14] 文书完整地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”“乡守”“视平”中均有涉及。

仓储是粮食管理中的重要一环，“廪”的设置与管理，是秦代仓储禀食的核心环节，也是秦代禀食制度的关键。廪在里耶简中颇为常见，并且多有名称，秦始皇二十六年至三十一年期间出现的廪至少有四处，分别是径廪、乙廪、丙廪、西廪。[7]42

从里耶简所载看，禀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[15]，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“廪”为“径廪”，其与“粟米”共现 20 次，而“廪”在里耶语料中共出现过 30 次，由此可推测在秦迁陵县辖区内，“径廪”曾是最常使用的一个粮仓，这与谢坤在对里耶中的“廪”进行分析时所得出的结论一致。[16]

官方的粮食管理机构向特定人员发放粮食的行为称为禀食，禀食过程需要遵循一定的程序。目前可知禀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”几个主要环节。[17-18] 而具体发放时，则需要多人参与完成。睡虎地《效律》中简 27-28 记载“某廪禾若干石，仓啬夫某、佐某、史某、禀人某”[19]，此简为秦代关于禀食文书记录的官方依据。

从上述词频共现结果可得，“粟米”“令史”“视平”“乡守”等关键词均指向禀食的管理以及记录，与上述几个关键词共现的禀人、史、佐、仓、仓守、乡守等均为禀食发放的管理者，经过整理，里耶秦简所见禀食文书记录也可大致总结为“某廪粟米若干石，某年月仓/仓守/乡守/田官守某、佐/史某、禀人某出禀某，令史/令佐某视平，某手”。由此记录可看出，禀食发放至少需要三人以上参与才行，并且分工明确，仓、仓守、乡守、田官守等负责主理，佐、史负责记录监督，禀人负责具体劳作。

结合简文，我们发现可以通过“佐/史某”推测禀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仓是史感禀人□　令史尚视平　”[7]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，其多次参与禀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，可推测其为秦始皇三十一年间粮食发放最重要的管理者。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禀食发放的重要管理人员。

综上，里耶秦简中主要为洞庭郡迁陵县的文书档案，包括其与上属洞庭郡府，同级酉阳县以及其下属启陵、贰春、都乡三乡的往来文书，且其中多为债务、粮食及徒簿类文书。可以看出从词频角度对简牍材料的文本内容进行分析具有一定的可行性。

(二) TextRank 文本摘要

根据上述研究框架抽取文本摘要，见表2，并从事件的角度对其内容进行分析。

表2 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 烦 故 为 公田吏 徙属 事 答 不备 分 负 各 石 少半 斗 直钱 百 烦 冗佐 署 迁陵 今 上 责 校券 谒告 迁陵 令 官计 者 定 钱 百 受 旬阳 左公田 钱计 问 可 计 付 署 计 为报 敢言 辛亥 旬阳 丞 滂 敢告 迁陵 丞 主 写 移 移券 可 为报 敢告主 兼 手 庚子 迁陵 守丞 敬 告 司空主 律令从事 言 懻 手 即 走 申 行 司空
摘要2	戊戌朔 癸卯 尉守 窃 敢 洞庭 尉 遣 巫 居贷 公卒 安成 徐 署 迁陵 今 徐 壬寅 事 谒令 仓贷 食 移 尉 展 约 敢言 癸卯 迁陵 守丞 臆 告 仓主 律令从事 逐 手 即 徐 入
摘要3	辛丑朔 丙午 司空 腾 敢言 阳陵 逆都 士 越人 赀钱 千百 越人 戍 洞庭郡 不智 何 县 署 今 为 钱校券 上 谒令 洞庭 尉 令 越 人 署 所 县 责 受 阳陵 司空 司空 不名计 问 何 县 官计 付 署 计 为报 已 訾 其 家 家贫 弗能 入 乃 移 戍所 报 署 主责 发 敢言 戊申 阳陵 守丞 厨 敢言 写上 谒报 报 署 金布 发 敢言 儋 手 癸巳朔 朔 阳陵 遫 敢言 至今 未报 谒追 敢言 堪 手
摘要4	甲子朔 癸酉 启陵 乡守 意 敢言 廷下 仓守 庆 书 言 令佐 赣 载 粟 启陵乡 今 已 载 粟 石 为 付券上 谒令 仓守 敢言 甲子朔 乙亥 迁陵 守丞 郫 告 仓 主下券 律令从事 壬 手 乙亥 旦 守府 卬 行
摘要5	戊申朔 癸亥 洞庭 叚守 昌 谓 迁陵 丞 迁陵 上 坐 反 适 罪 当 均 输 郡中 者 人 今 皆 输 迁陵 其 听 书从事 它如律令 新武陵印 行事 丁酉 迁陵 守丞 敦 狐 告 司空主 律令从事 夫 手 走 郤 即 行
摘要6	庚戌朔 丙子 司空守 樛 敢言 前 言 競陵 荡阴 狼 假 迁陵 公船 袤 丈 尺 名 曰 求 故荆 积瓦 未 归 船 狼 属 司马 昌官 谒告 昌官 令 狼 归 船 报 曰 狼 律在 复狱已 卒史 衰 义 所 今 写 校券 牒 上 谒言 卒史 衰 义 所 问 狼 船 存 所 其 亡 为 责券移 迁陵 弗 属 谒报 敢言 庚辰 迁陵 守丞 敦狐 卻 司空 自 叚 狼 船 何 故 弗 辟 今 而 誧 曰 谒 问 覆狱 卒史 衰 义 衰 义 事已 不智 所 居 其 听书从事 懻 手 即 令 走 行 司空

摘要1为政府与其官吏之间由于工作等原因而产生的具有负债或者赔偿关系的文书。大意为官吏“烦”在任时没有按要求备足粮食，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入账目，并且将记录这一债务的校券转移至其现居县。[20]129 摘要2描写了一个因欠债而在迁陵县戍守的外郡戍卒。籍贯为巫县的徐为外郡人，由洞庭郡尉派遣至迁陵，“徐”很可能是因欠债而在迁陵县戍守的外郡戍卒。摘要3为里耶简牍9-8，是一种洞庭郡行政事务内部协调的文书，根据词频统计结果可知，此摘要记录为跨县追债记录，完整地展现了跨县追讨债务的流程和细节。[21] 摘要4为粮食的运输记录，记载了秦代迁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。摘要5为一条刑徒管理记录，描述洞庭郡安排坐罪者六十六人的服刑地点，将这些人安排于迁陵服刑，由迁陵司空管理。[22] 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，之后发现船已丢失，于是将此事写于券上并交给地方政府，相信报告了时间经过。[20]128

根据抽取结果可以看出，里耶一、二卷文本中涉及大量的债务文书，主要包括居赀赎债。为无法

足额缴纳赀罚以劳役代偿称为居赀；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）。[23]

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，相比之下里耶一、二卷中的律令类内容则主要是通过刑徒管理（摘要 5）、仓储管理、邮驿津渡管理（摘要 6）等等管理类的记录来侧面体现，通过这些记录，勾勒出了秦代官府的日常运作及各类制度，鲜活地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、粮食运输管理、刑徒管理记录。相较于词频实验，文本抽取结果均为里耶材料中较为完整且具有代表性的事件，能够较好地涵盖简牍材料的主要内容，并且能够更为细致生动地展现秦代官府之间的沟通、管理的具体细节。

（三）LDA 主题建模

依据前述研究框架，本文采用困惑度与人工评价相结合的方式对模型进行评价从而选取最合适的主题数目，结果显示，困惑度在主题数为 12 及 15 时最小，结合人工分析判断，最终确定 12 个主题为一、二卷内容中最为合适的主题数量。

表 3　主题—关键词概率分布

主题	关键词
主题 0	手　守　急　如　史　已　食　毋　库　酉　心　少　平　丙午　爵
主题 1	令　署　书　言　者　下　其　敢言　县　手　行　上　守府　当　报
主题 2	迁陵　发　洞庭　敢言　廷　丞　司空　朔　邮　官　行　乡　金布　主吏　守丞
主题 3	过　可　尺　传　酉阳　笥　此　寸　田　未　迁　中　在　买　丈
主题 4	廷　陵　尉　少内　元　留　及　壬　黔首　丙申　乙亥　水刻　追 丁巳　刻下
主题 5	为　百　曰　所　皆　今　移　主　问　弗　敢言　各　见　具　计
主题 6	斗　石　佐　半　来　令史　出　正　入　升　视平　武　夫　敬　手
主题 7	人　仓　付　吏　捕　敢　其　徒　臣　取　尽　里　户　禀　隶妾
主题 8	邮行　洞庭　迁陵　行　邮　欲　乙丑　癸酉　癸　赐　辛　大女子　毋　手　枳
主题 9	甲　货　到　不　与　子　监　令佐　它　事　遣　旦　嘉　论　就
主题 10	迁陵　粟米　手　斗　士五　石　是　禀人　辰　等　仓守　令史　数　少半　出禀
主题 11	钱　受　得　不更　狱　千百　都乡　雨　牒　当　斤　辛未　课　校　殴

如表 3 所示，实验以 12 个主题数为基础，利用 LDA 模型得到主题—关键词概率分布，分别描述了每个主题的概率分布排名的前 15 个关键词，根据特征词提取方法，这些结果词大多具有高度语义，代表了里耶一、二卷中的重要主题。

主题 2 中“迁陵”“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，杨宗兵简文例中分析，“守”“丞”及“守丞”互见无别[24]；“发”为拆阅文书，“邮”传达文书的机构，其均为传递、收发文书时所用，所以主题 2 可以定义为洞庭郡迁陵县的文书邮传。

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，当仓、少内等机构进出粮食及财物时，需要有人监督，称为“视平”，或简称为“视”“平”“监”。秦简显示，监督人主要是令史，但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”“赀”，“赀”为有罪被罚款或罚物者，“赀甲”“赀盾”皆可以代偿赀款或物，“赀甲”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“令佐赣入赀钱七百七十六 元年八月庚午 令佐赣监”[8 70-71]，所以主题9可定义为赀罚问题。

主题10中“粟米”为粮食，“斗”“石”“少半”为粮食的计量单位，“禀人”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，卅一年八月壬寅仓是史感禀人堂出禀隶臣婴自槐廪，令史悍平，六月食，感手”[7 116-117]，其中出禀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；“数”为数量，可将主题10定义为出粮数量的管理及记录。

LDA主题模型从主题的角度进一步对里耶文本进行解析，凝练出了邮传、赀罚、粮食、债务者粮食发放等等主题，由于简牍文本的特殊性和简文的多义性，其处理难度极大，其余主题中的关键词较为散乱，目前难以给出确切的定义，需要进一步针对其中较为清晰的内容进行分析。

四、结论与展望

文本挖掘方法为人文学科中的计算机辅助解释提供了一个有效的方法。本文在近年来国内外相关研究的基础上，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、二卷中主要以洞庭郡迁陵县的文书档案记录为主，具体往来官府涉及洞庭郡、酉阳县、下属启陵、贰春、都乡三乡等地。其中涉及文书邮传、债务、粮食、徒簿、案件、刑徒管理等等内容，与目前校释中给出的结论基本一致。本文将简牍研究与文本挖掘技术相结合，可以有针对性地提高简牍材料处理效率及有效性，为简牍材料的研究提供新思路，证实了人文领域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。

尽管如此，本文的主题内容挖掘方法仍存在一些局限性。首先，虽然算法不存在人为的偏见性，但简牍中广泛存在较为严重的残断及集外字情况，会极大影响模型的识别以及对结果的分析。其次，抽取型文本摘要算法虽然可为我们提供较为完整的结果，但其在抽取时会倾向于抽取有较为完整语义的内容，而里耶简牍中存在较多的残断现象，在这方面会不可避免地存在偏颇。最后，由于简牍中存在较多的残断及未登录字，其相较于传世文献及现代汉语的主题拟合结果较为散乱，仅有部分主题可以为其定义标签。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，依次对其实现文本挖掘研究。

参考文献：

[1] CHEN J W. East Asian studies macroscope [EB/OL]. [2021-06-20]. http://macroscope.cdh.ucla.edu.

[2] ALLEN C, LUO H L, MURDOCK J, et al. Topic modeling the hàn dian ancient classics [J/OL]. (2017-10-12) [2021-06-20]. https://pdfs.semanticscholar.org/a837/5978c0a48ab19e79a399cbb9e48cf9603dad.pdf.

[3] NICHOLS R, SLINGERL 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.

[4] 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(7): 30-38.

[5] 李静. 浅析简帛文献的数据库建设 [J]. 江汉考古，2017(5): 128-130.

[6] 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.

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

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

[9] 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(2):407-410.
[10] DAVID M. B, ANDREW Y. N, MICHAEL I. J. Latent Dirichlet Allocation[J].Journal of machine learning research,2003(4/5):993-1022.
[11] 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(9):42-50.
[12] 沈刚.《里耶秦简》(壹)所见作徒管理问题探讨[J].史学月刊,2015(2):22-29.
[13] 朱圣明.里耶秦简所见秦代迁陵县公船相关问题研究[J].古代文明,2014(2):47-59.
[14] 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(4):26-35.
[15] 刘鹏.秦代地方禀食的几个问题[J].中国农史,2018(1):57-68.
[16] 谢坤.出土简牍所见秦代仓、廪的设置与管理[J].中国农史,2019(3):46-57.
[17] 谢坤.简牍所见秦代地方禀食原则再探[J].中国农史,2020(5):65-72.
[18] 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(5):133-140.
[19] 睡虎地秦墓竹简整理小组.睡虎地秦墓竹简[M].北京:文物出版社,1990.
[20] 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(6):128-132.
[21] 朱红林.里耶秦简债务文书研究[J].古代文明,2012(3):44-50.
[22] 王勇.里耶秦简所见秦代地方官吏的徭使[J].社会科学,2019(5):154-162.
[23] 刘鹏.秦简牍所见居赀赎债问题再探[J].北京社会科学,2021(8):44-56.
[24] 杨宗兵.里耶秦简县"守"、"丞"、"守丞"同义说[J].北方论丛,2004(6):11-14.

【责任编辑　刘亚楠】

Research on Qin and Han Bamboo Slips Text Retrieval from the Perspective of Digital Humanities, Taking Liye Bamboo Slips (Volume I and II) as an Example

ZHU Lin, FENG Huimin, LIU Ming, ZHANG Penglei, TANG Jie, CHEN Yiwen

(Advanced Research Institute of History of Science, Northwest University, Xi'an 710127, China)

Abstract: The Qin Bamboo Slips unearthed in recent years provided the rich and reliable resources for the study of the history from the late Warring States period to the Qin Dynasty. How to make more efficient use of the bamboo slips for historical research has gradually attracted the attention of the academic circles. Based on the first and second volumes of Liye Bamboo Slips that have been proofread and published at present, this paper excavates and analyzes the contents of bamboo slips from the three dimensions of word frequency, event and topic through text retrieval technologies such as feature word extraction, word frequency statistics, text summary and topic model. The experiment shows that the first and second volumes are mainly official documents, which involve document postal transmission, debt, food, notebooks, case, prisoner management and so on. It can better reflect the general content of Liye's first and second volumes, and is generally consistent with the subject content given in the first volume of proofreading. This paper provides a new idea for the research of bamboo slips, which has a certain theoretical significance and reference value for the text retrieval research in the professional field of bamboo slips, and proves that it is feasible to understand and interpret China's rich cultural heritage with the help of text retrieval method.

Key words: Liye Bamboo Slips; text retrieval; text summary; topic modeling