攻玉以石

历史文本的词汇标记及应用

项 洁 胡其瑞

摘 要 历史文本是历史学研究的基础素材,通过对文本内容的爬网,历史学家将文本中有意义的信息整理、拼凑并脉络化。历史学是一门研究人在时间中的活动轨迹的学科,在加入地理空间的概念之后,历史文本将变得更加立体。跳脱以往在纸本数据中的线性阅读,对信息时代的历史文本,通过技术的协助增添词汇标记,再利用对标记词汇的分析与可视化,鸟瞰并掌握历史文本中隐含的脉络。通过探讨历史文本中人物、时间、地名与对象词汇标记对历史研究的意义,描述各种标记的目的与特性,尤其指出词汇标记不只是辨识词汇,还需要达到"消歧"与"聚合"的功能。同时介绍两个自动标记工具——"码库思古籍半自动标记平台"(MARKUS)和"批次标记工具"(CT Tool)。这两个工具使得大量快速标记人、时、地、物成为可能。透过实际的研究成果案例,说明如何运用标记过的文本;透过时间、人物、地理与对象词汇标记的实际效益,说明历史文本中的词汇标记及其在历史研究中的应用。最后讨论事件标记的问题,指出事件标记与其他词汇标记本质上的不同。

关 键 词 词汇标记; 数字人文; 历史文本; DocuSky; MARKUS
分 类 号 K061

作者简介 项洁(通讯作者),台湾大学资讯工程学系教授,台湾大学数位人文研究中心主任,Email:jhsiang@ntu.edu.tw; 胡其瑞,台湾大学数位人文研究中心研究员。

0 前言

自从文史研究者佛朗哥·莫雷蒂(Franco Moretti)提出"远读"(distance reading)概念,试图将一个文体类型(genre)透过整体性的观察来分析、探讨文学发展的关联性以及文学形式转变的原因后,这种将大量文本融于一炉的想法似乎暗示只要通过大量的电子全文,辅以适当的数字工具,就能够在其中找到一种脉络关系用于分析[1]43,56,399。基于这种理念的颠覆性,许多数字人文相关文章纷纷引用莫雷蒂的想法,用以说明数字人文研究在"远读"的理念下获得实践。在数字人文研究者的诠释下,莫雷蒂所提出的"远读"概念被认为是在可能"牺牲信息的情况下看到超级大量文本中的趋势、结构等普通阅读无法发现的现象"[2],更有论者称"远读"就是数字人文的起点[3]。

莫雷蒂在其所撰的《标题公司:对七千个标题的反思》(*Style, Inc. Reflections on Seven Thousand Titles*)一文中利用对1740—1850年所出版的七千部小说的标题长短、标题词汇进行量化分析的结果,探讨这时期小说出

版的发展与写作风格的转变[4],给文学风格分析(literary style analysis)或作者判定(authorship attribution)等议题带来许多开创性的研究观点。以巨量的文本实践数字人文方式的探讨,确实突破了以往的研究路径。信息技术的协助使以往力有未逮的人文研究方法获得了突破性的发展。

历史文本和文学创作却有很大的差别。譬如小说的创作,它的时间、背景可以是虚构也可能为真实,地点也是一样,人物更可能是作者天马行空的创意。而不同的小说与小说之间,很难从"内容"或"人物"看出有脉络可循,除非是系列的创作如《名侦探福尔摩斯》(*Sherlock Holmes*),小说的主人公很少重复,剧情也很难发生跨小说文本的延续关系。历史文本却不同,同一个人物在同一时期的不同历史文本中往往重复出现,同样的事件也可以在不同文本中用迥然不同的角度描述。地点和时间更是如此,同一个地点可以在上千年的不同文本中扮演重要而不同的角色,同一个时间更是会在大量同时代的文本中被提及。而且不同时代、地域的历史文本,虽然出自不同的作者,但是书写的格式往往大同小异,以至如莫雷蒂希望的借由文本进行的风格分析在历史文本上似乎无用武之地。然而阅读大量文本本就是历史研究重要的工作,诚如王泰升所说,历史文本的论述必须基于史料,参考的史料越多,对历史论述的掌握就越发精确,因为"看得愈多,遗漏相关史实的可能性愈低,甚至对于某特定历史/社会现象之属常态与否的疑问,也可从发生频率见其端倪"[5]33。换而言之,历史研究的一个重要课题就是在文本中寻找有意义的脉络,但是当史料量大到单靠阅读无法快速且有效地找到脉络的时候,有没有可能运用一个像莫雷蒂建议的"远读"方法来快速掌握这些史料?要回答这个问题,首先须了解莫雷蒂提出的"远读"架构有一个隐藏的假设,也就是除了同属一个文体类型外,各个文本间并没有紧密的关联。历史文本则不同,在各个文本中不断出现的人名、地名、时间或重要的对象等具有特殊意义的词汇,有形无形地将大量文本串联在一起,甚至可以夸张地说,可以把一个历史文本的集合想象成一个单一的文本①,透过共同出现的词汇探索隐含在跨文本间的各种脉络。但是要做到这样的连接,就必须把相关的词汇做适当的标记。如何在历史文本中有效地提取与标记词汇,又如何运用它们在历史研究上萃取并观察脉络,就是本文要探讨的议题。

1 历史文本中的人、事、时、地、物

英国史学家卡尔(E. H. Carr, 1892 - 1982)在其《何谓历史?》(*What is History*?)一书中为"历史"一词做了定义:"历史就是史学家与史实间不断交互作用的过程,是'现在'与'过去'之间永无止境的对话"。[6]126这句话说得相当简洁,但是却把"何谓历史研究"这个问题作了言简意赅的诠释:所谓的历史研究,正是历史研究者在时间长流当中从现在的时空探索过去的史实。因为历史学是一门研究时间之流里所发生事件的学问,所以历史文本中人、事、时、地、物的记载,往往是史学家所关注的重要元素。

史学家布洛克(Marc Bloch)认为,历史学不仅仅是一个以人为研究对象的学科,而且是"在时间中的人的科学"。布洛克把时间比喻成为有机的生命实体,每一个历史事件,都是浸泡在"时间血浆"里的物质,而"历史学者在人与文明二者的生命航图上,描绘出精确的时刻时,才会感到他已给予一幅真确的图像"。[7]33-34"人"是事件的主角,而"文明"则是人在事件当中所创造出来的成果,历史学家的工作,就是将这些零散的信息尽力拼凑起来。

不过,人与时间只是历史研究中的两个元素,加入了"地理空间"信息后,历史研究便更为立体。中国传统学术中有所谓的"舆地学",即"沿革地理研究"。班固《汉书·地理志》以降的史书中,地理信息及其沿革可谓汗牛充栋,但这仅止于信息的整理与保存,学者更好奇的反而是这些地理信息的变化及其原因,也因此才有"历史地理学"的出现。历史地理学又可概分为"历史人文地理研究"与"历史自然地理研究"两大领域,其下

① 一千年前郑樵就做过"集天下书为一书"的想象,详见其《通志·总序》。

又可再细分更多独立的研究主题。[8]然而,探究一个空间点位上所发生的种种事件,其根本要素还是地理空间的定位。

近年来由于信息技术的发展,地理信息系统(GIS)被大量运用在历史研究中,历史事件被放上了地图,让原本只在文字中的史实,能够在空间中被观察;透过图层的数字化工作,许多以往只能在纸本上运用的地理图层,现今都能透过网络的共享取得。如中研院地理资讯科学研究专题中心(GIS 中心)搭建的"中华文明之时空基础架构"(CCTS)[9]、"台湾百年历史地图"[10]等,借由资源的整合,让历史上许多宝贵的地图图资在共通的标准下得以被研究者使用,历史文本的内容也能够跃然于地图上。

对象词汇在历史文本中相当广泛,举凡考古报告中的出土文物、朝贡国书中的进贡物品、田野调查的文物采集、地方志所记载的人造建物(宫庙、墓葬等),乃至于交易商品或是科仪文本中的宗教器物,都是在历史研究中会被注意到的研究对象。对象的记载,反映出人类文明发展的进程,在人类学研究中,常以"物质文化"(material culture)这个专有名词去探讨对象与创作者之间的关系;在历史研究中,对象在时间中的发展,则是探寻历史发展的重要脉络。除此之外,一些非物质性的词汇,如官名等,更是历史研究探索的重要主题。

以上人、时、地、物四个元素都是可以从文字直接识别出的,"事件"却是一个比较抽象而主观的概念,因为事件的名称往往是后人所给予,在事件发生的当时,历史文本中该名称不会出现。举例来说,1911 年的"辛亥革命""武昌起义"等事件名称,都是从革命党人的角度事后所赋予的。对清政府而言,这个事件是地方上的动乱,只是最后导致了政权的覆亡。若是对当时的报纸进行检索,不难发现,"辛亥革命"这个词最初是在 1913 年 1 月以后的天津《大公报》上出现的[11]。因此,事件词汇往往都是见诸事后,在事件发生当时的历史文本中并不会明确出现。

然而,历史的研究不是只有年表的排列和历史图层的呈现而已,更重要的是研究者在大量的历史文本中透过词汇来串连丰富的人、时、地、物信息,从文本当中寻找可供研究的脉络。这种寻找与提取词汇的过程在数字人文方法中就是所谓的文本词汇标记。

2 历史文本词汇标记的目的与功能

随着信息技术的进步,大批的历史文本被数字化成电子文本。举凡 1971 年所启动的"古腾堡计划"(Project Gutenberg)[12],以及台湾中研院于 1980 年代开始的"汉籍全文数据库计划"[13],将大量的历史文本通过扫描与全文缮打、文字辨识等方式制成电子文本,一些商业古籍库更是动辄数十亿字的全文;再提供全文检索,让史学研究的资料变得唾手可得。然而,大量的电子资源虽然降低了研究者在茫茫字海中寻找的困难,却也带来巨量的检索结果,以及许多数据上的噪音(noise),致使研究者反而陷入电子文本"求全"(recall)或是"求准"(precision)的困扰当中。因此,透过电子文本进行历史研究如果只是单纯仰赖电子全文,其实并不能完全满足历史学者的期待。也因此需要将不同类别的词汇进行注记与分类,以利计算机协助研究者进行文本的探勘与分析。而这种将词汇进行注记与分类的动作就被称为"标记"(tagging)。

从文本中提取词汇不是一个新的议题,信息学者透过"自然语言处理"(NLP)的技术,已经可以将文本中的词汇切分,让不同的词汇被撷取出来。大约在 1980 年代末期,利用机器学习(Machine Learning)建立的算法模型,已能让计算机在大量的中文文本中学习如何分析句子中的不同词汇类型;或是通过"断词",从一长串的文字分辨出主语、谓语、宾语等,甚至能够协助文本区分不同的词汇类别并进行语意分析。目前较广为使用的中文断词处理工具,如"结巴(jieba)中文断词程序"[14]及由中研院"中文词知识库小组计划"(CKIP Lab)所建置的"中文断词系统"[15],均是经历大量的机器学习过程之后所开放出来的公开工具,为现代中文的断词与分析带来许多的便利与技术上的突破。

但是中文历史文本词汇标记的需求和一般断词又不太相同。历史文本大多是古文,而古文的文法与白话

文差别甚大,为白话文设计的断词工具遇到文言文,效率通常大打折扣。但从另一方面看,历史文本的断词需求又比较简单,因为历史文本标记的目的不是分析词性或句子的构造,而是找出一段文字中的特殊词汇(通常是名词),这样的技术在信息工程领域叫做"命名实体识别"(NER)。这项工作通常被认为是自然语言处理的一个子问题(subproblem),在技术上比了解句子构造或语义分析要容易得多。然而历史词汇的标记不单单是找出词汇,还有其他的需求。以下将对历史文本中人、事、时、地、物的词汇标记需求做整体性的介绍。

2.1 时间的标记

"时间"当然是历史研究中最重要的信息之一。中国历史特殊的纪年方式使得读者在阅读时不易明白事件的先后,面对大分裂的时代(如南北朝)尤其如此。第一个明显指出时间排列重要性的是《史记》,太史公用《十二诸侯年表》,让阅读"自共和讫孔子"之"世家"和"列传"的读者获得一定的时间先后感。

标记时间的关键是用一个共通的方法把不同的时间放在同一个体系下做衡量,也就是需要一个共同的标准来进行时间的正规化。时间正规化通常是以公元作为标准①,各种不同的纪年方式与其对应,但需要强调的是,这并不是用公元时间取代原有的时间记录,而仅是用共同的公元时间当作时间对应的锚点(anchor)。举例来说,清咸丰年间,正值清政府与太平天国作战。太平天国有自己的"纪元",如咸丰八年四月六日是"太平天国八年四月九日",同时是日本孝明天皇的安政五年四月六日,也是朝鲜哲宗的九年四月六日。这样紊乱的时间信息,确实相当需要统一的时间规范。上述这一天是公元纪年的 1858 年 5 月 18 日,以"1858 - 05 - 18"这个锚点为标准,可以将上述四个不同的纪年进行串连,使同一天内东亚不同国家/地区发生的事情得以并列进行观察。

时间是一个绝对的观念,也是一个相对的观念。有时历史记录不会记载确切的时间,关心的反而是事情发生的先后。而且有些时间是无法考证的,如上古的诸多传说;或无法进行确认,如只记录"光绪十三年"而没有月份,甚至只有"光绪年间"。对这些情况,代表时间的标记最好均以"起""讫"来界定,如咸丰八年四月,标记时的"起"为"1858 - 05 - 13"(咸丰八年四月初一),"讫"则为"1858 - 06 - 10"(咸丰八年四月二十九日)。如果是咸丰八年四月初六日,则"起讫"均以"1858 - 05 - 18"标记。

对东亚时间标记研究比较早的是日本开发的"HuTime"[16],至于公开的中公历时间对应参考库,至少有法鼓文理学院提供的"佛学规范资料库"之下的"时间规范资料库"[17],以及可以查找更大时间范围的中研院"两千年中公历转换工具"[18]等。其中,法鼓文理学院的"时间规范资料库"甚至可以提供邻近的日本、朝鲜等国的纪年对应。台湾大学亦创建查找明代以降时间信息的"中西历对照查询系统"[19]。

2.2 人名的标记

除了辨识出某个字符串是否为人名外,历史人名的标记还有另外两个重要的目的,就是"消歧"与"聚合"。"消歧"的意思是确认同样的名字指的是不同的人。根据"中国历代人物传记资料库"(以下简称"CBDB")的记录,姓名为"王维"的人历史上就有 11 个之多,其中唐朝两人、宋朝四人、明朝五人,人名标记系统需要知道文本中出现的"王维"是哪个王维。"聚合"的目的恰恰相反,由于中国历史人物往往除了本名还有字、号或是别名,在一个文本当中,有时以全名记录,有时以字号称呼,有时甚至只记载姓名当中的一两个字。以《台湾郑氏纪事》的一段文字为例:

庆长十七年壬子(明万历四十年),明 郑芝龙 及祖官来谒幕府于骏府,幕府亲问以外国事。 芝龙 献药

① 在 1582 年教宗格里高利十三世(Gregorius PP. XIII)颁布"格里高利历"(Calendarium Gregorianum)之前,西方的历史记载大多采用的是"儒略历"(Julian calendar)。由于格里高利历与儒略历之间有着些许的误差,在进行两历的换算时需要一些工具协助,在准确的历法日期确认后方能进行时间信息的标记。通常做法是用儒略历做标准,但为说明简单起见,本文以格里高利历说明。

品(武德大成记、国史、武德编年集成。按祖官不详何人),幕府命馆之长崎(逸史)。芝龙字飞黄(郑成功传。本书曰:“小名一官”。按当时明人来我邦,率匿名称某官,盖一、二排行之类,官称爷若郎之类,犹唐人五郎、三郎称也;为小名者恐误,故不取焉),后号飞虹将军(武经开宗、华夷变态),泉州南安县石井巡司人也。父绍祖。芝龙兄弟四人,仲芝虎,叔鸿逵,季芝豹,伯为芝龙。芝龙生而姿容秀丽(郑成功传)……按芝龙至骏府与居平户,岁月前后不可得而详,姑系于此),称平户老一官(琉球事略)。后乘商舶数来往本邦(长崎夜话草)。[20]

上文中框内的词指的都是郑芝龙。如果只是针对相异词汇进行统计,“郑芝龙”“芝龙”“飞黄”“一官”“飞虹将军”“平户老一官”这六个不同的词就会被认定为六个不同的人物,词汇与词频的统计就会与事实不符。

以上例子说明,中文的人名词汇,不是单纯地只要能够识别就好,真正能够建立文本间脉络关系的,是将不同意义的词汇进行“消歧”,将相同意义的词汇予以“聚合”。处理消歧和聚合的关键在于给不同的人物一个不同的识别序号(Reference ID)。例如唐朝诗人王维在 CBDB 里的序号是 cbdb32174,明万历进士王维则是 cbdb212634,就像身份证号码一样。在词汇标记的时候将该人物所属的序号一并加入到人物词汇当中,就能够实现历史上众多同名者之间的消歧。聚合也可以通过使用同一个识别序号的方式实现。如将郑芝龙在 CBDB 的序号 cbdb59197 标记于其六个别号词汇后,这些别号就会获得同样的 ID,聚合在一起(图 1)。这样系统在进行词汇统计的时候,可以将这六个词汇的总数量合并统计在“cbdb59197”这个序号之下。

图 1　以识别序号实现人名词汇聚合示例(以 DocuSky 制图)

除了 CBDB 外,可以协助人名标记的历史人物参考数据库还有法鼓文理学院以佛教人物为主的“人名规范资料库”[21]。这两个数据库不仅仅提供人名与序号,还提供每个人名的相关信息,如字号、生卒年、亲属关系、籍贯等。这些信息可以为进一步的脉络萃取所用,且提供一些不易察觉到的关系,譬如在一个书信的数据集中,如果人名是用 CBDB 标记,则很容易可以看出哪些书信往来者有相同的籍贯。这些序号也能够协助我们向序号的提供者进行相关数据的取用,即为“资源参照”。

2.3　地名的标记

文本中地名词汇标记的挑战与人名类似,在识别文本中的地名之后,也有消歧和聚合的需求。这不难想

象,因为中国本来就有许多重复的地名,一个地方多次改名的例子更是屡见不鲜。地名标记比较特别的是,每个地名理论上都应该有一个地理经纬度坐标,如果能够把这个坐标标出,利用地理信息系统来呈现,会有强烈的可视化效果,所以通常研究者也希望标出地名的经纬度坐标。然而地名亦有时间的纵深,历史上同一个地名即使看似指代同一个地方,在坐标上差异可能很大。譬如浙江的永嘉县,自东晋建郡以来一直主要在瓯江以南,然而 1949 年后改至瓯江以北(原来瓯江南岸县治变成温州市市治),如果对照地图阅读史籍的时候忽略了这个变化,可能会产生很大的困惑。

和人名标记一样,地名标记的消歧、聚合与历史名称更迭的问题也可以通过赋予各个地理位置一个唯一的识别序号来解决。在这里需要说明,经纬度坐标本就是一个地理位置的唯一标志,再加一个识别序号似乎是多此一举,但需要多一个识别序号的原因在于不是每一个历史地名都可以找到坐标:有的考察不到,有的连存在与否都存疑,所以还是需要一个唯一的识别序号来标记。

有一些提供历史地名序号的参考数据库也同时提供了该地名的经纬度坐标,搭配适切的工具就能够从数据库中同步取得坐标信息,并将坐标转化成点位标记在地图上。目前公开的可以通过应用程序接口(Application Programming Interface,API)进行介接使用,且可提供目标历史地名的数据库至少有哈佛大学"中国历史地理信息系统"(以下简称"TGAZ")[22]、中研院 GIS 中心的"中华文明之时空基础架构系统"、台湾大学数位人文研究中心为台湾地名制作的"台湾历史地名坐标资讯库"(以下简称"TWGIS")[23],以及法鼓文理学院为佛学研究制作的"地名规范资料库"[24]。

2.4 对象的标记

对象的种类可以很多,在我们处理过的数据里至少有官名、草药、疾病、贸易商品、契书、墓葬、动物和植物。消歧的问题一般来说比较少见①,聚合的问题则十分常见,如《历代宝案》中记有琉球王国对中国进贡胡椒[25],但文书中有时写"胡椒",有时写"楜椒",后者或许是笔误,但也可能是"胡椒"的另一种称呼。若是单纯将上述两个"胡椒"独自标记,对于计算机而言就会变成两种不同的对象而分别独立计算。解决这些问题的方法也是唯一的识别序号,但与人名、地名不同的是,通常没有像 CBDB 这样现成的参考数据库可以使用并协助消歧与聚合,所以针对一个特定的对象标记需求,要创造一个需要标记的词汇列表,再赋予列表上每个代表不同对象的词一个识别序号。

事实上这个问题在其他的标记中也会发生。譬如,并不是每个出现在历史记录中的人物都在 CBDB 里,尤其是像地方契约文书中的小人物,几乎没有哪位会在其他的历史记录中出现,也不是每个地名都被某个地理数据库收入。遇到这种情形,需要用一些词汇萃取的方法去发掘未知的词汇。如前所言,命名实体识别是自然语言处理的一个子领域,虽然研究成果丰硕,但仍然没有一个完全自动的方法可以从文本中萃取出所有需要的词汇,尤其对于中文的文言文,效果通常不是很理想。文言文的人名萃取可使用主动学习(active learning)的方法[26],也可使用逐点互信息(PMI)的方法,即对研究文本进行断词,配合规则找出候选人名,然后针对文本中的人名特性,进行人名验证[27]。在一个用《资治通鉴》做文本的实验中,这个方法萃取了 18893 个人名、10085 个地名和 13888 个官职。[28]一个比较互动和有效的做法是词夹子,就是利用名词前后相关词的特性,"夹"出有相同性质的词汇[29],直观且具有高互动性,如在"DocuSky 数字人文学术研究平台"(以下简称"DocuSky")②中的"文本撷词工具 2020 版"即是以词夹子萃取词汇的工具[30]。

① 但也有例外,如有些药名在不同时代代表不同的草药,官名亦同。

② "DocuSky 数字人文学术研究平台",是由台湾大学数位人文研究中心、资讯工程学系数位典藏与自动推论实验室规划,项洁教授主持、杜协昌博士设计开发的以向人文学者提供数据转文件、上传、建置个人化数据库与运用数字工具进行文本分析服务为目的的数字人文研究平台,详见:http://docusky.org.tw。

2.5 事件的标记

事件标记的性质和前四种标记不同,前四种标记主要是对词汇的标记,事件标记则是对一段文本的注释,所以事件的标记往往会以元数据(metadata)的一个字段表示。自动标记事件的困难在于事件和文本的内容性质是息息相关的,也和文本的语义(semantics)相关,所以无法设计一个一般用途的事件标记方法。在"台湾历史数位图书馆"(以下简称"THDL"),我们为古地契文件集设计了一个"土地移转图"的事件标记,先设计自动方法找出同一块土地的上下手契关系,并用它们的传递闭包(transitive closure)①形成一个土地移转图,所代表的是那块土地在数据中集中显示的整个活动历史。[31]借助这种方法所形成的最大的土地移转图包含 103 张土地契约文书。[33]针对 THDL 中的明清台湾相关行政档案,我们也设计了"奏折/上谕引用关系图"的事件标记,将前后引用的奏折和上谕串联起来[34],其中最大的引用关系图与林爽文事件有关,有 153 份奏折和上谕互相引用。另外一个有趣的例子是关于巡台御史存废问题的,从引用关系图中可发现迭次引用的奏折/上谕有 28 件,而且时间从康熙六十年(1721 年)到道光十二年(1832 年),跨越 112 年。这种文件间的脉络用人工的方式几乎无法发现。蔡宗翰也设计了一种基于自然语言的处理方法并应用在《明实录》上发掘事件。[35]

因为事件标记的语义特性,目前并没有具有一般性的标记方法,为了特定领域而设计的做法也都只能以前处理(pre-process)的方式应用,本文的讨论部分将做进一步的说明。

3 标记文本词汇的数字工具

一个数字文本中需要标记的人、时、地、物的词汇很多,如果人工进行标记,显然旷日费时,所以需要自动的方法。本节介绍两种方法,一种是荷兰莱顿大学开发的"码库思古籍半自动标记平台"(以下简称"MARKUS")[36],一种是台湾大学开发的"批次标记工具"(Content Tagging Tool, 以下简称"CT Tool")。

MARKUS 的标记原理类似阅读文本时"画重点"的习惯,将不同类型的词汇用不同的颜色显示出来,而且大部分画重点的动作都能够以自动的方式协助使用者完成。在开始自动标记之前,用户可以分别选择用内建的 CBDB 或法鼓"人名规范资料库"标记人名,用 TGAZ 或 TWGIS 标记地名,用法鼓"时间规范资源库"标记时间及用法鼓佛学词汇库标记佛学词汇。其他研究所需的词汇类别也可用自建词汇表的方式处理。② 选择完后,只需要一键操作,MARKUS 就会把在各种词汇集里的词汇用不同的颜色标注。图 2 是对《新唐书·列传文艺》(卷 127)中关于王维的文字进行标记所得的结果。

MARKUS 是半自动而不是全自动的工具,因为在 MARKUS 标记完后,用户还是需要进行除错的步骤。除了要添加 MARKUS 没有标记到的词汇与修正标错的词汇外,使用者还须手动消歧,因为一个人名可能代表不同的人。在图 3 中出现的人名"裴迪"下面有打点的标记,这表示"裴迪"在 CBDB 中有多于一个的 ID,所以使用者需要手动决定这位"裴迪"是历史上的哪位裴迪,并在四个 ID 中选取适合的代码。

人名聚合的问题在 MARKUS 中可以比较自动地解决。如果一个人名是用 CBDB 标记,MARKUS 会将同一个人的姓名和其在 CBDB 中的别号、别名等用同一个 ID 标记。

地名的标记也相仿。如图 4 所示,借由 MARKUS 与 TWGIS 的介接,可以把"中庄仔"标记为地名之后,在 TWGIS 中查找出合适的 ID——twgis36813。在标记地名 ID 之后,对于数据库中所记录的该地名的坐标信息,也能够经由这个 ID 向数据库取用,并通过合适的工具在地图上显示位置。

① "传递闭包"是一个数学名词,在此处大意就是把相关的上下手契一个一个串起来,直到没有可串的为止。

② MARKUS 最近也内建了朝鲜相关的人、地、官名词汇表,详见:http://dh.chinese-empires.eu/beta/。

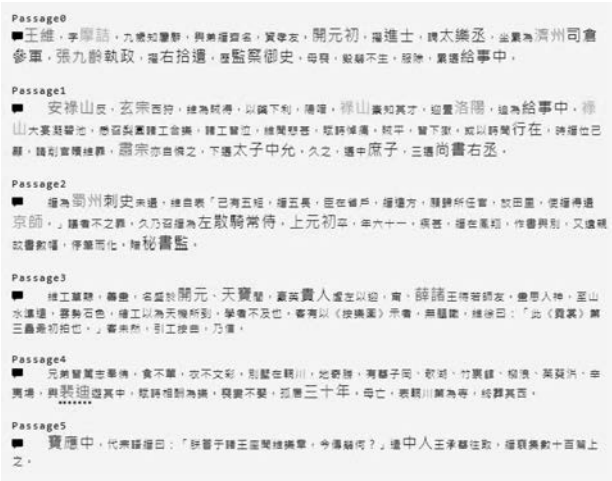

图 2　MARKUS 对《新唐书》“王维”段的标记

图 3　MARKUS 中所显示的 CBDB 人名 ID 信息

图 4　MARKUS 与 TWGIS 介接使用示例

MARKUS 的时间词汇标记参照法鼓文理学院的“时间规范资源库”，自动标记功能可以自动识别可能的中历时间（年月日须以中文书写）并进行标记，为较为明确的中历时间加上“公元年月日：朝代帝号”形式的时间权威信息，如图 5 所示。

图 5　MARKUS 的自动标记时间功能

MARKUS 有几个可以改进的地方。第一是目前每次只能标记一个文本,所以如果有许多个文本需要标记,同样的步骤需要重复很多次;第二是标记非内建词汇类别的流程有些繁复,对不同的对象种类需要建不同的词表。为了解决这些问题,我们在 DocuSky 系统里设计了 CT Tool。该工具的工作方法,是将需要标记的词汇以 EXCEL 表格进行整理,依照词汇类别(tagName)、标记词汇(tagVal)与权威控制词汇(@Term)填入。其中,tagVal 是可能在文本中出现的词,@Term 则是代表同样名词但不同写法的标记词汇(tagVal)的标准词汇(相当于上节所叙述的识别序号,只是不用一串数字表示,而是用文字)。表 1 是一个例子。

从表 1 中可以看出不同的标记类别可以在一个窗体里处理,如果用这个窗体标记一个文本,“蕃茄”和“西红柿”都会被标成标记类别“Udef_fruits”,它们的权威控制词汇都是“西红柿”。换句话说,聚合的目标就简单地达成了,而被标到的陈氏到底是郑进之妻还是王英之妻(消歧)则还是需要使用者以手动的方式决定。① CT Tool 的另一个重要功能是可以同时(批次)标记许多文本。

表 1　　用于批次标记工具中的 EXCEL 表格范例

tagName	tagVal	@Term
PersonName	陈氏	郑进之妻
PersonName	陈氏	王英之妻
PersonName	林满妹	郑进之妻
Udef_fruits	蕃茄	西红柿
Udef_fruits	西红柿	西红柿
Udef_goods	胡椒	胡椒
Udef_goods	楜椒	胡椒

MARKUS 和 CT Tool 各有优缺点。作为首个这种类型的标记工具,MARKUS 的功能与设计思维的细腻让人不得不敬佩,而且因为 MARKUS 没有和任何系统捆绑在一起,标记过的文本可以用不同的格式输出,所以很有弹性。但是弹性太大也是一个弱点,文本被标记好以后,我们希望知道下一步要做什么,因为标记文本这件事当然不应该是最终的目标。这个问题也是 DocuSky 在进行规划设计时想要回答的核心问题之一。DocuSky 最重要的功能,就是将标记完的文本一键建成一个可以全文检索,有检索成果后可进行分类、词汇分析(如词频、共现等)和拥有各种可视化呈现功能的数据库。[37]

数据库中通常有许多文件,让标记工具能够一次同时标记所有文件当然是比较有效率的做法,所以 CT Tool 是 MARKUS 的自然延伸。但是因为 CT Tool 初始的设计就是为了在 DocuSky 的环境下做批次标记,所以我们限定 CT Tool 输出的格式是 DocuXML(DocuSky 建数据库所用的格式)。但 CT Tool 的原理和做法均具有普遍性,因此不难想象一个结合 MARKUS 和 CT Tool 优点的工具。

4　历史文本词汇标记后的运用

在电子文本上标记了人物、地点、时间和对象之后,接下来就是思考如何利用数字人文的研究方法来运用这些标记。以 DocuSky 为例,标记过词汇的电子文本在上载进入个人的 DocuSky 账号后便会成为一个具有标记词汇功能的个人化数据库。被标记的词汇可以在 DocuSky 中以后分类的方式进行统计(如下页图 6),而具备序号(RefId)的标记词汇,则可以在此统计中完成聚合与消歧。

① 关于“批次标记工具”更详尽的使用说明,参见:http://docusky.org.tw/DocuSky/docuTools/ContentTaggingTool/index.html。

此外,在带入序号信息后,DocuSky 可以根据序号,通过权威数据库所提供的 API 向该数据库请求序号人物或地名的权威信息。人物的相关信息及地名的空间坐标,都可以透过 API 从权威数据库的提供者处取得并直接在画面上显示。借由这种阅读方式,使用者可以更容易地掌握文本当中人物的相关讯息,或是将地点在地图上进行空间的呈现,达到所谓“左图右史”相互参照的阅读体验。

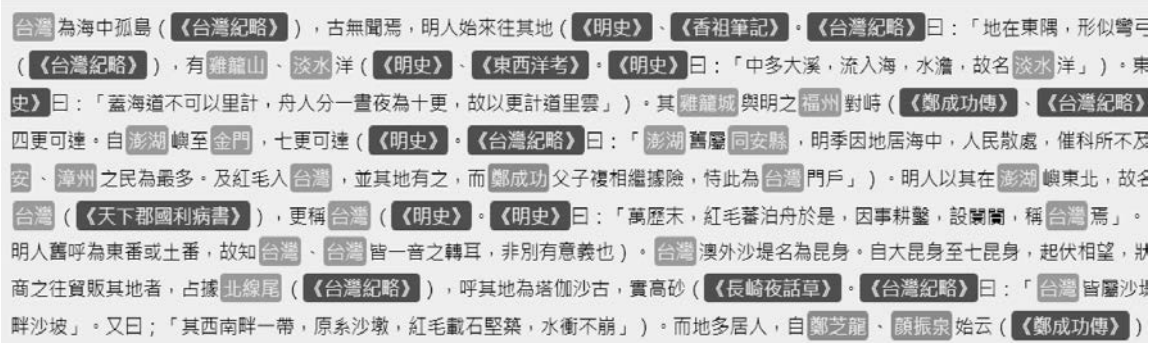

图 6　标记过词汇的文本在 DocuSky 中的显示

除了词频的统计、词汇的消歧聚合,以及词汇背后信息的延伸链接外,词汇标记最重要的功能,是将有关联的词汇借由标记链接起来,探讨文本之间的脉络关系,发掘以往线性阅读中无法探究的问题。诸如时间标记的对应、人物标记的关联,乃至于人、时、地、物标记的综合运用,都是在数字人文研究上重要的发展。

4.1　时间标记的运用

时间无疑是最直观的脉络之一。在台湾大学数位人文中心所建构的所有脉络分析系统中,时间都是最基本的检索后分类。透过可视化,观察检索成果的时间分布,可以对检索到的文件子集进行很清晰的时间鸟瞰。[38]

时间标记的另外一个运用是辅助阅读。许多文本(如编年体史书或日记)是根据时间来排列的,对时间信息进行标记之后,就可以时间作为锚点,在数字工具的协助下将不同文本中同一时间所发生的事情同时呈现出来。我们以赖思频所建置的“春秋三传对读系统”为例说明。[39]

《春秋》记录了春秋时期鲁国鲁隐公元年(公元前 722 年)至鲁哀公十四年(公元前 481 年)的历史。由于《春秋》的文字相当精简,因此后世有《公羊传》《谷梁传》与《左传》为之进行补充式的批注(以下简称“春秋三传”)。《春秋》是一部编年体的史书,受其影响,“春秋三传”也是以鲁国纪年作为段落区分,如此可以依据纪年的标记锚点来进行四部书之间的对应。

在“春秋三传对读系统”中,点选任一文本中的任一段文字,系统都会通过时间的锚点将其他三部书中关于同一时间的文字拉到一个画面当中。

如下页图 7 所示,点选《谷梁传》中的“鲁隐公元年五月”,此时,不但《谷梁传》中关于鲁隐公元年五月的段落成为红色,其他三部书的“鲁隐公元年五月”也会同步被带入同一画面当中。在文本中,我们可以看到《春秋》里只简短地记述了“夏,五月,郑伯克段于鄢”寥寥数字;在《公羊传》与《谷梁传》中,两位作者都评论了郑伯寤生杀其弟段这件事情正当与否;《左传》则花了更多的篇幅记叙了段叔背叛郑伯的全部经过,郑伯对其母所言“不及黄泉,无相见也”的誓言,以及颍考叔如何化解这对母子间的龃龉等历史。进一步往事件之前的《左传》探询,还可以看到作者如何描述郑伯、段叔与母亲姜氏之间的关系。

图 7　台湾大学“春秋三传对读系统”应用示例(一)

当点选《左传》的“鲁哀公元年三月”时(见图 8),会看到“三月,越及吴平,吴入越不书,吴不告庆,越不告败也”这段文字。这场吴国大败越国的战争,在《春秋》与《公羊传》《谷梁传》中皆未记载。而在此段落之前,《左传》所述正是我们所熟知的越王句践与吴王勾差之间的纠葛,然而,在《春秋》与其他二传之中,这段发生在东南沿海一隅的故事场被忽略了。

《公羊传》与《谷梁传》未曾记载,是因为《春秋》本无此事,但《左传》的作者却留下了这段记录。是《春秋》在撰写之时缺乏这段史料?又或者是刻意没有书写?比较作传者的叙述角度其实可以帮助我们更好了解《春秋》“微言大义”背后的历史事件与历史评价。《公羊传》与《谷梁传》的作者多半探讨的是《春秋》为何如此记载,《左传》的作者则偏好将事情的来龙去脉交代清楚,其他留给读者自行判断。古伟瀛就认为,诠释者的思维往往会烙印在被诠释的对象上,这样读者不仅可以对被诠释者更了解,同时也可以对诠释者本身的处境及内心思想有进一步的理解。[40]“春秋三传对读系统”确实也提供了观察上的便利。

图 8　台湾大学“春秋三传对读系统”应用示例(二)

至于前面提到中公历之间的对应,则反映出多种纪年文本之间的对应与对读需求。司马迁在撰写《史记》的《十二诸侯年表》与《六国年表》时已经看到这个问题。为了将不同诸侯国的历史并列呈现,司马迁以周王纪年为基准,让各国的历史可以与之对应。司马光在编写《资治通鉴》的六朝时期时也遇到同样的问题。他发现当时各国林立,“皆与古之列国无异”,若把任一国当作正统,岂不是把其余的国家当成了“僭伪”?为了避

免有所谓“正伪之别”，司马光“不得不取魏、宋、齐、梁、陈、后梁、后唐、后晋、后汉、后周年号，以纪诸国之事，非尊此而卑彼，有正闰之辨也”[41]。《资治通鉴》保留了诸国的纪年以一视同仁，却也把各国的历史变成独立而断裂的“国别史”。司马迁之所以能够完成年表，是因为春秋战国时期有周王室纪年可以作为权威参照，六朝时期却没有，当时也还没有公元纪年的概念。方佩雯的硕士论文《应用中历时间标准化于六朝正史对读》就是将六朝时期各国的纪年予以标准化，再利用前述“春秋三传对读系统”所使用的对读工具将纷乱的六朝历史并列呈现[42]，类似应用如图 9。

方佩雯的研究将法鼓文理学院的“时间规范资料库”作为参照，以系统性的方式提取出六朝各国的时间词汇，再透过“时间规范资料库”进行系统性的批次转换，使各国独立的纪年时间信息对应公历时间，如此，便能依照所取得的标准化时间信息进行史实的排序。除此之外，方佩雯将原有对读工具中以单一时间标记进行对应的方式改变为时间的区间对应。因为每段文本都会有零至多个时间词汇，单独对应反而会更加混乱，因此，方佩雯采取的是以时间区段作为对应，即将文本段落中的起迄时间作为标记，让起迄时间相互涵盖的段落可以透过对读工具连结起来。

方佩雯以陈高祖（霸先）与梁国大将王僧辩的冲突为例，分别在《北齐书·帝纪》《陈书·帝纪》与《梁书·列传》中进行探索。背景是南梁贞阳侯萧渊明在攻打东魏时兵败被俘，公元 550 年，东魏权臣高欢之子高洋篡位自立，建立北齐，是谓齐文宣帝；公元 554 年，梁元帝被西魏将领所杀，南梁陷入立帝之争。此时上场角逐帝位的，有陈霸先属意的萧方智与北齐文宣帝特意于公元 555 年送回南梁的萧渊明。通过方佩雯建立的对读工具，在《北齐书·帝纪》中可以看到，梁元帝死后“梁将王僧辩在建康，共推晋安王萧方智为太宰”[43]；而《陈书·帝纪》记叙同一时期的文字中提到立萧方智为太宰是陈霸先与王僧辩共同的决定[44]，之后，王僧辩立贞阳侯萧渊明为帝，陈霸先对此感到不悦，故“高祖居常愤叹”[44]。这两段记载反映出两件事情，其一，陈霸先不愿立萧渊明为帝；其二，北齐积极于立萧渊明为帝。但我们却无法得知王僧辩立萧渊明是否与北齐有直接的关系。不过，若审视《北齐书·帝纪》中关于王僧辩立萧方智为太宰的前后文，并透过时间区段的对读，便可在《梁书·列传》中找到北齐文宣帝高洋曾致书王僧辩，希望王僧辩能够支持立萧渊明为帝的记叙，而王僧辩却也因此招致杀身之祸[44]。

图 9　基于中历时间标准化的对读系统应用示例（南朝史书“帝纪”对读）

可见,散见于《北齐书》《陈书》与《梁书》的记载,因为时间的标记而能够以数字工具进行串连,使看似独立的历史事件全都被兜组在一起,原本只能线性阅读的史料有了更多元的探讨方向。

4.2 人物标记的运用

经过标记后的人名让研究者可以对不同人物出现的频率在数字工具中进行快速统计,也能够快速掌握所关注的人物在不同文本、章回或是档案中的出现。许多多(Xu Duoduo)等人所撰的《新加坡华人庙宇碑铭所载芳名录的数位人文探索》(*Chinese Singaporean Temples: Digital Humanities Approaches to Frequency Lists of Sponsors*)即是对《新加坡华文铭刻汇编 1819—1911》进行词汇标记后开展数字人文研究的作品。该文对新加坡华人庙宇中碑刻上的"芳名录"内人名进行了标记与统计,再利用 DocuSky 进行统计,便可看出哪些信徒在多间庙宇进行过捐款,也可借此理解哪些人物在新加坡华人信仰圈中可能有一定的声望与重要性。[46]

人物词汇标记过后,除了可进行单一人物的统计外,还可以进行人物与人物之间的关联运算。在 THDL 的明清档案文献集中检索"林爽文",可得 1666 笔资料,而排在文件出现次数(document frequency,df)第二位的是"福康安",出现于已检索出来的 1666 笔资料中的 636 笔。乍看之下,福康安与林爽文的关系应该最密切,其实不然。THDL 的词频分析工具中,除了文件出现次数外,还有一个"相关系数(t→q)"指数。如图 10 所示,"福康安"的相关系数是 0.370,代表 THDL 里所有出现福康安的文件中,只有 37% 也同时提到林爽文。的确,虽然福康安去台湾的主要任务是镇压林爽文,但在平定了林爽文之后,福康安还向乾隆提了一连串的建议,如屯番等[47],这些事迹对往后清代对台湾的治理产生很大的影响,但这些文件均与林爽文无关。相对之下,一些其他检索"林爽文"的共现(co-occurrence)人物如王芬(92.3%)、何有志(94.1%)、林劝(92.7%,林爽文父)等,相关系数都超过 90%,代表这些人与林爽文的关系较福康安高出许多。这些相关系数高的人,不是与林爽文有亲戚关系,就是与林爽文一同举兵反抗政府。所以在实际的关联性上,就可以透过共现频率和相关系数迅速掌握词汇与词汇之间的关系,而前提则是必须进行人名词汇的标记。

依 df 降幂排列

人名 Term t		df	t→q	回报
林爽文	生平	1666	1.000	
福康安	生平	636	0.370	
常青	生平	614	0.418	
李侍尧	-	543	0.311	■■■
柴大紀	生平	462	0.405	
黄仕簡	生平	253	0.267	
普吉保	生平	243	0.464	
任承恩	生平	237	0.572	
海蘭察	生平	212	0.596	
恒瑞	-	205	0.470	■■■
普爾普	-	92	0.652	■■■
和珅	-	90	0.279	■■■
王芬	-	84	0.923	■■■
魏大斌	-	84	0.344	■■■
梁朝桂	生平	82	0.412	
阿桂	-	81	0.280	■■■
邱能成	-	79	0.632	■■■
陳泮	-	76	0.826	■■■
楊起麟	生平	72	0.514	
黄奠邦	-	70	0.515	■■■

图 10 "台湾历史数位图书馆"(THDL)词频工具应用示例

标记后的人物词汇还可以进行可视化的呈现,比如可以利用斯坦福大学所开发的 Palladio 将多份文件中重复出现的人物联结起来,将共现关系制作成词汇关联图。这种在不同历史文本间找寻相同人物并进行可视化呈现的方法,有助于在庞大的历史文本数据中迅速锁定研究对象。例如,《淡新档案》的 11509 案中有一件关于竹北二堡红毛港庄教会的教民调查清册。若利用教民的姓名词汇在整份《淡新档案》文献集中查找,可以发现其中有几位教民也出现在其他红毛港庄一带相关案件中。因此可以利用 Palladio 将这样的关联性制作成关联图呈现(图 11)。

如图 11 所示,教民吕泉同样出现在丈量舞弊案中,而许标、许俊则是在货船抢案中出现。这些案件中,涉

案教民都是以“被告”身份出现。红毛港教会设立的时间约在 1877 年,11509 案的成案年为光绪十七年(1891 年),吕泉所涉案件发生在光绪十三年(1887 年),货船抢案约在光绪十一年(1885 年)前后。无论涉案教民是在案前或案后信教,多少也反映出地方恶徒成为教徒或者倚仗教徒身份进行犯罪的现象,因此有学者认为清末一般民众批判基督教信徒“靠番仔势”的形象可能不是空穴来风[48]。

由上述的例证可知,对于完成人名词汇标记的文本,我们可以进行多样化的分析与呈现。透过数字人文工具的协助,可以快速掌握人物在历史文本中的出现情形与关联性,从而进行更多研究上的探索。

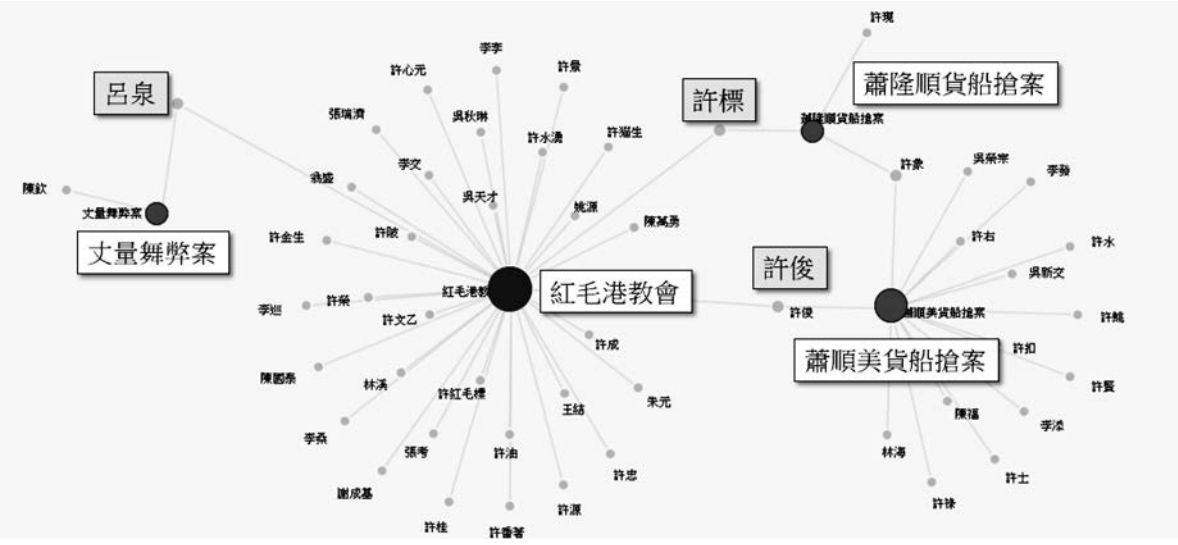

图 11　跨文本人名词汇关联图呈现示例(以 Palladio 制图)

4.3　人时地物标记的综合运用

地名词汇的标记,可以通过权威信息的带入在地图上显示历史文本中事件发生的地点。透过空间的观察,也能够探询作者在撰写文本时的空间知识,若加上时间因素进行排序,还能认识到历史事件在空间中的迁移。举例而言,《水浒传》中许多角色出场时,施耐庵都会提及他们的出生地,例如“九纹龙”史进是华阴县人,“及时雨”宋江是郓城县人,“旱地忽律”朱贵是沂水县人。若将这些籍贯标示在地图之上,可以看到施耐庵笔下人物的籍贯北不过河北,西不过陕西,大部分集中于长江以北,但也有祖籍琼州(海南岛)的孙新。由这个分布可以看出作者当时的“天下知识”到达哪些地方。

对象标记与时间标记的结合有助于理解某些特定对象在历史中出现或消失的变动。以收录琉球国与周边国家往来文书的《历代宝案》为例,“胡椒”曾是明代琉球国重要的朝贡物与交易商品,但是经过对象与时间的交叉比对,可以发现 1579 年后“胡椒”在文件中消失了好多年,直到清初才恢复了一些,之后又不再出现琉球国。依此对象在时间轴当中的呈现,可以推论在明末清初,胡椒在朝贡贸易上的重要性已经降低。这可能是因为中国已经出产足够的胡椒,又或者是有别的渠道供应了中国所需,还也许和西方人垄断香料贸易有关,以致琉球国不再需要向中国进贡或贩卖胡椒。

5　讨论

历史是研究人在时间中的活动轨迹的一门学问,但地理空间和各种对象不可避免地穿插其间,使得历史研究成为结合人物、时间、地理空间与对象的工作。传统的研究方法是通过在历史文本中爬网文字,记录人物、时间、对象与地理位置的关系,从中探寻在时间之流中曾经发生的史实。

然而在数字时代,研究者面对大量的文本,有效、快速地标记人、时、地、物就成了当务之急。本文先针对人、时、地、物标记的目的与需求分别做说明,指出对时间标记而言,最主要的工作是时间的正规化;对人名和地名,除了认出词汇外,标记还需要处理消歧和聚合两大问题,而地名更有标记坐标经纬度的需求;对象标记也有消歧和聚合的问题,但对象种类繁多,所以词汇辨识是比较大的挑战。

接下来本文介绍了两个工具——MARKUS 和 CT Tool,以及如何结合这两者与现有的参考数据库,如 CBDB、TGAZ,法鼓文理学院的人名、地名、时间规范及 TWGIS 等,达到快速标记的目的;并且介绍了如何通过标记序号,达到词汇的聚合、消歧,以及权威信息的延伸阅读。最后通过一些例子,展示标记能够提供的脉络观察与鸟瞰。

本文没有详细介绍的是事件标记的做法。谈到事件的标记,不能不提南宋袁枢写的《通鉴纪事本末》。袁枢将《资治通鉴》的部分内容标记成 239 个事件,而且无意中创造了一种新的历史书写方式——纪事本末体,对中国传统的历史书写产生很大的影响。然而《通鉴纪事本末》并没有涵盖所有《资治通鉴》的内容(后者有 294 卷,前者只有 42 卷),而且所解析出的事件,明显反映作者自己的偏好。钱穆批评《通鉴纪事本末》对战国只选取"三家分晋"和"秦并六国"两个事件,"把整个战国史都忽略了",又言"历史不能只管突发事件,只载'动'与'乱',不载'安'与'定',使我们只知道有变,不知道有常";他甚至严厉地说,"袁枢实当不得是一个史学家,他这书的内容也不能算是一部史学名著"。[49]248,250 姑且不论钱穆的批评是否过苛,他至少指出了标记事件的困难与主观的本质。

事件标记的第一个挑战是标记的全面性。假设一位研究者要对一位文本集当中的人物做标记,但如果只标记其中三分之一的文本,不标另外三分之二,那么这些标记从统计上来说是没有任何代表性或用处的。如果囿于人力或时间,无法标记所有的人物,那么可以选择一些特定的人物(譬如研究者特别感兴趣的人或所有巡抚以上官员等)去标记想要研究的文本集,而不是只标记一部分文本中的人物。钱穆对袁枢的批评多少和后者没有对《资治通鉴》做全面性标记的做法有关。无论《资治通鉴》本身取材有无偏见,这本书反映了司马光对治理国家的看法,然而袁枢仅仅取了其中三分之一做纪事本末,而且只取"乱",不取"治",完全达不到"鸟瞰"《资治通鉴》的目的,这样如何可以称为《资治通鉴》的纪事本末?

但是如何全面性地标记事件呢?事件是一个语义型(semantic)概念,而其他四个词汇类型(人、时、地、物)则是语法型(syntactic)概念,语法型的词汇通过字符串匹配(string matching)即可萃取,语义型词汇则困难得多。以前文提到的土地移转图为例,要判断两张契约文书是否为上下手契,必须先萃取出契约种类、买方、卖方、土地位置、四至(土地的东南西北界)、买卖时间等;如果是阄分契或隔代才交易的契书,就更复杂,在用自动方法从契书中取出这些信息后,还需再设计一个算法去判断是否为上下手契。[50] 即使如此,这个算法仍然会遗漏一些上下手契(我们的方法求全率是 71%)。但至少这一上下手契的计算和标记是针对 THDL 当时的整个古契书文件集做的,不是只拿一个子集合做实验,所以勉强具有全面性,也可以从中发掘出一些前人看不到的脉络与研究议题。无论如何,这个例子体现了标记事件的第二个挑战,也就是事件内在的语义性。

第三个挑战是所谓事件一定是关联着文件特性的。上下手契的关系不会在明清行政档案中出现,奏折/上谕引用关系也不会对地方契书有意义,所以针对不同的文件性质,需要设计不同的方法。而又因事件分析往往牵涉对文件集整体的分析,计算起来需要很多的资源,所以通常只能用前处理的方式找出事件关系,而无法做实时(real time)的标记。这使得事件标记的做法在思维和策略上均与其他的标记类别大为不同。

如果把现在的历史研究环境和 30 年前比较,大量电子文本的出现显然对研究方法造成很大的冲击,若要同时驾驭大量的历史文本,鸟瞰或从中发掘脉络,词汇标记应该是必不可少的步骤。MARKUS 与 CT Tool 这类工具的出现,固然简化了在大量文本中进行语法型词汇标记的工作,让研究者能够快速掌握人、时、地、物词汇之间的关联,但这些工具仍有改进的空间,尤其是面对语义型"事件"的识别与标记,还有许多理论和技术上的问题需要克服。本文希望能够抛砖引玉,让学界关注历史文本标记的议题,进而发展更完善的理论基础和更便利的数字工具,让历史文本的词汇标记可以更有效率,使数字人文的研究方法能够更加丰富。

参考文献

[1]Moretti F. Distant Reading[M]. London: Verso, 2013:43, 56. 赵薇. “社会网络分析”在现代汉语历史小说研究中的应用初探——以李劼人的《大波》三部曲为例[M]// 项洁. 数位人文:在过去、现在和未来之间. 台北:台湾大学出版中心, 2016:399.

[2]向帆,何依朗. “远读”的原意:基于《远读》的引文和原文的观察[J]. 图书馆论坛,2018,38(11):44-48+43.

[3]姜文涛,戴安德. “数字人文:观其大较”[J]. 山东社会科学,2017(9):31-32.

[4]杨玲. 远读、文学实验室与数字人文:佛朗哥·莫莱蒂的文学研究路径[J]. 中外文论, 2017(1):295-309.

[5]王泰升. 数字化历史数据库与历史研究——以明清档案、淡新档案、日治法院档案等数据库为例[M]// 项洁. 从保存到创造:开启数位人文研究. 台北:台湾大学出版中心, 2011:33.

[6]Carr E. 何谓历史?[M]. 江政宽,译. 台北:博雅书屋, 2009:126

[7]Bloch M. 史家的技艺[M]. 周婉窈,译. 台北:远流出版事业股份有限公司, 2020:33-34.

[8]葛剑雄,华林甫. 五十年来中国历史地理学的发展(1950-2000年)[J]. 汉学研究通讯, 2002,21(4):16-27.

[9]“中央研究院”. 中华文明之时空基础架构系统[EB/OL]. [2020-10-10]. http://ccts.sinica.edu.tw/.

[10]“中央研究院”地理资讯科学专题研究中心. 台湾百年历史地图[EB/OL]. [2020-10-10]. http://gissrv4.sinica.edu.tw/gis/twhgis/.

[11]不着撰人. 预备赔偿外人损失办法[EB/OL]. [2020-10-10]. http://tk.dhcdb.com.tw/tknewsc/tknewskm.

[12]Project Gutenberg. Project Gutenberg[EB/OL]. [2020-10-10]. https://www.gutenberg.org/.

[13]“中央研究院”历史语言研究所. 汉籍电子文献资料库[EB/OL]. [2020-10-10]. http://hanchi.ihp.sinica.edu.tw/.

[14]GitHub. 结巴断词工具[EB/OL]. [2020-10-10]. https://github.com/fxsjy/jieba.

[15]“中央研究院”. 中研院中文断词系统[EB/OL]. [2020-10-10]. http://ckipsvr.iis.sinica.edu.tw/.

[16]TatsukiSekino. HuTimeProject[EB/OL]. [2020-10-10]. http://www.hutime.org/.

[17]法鼓文理学院. 时间规范资料库[EB/OL]. [2020-10-10]. https://authority.dila.edu.tw/time/.

[18]“中央研究院”数位文化中心. 两千年中西历转换[EB/OL]. [2020-10-10]. https://sinocal.sinica.edu.tw/.

[19]台湾大学数位典藏与自动推论实验室,数位人文研究中心. 中西历对照查询系统[EB/OL]. [2020-10-10]. http://thdl.ntu.edu.tw/datemap/index.php.

[20]川口长孺. 台湾郑氏纪事[EB/OL]. [2020-10-10]. https://zh.wikisource.org/w/index.php·title=%E8%87%BA%E7%81%A3%E9%84%AD%E6%B0%8F%E7%B4%80%E4%BA%8B&oldid=1316886.

[21]法鼓文理学院. 人名规范资料库[EB/OL]. [2020-10-10]. https://authority.dila.edu.tw/person/.

[22]Fairbank Center for Chinese Studies of Harvard University&the Center for Historical Geographical Studies at Fudan University. China Historical Geographic Information System, CHGIS[EB/OL]. [2020-10-10]. http://sites.fas.harvard.edu/~chgis/.

[23]台湾大学数位人文研究中心. 台湾历史地名坐标资讯库[EB/OL]. [2020-10-10]. https://docusky.org.tw/DocuSky/docuTools/Geocode/map.html.

[24]法鼓文理学院. 地名规范资料库[EB/OL]. [2020-10-10]. http://authority.dila.edu.tw/place/.

[25]台湾大学数位人文研究中心. 历代宝案脉络分析系统[EB/OL]. [2020-10-10]. http://lidaibaoan.digital.ntu.edu.tw/.

[26]叶智豪, 王昱钧, 蔡宗翰. 历史文献的命名实体撷取——结合主动学习法之半监督式模型[M]// 项洁. 从保存到创造:开启数位人文研究. 台北:台湾大学出版中心, 2011:131-144.

[27]彭维谦等. 自动撷取中文典籍中人名之尝试:以PMI(Pointwise Mutual Information)断词于《资治通鉴》的应用为例[C]. 台北:第四届数位典藏与数位人文国际研讨会, 2012.

[28]彭维谦. 不同脉络中的历史文本之自动分析以《资治通鉴》《册府元龟》及《正史》为例[D]. 台北:台湾大学, 2013.

[29]谢育平. 同位词夹子:主题式分类词库萃取算法[M]// 项洁. 数位人文研究的新视野:基础与想象. 台北:台湾大学出版中心, 2011:133-162.

[30]杜协昌. 一个数位人文内容研究的文本撷词工具[C]. 台北:第十一届数位典藏与数位人文国际研讨会, 2020.

[31]Chen S, HuangY, Hsiang J, et al. Discovering land transaction relations from land deeds of Taiwan[J]. Literary and Linguistic Computing, 2013, 28(2):257-270.

[32]陈诗沛. 信息技术与历史文献分析[D]. 台北:台湾大学, 2011.

[33]涂丰恩. 善化地区的环境变迁、土地开发与地权纠纷(1890-1920)[C]//第六届台湾总督府档案研讨会论文集. 南投:“国史馆”台湾文献馆, 2010:504-505.

[34]Hsiang J, Chen S, Hou I, et al. Discovering relationships from imperial court documents of Qing dynasty[J]. International Journal of Humanities and Arts Computing, 2012, 6(1-2):22-41.

[35]Tsai T, Lu Y, Wang Y. Event Extraction on Classical Chinese Historical Texts: A Case Study of Extracting Tributary Events from the Ming Shilu[C]. Utrecht: Digital Humanities 2019, 2019.

[36]Hou Ieong Ho Brent, Hilde De Weerdt. MARKUS Text Analysis and Reading Platform[EB/OL]. [2020-10-10]. http://dh.chinese-empires.eu/beta/.

[37]Tu H, Hsiang J, Hung I, et al. DocuSky, A Personal Digital Humanities Platform for Scholars[J]. Journal of Chinese History, 2020, 4(2):564-580.

[38]涂丰恩, 杜协昌, 陈诗沛,等. 当信息科技碰到史料:台湾历史数位图书馆中的未解问题[M]// 项洁. 数位人文研究的新视野:基础与想象. 台北:台湾大学出版中心, 2011:21-44.

[39]赖思频, 项洁. 春秋三传对读系统[EB/OL]. [2020-10-10]. http://doi.org/10.6681/NTURCDH.DB_AR3C/Service.

[40]古伟瀛. 顾炎武对《春秋》及《左传》的诠释[J]. 台大历史学报, 2001(28):69-91.

[41]司马光. 资治通鉴[M]. 北京:中华书局, 2011.

[42]方佩雯. 应用中历时间标准化于六朝正史对读[D]. 台北:台湾大学, 2020.

[43]李百药. 北齐书[M]. 北京:中华书局, 1972.

[44]姚思廉. 陈书[M]. 北京:中华书局, 1972.

[45]姚思廉. 梁书[M]. 北京:中华书局, 1973.

[46]许多多,丁荷生,马德伟. 新加坡华人庙宇碑铭所载芳名录的数位人文探索[J]. 数位典藏与数位人文, 2020(4):37-71.

[47]台湾银行经济研究室. 台案汇录壬集[M]. 台北:台湾银行, 1966:1-8。

[48]胡其瑞. 数位人文视角下淡新档案中的教堂、教士与教民[C]. 彰化:2020年基督教宣教士文史国际学术研讨会, 2020.

[49]钱穆. 中国史学名著[M]. 台北:三民书局, 2019:248, 250.

[50]黄于鸣. 台湾古地契关系自动重建之研究[D]. 台北:台湾大学,2009.

Vocabulary Marking and Application of Historical Text

Hsiang Jieh　　Hu Chijui

Abstract Historical text is the basic material of historical research. By crawling the content of the text, historians organize, piece together and contextualize meaningful information in the text. History is a discipline that studies the trajectory of human activities in time. After adding the concept of geographic space, historical texts will become more three-dimensional. Instead of linear reading in paper data in the past, historical texts in the information age can add a lot of vocabulary tags with the assistance of technology, and then use the analysis and visualization of tagged vocabulary to take a bird's eye view and grasp the implicit context in historical texts. By discussing the meaning of person, time, place name and object vocabulary marks in historical texts for historical research, describing the purpose and characteristics of various marks, especially pointing out that vocabulary marks not only identify words, but also need to achieve "disambiguation" with the "aggregation" function. At the same time introduce two automatic tagging tools, "Code Library Semi-automatic Marking Platform for Ancient Books" (MARKUS) and "Batch Tagging Tool" (Content Tagging Tool, CT). These two tools make it possible to quickly mark a large number of people, times, places, and things. Illustrate how to use marked texts through actual research results; use time, person, geography, and object vocabulary to mark actual benefits to illustrate the use of vocabulary marking and application in historical texts and historical research. Finally, we discuss the issue of event markers, and point out that event markers are essentially different from other lexical markers.

Key words Text Annotation; Digital Humanities; Historical Text; DocuSky; MARKUS