

数字人文中的文本挖掘研究*

□郭金龙 许 鑫

摘要 概述数字人文的概念、研究内容和现状,指出文本挖掘方法是数字人文研究的一个研究热点与趋势。在综述文本挖掘在数字人文各个研究领域中的具体应用基础上,重点介绍欧美发达国家文本挖掘应用于数字人文研究的前沿实践,以期为我国人文学科研究方法与范式的转型提供借鉴。

关键词 文本挖掘 数字人文 人文研究

1 引言

近年来国内数字化的浪潮越来越猛,各种古籍数字化、文献数据库建设的项目层出不穷。计算机技术与网络技术已经逐步深入到人文研究的各个领域,有效提升了人文研究各个环节的效率。尽管数字技术与人文学科的结合让我们看到了数字技术在人文科学研究中的巨大价值,然而就目前的情况来看,人文研究中利用数字技术的手段还比较单一,大都停留在人文资源的简单组织、检索与利用,缺乏对现有资源的深度分析和挖掘。国外在数字技术与人文研究交叉结合的研究领域形成了一个新的学科:数字人文,所取得的成果值得重视。数字人文学科从一个新的理论高度看待数字技术与人文研究的关系及其带来的影响。它为人文研究者提供了新的工具和平台,对海量的数字资源进行组织、标引、检索、挖掘和分析利用。在这样的背景下,本文在简单介绍了数字人文学科的概念和现状后,论述了数字人文的一个重点研究趋势是对海量数据的分析,而新兴的文本挖掘方法由于其能有效处理海量文献,处理非结构化数据以及较高的研究品质而成为网络时代数字人文研究中很有前景的研究方法。文章接着重点分析了文本挖掘方法在数字人文研究中的典型应用,介绍了欧美发达国家把文本挖掘方法应用于人文学科的一些前沿的研究项目,以期加快我国人文学科研究范式的升级和转型,促进我国人文学科的飞跃发展。

2 数字人文:概念、现状与趋势

2.1 数字人文概述

数字人文(Digital Humanities),又称作人文计算(Humanities Computing),是一个将现代计算机和网络技术深入应用于传统的人文研究与教学的新型跨学科研究领域^[1]。其产生的背景主要是由于海量的图书、期刊、报纸、照片、绘本、音乐、视频等人文资料的数字化,尤其是数字图书馆的迅猛发展。面对海量的数字文本,研究者急需新的工具和平台对之进行组织、标引、检索、挖掘和分析利用。数字人文是一个典型的文理交叉领域,研究项目和研究团队常常既包括传统人文领域(哲学、历史学、文学、语言学、艺术学、人类学等)的研究者,还包括精通计算机技术和多媒体技术的专家学者。在这两类人员的协作下,诸如数字仓储、文本挖掘、多媒体出版、数字图书馆、信息可视化、虚拟现实、地理信息系统等多种信息技术开始在人文领域得到深入应用。数字人文研究的核心在于通过将研究对象数字化以支撑、保障和创新人文科学研究的内容、方法和模式,研究对象包括各种可计算的基础数据对象,如自由文本、格式化数据、图像、声音等。针对这些数据进行的计算包括文本分析与检索、地图可视化、音视频检索等。数字人文的方法随着信息技术的发展而不断变化,因此其研究边界相对不固定。

2.2 数字人文研究现状

2.2.1 国外

国外数字人文的研究早已形成非常清晰的学术

* 本文系2011年度国家社科基金青年项目“联合虚拟参考咨询系统的知识库研究”(11CTQ003)的研究成果。

共同体。各种数字人文研究学会和专门的研究机构在国外已经非常普遍。多项数字人文研究项目和成果获得了国家层面的资助与学界的广泛关注,并取得了较大的反响。现在该研究领域形成了一个国际数字人文组织联盟(The Alliance of Digital Humanities Organizations),由人文计算领域影响力最大的三个组织组成:文学与语言学计算协会(Association for Literary and Linguistic Computing)、人文领域计算机应用联合会(The Association for Computers in the Humanities)和数字人文学会(The Society for Digital Humanities / Société pour l'étude des médias interactifs)。该联盟每年召开一次数字人文年度大会,由北美和欧洲的高校和研究机构轮流举办。

从数字人文的实践来看,全球范围内的数字人文研究中心已经不下 50 家,主要集中在欧美日等发达国家,其中较为知名的研究机构有:伦敦国王学院的人文计算研究中心、美国斯坦福大学的人文实验室、马里兰大学的人文技术研究机构、伊利诺伊大学的科学与学术情报研究中心以及日本立命馆大学的京都数字文艺研究中心等。很多项目和研究机构都得到了国家级和专业机构的基金资助。

2.2.2 国内

“数字人文”这个概念在国内尚是一个新鲜的技术概念。据中国教育部社科司统计,大陆教育系统内有 729 家人文研究机构。但是到目前,专门的数字人文研究中心只有武汉大学 2011 年刚成立的一家。中国国家自然科学基金和国家社会科学基金及教育部人文社科基金尚未资助过以“数字人文”为主题的研究项目,目前也没有以“数字人文”为主题的连续性研讨会。

尽管没有使用“数字人文”的名称,许多高校和研究机构早已开展了“数字人文”研究,如中国社会科学院的地理信息服务平台系统^[2];南京师范大学虚拟地理环境实验室的华夏家谱 GIS 平台^[3];中国艺术研究院的西北人文资源环境基础数据库^[4];北京大学的中国基本古籍库^[5];首都师范大学国学传播中心的古籍电子定本工程^[6]等。这些数字人文项目为人文学科提供了基础数据保障和简单分析工具,大大方便了人文学科的研究过程。可见在数字图书馆建设浪潮的推动下,大陆的数字人文项目尤其是人文资料专题数据库建设取得了非常突出的成

绩,且大部分人文学科已经采用计算机技术和网络技术进行人文资料的数字化整理、分析和展示。

但是就现状来看,对相关信息技术的使用还处于比较初级的水平。大部分的项目还停留在传统文本的数字化、数字作品的保存和简单分析与展示层面,尚未有项目针对已有的海量人文资源,挖掘其中更深入的知识和模式。

2.3 数字人文的研究趋势

数字人文是一个快速发展中的文理交叉学科,已经在多个研究分支取得了丰富的成果。如在历史学方面的基于 GIS 的历史地理可视化;在文学方面的文本挖掘与 TEI 标准;在语言学方面的基于大型语料库的语料库语言学;在舞蹈方面的视频捕捉、运动分析与虚拟现实再现;考古学方面的图像分析、色彩还原和数字重建等^[1]。随着数字图书馆的发展和互联网数据的激增,每一个学科公开文献的品种与数量都越来越多。人们不但无法阅读一个学科或主题的全部文献,甚至无法浏览或知晓一个学科或主题的全部文献。面对海量文献,常规的阅读和分析已经无法应对内容稀薄而数量庞大的文献了。在这样的形势下,处理海量数字文献成为数字人文研究众多分支的共同目标。

在文学研究领域,面对越来越多的电子出版物,人们已经开始尝试采用字词统计、信息抽取等量化的研究方法,该方法在作者归属和文学风格与文学流派分析上已经显示出强大的威力,大大拓展了传统的基于单个作品分析的局限。在历史学研究领域,通过对大量历史文本进行命名实体和空间历史数据的抽取,可以对历史知识和历史事件进行静态和动态的可视化展示,以揭示社会经济系统在时空演变中的客观规律及相应的政策与规划手段。在语言学研究领域,基于大规模语料库的研究已经成为主流。“语料库语言学”的迅猛发展使得自然语言文本的采集、存储、检索、统计、自动切分、词性标注、语义标注等领域取得了一系列突出的成果。

由以上分析可知,数字人文研究的一个重要趋势是对海量文献的处理。无论是文学、历史学研究中新的挖掘工具与平台的需求,还是语言学研究中对海量真实语料分析技术的关注,核心都是对海量文献处理的技术,也就是文本挖掘技术。文本挖掘一方面可以充分挖掘现有数字化项目中累积的海量人文资源,发现其背后隐藏的模式;另一方面可以

利用海量的真实语料进行语言学的实验与研究。

3 文本挖掘技术

文本挖掘(Text Mining)是一个新兴的交叉研究领域,涉及到数据挖掘、机器学习、统计学、自然语言处理、可视化技术、数据库技术等多个学科领域的知识和技术。目前关于文本挖掘并没有统一的定义,关于文本挖掘的名称亦有“文本数据挖掘(Text Data Mining)”或“文本知识发现(Knowledge Discovery in Text)”等不同说法。一个比较广泛使用的定义是:文本挖掘是指为了发现知识,从文本数据中抽取隐含的、以前未知的、潜在有用的模式的过程^[7]。它是一个分析文本数据,抽取文本信息,进而发现文本知识的过程。这里的知识包括各种模式、模型、规则或趋势等。

文本挖掘技术包括一系列广泛的文本处理与数据挖掘技术,其完整过程包括预处理、模式挖掘、模式评价等多个步骤。最常用的文本挖掘技术包括文本结构分析、文本摘要、文本分类、文本聚类、关联规则、分布分析与趋势预测、可视化技术等。这些方法的综合运用,为从海量的文本数据中发现新颖、有趣的知识提供了技术手段。

文本挖掘方法目前在多个领域取得了广泛的应用。在自然科学研究领域,尤其是生物学、医学等,文本挖掘在基因、蛋白质及其相互关系的研究中发挥了巨大的作用。在商业领域中,文本挖掘技术被广泛应用于客户关系挖掘、用户使用偏好挖掘、竞争情报分析。近年来,由于互联网信息的剧增以及大量数字图书馆项目产生的海量数字资源,文本挖掘技术在人文和社科研究领域的应用得到了人们越来越多的关注。

4 文本挖掘在数字人文研究中的应用

近年来,国际数字人文大会(Digital Humanities)中,有越来越多的论文探讨文本挖掘技术的应用,并产生了一大批重要的实践项目。研究热点包括作者归属与风格分析、作品情感分析、人物关系挖掘、模式发现与可视化、人文学科领域本体构建等。

4.1 作者归属与风格分析

人文研究中,尤其是古典文学和历史学研究,常常会涉及到对身份不明的作品的鉴别。文本挖掘在作者归属研究中的应用,就是通过对文章风格特征

的分析来辅助鉴定作者身份。文章风格分析主要是利用计算机辅助技术对作者在用词、句式等方面的特点进行定量统计分析。通过文章风格的统计分析进行作者归属的研究是基于这样的假设,即每个作者的风格都有某些无法被有意篡改的独特特征。风格分析的目的正是识别一个作者区别于另一个作者的特征^[8]。

在古典文学研究中,关于这一应用最具影响力的研究,莫过于利用统计学的方法对《红楼梦》一书的作者归属进行的一系列研究。1980年6月,美国威斯康星大学陈炳藻教授在首届国际《红楼梦》研讨会上宣读了一篇《从词汇上的统计论〈红楼梦〉的作者问题》的文章,开创了该领域的先河。陈炳藻教授利用电脑对《红楼梦》前80回和后40回的用字进行了测定,并从数理统计学的观点出发,推断出前80回与后40回的作者均为曹雪芹一人的结论。然而,复旦大学数学系李贤平^[9]教授却给出了相反的回答。李贤平教授把120回看成120个样本,选取了47个虚词作为特征向量,进行了聚类分析,发现前80回聚为一类,后40回聚为一类,得出了《红楼梦》不是出自一人的手笔的结论。相关的研究和探讨并没有结束。施建军^[10]对李贤平教授运用聚类分析对红楼梦作者进行判断的方法进行了具体剖析,指出了方法的不严谨之处,认为以同一部作品为对象的聚类分析不能够进行作者的判定。同时以鲁迅和瞿秋白的18部作品为例,以131个常用汉字的使用频率作为两个作家的风格特征向量,进行了聚类分析实验。文章指出,运用聚类方法对古典文学作品的作者进行判断的可信度是值得商榷的。除了对《红楼梦》作者的考证外,相关学者如武晓春^[11]、年洪东^[12]等对其他作家的不同作品也做了研究和实验。

国外研究者除了利用各种分类和统计算法对作者归属进行了大量的研究^[13,14]外,还利用文本挖掘技术对海量文本中作家的性别特征进行了挖掘。Helma等^[15]对古希腊作品中的作者性别特征进行了分析与研究。Shlomo等对法国文学中男女作家的性别特征进行研究,得出了男女作家不同的写作风格的定量佐证^[16]。

4.2 作品情感分析

对文学作品的情感分析,是文学研究中的一个重要任务。传统的文学作品情感分析,只能对较短的文本进行人工分析或者简单统计。而利用文本挖

掘的方法,则可以对海量的文学作品的情感倾向性进行自动分析,并挖掘出总体的情感特征和相关的重要情感特征词。情感分析是内容相关的,即与文学作品的内容和主题相关,而前面的作者归属,即作品风格的分析则是内容无关的,是某一作者的个人的风格特征。国内外利用情感分析进行文学作品的研究产生了一大批重要的研究成果。

国内,在计算风格学领域,利用各种分类、聚类的算法,对文学作品的情感特征(如豪放还是婉约)进行自动判别的研究已经取得了丰富的成果,在计算机辅助的唐诗宋词研究方面的成果尤为突出。多篇硕博学位论文^[17,18,19]利用各种机器学习和数据挖掘算法对唐诗宋词的风格和情感进行了自动分类研究,达到了较高的检全率和检准率。

国外则对诗歌、小说等各种体裁的文学作品的情感特征利用文本挖掘方法进行了探索。Tom Horton 等^[20]对 19 世纪美国的小说(包括著名的《汤姆叔叔的小屋》等)进行了基于机器学习的情感分类研究。他把小说中每一章按照情感强度标记为高、中、低,然后进行机器学习,再对新的小说文本进行测试,发现了表征不同情感强度的具体场景和词汇。Plaisant 等^[21]对美国 19 世纪著名女诗人艾米莉·迪金森的书信进行了挖掘,通过文本分类探究其诗歌中色情特征的表现。该项研究通过与迪金森研究专家的交互共同完成,通过自动分类与人工判定相结合的方法,发现了其诗歌中表现色情特征的指示词,并且取得了较高的分类精度。

4.3 人物关系挖掘

人物关系挖掘指利用计算机辅助的方法自动从文学作品中抽取相关人物及人物之间的相互关系并挖掘其结构特征的方法,这主要涉及命名实体识别、关系抽取等技术。利用文本挖掘中信息抽取等方法可以对小说中的人物进行社会网络关系挖掘,使得文学研究者可以从大规模的文本中对人物关系和背景有一个系统而全面的认识,弥补了传统研究方法只能通过对单个文本进行深入分析的缺陷。这一新的研究视角可以对传统研究方法得出的结论进行新的佐证或者否定。

国外的研究者通过引入计算机领域的有关模型对文学作品中的人物关系的挖掘进行了较多的尝试。Elson 等^[22]对 60 部 19 世纪的英国著名小说和期刊进行了社会网络关系挖掘。文章通过识别由引

号标记的人物对话,然后指派给特定角色,构建了小说中人物的社会网络关系。文章研究了抽取的社会网络的特点,社会网络之间的关系以及社会网络与小说背景和特征的关系。结果表明这一时期的大部分小说不符合文学研究者通常给出的两个特征描述。因此,文章给出了对这一阶段社会网络特征的新解释,阐明了小说人物的数量与社会网络特征的相关关系以及小说的形式(如第一或第三人称)与背景(城市或乡村)对社会网络关系的影响。这种方法对传统研究的结论提出了新的挑战,并提出了新的视角。在另一篇文章中,Elson 等^[23]对如何从小说对话中抽取人物社会网络关系的方法进行了详细介绍。另外,Asli 等^[24]利用 ACTM(actor-topic model)模型对 19 世纪的小说进行了社会网络挖掘,文章同时对人物中隐含的社会网络关系进行了深入挖掘。

人物关系挖掘主要是针对小说这样的文学体裁,是一个比较前沿的研究方法。目前主要是将计算机领域的 ACTM 模型应用到文学研究领域,但是该方法涉及的技术比较复杂,因此还未有大规模的应用与推广。

4.4 模式发现与可视化

模式(pattern)是一个相对宽泛的概念,这里指隐藏在文本中的规律和特征。人文作品一般都有着独特的结构和语义特征,通过文本挖掘可以发现文本中隐藏的模式,并通过可视化技术加以展现。可视化的方法可以有效地帮助我们发现文本中一些固有的模式和特征,并通过多维的方式进行展现。

在文学研究方面,Bei Yu 等^[25]通过对文学评论文本集和数据挖掘论文集进行关键词挖掘,试图发现文学研究活动与数据挖掘活动可能的交叉点。挖掘的结果表明,文学研究者事实上也是“数据挖掘者”,只不过挖掘的知识和模式不同而已。文学研究者希望发现的是“叙事模式”、“婚姻模式”、“情节模式”等,而数据挖掘研究者关心的是更抽象的模式如“时间序列模式”、“关联模式”、“拓扑模式”等。这里的模式是更一般意义上的概念,主要是对文学文本集进行一些探索性的研究,而不是针对某些具体的主题如情感特征、人物社会网络关系等。

国外相关研究对文学作品中的模式发现和可视化进行了许多探索。Don 等^[26]利用自行开发的文本挖掘工具 FeatureLens 从高频词组、高频单词、高

频 n -grams 闭项集三个层面及篇章、段落多个粒度对不同类型的文本集合(两本书和演讲集)进行了模式挖掘和可视化研究,并对系统的可行性进行了用户案例研究。Clement^[27]对格特鲁德·斯坦因的 *The Making of Americans* 一书利用 MONK 项目提供的文本挖掘工具进行了重新解读,通过对文本中语言重复模式的挖掘与展示,发现了一些与传统文学研究者不同的观点。

在历史学研究方面,越来越多的学者开始借助 GIS 技术进行历史知识和历史事件的静态和动态的可视化展示研究。可视化技术可以发现历史事件随时间和空间的演变规律,并可直观地进行展示。历史领域的可视化一直是史学界的研究热点,通过与文本挖掘技术相结合,可以给历史研究者带来更大的便利。文本挖掘可以抽取海量文本中与地理和历史相关的命名实体,并可通过可视化的方式展现其随时间变化的趋势。

国外关于文本挖掘在地理历史可视化中的应用产生了许多有影响力的项目。美国国家人文基金(NEH)资助了北德州大学的一个历史领域文本挖掘与可视化应用的项目^[28]。该项目结合文本挖掘技术和 GIS 可视化技术对历史报纸进行挖掘和可视化展示。文本挖掘的作用在于从海量的数字化的历史报纸中发现散布在文本中的语言模式,可视化技术可以把这种语言模式按照时间和空间的转变进行可视化展示。另一个较有影响力的项目是由英国艺术与人文研究委员会(AHRC)资助的 Hestia^[29]项目。项目提供了研究古代世界空间概念的新方法,通过对希罗多德(希腊的历史学家)历史作品中提到的地点及其地理特征的描述进行信息抽取和可视化,提供了可以深入挖掘文本拓扑结构的可视化工具。

4.5 其他应用

除了以上几个典型应用外,文本挖掘在数字人文研究的其他领域也有着广泛的应用。如利用文本挖掘技术辅助人文领域的本体构建。在哲学领域,比较有影响力的项目有印第安纳大学承担的 InPho 项目。该项目开发了一个在线的哲学本体,并可经用户反馈动态地更新。相关的哲学本体项目还有 Josh Dever 维护的“Philosophy Family Tree”^[30]、Pasin 和 Motta 主持的“PhiloSURFical project”^[31]等。在历史领域中,比较有影响力的项目是武汉大学董慧教授主持开发的“国共合作”历史领域本体。

项目所构建的“国共合作”历史领域本体描述了从五四运动开始一直到前不久连战访问大陆的这段历史时期涉及的概念、术语、关系、个体^[32]。国外相关的项目还有欧盟资助的一个欧洲历史的本体项目^[33];

另外,在考古学、人类学领域文本挖掘的方法也逐渐得到了应用。在考古学领域,学者们尝试运用文本挖掘的方法对古代语言进行分析,哈佛大学从事宋史研究的著名学者包弼德认为,“古代汉语是一种非常有规律的语言,古代汉语文本甚至比英语文本更适合运用‘文本挖掘’技术提取信息。”^[34]英国人文与艺术委员会资助了 ADS(Archaeology Data Service)^[35]项目。该项目长期保存考古学相关的数字化数据,并为挖掘考古学数据提供了相关工具如 Archaeotools^[36]。

5 国外数字人文项目中文本挖掘的前沿实践

5.1 美国

美国的数字人文研究较为领先,研究团队众多,与文本挖掘相关的研究项目十分丰富。

美国国家人文基金(NEH)资助了多项人文领域的文本挖掘项目。2008年 NEH 资助了乔治梅森大学的历史与新媒体中心(CHNM)的一个历史领域文本挖掘项目。项目利用文本挖掘技术开发复杂的信息检索、信息抽取和信息分析工具帮助历史研究者利用各种微观(如词频分析)和宏观(如文档分析)分析方法对海量的数字化文本进行分析。该项目的案例分析探究是否如人们通常认为的那样,从维多利亚时代到 19 世纪末,英国人对科学、宗教、进步的观点发生了巨变。该项目得到了国家人文基金数字人文办公室(ODH)的追加资助,同时还是 Google 2010 年资助的 12 个国际人文社科文本挖掘项目之一^[37]。为了解决近几年各种数字化项目产生的海量文本的有效利用问题,ODH 联合英国信息系统委员会(JISC)、美国科学基金会(NSF)、加拿大社会科学与人文学研究委员会(SSHR)发起了“Digging into Data Challenge”项目^[38],资助了一大批人文和社会科学领域挖掘海量数字文本的创新性研究项目。

在私人基金资助的项目方面,梅隆基金(Andrew W. Mellon Foundation)资助了一系列人文领域的文本挖掘项目,较知名的项目有 SEASR、NO-RA、MONK 等。SEASR^[39]为应用前沿的数字人文

技术提供了研究和开发环境;NORA^[40]开发了一系列软件工具,从现有的数字图书馆海量人文资源中发现、挖掘重要模式并进行可视化。MONK^[41]则在NORA项目的基础上召集众多高校和研究机构,增加了更多社会软件功能并对在分布式环境下部署文本挖掘工具进行了探索。

有些技术公司也加入到数字人文项目的投资中来。Google于2010年资助了来自文学、语言学、历史学、古典文学、哲学、社会学、考古学和人类学8个学科的12个与文本挖掘相关的数字人文研究项目。到目前为止Google已经对涉及300多种语言的一千两百万本图书进行了数字化。Google研究项目旨在利用这些数字化的语料支持人文领域的文本挖掘研究。研究者可以利用谷歌图书的数据库,开发可以被其他研究者使用的数字工具和协议。

另外很多高校也承担了多项文本挖掘项目,开发了许多文本挖掘工具和平台,如芝加哥大学开发的PhiloMine^[42]文本挖掘工具融合了机器学习和多种文本挖掘技术。2010年12月16日,哈佛大学的研究者在在线期刊*Science Express*上发表了一篇文章,对500多万本1800年至2000年出版的英文书籍(约占所有出版书籍的4%)进行了文本挖掘,通过分析词语的共现频率和定量分析得出了如“我们每年都以更快的速度忘记过去”、“技术被采纳的速度越来越快”等相关结论^[43]。

5.2 英国

在利用数字技术辅助人文科学研究中,英国也走在了前列,成立了多个与文本挖掘有关的研究中心,并召开了多次国际会议和研讨会。英国信息系统委员会(JISC)已经资助了多个人文社科领域的大型文本挖掘项目,如HiTher^[44]、ASSERT、AS-SIST等。在大学方面,曼切斯特大学和伦敦国王学院是人文社会科学应用文本挖掘技术的研究重镇。

英国成立了第一个国家资助的文本挖掘中心——NaCTeM^[45]。NaCTeM的功能是根据英国学术界的需求提供文本挖掘服务。NaCTeM的活动与英国e-science框架、网格技术以及相关软件和中间件开发标准密切相关。NaCTeM第一阶段主要关注生物学领域的文本挖掘,承担了多项生物学、医学、化学等领域的文本挖掘项目。近年来该中心也承担了多项人文社会科学领域中的文本挖掘项目。如前面提到的ASSERT项目,就是NaCTeM

向人文社会科学领域拓展的一个项目。

在国际会议方面,英国召开了多次与人文领域文本挖掘相关的会议和研讨会。2006年7月在兰卡斯特大学举行了“历史领域文本挖掘”研讨会,2007年7月在格拉斯哥大学举行了“史学家的文本挖掘”研讨会。研讨会汇集了来自英语语言和文学、语言学、计算机科学、历史学等多个学科的研究者,对历史研究领域如何使用各种文本挖掘工具和技术进行了深入探讨。2010年5月在兰卡斯特大学召开了主题为“数字人文中的文本挖掘:概念历史、批判话语分析与语料库语言学”的研讨会^[46]。该研讨会把传统的历史领域研究方法和文本挖掘方法进行了更深入的结合。2010年国际数字人文大会7月在英国伦敦国王学院召开,大会设立了“数字人文中的文本挖掘”研讨会^[47]。研讨会比较了基于理论的人文研究与模型驱动的计算机科学研究的区别,并给与会者提供了亲自操作文本挖掘工具的机会。

5.3 欧洲其他国家

欧洲的数字人文研究起步较早。欧洲科学基金会人文常务委员会(ESF/SCH)对数字人文也给与了高度关注,并联合IEEE举办了多次e-humanities会议。欧盟资助的专门的数字人文项目包括公共语言与技术基础平台项目(CLARIN)和DARIAN项目。这些项目平台为进一步的文本挖掘应用提供了基础设施和语言工具。德国在有关项目的研究中处于领先地位,资助了一大批与文本挖掘相关的平台和项目。

德国的eAQUA^[48]项目是其中较具影响力的一个项目。该项目旨在从古代文本中产生特定领域的知识,并通过eAQUA门户提供相关知识,该项目实现了信息科学与古典文学研究的结合。项目从现有的数字化项目中获取数据,并提供标准的数据导入接口,研究的问题包括丢失作品的语义重建与分类;柏拉图作品的引用与传承情况;古文献的自动分类;铭文重要模板的抽取;拉丁喜剧的韵律度量;碎片文本的填充等^[49]。在2010年国际数字人文会议中,该项目开发的相关软件在“数字人文中的文本挖掘”研讨会中进行了展示。

另一个较有影响力的项目是TextGrid。TextGrid是由德国联邦教育和研究部资助的一个联合项目,由多个大学图书馆、研究机构及商业公司共同承担。项目于2006年启动,旨在创建一个虚拟研究

图书馆,在支持 TEI 编码^[50]的环境下开放源码和接口,为人文学者提供一个处理、分析、注释、编辑、链接和发布文本数据的网格工作平台,其目标是在一个动态而又虚拟的网络和跨学科环境下促进学术研究。TextGrid 项目首先在文学和语言学领域启动。在项目的第二阶段增加了人文艺术学科和相应的特殊功能,如为艺术史提供了图像数据的标注功能;为古典文学提供了注释编辑功能;为音乐提供了在 XML 编辑器中整合笔记的功能。该项目是德国第一个网格技术在人文学科中的应用项目,为推进文本挖掘技术在人文学科的应用提供了基础设施和应用平台,是德国 e-humanities 项目的一个里程碑。

此外,法国、意大利、西班牙等国的人文学者也在数字人文研究与实践的过程中对文本挖掘的应用进行了探讨。

6 结束语

随着武汉大学成功建立全国第一个数字人文中心,国内相关学者已经更深刻地意识到新的计算机和网络技术对人文学科的影响。随着 e-social-science、e-humanities 等相关运动的进一步深化,信息技术对人文科学的影响越来越大。各种 e-infrastructure 基础平台建设的进一步深入,网格技术在人文领域的大规模应用以及分布式文本挖掘等新的文本数据分析技术的发展将给人文学者带来更大的便利,对创新人文科学的研究方法和研究范式产生深远的影响。

因此,国内人文学界应该抓住这一机遇,在新的形势下,借鉴国外数字人文研究中的经验,从以下几个方面努力推进数字技术在人文研究中的应用和发展。

① 完善数字人文学科的理论体系,从新的理论高度总结和归纳数字技术与人文研究的关系。

② 加强数字人文工具的研究,数字人文工具和平台在文本挖掘研究中起着核心作用,应该借鉴国外的相关软件,开发适合中文文本挖掘的平台和工具。

③ 积极参与国际合作,鼓励我国的人文学者积极参加国际上的数字人文专题会议和组织联盟,争取与国外的人文学者共同开展相关的文本挖掘项目,共同推进我国人文研究的进步。

参考文献

1 王晓光.“数字人文”学科的产生、发展与前沿.[2011-11-24].

- http://blog. sciencenet. cn/home. php mod = space&uid = 67855&do=blog&id=275758
- 2 中国社会科学院. 中国社会科学综合地理信息服务平台.[2011-11-19]. <http://gis. cass. cn/>
- 3 南京师范大学虚拟地理环境教育部重点实验室. 华夏家谱 GIS 网.[2011-11-19]. <http://www. hxjiapu. com. cn/Index. aspx?&ThreadID=>
- 4 中国艺术研究院. 西部人文资源网.[2011-11-19]. <http://www. xbchina. com/>
- 5 清华大学图书馆.《中国基本古籍库》全文网络版.[2011-11-19]. <http://www. lib. tsinghua. edu. cn/database/jibenguji. html>
- 6 首都师范大学. 古籍电子定本工程.[2011-11-19]. <http://www. guoxue. com/zt/dzdb/>
- 7 Tan A H. Text Mining: The State of the Art and the Challenges//Tsinghua University. Third Pacific - Asia Conference on Knowledge Discovery and Data Mining, Beijing, 1999: 65-70
- 8 Diederich J, Kindermann J, Leopold E, et al. Authorship Attribution with Support Vector Machines. Applied Intelligence, 2003, 19(1/2): 109-123
- 9 李贤平.《红楼梦》成书新说. 复旦学报(社会科学版), 1987(5): 3-16
- 10 施建军. 关于以《红楼梦》120 回为样本进行其作者聚类分析的可信度问题研究. 红楼梦学刊, 2010(5): 318-335
- 11 武晓春等. 基于语义分析的作者身份识别方法研究. 中文信息学报, 2006(6): 61-68
- 12 年洪东等. 当代文学作品的作者身份识别研究. 计算机工程与应用, 2010, 46(4): 226-229
- 13 Gerritsen C M. Authorship Attribution Using Lexical Attraction[硕士学位论文]. Boston: Massachusetts Institute of Technology, 2003
- 14 Peng F, Schuurmans D, Keselj V, et al. Language Independent Authorship Attribution using Character Level Language Models//Hungarian Academy of Science. 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, 2003: 267-274
- 15 Dik H, Whaling R. Mining Classical Greek Gender. [2011-11-19]. <http://cybergreek. uchicago. edu/MiningGender. pdf>
- 16 Argamon S, Goulain J B, Horton R, et al. Text Mining Gender Difference in French Literature. [2011-11-19]. <http://www. digitalhumanities. org/dhq/vol/3/2/000042/000042. html>
- 17 胡俊峰. 基于词汇语义分析的唐宋诗计算机辅助深层研究[博士学位论文]. 北京: 北京大学计算语言学研究所, 2001
- 18 苏劲松. 全宋词语料库建设及其风格与情感分析的计算方法研究[硕士学位论文]. 厦门: 厦门大学, 2007
- 19 吴春龙. 宋词风格的计算机辅助分析研究[硕士学位论文]. 厦门: 厦门大学, 2008
- 20 Horton T, Taylor C, Yu B, et al. 'Quite Right, Dear and Interesting': Seeking the Sentimental in Nineteenth Century American Fiction//Paris-Sorbonne. Digital Humanities. France, 2006
- 21 Plaisant C, Rose J, Yu B, et al. Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces//ACM Press. Proceedings of the 6th ACM/IEEE - CS Joint Conference on Digital Libraries. Chapel Hill, NC[C]. New York, 2006

- 22 Elson D K, Dames N, McKeown K R. Extracting Social Networks from Literature Fiction. [2011-11-19]. <http://www.cs.columbia.edu/~delson/pubs/ACL2010-Elson-DamesMcKeown.pdf>
- 23 Elson D K, McKeown K R. Automatic attribution of quoted speech in literary narrative//AAAI, 2010
- 24 Celikyilmaz A, et al. The Actor-Topic Model for Extracting Social Networks in Literary Narrative[C]// NIPS Workshop: Machine Learning for Social Computing, 2010
- 25 Yu B. Toward Discovering Potential Data Mining Applications in Literary Criticism. [2011-11-19]. http://www.csdl.tamu.edu/~furuta/courses/06c_689dh/dh06readings/DH06-237-239.pdf
- 26 Don A, Zheleva E, Gregory M, et al. Discovering Interesting Usage Patterns in Text Collections - Integrating Text Mining With Visualization//ACM. 16th ACM Conference on Information and Knowledge Management, 2007, 213-222
- 27 Clement T E. 'A thing not beginning and not ending': using digital tools to distant-read Gertrude Stein's *The Making of America*. *Linguist Computing*, 2008, 23(3): 361-381
- 28 NEH. Mapping Historical Texts: Combining Text-mining & Geo-visualization to Unlock the Research Potential of Historical Newspapers. [2011-11-19]. <https://securegrants.neh.gov/PublicQuery/main.aspx?f=1&gn=HD-51188-10>
- 29 AHRC. Hestia. [2011-11-19]. <http://www.open.ac.uk/Arts/hestia/index.html>
- 30 Dover J. The Philosophy Family Tree. [2011-11-19]. <https://webpace.utexas.edu/deverj/personal/philtree/philtree.html>
- 31 Pasin M. PhiloSURFical project. [2011-11-19]. <http://philosurfical.open.ac.uk/index.html>
- 32 董慧等. 基于本体的数字图书馆检索模型研究(Ⅲ)——历史领域资源本体构建. *情报学报*, 2006(5): 564-574
- 33 European Union. VICODI. [2011-11-19]. <http://www.vicodi.org/about.htm>
- 34 中国社科院. 中国社会科学报刊网. [2011-11-19]. <http://sspress.cass.cn/news/13448.htm>
- 35 University of York. Archaeology Data Service. [2011-11-19]. <http://ads.ahds.ac.uk/>
- 36 University of York. Archaeology Data Service. [2011-11-19]. <http://ads.ahds.ac.uk/project/archaeotools/>
- 37 National Endowment for the Humanities. 'Text Mining'-Digging Through Digital Archives. [2011-11-19]. <http://www.neh.gov/news/archive/20101221.html>
- 38 National Endowment for the Humanities. Digging into Data. [2011-11-19]. <http://www.diggingintodata.org/>
- 39 University of Illinois Urbana-Champaign. SEASR. [2011-11-19]. <http://seasr.org/>
- 40 University of Illinois Urbana-Champaign. NORA. [2011-11-19]. <http://www.noraproject.org/>
- 41 University of Illinois Urbana-Champaign. MONK. [2011-11-19]. <http://www.monkproject.org/>
- 42 University of Chicago. PhiloMine. [2011-11-19]. <http://code.google.com/p/philomine/>
- 43 Michel J B, Shen Y K, Aiden A P, et al. Quantitative Analysis of Culture Using Millions of Digitized Books. [2011-11-19]. http://mf.uchicago.edu/publications/papers/Science_Culturalomics.pdf
- 44 Blanke T. Text Analysis in the Arts and Humanities. [2011-11-19]. <http://cnx.org/content/m31502/latest/>
- 45 Manchester Interdisciplinary Biocentre. The National Center for Text Mining. [2011-11-19]. <http://www.nactem.ac.uk>
- 46 Lancaster University. Text-mining in the Digital Humanities: The Interface between Conceptual History, Critical Discourse Analysis and Corpus Linguistics. [2011-11-19]. <http://ucrel.lancs.ac.uk/events/chcdac12010/>
- 47 King's College London. Text Mining in the Digital Humanities. [2011-11-19]. <http://dh2010.cch.kcl.ac.uk/academic-programme/pre-conference-workshops/workshop-2.html>
- 48 Universität Leipzig. eAQUA. [2011-11-19]. <http://www.eaqua.net/index.php>
- 49 Buechler M, Gerhard H, Sabine G. eAQUA - Bringing modern text mining approaches to two thousand years old ancient texts. [2011-11-19]. <http://asv.informatik.uni-leipzig.de/publication/file/123/IEEE09.pdf>
- 50 University of Virginia. TEI: Text Encoding Initiative. [2011-11-19]. <http://www.tei-c.org/>

作者单位: 华东师范大学商学院, 上海, 200241

收稿日期: 2012年1月2日

Text Mining Research in Digital Humanities

Guo Jinlong Xu Xin

Abstract: This paper firstly summaries the concept, content and current situation of digital humanities and then points out that the text mining method will become a trend of digital humanities research. It then focuses on the various applications of text mining in different areas of digital humanities and introduces the most advanced practices from developed western countries so as to provide reference for the development of domestic humanities research.

Keywords: Text Mining; Digital Humanities; Humanities Research