

· 电脑与史学应用 ·

古籍数字资料应用与史学研究

王 文 涛

(河北师范大学 历史文化学院, 河北 石家庄 050091)

[关键词] 古籍; 数字资料; 知识发现; 史学研究

[摘 要] 古籍数字化与当代先进的信息技术的结合日益紧密, 古籍数字资料在史学研究中的应用也日趋密切。检索结果和文献出处一次性输出为一个文件, 是优秀古籍检索软件应当具有的功能。对于检索到的数字资料, 如何进一步归类整理, 如何从海量信息中获取知识发现, 是数字资料应用中带有普遍性的新问题。逻辑概念的分类、归纳、排比、筛选和分析综合等, 非文字处理软件所长, 可以使用电子表格、数据库或其他文本软件与文字处理软件协同工作。当然, 检索不能代替研究, 将扎实的学术功底与先进的电脑网络知识相结合, 才能充分利用与享受信息技术带来的方便快捷, 提高学术研究的效率和质量。

[中图分类号] K061 [文献标识码] A [文章编号] 0583-0214(2009)01-0119-07

Digitizing Ancient Books and Historical Research

WANG Wen-tao

(History and Culture College, Hebei Normal University, Shijiazhuang 050091, Hebei, China)

Keywords: Ancient books; digitized materials; knowledge discovery; historical research

Abstract: The process of digitizing ancient books by integrating more and more advanced information technology is occurring more frequently in historical research. Suitable ancient books document retrieval software should be able to output the literature searching results with the corresponding reference. But the searching results beg further questions and possibilities such as how to further archive, and how to increase the relevancy of returned results to a search. This article gives some discussion about this topic in association with personal digitizing application experience, and try to throw out a minnow to catch a whale. The classification, generalization, parallelism, selection, analysis and integration of logical concept are not the strength of word process software. We may also use excel, database, and other text software to work with. However, document retrieval cannot replace research, to take advantage of the information technology and improve the efficiency and quality of academic research, we need to integrate the well-knit academic knowledge with the adept computer & network skills.

司马迁之后, 搜集和考辨史料, 作为研究历史的基础, 为绝大多数史学研究者所继承, 并逐渐形成考据之学。随着信息技术的发展, 古籍数字化热潮方兴未艾, 从零星制作到规模开发; 从初期的图形扫描到字符数码化; 从目录、文摘的制作到全文录入; 从制作单机版 CD 发展为网络版的数据库。随着 OCR 扫描技术的成熟、UNICODE 编码的统一、全文检索软件的完善以及 Web 技术的普及, 以网络为主要载体, 数字图书馆建设迅速发展, 古籍数字资料的应用与史学研究的结合日益紧密, 信息技术对人文社会科学的影响也在向深度和广度发展, 相关问题的应用与研究也越来越受到人们的重视。

中华民族创造了无比丰富的历史文化遗产, 古代典籍是中国历史文化遗产最为重要的物质载体, 是世界文化的重要组成部分。胡适先生认为传统的经史研究有很多优秀遗产, 但也存在范围太狭窄, 注重功力而忽略理解, 缺乏参考比较的资料等积弊, 故以清代三百年间第一流人才的心思精力, 都用在经学的范围内, 所获成果并不相称, 关键是缺少对古籍的系统整理, 又不注重学术成果的积累。针对清儒治学方法的缺陷, 近代以来学术界编纂了多种引得、通检、索引、汇编等工具书, 部分完成了索引式整理的任务, 为我们查阅古籍提供了诸多便利。但是, 中国古籍汗

牛充栋,经过系统整理的毕竟只是少数,方便的检索工具还是太少。即使是已有索引的古籍,用来解决具体问题时仍会感到种种不便。

史料是历史研究的基础。每一个从事史学工作的人都要搜集和积累史料,以前使用最普遍的方法就是抄录卡片,看见有用的资料就抄,并加以分类。这些卡片基本上是按照个人的需要辑录并供个人使用的,难以共享。人文学术研究是个性化鲜明的工作,一个人的精力和时间非常有限,将有限的时间和精力花费在浩繁、琐碎的翻检抄录工作中,繁累、枯燥自不待言,也影响学习和研究效率,学术进步自然也就困难了。因此,我们需要应用便捷、高效、准确的检索工具为人文学术研究服务。

搜集资料的工作麻烦而又艰巨,但十分重要。这个工作一定要做,因为这是进行学术研究的基础工作和必要手段。不亲自动手去做,就发现不了问题,而且,只有尽可能全面地占有资料,才可能对所研究的问题进行科学论证,得出正确的结论。要搞研究工作,就不能怕麻烦,要花大气力做搜集资料的工作。

进入信息时代,对于不同年龄的研究者来说,数字卡片正在和已经取代传统的纸质卡片,数字图书馆正在迅速丰富着个人的数字藏书。储存和检索数字资料,是史学工作者使用计算机和网络的主要用途。古籍数字资料的搜集与整理是一个完整的过程,它包括数字资料搜集范围的确定,数字资料的筛选与鉴定等。资料搜集不一定严格地按照上述先后次序,也可以同时进行,例如一边搜集,一边鉴别,视具体情况而定。数字资料搜集完成以后,必须进行认真的鉴别和核对,因为很难保证我们搜集到的数字资料都是可靠的,去伪存真,去粗取精,才能保证数字资料的可靠性。一般读书网站的数字图书错误较多,使用时必须要校对。由于这些数字图书都没有页码,直接查找纸质图书原文,费时费力,可以先用高质量的数字图书做初校,剩下的问题再与纸质图书核校。例如,香港迪志文化公司开发的文渊阁《四库全书》、北京书同文公司的《四部丛刊》以及“国学网”上的数字资料校对质量达到了图书出版的要求,用它们做初校,可以提高校对速度;“二十五史”校对可以使用南开大学陈永川开发的网络版“二十五史”全文检索系统,这个系统提供网络免费使用,其优点是页码与中华书局标

点本完全一致,便于和中华书局本核对。鉴别数字资料的方法很多,如对数字资料所存书籍年代的考证,作者和版本的考证,文字和方法方面的鉴定等。这是每一位史学工作者的基本功,在这里无须多谈。计算机作为现代科学技术发展的结晶,为存贮、检索、分析和处理大量史料提供了重要的技术保证。利用计算机把史料的整理归类工作做好,使用起来就方便多了。这就需要史学工作者熟悉、掌握基本的计算机信息技术,以便于和信息技术人员配合协作,把历史学数字资料建设和史学研究推向深入。

就现阶段中国大陆的古籍数字资料应用来说,有喜有忧。一方面,信息技术的迅速发展,为古籍数字化提供了充分的技术条件。信息高速公路将世界连接为一个名副其实的地球村,国际互联网提高了电子文献的检索效率,扩大了服务范围,便捷的信息传递节省了远程通信费用。新一代高性能计算机的海量存储和惊人的秒级运算能力,使我们再也不必为存储空间和运行时间的矛盾而苦恼。新的国际计算机信息处理标准的制定和实施,为建构全球统一的信息处理系统奠定了坚实的基础。通用 UNICODE 码包含 6 万多个汉字,为汉字信息处理的国际化和标准化开辟了道路。新的信息应用技术,如非键盘输入技术、中文数据库技术、多媒体压缩与传送技术、安全保密技术、自然语言理解技术等出现,为文献数字化事业提供了有力的支持。尤其是非键盘输入技术使文献载体转换方式发生了一场革命,自动识别输入技术(OCR)使海量信息输入的工作量大大降低,清华紫光公司所研制的非特定人手写识别软件仅用三个月时间就将一部 8 亿多字的文渊阁《四库全书》输入计算机,为同类工作积累了宝贵的技术财富。

另一方面,现有的网络古籍数字资料分布极不平衡,大部分网络古籍数字资料库集中在海外,大陆学者在使用上存在诸多不便:文字编码不统一,会员资格受限制,服务器连接也不够通畅。就目前掌握的情况看,尽管大陆的软件公司推出了各种版本的《廿五史》和常用古籍(重复开发多),还有《四库全书》、中国基本古籍库等煌煌巨制的检索系统,以及正在试运行的“龙语瀚堂典籍数据库”等,但缺乏整体规划和系统开发,真正投入 Internet 运营的只有“国学网”等寥寥数家。究其原因,除了服务器数据库运营技术复

杂,费用相对昂贵外,网络市场不够健全是主要因素,许多商家宁肯用上千以至数万的价格卖出光盘,或者以数十万的高价出售局域网版本,以求尽快收回成本。而大陆无论是学者个人,还是文史研究机构,经费相对拮据,没有能力支付软件费用,因而造成恶性循环。从长远看,借鉴海外网络运营经验,采用部分适当收费,部分免费开放的会员制运营,可能是一个比较有效的解决途径。

二

古籍数字化需要具备怎样的功能?数据应当怎样处理?使用何种技术才能保证古籍数据库具有我们需要的功能呢?这些都是我们在建设和应用古籍数字资料时经常遇到和思考的重要问题。

这里,我们所谈的数字化文献,不是为大众提供普及读本,而是为学术文化的繁荣奠定基础,这应当是我们制作和使用数字化文献的共识。数字化文献的功能不仅在于一般的信息查询,更重要的是古籍文献中的知识发现。它应当符合各种国际通用标准,具有开放性,可以在网络上传输,实现信息资源共享。古籍数字化的过程,基本上可以视为文献全文数据库的生成过程。一部古籍文献输入计算机,就形成了无标引的全文数据库,即半结构化的数据库。目前,以中国古籍为内容的电子读物多为此类产品,但这远不能发挥计算机的技术优势,也难以达到研究者的要求,其最大的缺陷就在于它不能像结构化数据库一样经由排序、筛选、分类和统计之类的管理过程产生再生资源,更谈不上知识发现。因此,对古籍中的数据进行充分的分析和处理,制作成结构化数据库,与半结构化数据库相结合,才是较为完美的方案。数字化古籍适合实现多途径排检功能,在确保信息查询的查准率和查全率的前提下,提供了实现海量信息中知识发现的可能。

古籍数字资料检索结果的输出,是关系到使用效率的一个非常重要的问题。目前,文渊阁《四库全书》全文数据库的检索结果必须通过阅读原文才能知晓具体内容,不能集中显示,给用户使用带来了一些不便。例如,“孔子”的检索结果多达23 757卷、111 641个匹配。有人做过统计,假定每个匹配的阅读时间平均为1分钟(加

上复制相关资料、标点等),每天8小时不间断地阅读,“孔子”的检索结果需要233天才能阅读、复制完。如果是通过网络阅读,受网络传输速度的影响,耗时将更多。

有的检索软件提供了将检索结果一次性全部输出的功能,省去了用户一次次复制、粘贴的重复性劳动,非常方便。例如陕西师范大学袁林先生主持开发的“汉籍全文检索系统”,收入文史哲类古籍文献2159种,共7.4亿字。使用该系统,成百上千条检索结果和文献出处可以一次性输出为一个文本文件,方便快捷。不过,该软件检索内容的输出以关键词所在段落为单位,不论这一段落是几十个字还是上千字。这就带来了新的问题,字数少的段落脱离语境之后缺少相关信息,需要回到原文中阅读,补充资料;字数多的段落,无用的文字信息需要删除。如果检索到几百条资料,一次性输出之后在word等文字处理软件中阅读、整理,删除无用信息,工作量也是相当大的。因此,有些用户认为,在全文检索系统中阅读、复制和在word中阅读、整理差别不大,没有将检索结果一次性输出的必要,因此很少使用该检索软件提供的一次性输出功能。其实,我们可以利用EmEditor、UltraEdit等文本工具提高对输出文本的阅读和处理效率,以弥补word在这方面的不足。

使用Emeditor处理一次性输出的文本,第一步是将无用的信息用查找替换功能删除。然后键入关键词查找定位,以删除无用的文字。一般说来,对用户有用的信息是以前设定的检索关键词为中心,只要能迅速找到这个关键词,就可以提高阅读、处理输出资料的速度了。“Emeditor”的优点是进行新的关键词的查找时,能够将当前检索到的所有关键词一直高亮显示,这一功能非常方便用户迅速定位到以关键词为中心的有用信息。这样做和在检索软件中一次次地复制、粘贴相比,速度快了很多倍。即使使用粘贴工具,工作效率也不如用EmEditor处理一次性输出文本。处理的资料越多,速度差别越明显。

输出文字少的段落,需要补充相关信息,方法是回到全文检索软件中读书,再次输入同一关键词,找到它所在的语境,补充所需信息。这样研读数字古籍比阅读同一纸质书籍的目的性更强,查找便捷,可以迅速了解相关信息,单位时间内获得的信息更多,有利于我们更全面准确地解

读史料。这样的读书方式是数字资料的优点,也是纸质书籍所不具备的,笔者将其理解为检索式的阅读。由于检索方便,很少有翻检之劳,研究者更愿意通过检索去研读相关信息,以便发现问题、分析问题和解决问题。这不是简单的以检索代替阅读,而是针对性更强、涉猎范围更广、更有利于钩沉索隐的读书研究。这种读书方式的变化,是数字资料带来的,它不同于以纸质书籍为对象的精读和泛读。将这种读书方式与传统的精读、泛读相结合,不仅可以丰富我们的读书和研究方式,还可以消除对“以检索代替研究”的诟病,提高我们研究工作的效率和质量。希望有更多的人关注以数字资料为对象的读书和研究方式,探究这种变化带来的深层次的影响。

三

数字资料搜集的方便快捷,带来了新的问题,搜集到的这些文字资料按照什么标准分类?用什么方式或技术手段做进一步的归类整理?目前,尚没有方便适用的程序和统一的方法。对于个人搜集的数字资料的二次归类整理,完全由研究者根据自己对资料的理解和研究需要而定。整理的方式有以下几种:去粗取精,按性质归类,按时间顺序排比,按研究问题分组等,以方便使用、省时省力为目的。

检索得到的数字资料经过初步整理之后,大体上可以分为两类:一是数据性资料,二是需要进行逻辑分类的文字。对数据性资料进行分析,是史学研究的重要内容。一般来说,这些历史数据都是离散的,对它们的分析应依据统计学的原则来处理。可以利用数据库软件来做统计分析,内容一般包括:平均数、近似值、相关分析、回归分析、时间序列、加权平均数和指数、分布规律的研究,等等。根据不同的情况,运用不同的统计方法就可以揭示出数据集合的整体特征,为我们认识这些数据的实质提供可靠的科学依据。大多数情况下,我们要获取的常用数据是平均数、近似值、时间序列、分布规律等,这些工作使用 Excel 之类的电子表格软件就可以实现,不用学习复杂的数据库软件。下面结合个人的使用体会,谈谈对于个人搜集的数字资料的二次归类整理和从海量古籍信息中获得知识发现的问题。由于无成例可循,个人的探索难免存在偏颇和不完善之处,笔者发表拙见,意在抛砖引玉,不当之

处,敬请方家赐教。

首先以秦汉时期自然灾害数据的统计分析为例,讨论数据性资料的整理。自然灾害的历史比人类更悠久,从不同角度、按不同灾种对秦汉时期的自然灾害进行研究,近年来已有不少成果面世,为我们的进一步研究奠定了良好的基础。笔者本想直接采用已有的秦汉自然灾害统计数据,但是,将搜集到的统计数据比较之后发现,大概是确定检索文献的范围和对史料解读的不同,研究者得出的统计结果差别较大。例如:杨振红统计,汉代共有 242 个年份发生了灾害,总计发生各种自然灾害 420 年次。^① 黄今言、温乐平统计,汉代自然灾害共 346 次,其中水灾 71 次,旱灾 48 次,蝗灾 42 次,地震 77 次,疫灾 18 次,风灾 21 次,淫雨霖雨 15 次,冰雹 20 次,霜雪 11 次,饥荒 23 次(如除去饥荒,则为 323 次)。^② 李辉统计,汉代有 292 个年份发生自然灾害,其中水灾 121 次,旱灾 106 次,地震 104 次,虫灾 62 次,疫灾 49 次,风灾 33 次,雹灾 35 次,低温类灾害 31 次,山崩地裂 39 次,共计 580 次主要自然灾害。^③ 陈业新统计,在两汉 420 余年中共发生了 529 次灾害,其中水灾 105 次、旱灾 111 次、地震 115 次、蝗灾 64 次、疾疫 42 次、风灾 37 次、雹灾 38 次、雪灾 16 次、霜灾 7 次、冻灾 14 次。^④

笔者一向尊重他人的研究成果,并将其作为个人进一步研究的基础和借鉴,但并不盲从。可能是受论著体例的限制,上述统计数据大多数只有总的灾害统计结果而无具体的分类灾害统计列表。根据同样的文献资料却得出不同的统计数据,而我们并不知道差异出在哪里。只看总的灾害统计结果,我们根本不可能弄清楚彼此的统计差别在哪里,疏漏的数据是什么?为什么会出现这种差异?是取舍的标准不同呢?还是无意的疏漏?由此自然便会产生“我该相信谁”的疑问。陈业新《灾害与两汉社会研究》一书的附录“两汉灾害年表”,省去了引用时的翻检之劳,是

① 杨振红:《汉代自然灾害初探》,《中国史研究》1999 年第 4 期。

② 黄今言、温乐平:《汉代自然灾害与政府的赈灾行迹年表》,《农业考古》2000 年第 3 期。

③ 李辉:《试论两汉时期自然灾害的主要特点》,《社会科学战线》2004 年第 4 期。如除去山崩地裂,则为 541 次。

④ 陈业新:《两汉时期灾害发生的社会原因》,《社会科学辑刊》2002 年第 2 期。

汉代自然灾害研究的新成果;不过,由于没有自然灾害分类表,当我们想了解各类自然灾害的详情时,仍然不得不重新检索核对,使用还是不太方便。统计数据的不准确虽然不会太多地影响我们对自然灾害本质的认识,但建构在数据统计基础上的量化分析结果的可信度就会打折扣了,对局部问题的分析则不可避免地会出偏差。有感于此,笔者没有直接引用已有的秦汉时期自然灾害的统计数据,而是不避烦难,逐一核对史料,认真考定,得出和已有成果不同的统计数据,整理出详细的“秦汉自然灾害分类表”和“秦汉自然灾害年表”。^① 这样的工作可以更清晰地反映出文献记载的实际,当然也更容易暴露研究中存在的问题。笔者的工作是在已有研究基础上的新进展,缺失可能仍然存在,因为这样的统计资料难以掩饰错漏,同样也便于有针对性地修订、补充、完善。

袁林主持开发的“汉籍全文检索系统”提供了三种选择文献的方式:四部(经、史、子、集)序、时代序和拼音序,我选用时代序“秦汉”部分,检索文献 61 种。选择若干个关键词检索,每个关键词的检索结果输出为一个文本文件。

使用检索软件一次性输出功能得到的秦汉时期自然灾害资料,可以分为两类:史料及其出处。如何对资料做进一步分类呢?仅仅使用文字处理软件显然是不行的。笔者在做灾害资料统计分析时,利用 word 和 Excel 协同工作。Word2003 的 office 剪贴板可以剪贴 24 项,将史料和出处分别剪切,一次剪切 12 条史料。(也可以使用剪贴工具来做)然后切换到 Excel,设置好“灾情和史料出处”两列表头,从“Office 剪贴板”逐项粘贴。粘贴完以后,再到 Word 中剪切,再切换至 Excel 粘贴。循环往复,直至处理完一个 Word 文档。只有灾情和史料出处两类信息,我们仍然无法排序归类统计,必须增加新的分类。后续工作都在 Excel 中进行,我添加了灾害发生的时间(“帝王纪年”、“公元纪年”、“季节”)、“灾害发生地点”、“赈灾措施”等分类,相关文字从“灾情”列中析出,缺少信息再检索补充。分类信息补充完成以后,就可以按照研究的需要进行不同的排序统计了。我用这种方法制作了 40 多个表,分析统计秦汉时期自然灾害的时空分布特点、灾害的频度、平均值、赈灾措施,等等。将分类灾害表汇总,按公元纪年排序,得到秦汉时期

自然灾害年表。

古籍中的数据性资料虽然不少,但更多的是文字资料。从学术研究的角度看,古籍中既有古代先贤完整表述思想体系的“撰述”,也有保存古代历史断片的“记注”。研究先贤的思想,当然要尊重其“撰述”的完整性及其内部的逻辑,在其时代语境之中作“同情之了解”;而每当我们把零星的断片(即史料)按照一定规则重新排列、组合以后,都会有一种豁然开朗的感觉,因为我们从中发现了那些资料在原有脉络之中难以解读出的字面之外的第二甚至第三重含义,以及它们之间的各种内在关联,我们对这些含义和关联作进一步的分析或综合,往往会有新的发现和解读,这就是史学研究的一般过程。这一过程在手工查阅纸本文献的时代,需要学者具有深湛的功力,否则很难得到完美的解决,因为纸本古籍大多缺少必要的索引,而且纸本检索工具不能按照读者的要求提供多种排检方式,可用性有限;此外,研究者对文献本身的认识是随着研究工作的深入而逐步清晰起来的,在工作初期往往难以明确提出与自己的研究题目完全切合的全部关键词,而是要在较大范围内进行模糊查询或渐进式查询,这更是纸本检索工具书所不能解决的。上述困难在信息时代变得容易多了,检索关键词可以不断地调整尝试,随心所欲,只要能想到,文献中又存在,就能检索到,成百上千条资料在一瞬间就可以搜集在一起,输出为一个文件。然后再使用与处理数据性资料相同的方法,Word 和 Excel 联合作战,对文字性资料进行逻辑分类,以便从中获得知识发现。下面谈两个具体的应用实例。

笔者以现代人习用的西方政治学所界定的“专制”为题,考察中国古代专制概念的含义和应用语境的变化。^② 对这个问题的研究以文渊阁《四库全书》(电子本)为史料范围,《四库全书》收录典籍 3460 多种。检索结果,在《四库全书》中“专制”出现 1800 多次,“颡制”75 次。由于《四库全书》的检索结果只能一条条地复制粘贴,为减少复制粘贴之劳,将检索搜集资料的工作分作两步进行。“汉籍全文检索系统”中有的文献在该

① 王文涛:《秦汉社会保障研究——以灾害救助为中心的考察》,中华书局 2007 年版,第 312~385 页。

② 王文涛:《中国古代专制概念述考》,《思与言》(台湾)2006 年第 4 期。

系统中检索输出;“汉籍全文检索系统”中没有的文献在《四库全书》中检索复制,用剪贴工具自动粘贴。对“专制”概念的分类、归纳、排比、筛选和逻辑分析等工作,主要使用 Excel 的排序和筛选等功能及其它文本软件完成。

综合分析了一千多条史料之后,我获得了一些解读少量史料难以得到的学术界无人提出的知识发现。例如,在《四库全书》中,“专制”用于君主并有确指对象的史料仅三见^①，“君主”和“专制”连用的情况只有一例。^②王莽、孙权、司马炎和武则天,都是由人臣变为“九五之尊”的,他们在称帝前均曾被指责有“专制”行为,而称帝之后的所有言行,不论是当时还是后世,在近代以前均无人再以“专制”称之。中国古代的“专制”一词广泛用于人臣,就身份而言,有后妃、外戚、宦官、佞臣、权臣、诸侯、藩镇、悍将,等等。从先秦至晚清,“专制”的应用语境基本没有变化,就是说,“专制”含义的稳定性,与封建专制制度的长期性、稳定性是相一致的。“专制”有四个义项:独断专行、越权自作主张、控制掌管和君主独掌政权。现有辞书收了三个,“越权自作主张”是笔者的观点。在鸦片战争前的中国社会,前三个义项基本上不用于君主,广泛用于人臣,人臣专制的实质是专制王权的变异和向专制王权的回归。“独断”与“独揽”是描述中国古代君主专制现象的词语,权力独揽、决事独断是专制帝王的权力特征,也是中国古代社会的普遍观念。作为国家政体或社会制度的“专制”在西方列强打开中国大门以后才产生,“君主”与“专制”也越来越多地联系在一起。“专制”含义的丰富和应用语境的变化,反映了中国政治思想在近代的剧烈变动。

上述工作如果在文字处理软件中做,费时费力而且容易出错。据笔者了解,目前,还有研究者采用这样的方法处理检索得到的资料数字,对每一条资料加注分类信息,然后打印出来,用剪刀逐条剪开,再按照分类信息归类。这种方法不仅速度慢,而且不便进行再次分类,希望笔者的做法能够给他们提供帮助或借鉴。

另一个应用实例是“考察赈济类词语的变化与汉代社会救济的关系”。研究词语的含义及其演变,是语言学分支学科语义学的范畴,将语义学研究引入历史学研究,也是历史学多学科研究的需要。语言随着人类的产生而产生,伴随社会

的发展而发展,这是语言发展的一般性规律。汉代文献多于先秦,同类词语自然应该多于先秦,这无疑是正确的符合语言发展规律的认识。但是,如果停留在这种一般性认识上显然是不够的,我们需要具体而明晰地了解历史时期语言的发展情况及其与社会发展的关系,深入准确地认识我们的研究对象。通过考察赈济类词语的数量和语义的变化来认识汉代社会救济的发展,就是一种探索性的尝试。透过词语的增减变动、产生消亡来认识社会的进展和人类思想语汇的丰富,是信息时代提出的一种更加精致的研究历史的方法,也是历史学多学科研究的方法之一。传统的史学研究手段对于分析处理海量信息的研究方法基本上是望而却步,不敢问津,日益丰富的全文检索古籍数字资料库和计算机处理海量信息的强大功能,为这种研究方法提供了科学的便捷高效的研究手段。从海量信息的统计分析中获取知识发现,是计算机之所长,也是传统的史学研究不曾做或极少做过的工作。拜信息技术发展之所赐,我们有幸能够运用现代化的利器去研究古老的历史学科,用先哲前贤没有用过的方法从一个新的角度去研究我们的历史。人们常说,信息技术的发展正在日益深入地影响和改变着我们的生活和工作,笔者也经常在思考信息化和史学研究的关系,以及如何从理论上把握现代信息技术对传统学科历史学的影响等问题。这是一个很大的新课题,笔者所做的尝试就是利用计算机处理海量信息从而获得知识发现的个人体验,研究中的诸多知识发现都是建立在分析处理大量史料(几乎穷尽同时期的资料)的基础上,和传统的考据方法类似而不相同,具体方法参见前文所述。例如,在统计了大量资料之后发现,两汉文献中“赈济”类词语从先秦的5个发展到23个,增加了18个;作“救济”解的“赈”在汉代已与“振”通用,由“赈”组成的合成词先秦未见,始于汉代;“振恤”和“振除”均不见于汉代文

① 检索范围为文渊阁《四库全书》电子版,香港迪志文化公司制作,上海人民出版社,1999年版。“专制”最早见于《国语·楚语》,春秋楚国大夫子张借称赞商王武丁劝楚灵王虚心纳谏,第二条见于《周书·宇文孝伯传》。第三条见于《宋史·隐逸传中》。另有几条为泛指。

② 严遵《道德指归论》卷六:“君主专制,臣主定名。君臣隔塞,万物自明,故人君有分,群臣有职,审分明职,不可相代。”

献,“振恤”已为“赈恤”所取代;现有辞书中的“赈施”和“赈粟”书证晚出;汉代已经出现的“振贫”、“赈贫”和“振给”辞书未收;“稟”在汉代有“受谷”义,而《汉语大字典》和《汉语大词典》未收录;《辞源》和《汉语大词典》将“赈贷”解作“救济”并不十分准确,包括有偿和无偿两种“赈济”;等等。

四

越来越多的史学研究者对应用计算机的认识已经从简单的文字处理发展到文献资料的检索查询,这无疑是一种很大的进步。从提高研究效率和质量来说,满足于这种进步是不够的,我们应该努力提高利用计算机综合分析处理文字信息的水平,丰富研究手段,从检索查询的一般性应用向归类分析、综合统计的高级应用发展,增强从海量信息中发现知识的能力,积极主动地参与信息化与史学关系的方法和理论研究,深入挖掘蕴含在中国汗牛充栋的文献典籍中的知识宝藏。

运用古籍数字资料必须要重视这样一个问题,就是摆正系统读书和按需搜集资料之间的关系。要论述某一个问题,即使将资料都搜集齐全了,并经过排比和取舍,最后得出了结论,这个结论也不能说准确无误。因为历史研究是一项系统而全面的学术活动,它不仅要有点,还要有面,有整体。如果我们不是从全局的角度去理解某一点,把视野仅仅局限于这一点上,就可能犯“一孔之见”的错误。外国人研究中国历史在搜集史料上用力甚勤,收获很大,但在对许多问题的论述上,往往达不到“鞭辟入里”的精深境界。为什

么呢?因为他们对中国历史缺少建立在大量感性知识上的理解。

我们现在研究古代历史,也存在着这个问题。所以,我们应当对古代文献中最基本的史料有比较正确、全面的理解,在此基础上搜集资料,去发现问题和研究问题。这样,我们可能和别人搜集的资料一样,但因为有了整体的历史观念,在论述问题时,我们所站的高度不同,得出的结论就有差别了。目前,学术界仍然存在着先有观念、再找资料的错误倾向,这是应该杜绝的。有些人以检索代替研究,不核原文,不审背景,错谬频出。古籍数字资料全文检索库,是文史研究的丰富学术宝藏,入宝山而不取固然可惜;但要加工提炼,如果胡掘乱采,不仅浪费学术资源,而且破坏学术环境。浮躁和功利化倾向对文史研究质朴、严谨的学风是一种伤害,应当也必须进行纠正。纠正的方法应当是批评与引导相结合,正确的态度是,掌握先进的电脑网络知识,充分利用和享受信息技术和古籍全文检索资料为我们带来的方便快捷,把节省下来的大量时间用在资料的考订、分析和历史问题的思考上;扎实的学术功底非常重要,只有博闻强记,并加以融会贯通,才能提高我们的研究质量和效率,否则再先进的检索系统也只能是无的放矢。

收稿日期 2007—12—20

作者王文涛,历史学博士,河北师范大学历史文化学院教授。

【责任编辑 殷 铭】