

·信息工作·

面向海量文献的数字化系统研究

苏 云 张庆来 (兰州大学管理学院 甘肃兰州 730000)

摘 要 文章针对海量的文献资料如何快速录入计算机的方式方法提出了解决方案,首先通过对键盘录入、手写录入、听写录入和扫描录入四种文献数据采集方法的比较,提出了扫描录入是海量信息处理的唯一选择;其次,对扫描录入技术进行了历史回顾和现状分析;最后,提出了OCR数字化处理工厂的一揽子解决方案,即通过文字自动录入、流水线管理、质量控制和员工管理、系统管理四大功能实现海量文献的数字化。

关键词: OCR 技术 扫描录入 文献识别 文献数字化

中图分类号: G203

文献标识码: A

文章编号: 1008-6938(2010)02-085-05

Solutions for Mass Literature Digitization

Su Yun Zhang QingLai (School Of Management, Lanzhou University, Lanzhou, Gansu, 730000)

Abstract: In this paper, a solution is proposed to input mass literature quickly into computer. First of all, through the comparison among the four kinds of data collection methods----keyboard entry, by-hand input, voice dictation, and scan input---scan input is found to be the only choice of mass information processing. Second, it is a review of the OCR technology and analysis of current situation. Finally a solution of OCR digital processing plant is put forward, that is to say, mass literature digitization can be perfected by text automatic input, pipeline management, quality control, personnel management, and system management.

Keywords: OCR technology; scan input; literature identification; literature digitization

CLC number: G203

Document code: A

Article ID: 1008-6938(2010)02-085-05

1 引言

五千年的中国文化遗留下极其丰富且数量庞大的历史文献,这些文献主要保存形式以甲骨、简牍和纸张作为载体,通过编纂引得、通检、索引和汇编等工具书达成文献整理和查询的目标,由于文献数量巨大和人力有限的矛盾,经过系统整理和方便的检索工具非常稀缺,加之受存储空间的限制,许多年代久远的孤本书、善本书已出现了纸张脆弱、字迹变色、书页脱落和破损发霉等现象,很多出土的甲骨、简牍和纸张也出现了腐蚀和霉烂的状况,严重影响了文献的使用和保存寿命,文献的数字化迫切性已成为信息工作者的当务之急,图书馆和档案馆应该积极顺应网络时代的潮流,运用计算机相关的数字化技术,对文献进行加工和处理,建立书目数据库、全文数据库和综合检索系统,并通过光盘和网络等途径进行信息的传播。本文针对该问题提出了面向海量文献信息数字化的处理解决方案,尤其对文字的批量识别提出了系统化的解决途径。

2 海量文献数字化处理的现状

2.1 传统的海量文献数字化技术及比较

如何将海量的文献资料快速录入计算机是文献数字化研究工作的重要内容,而文献数字化的瓶颈就在于如何将海量的文献录入计算机的方式方法,就传统处理技术而言,数据的录入方法有键盘录入、手写录入、听写录入和扫描录入。

(1)键盘录入法。键盘录入法有阴阳码输入法、郑码输入法、形象码输入法、汉码系列输入法、智能二笔输入法、双笔码输入法、汉正码输入法等,总共不下几十种,最常用的是各式各样的五笔字型 and 拼音输入,其中五笔输入法常用的是王码五笔、陈桥五笔、念青五笔和极点五笔等,任何一种五笔输入法只要掌握文字的拆分规则就能使用,拼音输入法常用的有智能ABC、拼音加加、紫光拼音、搜狗拼音、中文之星智能狂拼、三好拼音、极点拼音、

基金项目: 本文系甘肃省科技厅资助项目“信息管理词霸电子词典的研究”(编号: 0804GKCA045)研究成果之一。

收稿日期: 2009-09-26,责任编辑: 魏志鹏

五万拼音、递推联想拼音等,只要会拼音就会输入,这两者录入速度不分伯仲,关键在操作人员的熟练程度。国际专业录入师的打字速度是在 240 字/分钟左右,一般打字员的速度是 50~70 字/分钟,这种录入速度相对海量的文献资料是一种效率极低的信息数字化处理方式,不但费时费力,而且资金耗费巨大,会造成大量文献资料的积压。

(2)手写输入法。手写输入法亦称为手写笔输入法,主要有台湾的蒙恬系列手写笔、大恒笔才子手写笔、汉王大将军手写笔和紫光绘写大师等,手写笔是由硬件和软件两部分构成,硬件部分包括电子手写笔和写字板,软件部分是汉字识别系统。手写输入法的使用比较简单,录入员只需用手写笔在写字板上书写笔划清晰的汉字,写字板中内置的高精密的电子信号采集系统,就会将汉字笔迹的信息转换为数字信息,然后传送给软件系统进行汉字识别。汉字识别系统的作用是将硬件部分传送来的信息与事先储存好的大量汉字特征信息相比较,从而判断写的是什么汉字,并通过汉字系统在计算机屏幕上显示出来,手写输入系统的难点在于汉字笔迹的识别,因为每个人的手写字体不一样,所以汉字笔迹比较系统就必须能允许一定的模糊偏差,才能做到较高的识别率,但是手写笔的最快录入速度仅有 20~40 字/分钟,显然不适合海量文献信息的录入,但对录入手绘图形图像十分有效。

(3)语音输入法。语音录入就是听写输入法,较之键盘和手写输入,既快速又方便,速度是键盘输入的 2~3 倍,是手写输入的 6~8 倍。硬件设备只需带声卡的多媒体计算机和无噪音的麦克风。从技术层面而言,实现语音识别就是让计算机识别和理解人类语言的过程,是把自然语音信号转变为相应的文本。在语音识别过程中,首先要将人类说话的声音由模拟的语音信号转换为数字信号,然后从信号中提取语音特征,同时进行数据压缩,输入的模拟语音信号要进行预处理,建立识别基本单元的声学模型和进行文法分析的语言模型,计算机根据识别系统的类型选择能够满足要求的识别方法,采用语音分析方法分析出这种识别方法所要求的语音特征参数,按照一定的准则和测度与参考模式库中的模型进行比较从而得出识别结果。语音输入法已经出现十余年,但由于计算机处理速度的限制,并没有形成气候,后来随着计算机 CPU 主频的不断提高,出现了许多以 IBM 语音录入为内核的软件,例如 Windows Vista 就内置了语音录入软件,但要快速高效能满足海量文字录入的要求,就必须使用专业的语音输入软件,例如 IBM 公司的 ViaVoice 语音识别专业软件、Scansoft 公司的 Dragon Naturally Speaking Preferred

语音识别软件等。这里需要强调的是语音录入法对录入员的标准普通话水平的要求很高,由于中国是方言非常丰富的国家,这种录入法很难推广,而且语音录入时周边环境不能有噪音,即对环境要求过于苛刻。

(4)扫描录入法。键盘和手写录入面对的是漫长和繁重的工作,听写录入受到方言和周边环境噪音的影响,就现有技术而言,海量文献录入的唯一的選擇就是扫描录入法,速度可以达到每分钟 6000 字,具有其它录入方法不可比拟的优势。扫描录入的英文缩写是 OCR(Optical Character Recognition),就是让计算机认字和实现文字自动输入。它的工作原理是通过扫描仪或数码相机等光学输入设备获取文献纸张上的文字图片信息,利用各种模式识别算法分析文字形态特征,判断出汉字的标准编码,并按通用格式存储在文本文件中,是一种快捷、省力和高效的文字输入方法。具有以下三大优势:时间上,扫描录入法速度快和效率高,是人工录入的数百倍,甚至数千倍;经济上,扫描录入法节省了大量人力资源开销和降低了录入成本;准确性上,扫描录入法的录入准确率远高于其它人工录入法。

文献扫描录入的必备的硬件设备是扫描仪,主要有平板扫描仪、多功能一体机、高速扫描仪和网络扫描仪等,常用品牌有全友(Microtek)、爱克发(AGFA)、清华紫光(Uniscan)和惠普(HP),常用幅面是 A4、A4 加长、A3 等三种,如果扫描报纸、地图等,就需要 A1、A2 幅面的平板扫描仪,数据海量处理最常用的是高速滚筒式扫描仪,而高速扫描技术是依赖计算机 CPU 的性能来提高识别率和识别速度。最为常用的中文识别软件有清华紫光文通信息技术有限公司开发的 TH-OCR(TH 是 TsingHua 的缩写)、北京汉王科技股份有限公司研制的汉王文本王;其它优秀的识别软件还有:以我国战国时期“掌章奏文”官职命名的尚书 OCR 软件、以我国古代绘画颜色命名的丹青 OCR 软件、发明毛笔的古代大将命名的蒙恬 OCR 软件等等;^[1]外文识别软件的普遍功能要比中文识别软件要强大,尤其对书籍、报刊的版面还原技术要强大得多。常用的有俄罗斯软件公司开发的 ABBYY FineReader Professional、美国 IGS 公司研制的 ReadIRIS Pro,此外还有能够识别 114 种语言文字的 Recognita 软件、能够将识别文字发音朗读校对(Text-to-Speech)的 OmniPage 软件、发明复印机的施乐公司推出的复印和识别一体化的 XEROX TextBridge 软件。

2.2 基于 OCR 扫描和识别的海量文献数字化处理现状与分析

让机器代替人认字并记录,是人类很久以来的梦想。

早在20世纪20年代,西方就开始了字符自动识别的研究。有文献可考的最早机器字符识别系统是德国的科学家陶杰克(Tauscheck)的“阅读机”,1929年这项发明获得了德国专利;几年后,美国科学家汉德尔(P·W·Handel),也提出了利用技术对文字进行识别的想法,研制了“统计机”的类似机器,也获得了美国专利。自此之后,人类经过几十年的不断努力,使得OCR技术渐渐成熟,从最初的机械识别模式一直发展到今天利用抽取图像的数字化特征进行识别的电子模式。

相比英文OCR识别,汉字的识别要困难许多,这是由于英文是由几十个字母符号组成的文字,而“方块符号”的汉字字库要比英文字母表庞大近千倍,难度可想而知。我国在上个世纪70年代末就开始了这项技术的研究,至80年代中期,可识别上万汉字,识别率在90%左右,尤其是1987年《汉字识别的特征点方法》的问世是一个里程碑,这种方法是“以汉字字形结构的统计特征划分为汉字笔划上的特征点和背景处的关键背景点,并基于这个理论,推出了‘印刷体汉字文本识别系统’”,这个系统的研制成功标志着我国在印刷体汉字的识别技术研究方面已取得了实用化的突破,进入90年代之后,随着863项目在內的汉字识别系统逐渐成熟,不少研究单位相继推出了中文OCR产品,主要有清华文通(TH-OCR)、北信(BI-OCR)、中自(ICR)、沈阳自动化所(SY-OCR)、北京曙光公司(NI-OCR)等,这些系统均可以实现中英文混排、宋体、楷体、黑体、仿宋体、繁体等多字体、多字号的混排识别,文字识别率可达到95%以上。特别是21世纪的近十年,OCR识别技术随着扫描仪的普及得到了飞速的发展,扫描和识别软件的性能不断强大并向智能化升级发展。^[2]

一般说来,传统的OCR扫描和识别软件主要功能是通过以下六大过程来实现,即影像获取、影像前处理、文字特征抽取、比对识别、人工校正和结果输出。其中①影像输入就是将需要OCR处理的文献资料通过光学仪器(扫描仪、数码相机等)录入计算机;②影像前处理是OCR系统中,须解决问题最多的阶段,从得到一个不是黑就是白的二值化影像,或灰阶、彩色的影像,到独立出一个个的文字影像单元的过程,都属于影像前处理,这其中包含了影像正规化、去除噪声、影像矫正等的影像处理,及图文分析、文字行与字分离的文件前处理;③文字特征抽取可以说是OCR系统的核心,用什么特征、怎么抽取,直接影响识别质量的好坏;④比对识别是指当文字特征抽取结束后,不管是用统计或结构的特征,都须有一比对数据库或特征数据库来进行比对,数据库的内容包含预先对所有欲识别的文字的集合中元素采用文字影像单元一样

的特征抽取方法抽取特征所得的特征。通过比对,从而确定文字影像单元所对应的文字。由于OCR的识别率不可能达到百分之百,为了提高识别的准确度,字词后处理过程就必不可少,它利用比对后产生的识别文字与其可能的相似候选字群,根据上下文的识别文字找出最合乎词义的词,对识别结果进行更正,例如识别出“找们”,在词库中找不到这个词,而“我”是“找”的相似候选字,因此很自然的将“我”取代“找”,而成“我们”;⑤人工校正是保证OCR质量的最后阶段,也是最有效、最直接的阶段,在这个阶段要求录入人员花费精力和时间,去直接更正甚至寻找可能是OCR出错的地方。一个好的OCR软件,除了有一个稳定的影像处理及识别核心,以降低错误率外,合理、有效和便捷的人工校正的操作流程及其功能,也很大程度影响着OCR的处理效率和准确性;⑥结果输出就是将OCR产生的结果将按照要求提交给用户。^{[3][4]}

基于以上的过程,这种传统的处理方式一般采用一台告诉扫描仪和多台计算机相连接,把扫描的文献资料分派到不同空闲的计算机上进行识别处理,再将识别结果返回整理。这是一种串行的工作方式,虽然在一定程度上提高了扫描和识别效率,但扫描和识别协调同步很难实现,而且辅助工作量极大。

3 面向海量文献的数字化处理系统设计与分析

为了满足书籍、报纸期刊、报表票据、历史档案等文字录入的需求,也为了满足资源性网站和数据库开发对数据的需求,针对银行、税务、工商、医院等行业尤其是图书馆、档案馆对文字识别的需求,本文提出了OCR数字化处理工厂的一揽子解决方案。

3.1 系统的总体设计

本文提出的数字化处理工厂系统是应用OCR技术、实现工业化流水线管理方式的大型Internet系统设计。该系统设计通过强大的网络功能实现流水线方式的数据加工,并通过网络供千千万万个用户享用。实现数字化处理工厂系统硬件需要:一台小型服务器作为数据服务器和主域控制器,管理多台终端;高速扫描仪和微软的操作系统,大容量硬盘或磁盘阵列的存储设备(视加工规模选用);磁带库或光盘库的备份设备(选用)。实现四大功能,即文字自动录入、流水线管理、质量控制和员工管理、系统管理(见图1)。

3.2 系统功能与模块介绍

整个系统围绕两个互相联系的员工管理和OCR扫描文件数据库展开工作。员工管理数据库由员工信息表、工种信息表、员工考勤表、员工工作量表、班次表、工资管理

表等构成一个完整的员工资料库。员工依流程指定的步骤登录、考勤、申请工作、执行操作并接受管理监督。管理人员通过简明友好的系统管理界面可以方便地查询数据、备份数据和系统维护。该系统还提供安全日志供管理人员查询。OCR 录入资料数据库经由扫描录入、图像处理、版面分析、识别、纵校、横校、版面还原等工序处理最终形成。其中在信息传输上采取申请任务方式与分配任务方式相结合使用。其中申请任务方式是用户完成一件工作包的同时查看是否有已经分配的工作包,如没有,则申请另一个工作包;分配任务方式是由管理员分配工作包给每一个员工,为对此流程进行有效管理,建立了原始工作包表、工序跟踪表、返工单表、员工工作分配表、工作包表等。

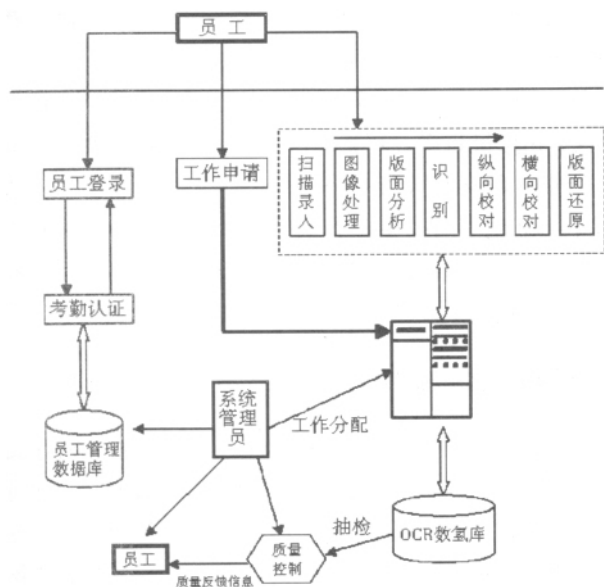


图1 汉字处理工厂系统逻辑模型图

(1)文字自动录入功能模块。采用在OCR领域领先的成熟文字自动录入技术,实现包括横版、竖版、简体、繁体各种版式的古籍、报刊杂志、公文档案、报表或票据和现代书籍的自动图像预处理、版面分析,能识别中文简体、繁体、英文及混排和多字体多字号文档。

(2)流水线管理功能模块。采用生产流水线管理方式,根据OCR技术和操作的特点,将生产过程划分成以下几道工序:①文献资料整理:为了便于扫描和以后的查询、检索而进行的文献分类、拆装、命名、编号等。②扫描:扫描是将纸质文献图像输入计算机的过程。一般把相关按文献页码顺序扫描,在扫描质量控制程序自动检测并修正后,自动保存到数据库中。③图像处理:为了提高识别率,对图像进行“消蓝去污”的处理,即去掉图像上影响识别率的噪音如麻点、下划线等,图像质量控制程序自动

监测图像处理质量。④版面分析:能自动进行版面理解并定位,判别划框区域是横排文本区、竖排文本区、表格区还是图像区,并对不同属性的区域以不同颜色的线框标识出来。自动版面分析在后台运行,操作人员可在前台进行确认,并对自动版面分析结果加入手工干预。⑤识别:把文字图像转化为计算机文字内码,可以识别印刷体和手写体中文(包括简体字和繁体字)、表格、中英文混排,识别出来的文字内码可以是GB码、BIG5码、GBK码或者Unicode码。识别过程在后台运行。⑥纵向校对:具有很强的查错、纠错能力。纵向校对是将一个图像或若干个图像中识别成同一个字的文字图像列在一起显示,并以突出颜色标出可疑字,便于操作人员发现错误和修改。⑦横向校对:这是传统的人工校对方法,操作人员直接对比识别结果文本和图像,以发现识别错误文字。系统自动调出文字对应的图像,进行比对。同时,以醒目的颜色标出识别可信度不高的文字。⑧版面还原:将识别并修改好的文本还原成跟扫描文稿版面的布局一样,可供计算机阅读和查询检索的RTF、PDF、HTML、SGML/XML格式的数字文档。⑨数据入库:版面还原数字文档的保存。

(3)质量控制和员工管理功能模块。质量控制是为了保证和控制系统的录入质量而采取的一整套方法与措施。主要是在各工序中加入对员工工作完成情况及差错量的监控和工作量的计算,以求将整体差错量控制在万分之二以内。员工的工作态度将会直接影响到数据录入的质量和工作效率,要使员工保持一种积极的工作态度,必须有好的管理制度和客观的评价标准和依据。该系统可以详尽地提供员工考勤情况和工作质量数据,并对员工的工作情况给予公正的评估。员工管理系统在整个系统中处于支配地位。该部分由考勤管理、工资管理、质量控制、工作分配、返工单管理和建立用户等几个模块组成。其中考勤管理是记录各员工的出勤、缺勤状况,岗位管理是记录各岗位的工作分配和员工的工作量、差错量(质、数量的差错要求控制在万分之五以内)的状况;工资管理是根据员工的考勤、工作量和差错量的情况,发放员工的工资并列明明细账目表。

3.3 系统功能优势与创新分析

本文提出了一个基于大型的Intranet网络系统实现系统框架,可将汗牛充栋文献进行数字化录入识别处理,是一个包含成千上万的加工数据资料和员工详尽的工作信息数据库系统。这样的创新,将单独的扫描识别通过整合方式组成了完善的数据加工生产工厂。

(1)采用生产流水线管理方式,改进了传统的串行的工作方的效率低下,将冗长、复杂的数据生产过程合理地

划分成若干道工序,每道工序操作简便,合理安排工作岗位,并行操作,生产效率和质量得到了3~4倍的提高。并且可以任意确定工艺流程操作顺序和组合,适于不同种类和不同要求的数据资源加工,实现了单机资源数字化过程和机群间高效率的相互协同作业。

(2)采用分布式操作,管理员可通过计算机网络实现对系统的远程管理,大大增加了管理员对数字化加工系统进行管理的灵活性。加之服务器对客户端的消息响应采用队列式管理,服务器运行会更加稳定和可靠。

(3)文字自动录入可实现批量扫描和识别,在不点击鼠标的前提下,实现数据自动命名、自动存盘、自动识别和自动校对,并将处理文件自动纠偏、去噪、OCR和压缩存储,极大的节省了人力资源。

(4)人工操作与后台自动运行相结合,把一些可由计算机自行处理的工序设置为后台自动运行,从而减少了人为造成的错误。

(5)数据质量得到了大幅度的提高,实现了数据检查、监督和协调的自动化,完善了系统权限管理和数据安全,员工工作效率得到了公正的统计和评估。

4 结语

概而言之,本系统的设计为数字图书馆、档案馆、政府机关等不同机构的大量文字、图表的自动录入提供了一种切实可行的处理方案,更适应网络时代建设网站过程中对文字和图像的需求,具有巨大的社会效益和经济效益。

参考文献:

- [1]张焱中.汉字识别技术[M].北京:清华大学出版社,1992.
- [2]任永芳.中文OCR与图书资料的再制作[J].高校图书馆工作,2001,(3).
- [3]迟春佳.OCR技术及其在高校图书馆信息资源数字化建设中的应用[J].中国科技信息,2007,(7).
- [4]王桂敏,齐凤河.OCR软件使用经验浅谈[J].科技信息,2006,(5).

作者简介:苏云,男,兰州大学管理学院副教授;张庆来,男,兰州大学管理学院讲师。

·简讯·

沙勇忠教授新著《信息分析》出版

信息分析是针对特定的需求,对信息进行深度分析和加工,提供有用的知识和情报。本书用新的视野、新的框架和新的发现,系统阐述信息分析的相关理论、技术方法和最佳实践,包括信息分析理论、信息分析工作框架、信息分析建模、信息分析方法、计算机辅助信息分析、科技信息分析、经济信息分析、社会信息分析、信息分析项目与机构管理等内容。除注重基础知识和新思想、新观念、新方法的介绍外,突出信息分析中解决问题的方法思路和方法应用,注重反映信息分析实践领域的最新进展和国际著名信息分析机构的经典工作,使用大量案例来阐述问题,为从信息分析角度提升读者的竞争素质和基础能力提供指南和帮助。全书共有图203个,表164个,案例27个,为方便读者扩展阅读,还以“相关链接”的形式提供了一些资料。

本书可作为高校信息管理与信息系统、工商管理、图书情报档案、新闻传播、社会学以及经济学等专业的教材和教学参考书,也可以供企业市场研究部门、政府政策研究部门、信息产业部门相关从业人员使用参考。

(沙勇忠,牛春华编著.信息分析.北京:科学出版社,2009)