广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(4):277～278
Vol. 23,No. 4 November 2007

一种通用的印刷体档案条目信息采集方法
A Universal Method for the Printed File Entry Information Collection

罗 维,卿 旭,龙 波,曾锻成
LUO Wei,QING Xu,LONG Bo,ZENG Duan-cheng

(广西计算中心,广西南宁 530022)
(Guangxi Computing Center,Nanning,Guangxi,530022,China)

摘要:介绍利用OCR光学字符识别的技术研发的一种通用的印刷体档案条目信息采集方法。该方法可以自动录入已经是印刷体的条目,最终信息从tif文件采集,可以利用各种OCR系统或人工智能方式进行识别获取,具有一定的适应性。

关键词:自动录入 档案条目 印刷体
中图法分类号:TP317 **文献标识码**:A **文章编号**:1002-7378(2007)04-0277-02

Abstract: This paper introduces a common method for the printed file entry information collection by means of OCR optical character recognition technology. The method can automatically input the printed entries, the ultimate information which is available and recognized through different kinds of OCR systems or artificial intelligence.

Key words: automatic input, printing fonts, archives entries

档案数字化需要对大量的纸质历史资料进行文本采集,录入相关数据库中进行规范管理。为了减轻加工人员手工输入的劳动强度,尽可能利用计算机智能录入,我们利用OCR光学字符识别的技术研发了一种通用的印刷体档案条目信息采集方法。

1 档案条目结构分析

一般一盒档案需要采集的信息有档案盒上的案卷目录和档案盒内的卷内目录。案卷目录包括档号、案卷标题、编制单位、编制日期、保管期限、密级等,卷内目录包括保管期限、件号、责任者、文号、题名、日期、页数等。案卷目录的条目顺序相对比较固定的,卷内目录是规则的表格。

2 流程设计

OCR采集的第一步必须先进行扫描,图像预处理,其次是OCR识别与人工纠正和条目入库,详见图1。

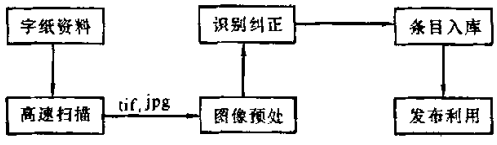

图1 档案资料的文本采集流程

2.1 扫描

扫描参数采用黑白二值扫描,扫描分辨率200DPI,存储为TiF G4格式就能达到速度与正确率的平衡点。如果扫描存储为JPG彩色格式文件,将需要进行二值化预处理。这些参数设置符合市面上多数高速扫描仪,能够保证高效加工和正确的OCR识别率。

2.2 图像预处理

预处理工作一般包含倾斜角度纠正、污点去除、黑边去除、二值化等。由于部分OCRSDK开发包已经集成图像纠正功能,角度纠正可以视实际情况除去,让OCR模块进行处理。由于扫描通常会出现部分小的污点,这部分污点可以利用污点大小进行选择性自动去除。污点的大小通常在一批资料中有一个比较适合的大小,可以根据经验设置。黑边是由于

收稿日期:2007-09-29
修回日期:2007-10-15
作者简介:罗 维(1980-),男,助理工程师,主要从事OCR技术应用研究。

扫描仪扫描纸张边缘形成的大面积的污点，严重时可能影响自动纠偏的正确性，必须去除。去除可以采用固定像素进行自动删除，也可以采用四边邻接像素进行选择去除。

2.3 识别纠正

应用汉王 OCR、清华文通 OCR、文萃 OCR，还有市面上很多基于以上技术的 OCR 系统都能分析出图像的版面信息，并可以根据版面还原出原有版面的 RTF 格式文件。RTF 格式是一种通用的格式，类似 Word 的 doc 格式，可以用 Word 进行打开编辑。为了能利用以上资源，我们要求识别结果保存为具有格式信息的 rtf 格式。

人工纠正主要是利用人工进行少数识别错误的文字进行更正操作，这部分是为了保证输入的正确性所必须的。当文字倾斜时，OCR 识别软件会将水平笔画当作斜笔画处理，识别率会下降很多。如果扫描后的文字图像倾斜角度超过15°，倾斜校正会产生较大的失真和误差，从而严重影响识别率，这种情况建议摆正原稿重新扫描。目前清华文通 TH－OCR V9.0、尚书7等常见的 OCR 软件都集成了较为人性化的直观校对软件纠正。

2.4 条目入库

识别纠正得到具有一定格式的 rtf 文本格式文件后，要对 rtf 文件格式进行编程提取每个数据库字段对应的文本信息。可以利用 Microsoft Office Word 对 rtf 文件进行提取。最强大的功能之一就是其组件，它以组件对象模型（COM）接口的形式公开其功能。比较常用的方法是搜 Tables 的操作函数。

案卷目录上一般都标识有字段名称，利用 Word 的索函数 Selection.Find 可以找到开始位置，结束位置一般是下一个字段开始位置，这样我们就能可以获取文本，然后逐一插入到数据库中。

卷内目录是一个表格，第一行标识表格中的字段名，第二行开始就是一条条的记录。每页一般只有一个表格，所以可以先根据 word 获取的文本表格数对每个表格进行单独提取行信息，并逐一入库。

这一步骤由于利用到编程手段，并且使用到 word 的 Com 接口，可以利用 Word 的宏功能进行编程前的脚本学习，这样可以加快编程速度。编程中应注意对象的销毁操作，以免造成内存错误。

3 系统环境与开发工具

该方法只适合 Windows 环境和网络数据库，入库机器必须正确安装 Microsoft Office Word 2000或以上版本。

由于以上方法使用到 Word Com 编程，所以开发工具最好选择 Microsoft Visual Basic? 6.0或拥有托管扩展的 Microsoft .NET 和 Microsoft C#？或 Microsoft Visual C++?进行入库部分的编程开发，这样能获取较多编程上的技术参考资料，方便实现程序开发。

4 结束语

本文提出的印刷体档案条目信息采集方法可以自动提取已经是印刷体的档案条目但是对手写体条目还有待人工智能技术的发展和对手写体识别的准确性提高。该方法最终信息是从 rtf 文件进行采集的，可以利用多种多样的 OCR 系统或人工方式进行识别获取，具有一定的适应性。

目前 OCR 技术在数字化工程以及专利数字化方面广泛应用，它为我国图书档案资料，以及其他文献资料的数字化提供先进的技术手段。数字化是信息社会的技术基础，利用先进的 OCR 技术可以实现数字化推进信息化建设，促进社会发展。

参考文献：

[1] 王顺兴，颜美代. 精解 Word 2000 VBA 与范例解析[M]. 北京：北京大学出版社，2001.

[2] 陈明，丁晓青. 复杂中文报纸的版面分析、理解和重构[J]. 清华大学学报，2001，41(1)：29-32，59.

[3] 吴佑寿，丁晓青. 文字识别与智能信息处理论文集(第七辑)[C]. 北京：清华大学智能图文信息处理研究室，2002.

（责任编辑：邓大玉）