

一种水文年鉴数字化方法及其应用

杜红娟¹, 刘九夫², 田亚奇³, 严小林²

(1. 水利部南京水利水文自动化研究所, 江苏 南京 210012; 2. 南京水利科学研究所, 江苏 南京 210029;
3. 南京扬子水利自动化技术开发总公司, 江苏 南京 210012)

摘要 阐述水文年鉴数字化技术中的表格、数字识别、版面还原和校核等问题, 利用泰比(ABBYY)软件公司开发的 Recognition Server 2.0 OCR 和 excel 的内置宏探讨出一种解决水文年鉴数字化可行高效的实施方法, 为水文工作者处理水文年鉴数据提供了一种可行的方法。

关键词 水文年鉴数字化; ABBYY 软件; 识别率; 版面还原

中图分类号 P333.9 **文献标识码** B **文章编号** 1006-764X(2011)S1-0081-02

水文数据是水文工作中不可缺少的信息资源。1988 年之前的水文数据都是在年鉴上刊印出来的, 1988 年及之后才有了电子文档, 如何省时、快速、准确地把水文年鉴数字化就成了目前迫切需要解决的问题。

ABBYY FineReaderProfessional 是一款专业的 OCR 软件, 它不仅支持多国文字, 还支持彩色文件识别、自动保留原稿插图和排版格式以及后台批处理识别功能, 能够直接在 MSWord, MSExcel, WordPerfect 及 WordPro 中扫描和读取文件、信件或各类表格, 并且能存成 RTF, TXT, DOC, CSV, XLS 或 HTML 等格式, 它能保持表格与图片中原始的多栏页面设计^[1]。

ABBYY Recognition Server 2.0 有 6 部分组成: 服务管理器、处理系统、校核系统、远程管理中心、基于应用程序的 COM 接口、网络服务, 这 6 部分可以安装在同一台计算机上也可以装在不同的计算机上。其中服务管理器是主要组成部分, 它起到给其他各个部件分配工作的作用, 没有服务管理器 ABBYY Recognition Server 2.0 将不能工作。服务管理器不执行识别任务, 而是把接收到的图片文件传给处理器, 由处理器来识别文件, 识别完成后传给服务器, 然后传给校对系统。

以往数字化方法是采用人工敲键录入的方法手工录入, 校对方法是背靠背同时录入校对, 注重 2 人录入一致, 缺少合理性检查, 费时, 工作量大。本文利用俄罗斯 ABBYY 软件公司开发的 ABBYY Recognition Server 2.0 OCR 和 excel 的内置宏来解决水文年鉴数

字化涉及的表格、数字识别率、版面还原等关键技术, 并提出了一种可行高效数字化的实施方法。

1 方法介绍

ABBYY Recognition Server 2.0 解决了识别水文年鉴数据错行和缺少表格框线的问题, 能够保持观测时间和观测值的一一对应, 特别是在 excel 表格中能够把观测时间和观测值划分在同一行的单元格中, 不错行, 保证表格结构按照原行原列输出, 更有利于校核逐日降雨量和流量的识别结果。

水文年鉴数字化基本流程共有 4 步: 扫描、PS 处理、识别、校核, 其中最重要也是最关键的就是识别问题。目前, 常用的识别软件有汗王、尚书 7 号、清华紫光等, 这些识别软件大都针对文本格式的文档识别效果较好, 不能把表格结构的数据按照原版格式输出。应用 ABBYY 软件进行识别, 并应用 excel 中的内置宏进行校核, 可以高效、快速地解决以上问题。对于水文工作者来说, 资料处理是一项琐碎耗时的工作, 能够找出一种准确、高效的批处理水文年鉴数据的方案, 无疑会大幅度提高工作效率。下面就水文年鉴数字化过程中的重要环节逐一陈述。

1.1 批处理

ABBYY Recognition Server 2.0 对于图片文件的处理是由处理器完成的, 处理器可以安装在多台计算机上并与服务管理器连接, 它的批量识别功能可以一次识别多页图片文件, 用户可以根据需要自己控制导入页数。ABBYY 软件能够在多台机器上分

别处理文档,工作者只要把批量图片文件导入 input 文件夹里,剩余的工作完全由软件自动完成。处理器处理完毕就会将处理结果传给服务管理器,服务管理器再将其传给校对系统,工作者通过校对系统校对识别后的文档是否有误。整个工作过程充分利用了机器,从而大幅度提高了处理速度。2 个雨量观测站全年共有 730 个逐日观测值,如果手工录入,即使不看年鉴,在 excel 中随意输入 730 个数值至少也要 300 s,而 ABBYY 软件中识别 1 页印有 2 个观测站全年逐日观测值的 A4 文件只需要 13 s,因此,大幅度提高了水文年鉴的录入速度。

1.2 识别率问题

ABBYY Recognition Server 2.0 对于纯数据表格识别的错误率较低,仅为 17.8%,而对于包含有数字和文字的表格识别错误率稍高,为 22.3%,因为该软件不识别中文,这也是有待改进的不足之处。ABBYY 软件优点之一就是可以对自动识别后不满意的识别结果进行再次识别,再次识别错误率就会降低到 4.1%。水文年鉴大都是数据表格,可以把自动识别的结果删掉,然后划出自己所需要的数据结构部分再次识别;ABBYY 软件具有版面自动分析功能,它会根据所划分的部分自动将表格的框线先识别出来,再对表格中的数字进行识别,对于被划分区域内的数字有很高的识别率,而且速度很快。识别完成后可以检验是否有漏掉或多加的表格框线,如果有漏掉或多加的表格框线可以增加或删除框线,横线、竖线都可以添加或删除,具体使用应用实例中有陈述。校正过程只需要校正表格结构的框线,不用校正数字部分,数字部分可以在输出后的 excel 中用 VBA 内置宏来校核,同时还可以检查出年鉴上的刊印错误;当然在该软件里也可以像其他软件一样逐一校正数字部分,但是耗时费力。

1.3 版面还原

校正完后只要点击界面上的“Accept”按钮,就会自动以指定的格式输出到“Output”文件夹里,水文年鉴以 excel 格式输出较好。ABBYY 软件能够将原图片文件表格结构中的内容按照原行原列输出,即使像水文年鉴中只有列边框线没有行边框线的表格结构,ABBYY 软件也能根据表格内容将行边框线自动加上,保持观测时间和观测值相对应,输出在同一行的不同单元格,保证不错行。

1.4 校核

以往水文年鉴采用手工录入,缺少合理性检查,利用 ABBYY 软件识别的数据可以在 excel 里用内置宏把一些错误的乱码信息过滤掉,只留下数据部分校核以验正误。通过运用 excel 的内置宏编程,用逐

日的观测值计算出月值,与实测的月值进行求差计算,如果差值为零则表明识别正确无误,差值不为零则用红色字体凸显出来,同时还把每 5 d 的值用不同颜色填充,方便有错误时校核。用 excel 内置宏校核水文年鉴数据可以一次打开多个 xls 文件,执行宏命令,快速省时。

2 应用实例

ABBYY Recognition Server 2.0 具有简单清晰的界面,操作简便,以逐日流量表的识别校核过程为例,说明 ABBYY Recognition Server 2.0 的操作步骤。把 Ps 处理过的图片文件放到创建的 input 文件夹里,经过大约 13 s 就可以在 Verification Station 里看到自动识别过的数据表格,由于 ABBYY 软件不支持中文,因此自动识别出的表格会有乱码现象产生。如果图片有水平或颠倒情况,可以用工具栏中的工具来进行旋转。自动识别出的表格如果不满意可以全部删除,然后把测站站号作为文本格式,逐日流量值和月值作为表格格式按先后顺序选中再次识别,在识别过程中会有短暂的停留时间,这时就要快速浏览一下哪个单元格有断线情况,如果有断线的表格可以对其进行修改,否则当年鉴上数据不在同一条直线上时,观测值与日期会出现错位;再次识别后的结果就会在 Output 文件夹里以 excel 格式出现。

ABBYY 软件最大的优点就是能够根据划分的表格数据结构固定行数和列数,例如逐日资料识别出的结果有固定的 32 行、13 列,而不会像传统录入资料那样多一行或少一行。本方案校核利用 excel 的内置宏,用一个快捷键就可以对打开的多个文件进行校核。

3 结 语

水文年鉴数字化技术中需要解决的问题主要有 3 个:①可选的识别字库是提高水文年鉴数字识别率的关键;②表格分析功能是水文年鉴数字化的基础;③实现水文年鉴的高效数字化。ABBYY Recognition Server 2.0 软件已经基本具备了以上 3 方面的技术要求,能够较好地解决水文年鉴数字化的问题,虽然该软件不能识别中文,但亦能满足水文年鉴处理的要求,比起其他 OCR 软件和传统录入方式已是一个很大的进步。

参考文献:

- [1] 陆文灏,张卫星,张峥.基于 OCR 软件的刀具信息数据库开发与应用[J].计算机应用,2010(3):63-67.

(收稿日期:2010-04-15 编辑:方宇彤)