采用双层 PDF 形式将方正书版文件制作为可检索式 PDF 文件

周 雪 莹

《烟台大学学报(自然科学与工程版)》编辑部,264005,山东烟台

摘 要 以方正书版文件转换所得的几类常见的 PDF 文件为素材,基于 OCR 技术和 PDF 文件编辑技术,探索出 2 类制作可检索式双层 PDF 文件的方法。用 Readiris 法制作的 Image-Text 型双层 PDF 操作简便、文件很小、可生成索引书签;用 Foxit PDF Editor 法制作的 Graphic-Text 型双层 PDF 清晰度高、文本精准。这 2 种双层 PDF 文件均可以很好地满足网络期刊文献检索的需要。

关键词 双层 PDF 文件;检索;OCR 技术;Readiris;Foxit PDF Editor

Making searchable network periodicals from Founder Bookmaker documents in the form of double-layer PDF // ZHOU Xueying

Abstract Using different types of PDF documents converted by Founder Bookmaker documents, two methods of making searchable double-layer PDF documents are explored based on OCR and PDF editing techniques. The Image-Text double-layer PDF made by Readiris is fairly small in size, the process is simple and convenient, and index bookmarks can be created simultaneously. The resolution of Graphics-Text double-layer PDF made by Foxit PDF Editor is exceedingly high, and the correctness of the text is perfect. Both of the two kinds of double-layer PDF can satisfy document indexing of network periodicals.

Key words double-layer PDF document; search; OCR ; Readiris;Foxit PDF Editor

Author's address Editorial Department of Journal of Yantai University (Natural Science and Engineering Edition), 264005, Yantai, China

目前网络期刊所提供的文献格式大都为 PDF,而期刊编辑部广泛使用的方正书版排版软件无法直接生成 PDF 文件,由各类软件转换生成的 PDF 文件或不适于检索,或版面显示效果不佳,不宜作为网络期刊的文献资源,这就极大地制约了网络期刊的发展。

本文基于 OCR 技术和 PDF 文件编辑技术,探索出将方正书版文件制作为可检索式 PDF 文件的 2 类方法,以期对网络期刊建设提供技术参考。

1 双层 PDF 文件简介

"双层 PDF" 又称为"可检索式 PDF"(searchable PDF)或"透明文字 PDF",一般将其定义为"底层是扫描图像(Image)层,上层是透明文字(Text)层的 PDF 文件"[1],阅读时看到的是与纸样一致的底层扫描图像,搜索或用光标选取时又可直接对上层文字进行操作。因其具有文字可检索的性质,可以在网络上进行在线检索,并通过建立索引数据库进行科学的管理,所以非常适用于网络期刊。万方和维普期刊数据库对用方正书版排版的期刊就是将样刊扫描后,经 OCR 识别生成 Image-Text(图像-文本)型双层 PDF 文件以供下载。

本文所讲的双层 PDF 不仅包含了上述定义,还将其范围进行了扩展,即底层除了为光栅模式的图像以外,还可以是矢量模式的曲线图形。也就是说,除了人们广泛接触的 Image-Text 型 PDF 外,双层 PDF 还包括 Graphics-Text(图形文本)型。

2 常用的书版文件转 PDF 方法的优缺点

双层 PDF 制作是在已有的单层 PDF 文件基础上进行加工,因此,需要对常用的书版文件转 PDF 的方法进行分析,以筛选出适于用来制作双层 PDF 的文件素材。

2.1 书版文件直接转换为 Text 型矢量 PDF

2.1.1 方正 PDF Creator 法 方正 PDF Creator 软件可将方正书版生成的 PS、S2、S72、PS2、EPS 等文件转换为 PDF,版面效果非常好,文本可复制;但所有文本都是全角形式,而文献检索时对英文、数字的检索都是 ASCII 码的半角形式,所以无法检索其中的英文和数字,大大影响了文献检索的命中率。

2.1.2 PS22PDF 法 ccyt(长城云天)PS22PDF 软件可将方正书版大样文件转换成 PDF,文本正确率高;但英文、数字也是全角形式,无法检索,且其字库不是方正特有的 CID 字库,使得页面显示效果欠佳,如字体发生改变、英文及符号之间间距不均,加之页面尺寸与文本字号都比原文件增大约 30%,不宜用来作为网络期刊的文献资源。

2.1.3 S2toPDF/PStoPDF 法 Nooog(龙谷)S2toPDF 和 PStoPDF 软件分别可将方正大样文件和 PS 文件转换成 PDF 文件,二者生成的 PDF 文件的文本正确率均极高。但 S2toPDF 生成的 PDF 文件中英文和数字也是不能用于检索的全角形式;PStoPDF 生成的 PDF 文

件中原方正仿宋变为了宋体,原斜体字符变为了正体,不能完美地反映排版效果。

2.2　文杰打印机虚拟打印生成 Graphics 型矢量 PDF

利用方正文杰系列打印机将方正书版文件虚拟打印为标准 PS 文件,再用 Adobe Acrobat 或 GSviewer 转换生成 Graphics 型矢量 PDF[2-3],清晰度高,版面效果与印刷版完全一致,适于校对和出片印刷;但其中的文字为转曲的图形,虽能用光标选中却不能提取文本,无法用于检索。

2.3　PSPPRO 虚拟打印法生成 Image 型光栅 PDF

利用方正 PSPPRO 的虚拟打印功能,通过 pdfFactory、FinePrint 或 Adobe PDF 将方正大样文件转换成 Image 型光栅 PDF 文件[4-5],其版面效果虽忠实于印刷版,但清晰度不如矢量 PDF,文本也无法提取,不能进行检索。

3　双层 PDF 的制作方法

3.1　Image-Text 型 PDF 的制作　由 2.1.1、2.2 和 2.3 节所生成的单层 PDF 文件版面效果与印刷版完全一致,可对其用 OCR 方法进行文本识别,可生成底层为光栅图像、上层为透明文字的 Image-Text 型双层 PDF。

1)利用 ABBYY Finereader(简称 ABBYY)9.0.0.882 以上版本。该软件可识别 2.1.1、2.2 和 2.3 节生成的各类 PDF 文件,在打开文件的同时就将矢量 PDF 进行光栅化处理。在 ABBYY 界面中,选择“页面”的“文档语言”为“简体中文;英文”,打开单层 PDF 文件,即开始逐页进行文本识别选择转换识别,点击“编辑图像”,对图像分辨率进行选择或设定,一般默认为 300 dpi。识别结束后将文件另存为“PDF/A 文档”,即为双层 PDF 文件。

2)利用 Readiris Corperate(简称 Readiris)10 以上中文版本。Readiris 中文版带有亚洲识别模块,对中文识别准确,可处理光栅 PDF 文件和 Graphics 型矢量 PDF 文件,在对后者进行识别时将其光栅化,得到的是底层图像为 300 dpi 的双层 PDF。打开软件,将“字符识别向导”中的图像来源选择为“图像文件”,语言为“中文(简体)”,次要语言为“英式英语”和“英语(美国)”,格式输出为“发送到 Acrobat/Reader 图像-文本”。打开单层 PDF 文件,点击“识别+保存”,即生成双层 PDF 文件。在“格式”的“PDF 选项”中勾取“制作书签”,便会随文件生成用页码和标题作为索引的书签,便于查找文中内容。

3)利用 Adobe Acrobat 8.0 以上版本。Acrobat 8.0 以上版本具有 OCR 功能,但只能处理光栅 PDF 文件。在 Acrobat 中打开 PDF 后选择“文档”—“OCR 文本识别”—“使用 OCR 识别文本”,保存或另存即生成双层 PDF。

上述 3 种方法各有其特点,从近期《烟台大学学报(自然科学与工程版)》中抽取 15 篇文章进行效果测试,结果见表 1。

表 1　3 种基于 OCR 的制作方法比较

方法	可使用的单层 PDF 范围*	识别前后的 PDF 文件平均大小(以 300 dpi 为例)/kb			转换速度	文本识别准确率
		Text 型 PDF(识别前 477 kb)	Graphics 型 PDF(识别前 291 kb)	光栅 PDF(识别前 1 499 kb)		
ABBYY	Text 型 PDF;Graphics 型 PDF;光栅模式 PDF	2 838	2 124	3 851	较慢	高
Readiris	Graphics 型 PDF;光栅模式 PDF	—	194	268	快	高
Acrobat	光栅模式 PDF	—	—	2 420	快	较低

* 表中 Text 型 PDF、Graphics 型 PDF 和光栅模式 PDF 分别对应方正 PDFCreator 法、文杰打印机虚拟打印法和 PSPPRO 虚拟打印法。

由表 1 可知,这 3 种方法中,只有 Readiris 处理后的文件比处理前大大减小,其他方法处理后文件均增大。ABBYY 可识别的 PDF 文件范围最广,文本识别准确性高,可满足检索的需要,但转换速度较慢、生成的双层 PDF 文件偏大;Acrobat 可识别的 PDF 文件范围最小,而且文本识别准确性较低,不适于用来检索;Readiris 转换速度快,不仅生成的 PDF 文件很小,文本识别准确率也很高,而且可以生成索引书签,最适用于双层 PDF 文件的制作。对于平时使用方正 PDF Creator 法的编辑部,可在日常校对和出片印刷时使用 PDF Creator 生成的单层 PDF 文件,在制作网络期刊时可考虑使用 Readiris 法或 ABBYY 法制作双层 PDF 文件。

3.2　Graphics-Text 型 PDF 的制作　由上文可知,PStoPDF 法生成的 Text 型 PDF 的文本准确率高,而且英文和数字均可检索,但字体显示效果不佳;而文杰打印机虚拟打印生成的 Graphics 型 PDF,其版面显示佳且清晰度高,但不含有文本信息:因此,可采用 PDF 编辑软件将这 2 类各具优点的 PDF“合成”在一起,即提取 Text 型 PDF 中的文本,以隐藏形式覆盖在 Graphics 型 PDF 上,得到 Graphics-Text 型双层 PDF。本文利用 Foxit PDF Editor 软件来进行“合成”,具体步骤如下。

1)在 Foxit PDF Editor 中打开 Text 型 PDF,点击“编辑”—“全部选择”,即选中当前页的文本;

2)按住 Ctrl 并用鼠标单击半字线、分号等横线,以将其从选中的文本中排除(因横线不能隐藏),点击“复制”;

3)在 Foxit PDF Editor 中再打开 Graphics 型 PDF,翻至相应页面,“粘贴”,即可将复制的文本覆盖在其上;