第 28 卷 第 3 期
2020 年 6 月

厦门理工学院学报
Journal of Xiamen University of Technology

Vol. 28 No. 3
Jun. 2020

doi: 10. 19697/j. cnki. 1673-4432. 202003012

PDF 文档表格信息的识别与提取

田翠华，张一平，胡志钢，高静敏，李西雨

(厦门理工学院计算机与信息工程学院，福建 厦门 361024)

[摘 要] 为编辑 PDF 文档中的表格信息，设计一种基于 Python 平台的，包含文件选取与文件转换两大功能模块的信息提取软件。该软件利用 Python 内置库，针对 PDF 中不同表格的结构设计算法，识别表格内的文字信息与表格结构，将得到的表格结构还原至 Word 与 Excel 文档中，文字信息同样复原至对应单元格内。实验表明：开发的软件完整快速地提取了 PDF 中的表格信息，并将其转化为易于编辑的 Excel 和 Word 文档，达到了预期目的；其转换速度与收费软件 WPS 相当，转换速度明显快于其他免费转换软件；识别精确度与迅捷 PDF 转换器、Smallpdf 等相比有所提高。

[关键词] PDF 文档；表格信息；信息识别；信息提取；Python 开发平台

[中图分类号] TP317 [文献标志码] A [文章编号] 1673-4432 (2020) 03-0070-07

PDF 是将文字、图像、表格等封装在一起的一种跨平台便携式文件格式。据统计，在世界主要数据门户网站中，大约 13%的已发布文件都是以 PDF 格式存储数据的[1]。其设计目的是能在不同应用程序、不同平台、不同硬件上以完全相同的样式输出内容[2]。为此 PDF 文档将其结构信息进行了封装，使其可编辑的信息减少，但这也使 PDF 文档中的逻辑结构或者语义结构均无法直接获得，从而在一定程度上妨碍了正常的修改。于是如何让 PDF 格式的文档转化为可编辑的 Word 或 Excel 文档就成了研究的热门话题。提取 PDF 文档中文本、图片、公式等信息的研究已有很多[2-6]，表格作为在金融领域广泛使用的数据存放格式，也是 PDF 文档的重要组成部分[7]。窦方坤等[8]基于 pdf2xml 对药学文献 PDF 中的文本元素进行抽取，此方法对识别表格线缺失的表格有着良好的识别效果。唐皓瑾[9]通过表格轮廓的栅格，将表格处理为便于识别的二维数组结构，提高了全框线表格的识别准确率。陆锦鹤[10]使用了 Excel 中的 VBA 编程功能，提出了一种从 PDF 文档里提取所需信息的算法，并自动排版，然后生成表格。但现有工具，将 PDF 转化为 Excel 或 Word 格式的效果并不好，通常不能正确识别带有合并单元格的表格，也难以一次性转换拥有多页表格的 PDF 文档。为更加快速有效地识别、提取 PDF 中的表格信息，本文利用 Python 开发平台内置库，设计出一种信息提取软件，识别、提取 PDF 中的表格，并按需求将提取的表格转换为 Excel 和 Word 这两种最常用的格式。

1 软件设计与模块功能的实现

1.1 软件设计思路

Python 具有简洁性、易读性以及可扩展性，在编程中既支持面向过程的编程也支持面向对象的编

收稿日期：2020-02-09 修回日期：2020-06-01

基金项目：厦门市科技计划项目 (3502Z20193058)；厦门理工学院优质研究生课程及案例库建设项目 (YG20190303)；厦门理工学院创新创业实验班校企合作项目 (2019SYB07)

通信作者：田翠华，女，副教授，博士，研究方向为云计算、物联网，E-mail：2010110711@xmut. edu. cn。

引文格式：田翠华，张一平，胡志钢，等. PDF 文档表格信息的识别与提取[J]. 厦门理工学院学报，2020，28(3)：70-76.

Citation: TIAN C H，ZHANG Y P，HU Z G，et al. Recognition and extraction of table information from PDF documents [J]. Journal of Xiamen University of Technology，2020，28(3)：70-76. (in Chinese)

程，能够更好地满足互联网快速迭代的需求[11]。作为时下广泛应用的编程语言，其以高度的封装性、灵活性和丰富的第三库资源，在数据分析领域中占据着重要的地位[12]。本文基于 Python 语言开发，使用图形界面库 Tkinter，微软文件操作库 Pdfplumber，利用库中所封装的操作来实现对 PDF 表格信息的提取与识别。该软件的功能主要包含文件选取模块和文件转换模块。软件设计流程如图 1 所示。

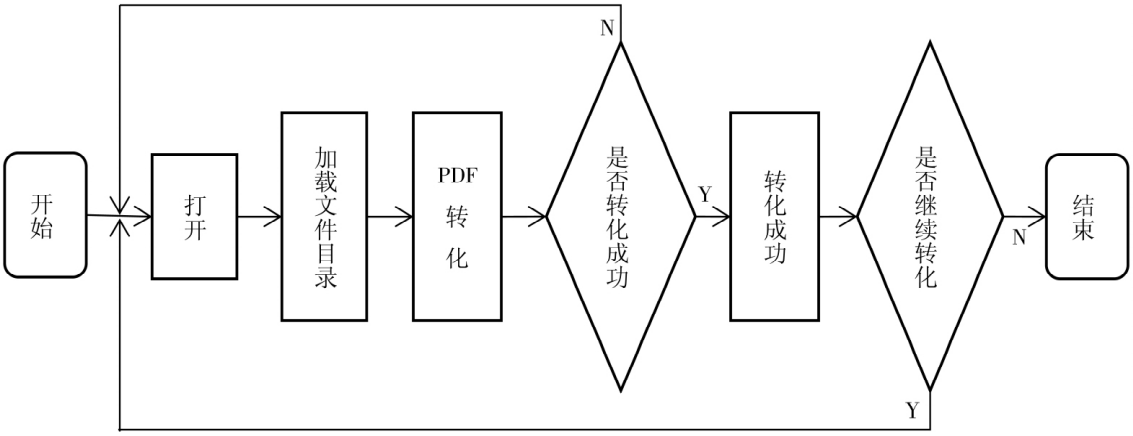

图 1　软件设计流程图

Fig. 1　System flow chart

1.2　文件选取模块

文件选取模块能够方便用户选取想要转化的 PDF 文件，该功能的实现，是基于 Python 中用于图形开发界面的 Tkinter 库。该库可以在大多数的 Unix 平台下使用，同样可以应用在 Windows 与 Macintosh 系统。使用 Tkinter 进行 GUI 界面开发，主要包括窗口实例化、控件定义、控件属性设置及控件布局等步骤[13]。导入 Tkinter 模块后，使用 root = Tkinter. Tk( ) 与 root. withdraw ( ) 便可创建一个文件目录弹出框，用户点击确定后，生成文件绝对路径，存入变量，由此截取变量中的文件名，将截取到的文件名与不同的扩展名连接，用以获取生成的文档名，待下一步使用。

1.3　文件转换模块

模块的主要功能是识别、提取 PDF 文件中的表格，并将识别结果导出为可编辑的 Word 与 Excel 格式。PDF 文档由文件头、文件体、交叉索引表、文件尾 4 部分组成。该模块根据 PDF 文件结构，采用 Pdfplumber 库来识别表格线，针对多样的表格结构设计算法，对 PDF 中的表格进行分析提取。图 2 为 PDF 文档解析过程。

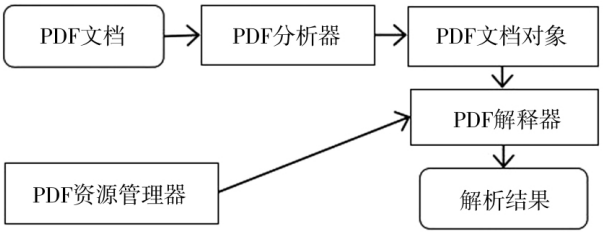

图 2　PDF 文档解析过程图

Fig. 2　PDF document parsing process

如图 2 所示，首先，创建一个文档分析器与 PDF 文档进行关联，提取 PDF 文档的结构信息，将提取到的信息保存至 PDF 文档对象中；其次，创建 PDF 资源管理器，存储字体或图像等共享资源；然后，创建解释器，将 PDF 文档对象编码为 Python 可识别的格式；最后，将解析出的页面数据存入 Layout 中。Layout 用以存放 PDF 文档中所有页面的信息，如单行文本、图片对象、线段信息等。在实际使用中，表格边缘线有可见边与不可见边两种。对于可见边，Pdfplumber 是基于 Pdfminer 进行解析

的，他在获取可直接识别的线段后，将重复的边进行合并操作；将线之间垂直距离相距较小的边对齐，使之位于同一直线；将线之间端点相近的边合并成一个线段。对于不可见边，Pdfplumber 根据每一页所解析出的字符对齐情况，来判断垂直和水平方向上存在的线，再将这些字符的顶部位置进行聚类，由此找到同行或同列的文本，最终将文本的顶部和底部的边缘线作为识别到的线段。

模块根据识别到的边框线确定表格线的交点，通过交点得到的最大矩形以确定表格内容所在区域，从而找到交点围成的最小矩形以确定表格的具体单元格，最终生成一个表格对象 Table。Table 类的 extract 方法能够抽取出每个单元格内部的文本。首先，抽取出 Table 对象；其次，按照行提取出表格每一行的文本信息，转化为 DataFrame 格式；最终，调用 to_excel( )，将 DataFram 形式的数据输出为 Excel。Excel 电子表格软件是应用最广泛的办公软件之一，可以进行各种数据的处理、统计分析和辅助决策操作，广泛地应用于管理、统计财经等众多领域[10] 133。

Microsoft Office Word 作为当代各行各业日常办公中不可或缺的文本编辑工具，已经基本取代了手写的办公模式[14]。因此有必要将提取的表格转化为 Word 文档。首先，创建一个空白 Word 文档，即 Document 对象；其次，调用 openpyxl 库中的 load_work( ) 处理带公式、带宏的 Excel，通过 max_row，row_column 获取表格行数与列数，在空白文档中创建表格；最终，将 Table 中的文本信息从上至下，从左至右依次填入空白表格内，保存新增的信息，生成可编辑的 Word 文档。文件转换流程如图 3 所示。

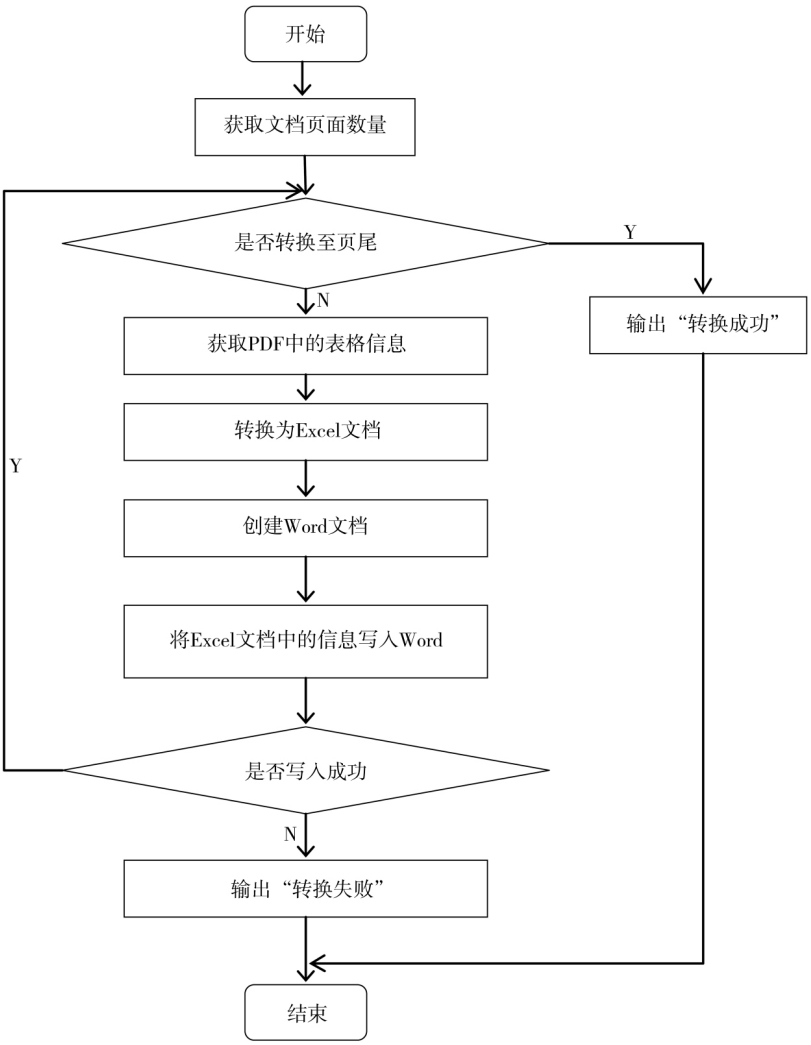

图 3 文件转换流程图

Fig. 3 File conversion process

文件转换的具体实现代码：

```
NewFilepath = Filepath [-3] +'xlsx'
print( NewFilepath)
for page in pdf. pages:
  print( page. extract_text( ) )
  for table in page. extract_tables( ) :
    for row in table:
      rowlist = str( row) . replace( " [',"",) . replace( "","") . replace( "","") . replace
      ( "\\n","") . split( ",")
      ws. append( rowlist)
      print( page. extract_text( ) )
    tb = pd. DataFrame( ws. values, index = None)
  tb. to_excel( NewFilepath, index = False, header = False)
pdf. close( )
```

将 Python 生成的 . py 文件转换为 . exe 上传网站 https: //cf20240924. icoc. vc，供人免费使用。

2　软件测试

运行文件转换软件，验证软件是否实现了之前所设计的所有功能，并与当前常见的一些 PDF 转换工具做对比。

2.1　实验工具

准备一组 PDF 文档，其中包含 5 类不同表格类型的 PDF 文档：Test1 包含带有边框的表格，表格内不含有合并单元格，如图 4 所示；Test2 包含带有边框的表格，表格内含有合并单元格；Test3 包含无边框的表格，表格内不含有合并单元格；Test4 包含无边框的表格，表格内含有合并单元格；Test5 具有 4 页，包含有框表格，表格内不含合并单元格。对比转换工具为：WPS 中 PDF 转换工具，迅捷 PDF 转换器，Smallpdf，以及本文的 PDF 转换软件 PDFtoWAE。实验在 WPS 中展示原 PDF 文档，在 Excel 和 Word 中展示提取的文档。

序号	名称	材质	商家名称	商家报价
1	保安过滤器	不锈钢 304	深圳市益家源环保科技有限公	12500
2	保安过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500
3	袋式过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500
4	袋式过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500
5	袋式过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500

序号	名称	材质	商家名称	商家报价
1	保安过滤器	不锈钢 304	深圳市益家源环保科技有限公	12500
2	保安过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500
3	袋式过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500
4	袋式过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500
5	袋式过滤器	不锈钢 304	深圳市益家源环保科技有限公	19500

图 4　第一类测试表格 PDF 原文档

Fig. 4　Test 1 PDF original file

2.2　实验结果

点击 https: //cf20240924. icoc. vc 下载并运行软件。首先，转换 Test1 类文档，打开软件后根据文

件选择对话框选择想要转换的文档，PDF 原文档如图 4 所示. 若转换成功则弹出 “转换成功”，并转换生成 Excel 和 Word 文档 2 个文件，分别如图 5(a) 和图 5(b) 所示。采用同样的步骤测试 Test2、Test3、Test4、Test5 类文档。

	A	B	C	D	E
1	序号	名称	材质	商家名称	商家报价
2	1	保安过滤器	不锈钢304	深圳市益家源环保科技有限公	12500
3	2	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500
4	3	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500
5	4	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500
6	5	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500
7	序号	名称	材质	商家名称	商家报价
8	1	保安过滤器	不锈钢304	深圳市益家源环保科技有限公	12500
9	2	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500
10	3	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500
11	4	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500
12	5	保安过滤器	不锈钢304	司深圳市益家源环保科技有限公	19500

(a) Excel文档

序号	名称	材质	商家名称	商家报价
1	保安过滤器	不锈钢 304	深圳市益家源环保科技有限公	12500
2	保安过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500
3	袋式过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500
4	袋式过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500
5	袋式过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500
序号	名称	材质	商家名称	商家报价
1	保安过滤器	不锈钢 304	深圳市益家源环保科技有限公	12500
2	保安过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500
3	袋式过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500
4	袋式过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500
5	袋式过滤器	不锈钢 304	司深圳市益家源环保科技有限公	19500

(b) Word文档

图 5 转换后的文档

Fig. 5 Converted documents

对 Test1 ~ Test5 组 PDF 文档，使用如今较为流行的几款 PDF 转换工具进行实验，将实验结果与本软件的转换结果做对比。表 1 为不同转换工具下 PDF 文档转为 Excel 文档的转换效果对比表。

表 1 不同工具下 PDF 文档转为 Excel 文档的转换效果

Table 1 Conversion effect of PDF to Excel documents under different tools

工具	测试组				
	Test1	Test2	Test3	Test4	Test5
PDFtoWAE	全部转换，内容无错	全部转换，部分单元格未正确识别	全部转换，内容无错	全部转换，部分单元格未正确识别	全部转换，内容无错，结果位于同一界面
WPS	全部转换，内容无错	全部转换，部分单元格未正确识别	部分合并单元格内容缺失	全部转换，可生成合并单元格	全部转换，内容无错，结果位于不同界面
迅捷 PDF 转换器	全部转换，内容无错	全部转换，首行部分单元格未正确识别	全部转换，内容无错	全部转换，部分单元格未正确识别	全部转换，内容无错，结果位于不同界面
Smallpdf	全部转换，内容无错	全部转换，首行部分单元格未正确识别	首行合并的单元格未正确识别	全部转换，部分单元格未正确识别	全部转换，内容无错，结果位于不同界面

表 2 为不同转换工具下 PDF 文档转为 Word 文档的转换效果对比表。

表 2　不同工具下 PDF 文档转为 Word 文档的转换效果

Table 2　Conversion effect of PDF to Word documents under different tools

工具	测试组				
	Test1	Test2	Test3	Test4	Test5
PDF to WAE	全部转换，内容无错	全部转换，部分单元格未正确识别	全部转换，内容无错	全部转换，部分单元格未正确识别	全部转换，内容无错，结果位于同一界面
WPS	全部转换，内容无错	全部转换，部分单元格未正确识别	部分合并单元格内容缺失	全部转换，转换无错	全部转换，内容无错，结果位于不同界面
迅捷 PDF 转换器	全部转换，内容无错	全部转换，部分单元格未正确识别	全部转换，部分单元格未被识别	全部转换，部分单元格未正确识别	全部转换，内容无错，结果位于不同界面
Smallpdf	全部转换，内容无错	全部转换，部分单元格未正确识别	全部转换，部分单元格未正确识别	全部转换，部分单元格未正确识别	全部转换，内容无错，结果位于不同界面

由表 1、2 可见，本文设计的软件完成了设想的主要功能，完整地将 PDF 表格信息提取出来，并转化为 Excel 和 Word 文档格式，方便用户使用。该软件在不同的情况下都能完整地转换 PDF 表格，并能够将所有表格转换在文档的同一界面内。对比其他工具，转换效果良好，但是单元格的合并效果不如 WPS。

用不同转换工具提取同一 PDF 文档中表格，并将表格转换为 Excel、Word 文档所需时间，如表 3 所示。

表 3　不同工具下 PDF 转为 Excel、Word 文档所需的时间

Table 3　Time required for PDF to Excel or Word documents under different tools

工具	转为 Excel 所需时间/s					V	工具	转为 Word 所需时间/s					V
	Test1	Test2	Test3	Test4	Test5			Test1	Test2	Test3	Test4	Test5	
PDF to WAE	1.43	1.67	2.21	2.17	1.39	1.00	PDF to WAE	1.43	1.67	2.21	2.17	1.39	1.00
WPS	1.56	1.61	1.77	2.31	1.34	0.98	WPS	1.46	1.55	1.86	2.31	1.34	0.97
迅捷 PDF	7.28	8.46	8.75	8.79	8.56	4.39	迅捷 PDF	8.65	9.00	9.55	10.79	8.72	5.40
Smallpdf	3.42	4.46	5.41	5.73	4.26	2.64	Smallpdf	3.26	3.56	5.26	4.74	4.32	2.42

表 3 中，V 是各转换工具转换所需时间与本软件转换所需时间比值的均值。可见，用 WPS 转换时，V 值接近 1，说明本软件与 WPS 的平均转换速度接近；用 Smallpdf 转换时，V 值超过 2，说明本软件比 Smallpdf 的平均转换速度快 1 倍以上；用迅捷 PDF 转换时，V 值达到 4~5，说明本软件比迅捷 PDF 的平均转换速度快 3~4 倍，其中在转换单行有边框表格的情况下速度最快。

在收费方面：WPS 在转换时需要用户充值，迅捷 PDF 转换器以及 Smallpdf 限制了用户的免费转换量。而本软件完全免费，提高了用户的使用体验。

3　结论

本文基于 Python 平台，开发出一种 PDF 表格信息提取识别软件。该软件利用平台的内置库，根据多样的表格结构设计算法，实现了 PDF 表格信息的识别、提取，并根据用户需求将提取的表格转换为 Excel 和 Word 格式。与已有的 PDF 文件表格转换工具的对比实验结果显示：所设计软件的转换时间与收费软件 WPS 相当，转换速度明显快于其他免费转换软件；识别精确度上与迅捷 PDF 转换器、Smallpdf 等相比有所提高。本软件已经打包为可直接执行的 .exe 文件，并上传至网站，便于用户下载使用，可进一步提高用户的使用体验。

但是该应用软件还存在一定的问题，需要进一步的完善，例如可以增加功能，使用户可以转换指定页等。在实际应用中，PDF 文档会有成百上千页，对于这种大数据量的文档，其转换时间较长，如

何改进代码以实现快速提取，也是接下来进一步研究的方向。

参考文献

[1] ANDREIWID S,ZANDER P.Unleashing tabular content to open data: a survey on PDF table extraction methods and tools [C] //Proceedings of the 18th Annual International Conference on Digital Government Research.New York: Association for Computing Machinery,2017: 54-63.

[2] 于丰畅,陆伟.基于机器视觉的 PDF 学术文献结构识别 [J] .情报学报,2019,38(4):384-390.

[3] LOVEGROVE W S,BRAILSFORD D F.Document analysis of PDF files: methods,results and implications [J] .Electronic Publishing-Origination,Dissemination and Design,1995,8(3): 207-220.

[4] HASSAN T.Object-level document analysis of PDF files [C] //Proceedings of the 2009 ACM Symposium on Document Engineering.Munich: Association for Computing Machinery,2009: 47.

[5] ZHANG X,GAO L,YUAN K,et al.A symbol dominance based formulae recognition approach for PDF documents [C] //IAPR International Conference on Document Analysis and Recognition (ICDAR).New Jersey: IEEE Computer Society,2017: 1 144-1 149.

[6] CHEN J,GAO L ,TANG Z.Informationextraction from resume documents in PDF format [J] .Electronic Imaging,2016(17): 1-8.

[7] 马晶晶.金融领域信息的自动抽取与分析方法 [D] .哈尔滨: 哈尔滨工业大学,2013.

[8] 窦方坤,曹皓伟,徐建良.基于文本元素的 PDF 表格区域识别方法研究 [J] .软件导刊,2020,19(1): 113-116.

[9] 唐皓瑾.一种面向 PDF 文件的表格数据抽取方法的研究与实现 [D] .北京: 北京邮电大学,2015.

[10] 陆锦鹤.Excel 软件 VBA 功能使用案例一则: 从 PDF 文件中提取出的信息中挑选需要的信息并重新排版形成可用的 Excel 表格文件 [J] .智库时代,2018(39): 133-134.

[11] 牛作东,李捍东.基于 Python 与 flask 工具搭建可高效开发的实用型 MVC 框架 [J] .计算机应用与软件,2019,36(7): 21-25.

[12] 徐玉芳,苏斌.Python 语言特点及其在机器学习中的应用 [J] .计算机产品与流通,2019(12): 142.

[13] 张喜红,王玉香.基于 Python Tkinter 课堂手机监管系统的设计 [J] .中州大学报,2019,36(2): 125-128.

[14] 刘艳茹,孙维耕,封平安.Word 文档模板的制作及其应用 [J] .科学技术创新,2019(25): 72-74.

Recognition and Extraction of Table Information from PDF Documents

TIAN Cuihua, ZHANG Yiping, HU Zhigang, GAO Jingmin, LI Xiyu

(School of Computer & Information Engineering,Xiamen University of Technology,Xiamen 361024,China)

Abstract: An information extraction software to edit the table information in PDF documents was designed on the Python platform,which contains modules of file selection and file conversion.The software uses the Python built-in library to design algorithms for structures of different tables in PDF,recognizes the text information and table structure in the table,and restores the obtained table structure to Word documents or Excel documents and the text information to the corresponding cells.Experiments show that the software developed extracts table information from PDF documents completely and quickly,and converts it into easy-to-edit Excel and Word documents,achieving the expected results.It has a conversion speed equivalent to that of paid software WPS and significantly faster than those of other free conversion software,and its recognition accuracy is better than Xunjiepdf,Smallpdf and the likes.

Key words: PDF document; table information; information recognition; information extraction; Python development platform

(责任编辑 宋 静)