基于PDF文本元素的表格信息提取方法

鲁娇[1]and 秦文[2]

[1] 北京科技大学，北京，100083
[2]洞见时代（北京）信息技术有限公司，北京
lujiaoff@163.com
qinwen@insight-time.com

摘要. PDF格式的文件因其独特的跨平台便捷性优势，成为了当前主流的文档格式，在金融领域，受监管要求和行业特性影响，会分析、处理、输出大量的PDF 文档，为了对上市公司年报公告中的表格进行正确识别与提取，减少表格抽取过程中的表格无关元素对提取结果的干扰，并实现纯结构化数据的生成，本文基于PDFMiner对上市公司公告PDF文本进行文本解析，提取其中的各种文本元素，并根据文本元素中线与文字的环绕位置进行表格的识别，从而实现表格的提取与重构，形成结构化的信息，便于存储。通过在上市公司年报公告中中提取财务报表信息，验证了本方法的有效性与实用性。

关键词: 文本元素，表格识别与提取

1 引言

PDF（Portable Document Format）由Adobe公司开发并推广，是一种独特的跨平台的便携文件格式。它的跨平台特性使得文件可以广泛的运用于Windows、Linux、Mac OS等当前主流的操作系统中，并使其成为电子文档发行和数字化信息传播的理想文档格式。随着互联网的普及发展与信息的爆炸式增长，越来越多的网络资料、科技文献、公司年报等都开始使用PDF格式作为电子文档的首选格式。

随着PDF格式的普及，大量的信息都以PDF文档的格式进行输出，因此从PDF文件中提取有价值的内容就成了一项非常有意义的任务。但是由于PDF文档结构比较复杂，PDF 解码后文本信息呈现“多元素”特征，各文本元素呈现出分散的字或线的集合，没有明确的逻辑关系。它不像word、excel文件那样只保存文字或结构化数据信息，PDF中包含文本、图形、表格各种格式的内容都是视觉上的，而不是真正的文字、图形、表格，因此抽取其中内容相对来说比较困难，尤其是其中的表格信息。文本呈现出表格只是基于视觉的，也就是说，在该类文档格式中本不存在表格格式，只存在一团团的文字和一些穿插其中的图像线，我们一般只能直观地从显示结果看到表格，而无法直接从文档格式中获取表格信息，对于这种类型的表格 的提取，我们需要根据文本元素的特征，重新构建表格的架构，从而实现表格的识别与提取。

文字流结构是一种由文字信息节点组成的链式结构，每个节点包含一个原子字符串以及相关信息，如位置坐标、包围盒信息、文字信息等等。基于文本元素的表格识别就是将这些一团团的以文字流节点与纵横交错的线条元素，通过一定的识别算法组织起来，成为具有表格编码特征的通用形式，进行表格重现，方便复用和编辑。

与传统基于图像表格呈现方式和存储结构迥然不同，以往的经验很难应用到 PDF 文档表格的识别中。传统的表格提取方法如OCR识别尽管效果较好但只是在视觉上完全还原元文档中的结构，并没有实现表格信息的结构化，像同一单元格中的换行问题还需要设置新的规则使其成为结构化表。

对 PDF 表格信息提取与再利用不仅可以提高日常办公和学习效率，也可以为日后进行数据挖掘和语义分析提供良好的研究基础。研究日常使用的表格信息特征具有很强的实用性。

在此背景下，本文提出来一种针对富含多表格的PDF格式文件根据文本元素的特征进行表格识别与提取的方法，用在上市公司年报财务表格数据的提取上，具有较好的效果。

2 数据来源

本文的课题来源于CCKS2019公众公司公告信息抽取评测任务，其中任务一为表格中的信息点提取，使用上市公司的年报文件，提取其中财务报表部分的多个表格。作为知识图谱构建的基础，结构化数据是必不可少的，因此，通过自动化的技术从公告中提取关键信息，将非结构化文本转化为结构化数据是本课题的目标。本文的主要内容是提出一种表格识别的方法，从年报中提取出想要的表格内容，并使用训练数据验证方法的有效性。

本文的数据来源于评测任务给出的893个上市公司年报PDF文件，其中包括596个半年报PDF文件，297个年报PDF文件，上市公司的半年度报告、年度报告包括从公司概况和经营情况到具体的财务数据及附注内容，其中的财务报表内容是人们关注的重点。除了PDF文件之外，评测方还给出了已经提取出的每个年报对应的提取出的JSON格式的结构化财务报表信息，用以对提取结果进行校验与评价。

3 工具介绍

PDFMiner是一个可以从PDF文档中提取信息的工具。与其他PDF相关的工具不同，它注重的完全是获取和分析文本数据。PDFMiner允许获取某一页中文本的准确位置和一些诸如字体、行数的信息。

首先需要解析PDF 文件，PDFMiner包含的类及其功能如下：

PDFParser：从一个文件中获取数据；

PDFDocument：保存获取的数据，和PDFParser是相互关联的；

PDFPageInterpreter：处理页面内容；

PDFDevice：将其翻译成你需要的格式；

PDFResourceManager：用于存储共享资源，如字体或图像。

使用PDFMiner进行文档的解析，可以完整的解析整个文件，识别出其中的文本框、表格等结构，PDFMiner中各个类的工作原理如下图所示：

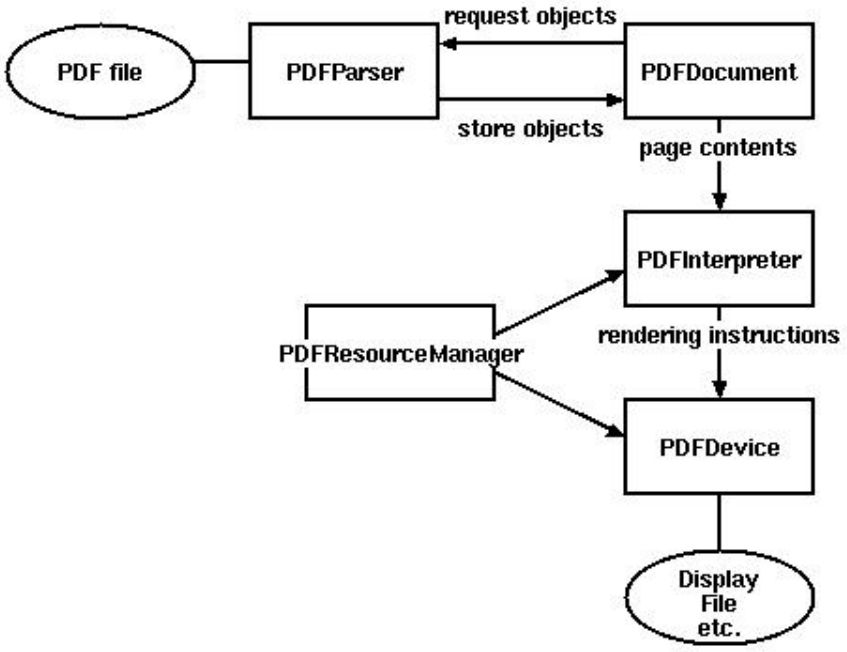

Fig. 1. PDFMiner中各个类的工作原理

由于解析PDF是一件非常耗时和内存的工作，因此PDFMiner使用了一种称作lazy parsing的策略，只在需要的时候才去解析，以减少时间和内存的使用。要解析PDF至少需要两个类：PDFParser和PDFDocument，PDFParser从文件中提取数据，PDFDocument保存数据。另外还需要PDFPageInterpreter去处理页面内容，PDFDevice将其转换为我们所需要的.PDFResourceManager用于保存共享内容例如字体或图片。

在解析PDF文件时，比较重要的是页面布局Layout，主要包括以下这些组件：

LTPage：代表整个页面。可能包含子对象，如LTTextBox，LTFigure，LTImage，LTRect，LTCurve和LTLine。

LTTextBox：表示可以包含在矩形区域中的一组文本块。请注意，此框由几何分析创建，不一定代表文本的逻辑边界。它包含LTTextLine对象的列表。get_text（）方法返回文本内容。

LTTextLine：包含表示单个文本行的LTChar对象列表。根据文本的书写模式，字符可以水平或垂直对齐。get_text（）方法返回文本内容。

LTChar：代表每个字符

LTAnno：将文本中的实际字母表示为Unicode字符串。请注意，虽然LTChar对象具有实际边界，但LTAnno对象不会，因为这些是“虚拟”字符，由布局分析器根据两个字符（例如空格）之间的关系插入。

LTFigure：表示PDF表单对象使用的区域。PDF表单可用于通过在页面中嵌入另一个PDF文档来呈现图形或图片。请注意，LTFigure对象可以递归显示。

LTImage：表示图像对象。嵌入的图像可以是JPEG或其他格式，但目前PDFMiner并不太关注图形对象。

LTLine：代表一条直线。可用于分隔文字或图形。

LTRect：表示一个矩形。可用于构图其他图片或图形。

LTCurve：表示通用贝塞尔曲线。

下图中展示了一页PDF的页面布局Layout：

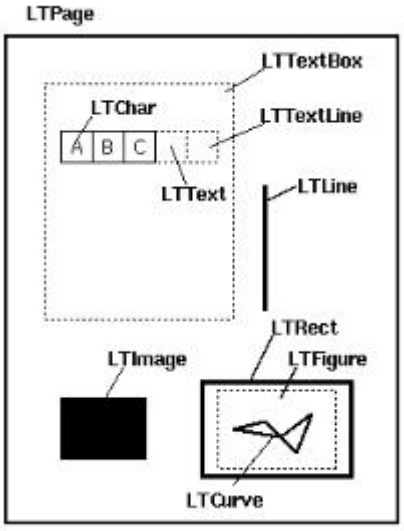

Fig. 2. PDF中的页面布局Layout及其名称

4 表格提取方法

4.1 PDF文本元素识别

借助PDFMiner可以对PDF本身的文本结构进行解析，解析后的PDF文件结构自顶向下的层次如下图所示：

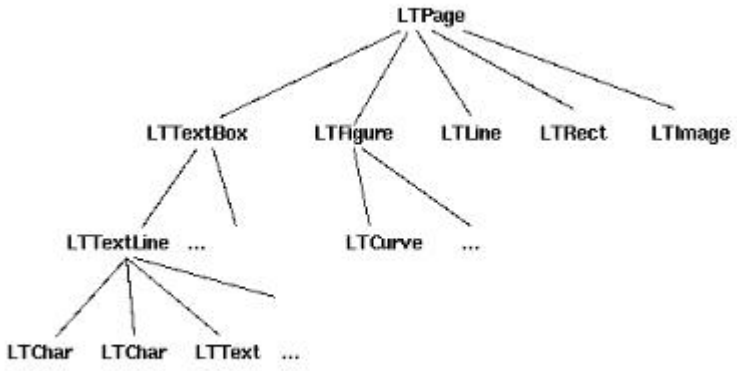

Fig. 3. PDF中的文本元素结构树状图

对于不包含图片的PDF，其中LTTextBoxHorizontal为文本段，LTTextLineHorizontal为文本行，LTCchar为单个的文本字符，将LTTextBoxHorizontal, 展开为LTTextLineHorizontal的集合，再将LTTextLineHorizontal展开为LTChar的集合，就得到了每一个字符的信息，他们都包含bbox，font等属性，表示其位置信息和字体等信息；对于PDFMiner来说，它将每一条线都看做一个rect对象，因此我们需要对每一个LTRect对象做处理使其成为单元格的一条线，若该对象的长大于宽，则看作是横线，反之，将其看做竖线，并增加其横竖向信息作为其新的位置特征。

这样，我们就得到了一个页面所有字符流和线条流的信息，由于进行表格识别与提取需要使用每个字与线的绝对位置信息来判断单元格位置及单元格内容，因此只保留bbox信息来作为位置特征。

4.2 单元格识别与数据填充

借助PDFMiner，可以获得每个字的位置信息以及所有的线条信息，下面将介绍如何利用这些基本的特征信息还原出真实的表格。

将TextBox中的文字全部解析为单个LTChar，其中每个LTChar都使用其位置信息bbox，在视觉上我们可以根据线的位置判断其为表格，但在解析文本的过程中发现，有些在视觉上相连的线，其位置坐标并不一定是首尾相连的，甚至有的表格是缺少完整的单元格的线的，因此，仅利用所有线的信息合成表格再填充文本具有一定的缺陷，容易造成表格出现缺行或者缺少文字的问题出现，需要使用新的方法来进行识别与提取。

与先定位线再定位文字相反，本文使用用字符位置确定其所在单元格边界的方法，从而实现了逻辑上单元格的定位，同时也实现了内容的填充。

该方法首先对每个字符进行其外围表格线条的查找与定位，在X轴方向，寻找位于其左边的竖直线条的最大横坐标和位于其右边的竖直线条的最小纵坐标，在Y轴方向，寻找位于其上方的最水平线条的最小纵坐标和位于其下方发水平线条的最大纵坐标，这样，该字符所在的单元格也就是一个矩形就被定位出来了。

在对每个字符进行外围表格线条查找时，由于每个字的位置信息其所在位置的左下角与右上角的坐标，及该字符的包围盒，以三个点对该字符进行边界寻址：

（1）以该字符包围盒左下角的坐标为基准；

（2）以该字符包围盒右上角的坐标为基准；

（3）以该字符包围盒的中心点位置坐标为基准；

若以三种边界寻址方式寻找的结果若相同，则可以准确定位该点所在的单元格。

除此之外，在少数情况下，有的PDF是没有线条信息的，因此需要进行栅格化处理，对每行字符进行排序，并计算相邻字符的间隔，对于间隔超过阈值的情况，将两个字符划分为不同的单元格中，同时，也就确定了单元格的位置。

以这种方式对所有的字符寻找其所在的单元格，遍历完之后，我们就可以得到所有字符与其所在单元格的一一映射，通过合并相同单元格的信息，可以得到该页中的所有单元格信息，其中每个单元格会对应多个字符的集合，将这些字符集合按照坐标位置进行排序、合并即得到该单元格的文字或数据信息。

4.3 表格分离与跨页表格合并

通过单元格识别与数据填充，我们得到了每一个单元格与内容的对应，但这时的所有单元格还是无序的，各个单元格之间还是没有关联的无结构数据，我们需要根据单元格的位置信息找到其的结构化关系。

每个单元格的纵坐标的信息代表了每个单元格行信息，横坐标的位置代表表格的列信息，根据行列的位置，我们可以将在同一个表中的单元格进行行列的排序，形成结构化的表。

对于同一个表中的单元格，行与行之间是连续的，即每行的单元格的纵坐标是连续的，在对所有表格的纵坐标信息进行排序后，我们只需要寻找出现断点的纵坐标所在的位置，即可以分割属于不同表格的单元格。

对于某些类型的公告文件，其中会包含较多的大型表格，如财务报表，对于跨越页的表格，首先，我们在解析每一页的时候都可以获得该页的位置信息，根据解析出的所有元素的位置坐标，我们可以或者页受第一行的坐标与页尾最后一行的坐标，对于每一页，若第一行非LTRect元素，则按照流程提取表头，并提取表格信息进行存储；若其顶端的第一行单元格信息与第一行元素的位置信息重合，则将其链接到上一页，若上一页最后一行为表格则将其存储到上一页表格的后方。

4.4 表名提取

对于每一个表格，都有对应的表名，可以方便读者找到表格所在的位置。在我们分割表格时，根据出现断点的纵坐标信息，就可以找到表名所在的纵坐标点的位置，并将表名提取出来，这里要注意点的是，表名是表格外的内容，需要寻找表格开始的纵坐标上方的最近的一行，即纵坐标差距最小的一行，这样就可以提取出该表对应的表头了。

根据提取出的表名与表格信息，将其存储为格式化数据，对于一个PDF文件中的所有表格，本方法都可以寻找到每一个表格，并与表名对应，使得信息保存与后续提取使用更有效率。

5 实验结果

本文使用上市公司年报作为待提取的PDF文件，通过本文方法，可以将所有的表格进行提取，在上市公司的年报中，均包含财务报表信息，根据表名，对资产负债表、利润表、现金流量表进行提取，在所有的PDF中均得到了完整的表格信息。

在CCKS的评测任务中，使用20篇年报进行特定表格的信息点的提取，对于提取的结果的评价，任务采用正确率、召回率和F1值作为评价指标，内容如下：

正确率 = 提取出的正确信息点数 / 提取出的总信息点数

召回率 = 提取出的正确信息点数 / 样本中的总信息点数

F1值 = 2 * 正确率 * 召回率 / (正确率 + 召回率)

最终，任务使用F1值作为最终的评测标准，本文方法提取结果的F1值为92.7%。

6 总结与展望

本文提出的基于PDF文本元素的表格信息提取方法，通过借助pdfminer解析文本元素，得到页面中的文字内容、文字位置、线条方向、线条位置与长度等信息，能够有效的对PDF 文本中的表格进行识别与内容填充，即使出现线条缺失的情况也可以使用字符的位置信息进行逻辑上单元格的生成。同时，该方法还实现了跨页表格的合并与表名的提取，可以将提取的表格存储成结构化形式，实现了非结构化数据到结构化数据的转化，使得表格提取准确性更高。

通过在上市公司年报中提取财务报表数据，证明了本方法的有效性和实用性。

参考文献

1. ZHANG Y,YU S. Extraction and removal of frame line in form bill［J］. Journal of Computer Research and Development,2008（45）：1000-1239.
2. ZHOU S,ZHAO J. Rapid form frame-line detection with arbitrary skew angle［J］. Computer Engineering，2008（34）:1000-3428.
3. 中国农业气象编辑部. Word 文档中三线表的处理技巧［J］. 中国农业气象（28）:274-354.
4. 张伯. 基于 PDF 文字流的表格识别技术的研究［D］. 北京：北京工业大学,2010
5. GIULIA SAVIO. Tabula-py［EB/OL］. https：//github.com/chezou/ta-bula-py
6. 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.