北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

版式电子文档表格自动检测与性能评估

房婧[1]　高良才[1]　仇睿恒[1,2,3]　汤帜[1,†]

1. 北京大学计算机科学技术研究所，北京 100080; 2. 数字出版技术国家重点实验室，北京 100080; 3. 中关村科技园区海淀园企业博士后科研工作站北大方正集团公司分站，北京 100080; † 通信作者, E-mail: tangzhi@pku.edu.cn

摘要　针对版式电子文档的特点，提出一种表格线分割符和表格文本的布局特征相结合的表格定位方法，并且对中英文档均有效。此外，针对缺少表格定位自动评估体系，构建了一个初具规模的公开数据集，由中英文版式页面等比例组成，对其标注基准结果，并针对移动阅读应用场景提出一套评估准则。实验部分通过和现有两个开源表格定位项目的比较，验证了新提出的表格定位方法的有效性和评估体系的实用性，特别是对中文数据集获得了较好的结果。

关键词　版式文档；表格定位；表格检测；自动性能评估

中图分类号　TP391

Automatic Table Boundary Detection and Performance Evaluation in Fixed-layout Documents

FANG Jing[1], GAO Liangcai[1], QIU Ruiheng[1,2,3], TANG Zhi[1,†]

1. Institute of Computer Science & Technology, Peking University, Beijing 100080; 2. State Key Laboratory of Digital Publishing Technology, Beijing 100080; 3. Founder Group Substation, Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing 100080; † Corresponding author, E-mail: tangzhi@pku.edu.cn

Abstract　The authors propose a novel and effective table boundary detection method via visual separators and geometric content layout information, which is effective for both Chinese and English documents. Additionally, due to the lack of automatic evaluation system for table boundaries detection, the authors also provide a publicly available large-scale dataset, makes ground-truth and proposes mobile reading oriented performance measurements. Evaluation and comparison with two other open source table boundary detection projects demonstrates effectiveness of the proposed method and practicality of the evaluation suit.

Key words　fixed-layout document; table location; table detection; automatic performance evaluation

近年来，随着国内数字出版行业的迅速发展和移动终端的日益普及，移动阅读已经成为数字阅读的最重要渠道之一。版式文档(如 PDF, CEB, PS 等)由于呈现效果具有平台独立性，已作为多数数字出版物的最终发布形式。也正是由于该特征，为移动数字出版技术带来了新的挑战：移动终端屏幕较小，原文档大小的版面直接阅读文字过小，或需要放大后反复拖动屏幕，阅读体验较差，这就对版式文档提出“流式化”的需求。流式文档在不同设备上重新排版时，通常以逻辑部件为单元，如文本段落、表格以及图形等，而传统的版式文档缺少这些逻辑结构信息。因此，版式电子文档的结构信息抽取，对移动数字出版行业有重要意义。

表格作为一种非常重要的结构化页面元素，以紧凑的二维结构存储和展现数据，遍布于科技文献、财经报告等各类文档中。在版式电子文档中，表格内容以基本的文本字符、图像元素按照坐标信息排布在页面上。如果不对表格区域进行识别，仅按

国家重点基础研究发展计划(2012CB724108)资助
收稿日期: 2012−06−03; 修回日期: 2012−08−15; 网络出版时间: 2012-10-26 17:04
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.008.html

照一般的文本段落抽取[1], 其单元格文本内容将被识别为众多小的段落参与到流式重排中, 显示在阅读设备上, 既影响正文的阅读连贯性, 又不能直观地体现表格的丰富结构信息。移动阅读对版式电子文档中表格识别的需求主要体现在两方面: 首先, 正确定位表格区域, 将整个表格显示到阅读终端, 允许整体放缩; 更近一步地, 识别行、列、单元格的布局和逻辑关系, 能够对大的表格在保持逻辑不变的前提下拆分折叠, 显示或隐藏某(些)行和列。其中表格区域定位是第一步也是最重要的一步。目前仅有少量文献针对版式文档提出相关方法, 并且都不能直接适用于中文文档。因此, 本文针对版式文档的特点, 提出一种表格线分割符和表格文本的布局特征相结合的表格定位方法, 并且对中英文档均有效。

此外, 随着表格定位方法的不断提出, 如何自动、客观的评估和比较不同算法在不同应用场景下的优劣同样具有重要意义。目前表格识别自动评估主要存在三方面问题。1) 缺少公开标准数据集, 特别是中文文档数据集: 现有文献在实验结果部分通常各自选择小样本数据集进行测试, 不同算法难以在相同数据集上比较结果。2) 缺少评测基准: 现有方法大多通过人工评测, 存在一定的主观性, 且耗费时间和精力, 因而需要标准数据集有正确的输出结果作为基准, 便于自动化测试。3) 缺少适用的评估准则: 现有方法大多采用信息抽取领域最常用的准确率和召回率作为评估准则, 然而表格定位错误类型不能严格意义的归类为假阴和假阳, 还包含区域的部分被定位等, 需要更细粒度的评估准则。因此, 本文构建一个初具规模的公开数据集, 对其标注基准结果, 并针对移动阅读应用场景提出一套细粒度评估准则, 整个评估体系提供免费下载(http://www.founderrd.com/marmot_data.htm)。在实验结果部分, 采用该评估体系对本文提出的表格定位方法和另外两个开源的PDF表格定位方法进行评估和比较。

1 相关工作

1.1 表格定位

近二十年来, 研究人员在表格定位领域发表了大量文献(见 Zanibbi 等[2]和 Silva 等[3]的综述文章), 然而大多方法是针对像素图像文档提出的, 代表性的有 Kieninger 等[4]的先驱研究工作 T-Rect 系统; 针对中文图像文档的有郑冶枫等[5]和史广顺等[6]分别提出的图像文档基于有向单连通链的表格框线检测算法和基于数据分隔符、线条连通区域的表格结构定位算法。少量针对版式文档的表格定位方法出现在 2005 年之后, 包括: Yildiz 等[7]提出的 pdf2table 系统, 通过检测和合并含有多个文本段的候选表格行实现定位; Oro 等[8]将页面的文本行根据所含文本段的数目分为 3 类(文本行、表格行和待定行), 合并连续的表格行和待定行得到表格区域; Liu 等[9]提出的表格检索系统 TableSeer 包含定位功能, 合并表格标题之后出现的连续“稀疏行”, 该方法基于表格一定含有标题的假设, 在提高准确率的同时, 由于无标题的大量表格无法被检测到而降低了召回率。上述方法思想类似, 都基于表格行中存在多个被空白隔开的单元格的布局特征, 局限性主要体现在: 1) 前两种方法只能处理单栏页面表格; 2) 文本段的划分受到空白大小阈值的影响; 3) 当同一页面中存在多个临近表格时, 难以互相区分; 4) 布局不规则的表格(如单元格内容稀少的表格)检测效果不佳; 5) 多出现表格被部分定位的情况。更重要的是, 上述方法均不能直接应用于中文版式文档的处理。

1.2 数据集

尽管表格检测得到了近二十年的研究和关注, 但是公开公用的数据集仅有 UW-3[10], 它是页面布局分割领域最常用的公开数据集, 其中包含一部分表格区域的基准, 共有 110 个页面, 包含 215 个表格。该数据集并非针对表格识别领域专门提出, 因此可用部分数量较小, 且仅针对图像页面。目前还缺少版式文档表格定位领域公开的数据集, 特别是中文文档的数据集。

1.3 评估准则

现有的大多表格定位研究均以准确率和召回率描述结果, 并且一个表格区域定位的正确性与否由测试人员人为判定。考虑到表格定位结果存在部分定位的情形, 使得人为判定不可避免存在主观性和难以复现性。已有研究人员提出更具针对性的评估准则。Hu 等[11]提出基于编辑距离的方法, 以“插入”、“删除”和“替换”分别描述表格区域误识别、未识别、拆分和合并错误, 用被操作的行数表示代价函数。但该方法的局限性在于, 最后的输出只有唯一的代价值, 不便体现各种错误类型发生的情况。Silva 等[12]提出的完全性(C)和纯粹性(P)的评估准则存在同样的问题。Li 等[13]通过计算检测到的表格区

域与基准中表格区域面积重叠的大小进行评估, 但不同算法对表格区域定义可能存在差异(例如是否包含表格标题, 是否紧贴文本内容或包含表格框线), 从而造成面积的差异, 影响评估结果。Wang等[14]采用检测到的表格区域与基准表格面积的交集分别与两个表格面积的比率作为评估准则, 同样存在面积多义性和不便反映错误类型发生细则的问题。

2 表格定位方法

本文提出的表格定位方法主要包含 4 个步骤: 1) 页面布局分析(分栏分析), 意在处理分栏页面的表格定位问题; 2) 表格线抽取, 解析 PDF 获取图形内容流, 分析表格线; 3) 依据文本坐标信息分析布局特征; 4) 表格定位, 预先将步骤 2)和 3)中获得的表格线和文本块互相验证, 定位后在后处理过程中去除可能被误识别为表格的区域。

2.1 页面布局分析

页面布局分析对表格定位的重要性体现在两个方面: 1) 对于双(多)栏布局的页面, 栏间空白与表格列与列之间的空隙表现出相似的布局特征, 使得整页内容容易被识别为一张表格; 2) 表格既可能位于某一栏内, 也可能贯穿多栏, 需要保证一个跨栏的表格不会被分栏拆成两(多)部分, 而单栏内的表格不与其他栏的正文文本合并到一起。本文利用单页页面上的前景空白和多页文档的分栏位置相似性来识别页面分栏。具体地, 采用 Breuel 等[15]提出的页面空白分析方法, 以页面的字符元素作为障碍寻找空白条, 并定义约束阈值过滤出较长的竖直空白。考虑排版习惯, 多页文档的连续页分栏特征相似, 因此我们采用统计的方法, 从连续页的竖直空白中确定页面分栏, 具体地, 选择当前页的前后连续 N 页(本文 N=5), 统计所有竖直空白在 X 方向的累积高度, 选择明显的峰值作为页面分栏位置, 仅保留各页面中在这些位置的纵向空白条。

2.2 表格线抽取

版式文档的表格线多数由图形元素构成, 每个图形对象包含一组绘制路径信息, 例如 m (移动至)、l (划线至)、re (绘制矩形)及 c, v, y (三次贝塞尔曲线), 每一步绘制指令都有精确的坐标描述, 可以获得精确绘制轨迹以及填充、颜色、端点样式等属性[16]。绘制结果有直线、贝塞尔曲线、矩形等。由于表格线通常为直线段, 我们首先过滤出直线绘制指令, 并将矩形对象拆成 4 条直线对象, 然后通过坐标拼接和聚类, 将绘制序列拼接成视觉的直线段。考虑到页面元素的绘制受到裁剪区的限制[16], 因此去掉位于裁剪区之外的图形元素, 同样去掉白色填充或勾边的页面不可视元素, 避免干扰。由此得到的表格线比图像文档通过霍夫变换等得到的表格线更加精确。

2.3 表格文本布局特征分析

通过 PDF 文档解析引擎, 可以得到页面的文本内容流, 具体到单个字符, 可以获得文本、最小包围矩形、字体和坐标等属性。首先将这些基本的页面字符元素按照竖直方向上包围矩形的交叠, 在页面内(或者多栏页面的单栏内)划分成文本行, 在单行内根据字符间距设定的阈值形成小的文本段。如图 1(a)中的黑色矩形框所示, 正文内容在页面宽度范围内(或者栏内)将形成完整的文本行, 而表格的内容由于单元格之间的空白间隙较大, 一行将被划分为多个小的文本段。

表格的文本布局, 最显著的特征就是单元格在行、列上按照某种方式对齐, 呈现二维格状结构。特别是表格列, 每一列上的单元格不管按照何种对齐方式, 单元格之间都有水平方向的交叠, 并且列与列之间互不干扰, 由空白分隔开, 即不同列上的单元格之间没有交叠。考虑到表格和页面正文内容的排版实际上遵从相同的规则, 即向右向下的顺序排版, 向右成行、向下成列, 本文提出一种基于图深度搜索的方法分析页面内容, 以形成正文文本块和表格文本块。具体地, 以页面中的文本段构造邻接矩阵, 按照行间向下, 行内向右的顺序深度遍历文本段, 在遍历过程中, 当不满足如下两个条件之一时, 中断聚合大的文本块, 以新的未被标记过的小文本段为起点继续遍历。记当前文本段为 A,下一行中的文本段为 B, 两个条件是: 1)A 与 B 字体相同; 2) A 与 B 水平方向存在交叠, 并且 A 不与 B 所在行的其他文本段交叠, B 也不与 A 所在行的其他文本段交叠。图 1(a)中的黑线是绘制的遍历示意图, 每次中断后形成页面上的文本块如图 1(b)所示。

2.4 表格定位

我们首先对页面中的直线段与文本内容互相验证, 得到候选的表格线和表格文本块, 这是因为 2.2 节抽取的直线既可能是表格线, 也可能是其他逻辑元素的一部分, 例如图形、公式、页眉页脚线; 而预先过滤正文文本内容, 可以使得表格检测目标更

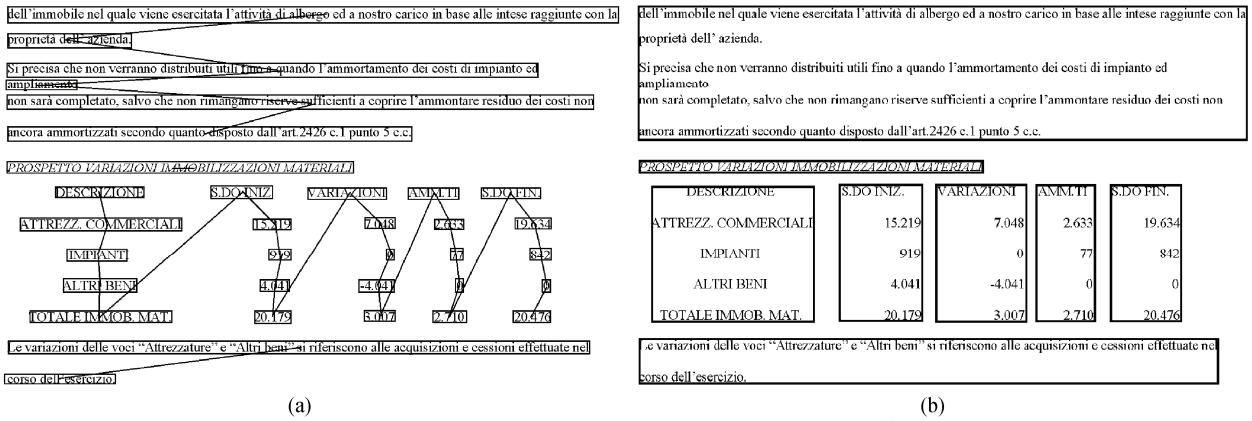


图 1　表格文本布局分析示意图

Fig. 1　Table content layout analysis illustration

加明确，很大程度上去除干扰因素，减少不必要的检测。在此基础上，检测表格区域的边界，最后在后处理过程中去掉部分误识别的表格。

2.4.1　互校验

按照 2.3 节描述的文本块聚合方法，页面中的表格内容将被聚合成不完全的表格列，而正文内容则被聚合成大块文字段落，如图 2 所示。我们首先根据如下特征去掉明显的正文文字块：正文段落的左边界和右边界或者与页面边框直接相邻，或者与分栏空白相邻；而表格列由于数量较多(大于等于 2)，不满足该规则(见图 2)。在去掉正文文字块的基础上，可以去除远离留存文本块的直线(页眉页脚线等)，并且考虑到表格线与表格文本的关系：水平表格线的上下方通常分别存在多个文本段，并且相应的小段水平方向存在交叠，该特性同样能够对水平直线起到校验作用。

2.4.2　检测表格区域

首先将检测的水平表格线按照长短排序，从最长的开始，判断它是否与多条竖直表格线相交，如果是则将相交并且最长的竖直表格线作为表格区域的高度，水平线的长度作为表格宽度，形成表格区域，位于该区域内的其他水平线和竖直线从待判断的表格线中删除，避免形成交叠或嵌套的表格。该过程的结束条件是所有的水平表格线都经过同样遍历。对于无线表格，则在栏内横向贯穿合并候选表格列文本块；对于仅有水平线的表格，在无线表格区域附近寻找位置相交或者距离相近的水平线，以确定表格区域的宽度，当多条水平线存在时同样可以确定表格区域高度。3 种情形对应的样例执行结果如图 3 所示。

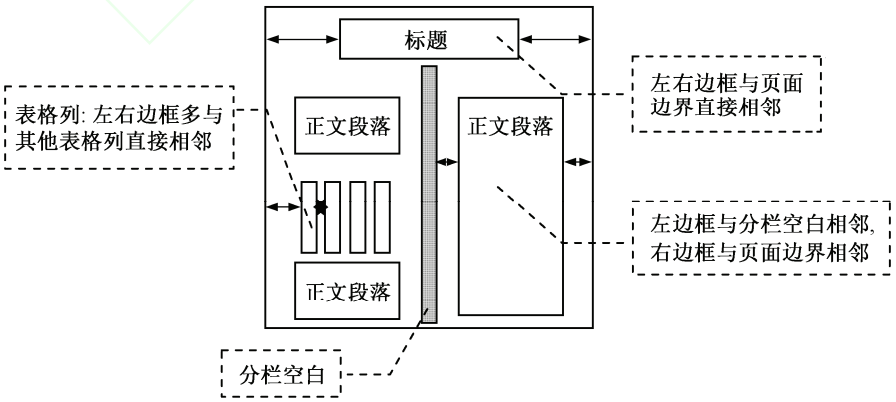
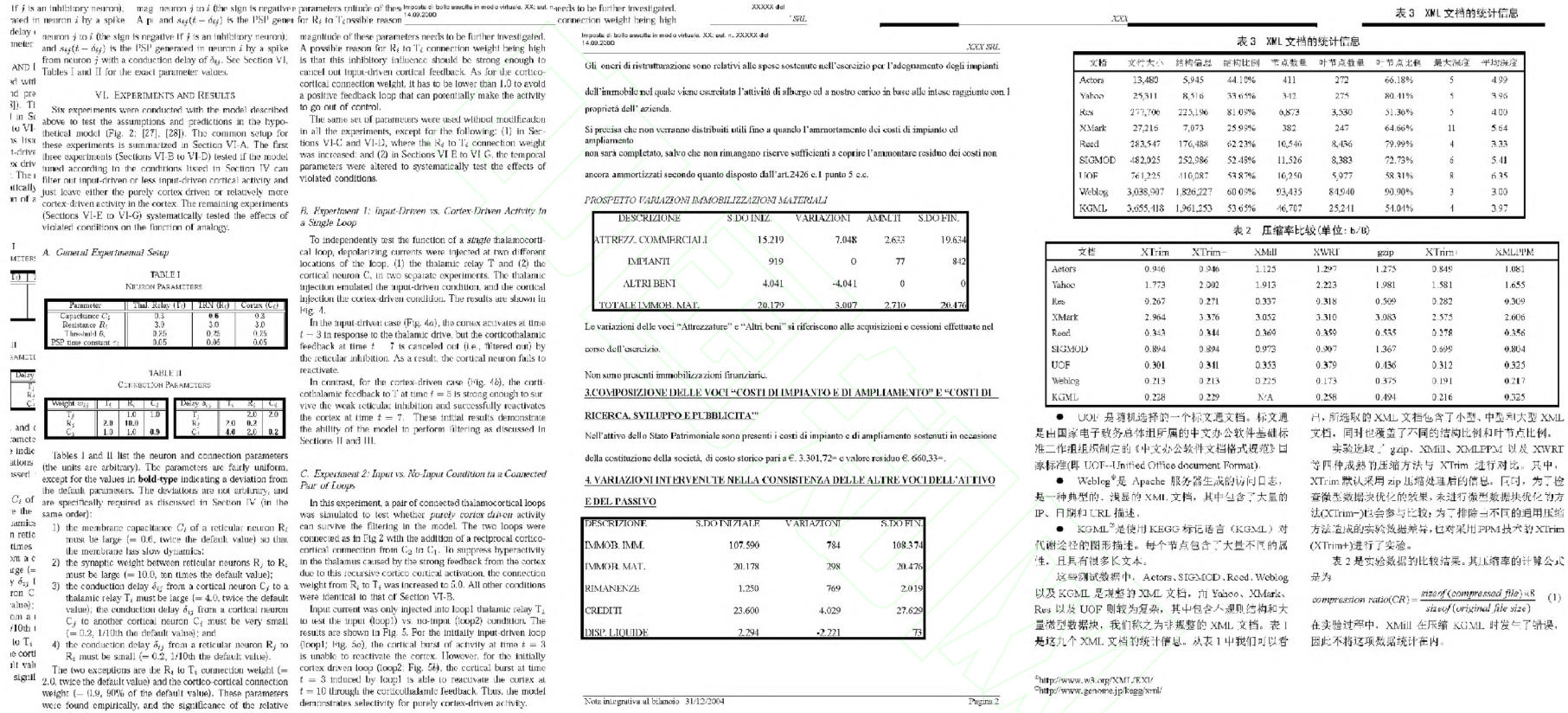


图 2　正文段落文本与表格区域文本特征

Fig. 2　Features of the paragraph content and table content



(a)

(b)

(c)

图 3　全线表、无线表和三线表区域检测样例

Fig. 3　Features of the paragraph content and table content

2.4.3 后处理

由于实际文档中表格布局的多样性, 不可避免的会造成一些非表格的部件被误识别为表格, 经过前述步骤, 文档中一些矩阵或者数据图可能被误识别, 因此需要根据表格的格状布局对误识别情况进行排查。后处理规则主要包括以下几点: 1) 表格区域内至少包含两行两列; 2) 区域内不包含曲线图形元素(可能出现在其他图形部件中); 3) 区域内不会只包含竖直方向的直线(排除矩阵)。

3 自动评估

3.1 数据集

本文已经构造了一个初具规模的数据集, 包含2000 个 PDF 页面, 中英文按照 1∶1 比例构成, 中文样例选自方正阿帕比数字图书馆随机选取的 120 本电子书; 英文样例选自于1500篇网络提取的会议和期刊论文, 发表年限介于 1970—2011 年之间。所选样例页面布局多样, 中文以单栏布局为主, 英文则兼顾单双栏布局。表格本身的布局也具有多样性, 从有线表到无线表, 横向表到纵向表, 栏内表到跨栏表, 等等。对于中英文样例, 又分别按照 1∶1 的“正例”和“负例”构成, 前者指的是至少含有一个表格的页面, 后者则是不含表格的页面, 以便充分挖掘误识别错误类型。数据集的基准结果由 15 人使用我们开发的半自动化标注工具进行标注, 并且每个页面的标注结果都经过 2 人校验通过。页面基准由3 部分组成: 1) 以 XML 格式描述的被标注基准结果; 2)600 dpi 的页面原图像; 3) 页面基本对象(字符、图形、图像元素)的 XML 描述。其中, 第一部分基准结果只包含了评估比较所必要的信息, 如对象编号和包围矩形, 具体的对象属性等信息则可以在第三部分中根据唯一的编号获取, 而第二部分的页面原图一是为了方便针对图像文档的提出算法使用, 二是可以直观地显示页面内容。

3.2 基准描述

基准由 XML 格式描述, 我们定义一个表格包含三部分内容: 表格标题、表格体和表格脚注, 在 XML 中分别对应标签 table caption, table body 和 table footnote, 其中 table caption 和 table footnote 为可选内容。这 3 个逻辑部件均由 textline 标签构成, table body 的 textline 指的是单元格粒度的小文本段, 其他两者的 textline 则为文本行粒度。textline 由 character 标签构成。因此基准总体呈现树形层状结构, 父节点和子节点的关系通过 ID 维护。此外, 页面上的其他逻辑结构也类似地经过标注和描述, 如正文段落、图像、公式等, 当表格区域被误识别时, 可以找到内容所对应的真实逻辑类型。

3.3 评价准则

为了提供细粒度并面向应用的评价准则, 本文首先定义 6 种表格定位错误类型: 1) 误判(fake), 将页面其他非表格部件检测为表格; 2) 遗漏(missed), 真实表格没有被检测到; 3) 合并(merged), 多个表格由于位置临近且布局相似而被错误的合并; 4) 拆分(splitted), 一个表格被拆成多个部分; 5) 部分(reduced), 表格只有一部分内容被检测出来使得区域范围比实际小; 6) 扩张(amplified), 临近表格的其他页面对象被错误地扩张到表格区域内使得范围增大。在此基础上, 以移动阅读应用场景为例, 由于该应用重视内容的完整性和连贯性, 被误判的对象如果是列表、矩阵或是其他与表格特征相似的逻辑部件, 严重程度会降低。再如, 一个表格被水平拆分比竖直拆分的严重程度小很多, 因为在重排时并不影响阅读的连贯性, 等等。因此我们将上述 6 种错误类型细化为 13 种错误标签, 如表 1 所示, 并赋予[0,10]之间的惩罚分值, 数值从小到大表示严重程度由低到高。错误标签和惩罚分值均允许根据不同应用需要修改和补充。

此外, 对一个表格的评价不但考虑错误类型的影响, 同样需要对该错误类型进行定量描述。例如, 尽管误判的影响比扩张要严重, 但如果只是页面上一块很小的区域被误识别为表格, 相比于一个虽然正确包含了表格区域, 但也囊括了大量其他页面内容的情况, 前者的影响要小。因此, 本文提出如下定量描述作为各个错误类型的系数。定义 N 表示数量,

表 1 针对移动阅读应用场景定义的错误标签
Table 1 Table detection error types for mobile reading application

6 种错误类别	针对移动阅读错误类别细化
fake	fake_figure; fake_matrix; fake_list; fake_mix
amplified	amplified_tabaccessory; amplified_matrices; amplified_mix
splitted	splitted_horizontal; splitted_vertical
merged	merged_horizontal; merged _vertical
reduced	reduced
missed	missed

PO 表示页面对象, G 表示基准集合$\{G_1, G_2, \ldots, G_i, \ldots, G_n\}$, A 表示实际检测的表格集合$\{A_1, A_2, \ldots, A_j, \ldots, A_m\}$, 则有: 对于 fake 和 missed 类型, 误判或者遗漏的表格中对象占页面所有对象的比例 $\mathrm{coe}_{\mathrm{fak}} = \frac{N_{A_j}}{N_{\mathrm{PO}}}$, $\mathrm{coe}_{\mathrm{mis}} = \frac{N_{C_i}}{N_{\mathrm{PO}}}$, 对于 reduced 和 amplified 类型, $\mathrm{coe}_{\mathrm{red}} = \mathrm{coe}_{\mathrm{amp}} = \frac{N_{A_j} \cap G_i}{N_{A_j} \cup G_i}$; 对于 splitted 和 merged 类型, N_s 表示一个表格被拆分成的个数, N_m 表示被错误合并到一起的表格数目, $\mathrm{coe}_{\mathrm{spl}} = \frac{1}{N_s}$, $\mathrm{coe}_{\mathrm{mer}} = \frac{1}{N_m}$。

最后, 评测标准包括两部分: 1) 由每个表格分别命中的错误类型(可能对应多种), 统计每种错误类型被命中的表格总数, 即可以评价各种错误类型发生的多少, 发现算法的问题; 2) 每个表格都有一个综合惩罚分值(各错误类型惩罚分值的最大值), 比如认为分值小于 3 的为可接受结果, 最后对整个文档统计可接受表格的数目, 再计算新意义下的准确率和召回率, 如表 2 所示。

表 2　根据可接受表格计算准确率和召回率

Table 2　Notations for evaluation metrics

符号	注释
NR	真实表格的数目
NM	遗漏表格的数目
NA	可接受表格的数目
NFA	误判, 但可接受表格的数目
NFU	误判, 且不可接受表格的数目
准确率	NA / (NR + NFA + NFU − NM)
召回率	NA / (NR + NFA)

4　实验结果

我们使用上节提出的评估系统, 评价和比较本文提出的表格定位方法、pdf2table[8]和 TableSeer[10]中的定位方法在面向移动阅读应用的效果。其中, pdf2table 和 TableSeer 本身并不能直接应用于中文 PDF 文档的处理, 但由于是开源项目, 我们将其文档内容流解析部分替换为我们开发的解析引擎, 一方面能够支持中文处理, 另一方面也保证在输入信息相同的条件下, 比较输出结果更加公平可靠。

表 3 统计了综合评测值和比较结果, 图 4 和 5 分别对应于中文和英文测试集中的细粒度错误统计。可以看出, 3 种算法各有优势, pdf2table 遗漏掉的表格数目最少, TableSeer 检测到最少的非表格区域, 而本文提出的方法, 在移动阅读领域表现最优, 优势主要体现在以下类型表格的定位: 1) 文本布局复杂, 但是具有表格线的表格; 2) 分栏页面中的表格, 跨栏表或者栏内表; 3) 没有标题的表格。而发生错误的主要原因如下: 1) 如果页面分栏恰好和表格列之间的空白条水平方向重合, 表格将被分成左右两部分; 2) 一些呈现格状结构的图形会被错误的检测为表格, 例如流程图, X–Y 数据图等; 3) 无线表有被部分识别或者扩张识别的错误。从表 3 的数据中还可以看出, 中文数据集的最终评测结果要明显优于英文数据集, 这是因为大量英文样例的表格线并非由图形元素构成, 包含着图像元素, 而中文样例的表格线无论从元素组成, 或者在表格中被使用的频率, 都比英文样例稳定, 也更加说明了结合表格线和文本布局特征的方法在处理中文文档上的优势。

在不同的应用场景下, 如果采用的错误类型和

表 3　测试结果比较

Table 3　Experimental results and comparison

方法	英文测试集			中文测试集		
	文献[7]	文献[9]	本文	文献[7]	文献[9]	本文
NR	667	667	667	682	682	682
NM	51	208	150	63	249	91
NA	261	232	374	223	192	547
NFA	22	1	37	5	0	4
NFU	111	27	21	18	8	19
准确率	0.35	0.48	0.65	0.35	0.44	0.89
召回率	0.38	0.35	0.53	0.34	0.28	0.80

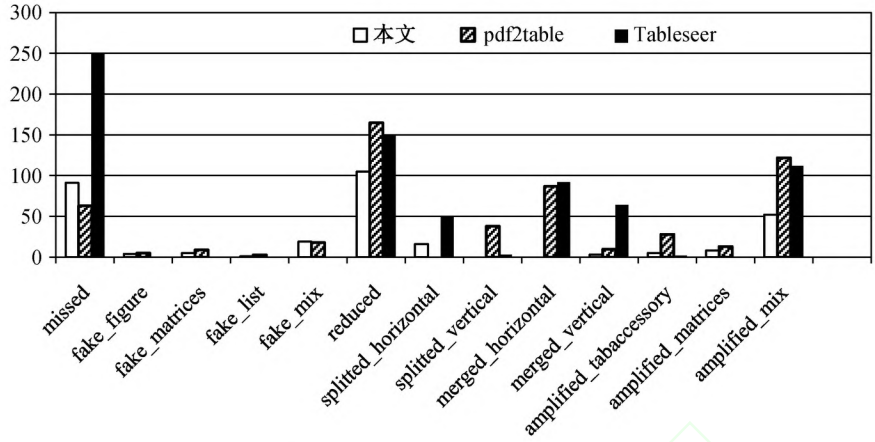


图 4　中文测试集错误统计

Fig. 4　Chinese dataset error statistics

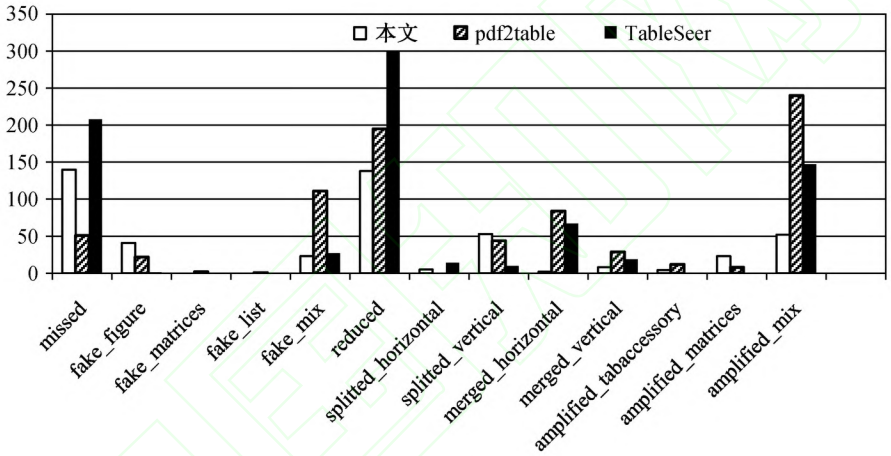


图 5　英文测试集错误统计

Fig. 5　English dataset error statistics

对应的分值发生变化，则可能得到不同的结果。从本实验结果可以看出，目前各个算法都未能达到满意的结果。本文评估系统中的标注基准，以及实验结果能够对开发或研究人员展示算法的错误类型和量化统计，从而使他们认识到算法的不足，得到改进的方向。

5　总结

本文针对版式电子文档，结合表格线文本布局特征，提出了一种中英文均适用的表格区域自动定位方法。由于文本布局越复杂的表格，通常表格线出现的可能性越高，提出的方法较已有的仅依赖文本布局特征的方法优势越明显。本文还提供了一套表格定位自动评估方案，包括等比例中英文数据集、基准结果及面向应用的细粒度评估准则，所有内容均为公开资源。最后采用该评估体系对本文提出的方法和另外两个开源项目进行评测比较，实验结果显示，本文的方法在移动阅读应用领域能够获得较满意的结果，特别是对于中文数据集的优势更为明显。在未来的研究工作中，我们将在表格区域检测结果的基础上，分析表格内部的物理和逻辑结构，并且提出相应的评估方案。

参考文献

[1] Fang J, Tang Z, Gao L. Reflowing-driven paragraph recognition for electronic books in PDF // Proc SPIE 7874, Document Recognition and Retrieval XⅧ, 78740U, Vol 7874SPIE. San Francisco, 2011: 1–10

[2] Zanibbi R, Blostein D, Cordy J. A survey of table recognition: models, observations, transformations,

and inferences. IJDAR, 2004, 7: 1–16

[3] Silva A C e, Jorge A M, Torgo L. Design of an end-to-end method to extract information from tables. IJDAR, 2006, 8: 144–171

[4] Kieninger T, Dengel A. A paper-to-html table converting system // Proceedings of the 3th IAPR International Workshop on Document Analysis Systems (DAS1998). Nagano, 1998: 68–75

[5] 郑冶枫, 刘长松, 丁晓青, 等. 基于有向单连通链的表格框线检测算法. 软件学报, 2002, 13(4): 790–796

[6] 史广顺, 骆春妹, 王庆人. 普通文档图象中表格结构的自动定位. 中国图象图形学报, 2003, 8(增刊 1): 228–231

[7] Yildiz B, Kaiser K, Miksch S. pdf2table: A method to extract table information from PDF files // Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI 2005). Pune, 2005: 1773–1785

[8] Oro E, Ruffolo M. PDF-TREX: an approach for recognizing and extracting tables from PDF documents // Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009). Barcelona, 2009: 906–910

[9] Liu Y, Bai K, Mitra P, et al. TableSeer: automatic table metadata extraction and searching in digital libraries // Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2007). Vancouver, 2007: 91–100

[10] Phillips I T. User's reference manual for the UW English/Technical document image database Ⅲ [R]. Washington: Seattle University, 1996

[11] Hu J, Kashi R S, Lopresti D, et al. Evaluating the performance of table processing algorithms. IJDAR, 2002, 4: 140–153

[12] Silva A C e. New metrics for evaluating performance in document analysis tasks_application to the table case. IJDAR, 2007, 1: 481–485

[13] Li F, Shi G S, Zhao H. A method of automatic performance evaluation of table processing // Proceedings of the 21st Chinese Control and Decision Conference (CCDC2009). Guilin, 2009: 4021–4025

[14] Wang Y, Phillips I T, Haralick R M. Table structure understanding and its performance evaluation. Pattern Recognition, 2004, 37: 1479–1497

[15] Breuel T M. Two geometric algorithms for layout analysis // Proceedings of the 5th IAPR International Workshop on Document Analysis Systems (DAS 2002). Princeton, 2002: 188–199

[16] ISO 32000–1: 2008. Document management — portable document format part1: PDF 1.7 [S]. Geneva: International Organization for Standardization, 2008