

# 复杂表格数据化中的单元格语义关系识别研究<sup>\*</sup>

林鑫<sup>1,2</sup> 余华娟<sup>1</sup> 闫奕臻<sup>1</sup>

(1. 华中师范大学信息管理学院, 武汉 430079; 2. 湖北省数据治理与智能决策研究中心, 武汉 430079)

**摘要:** 复杂表格能够以简单、直观的方式描述数据, 被广泛应用于各行各业, 然而, 复杂表格具有结构复杂、单元格类型多样、表格文档构成方式不一等问题, 需要进行数据化处理后才能实现共享与复用。因此, 本文构建一种基于无监督学习的单元格语义关系识别模型来实现复杂表格数据化, 首先利用机器视觉技术实现复杂表格分割, 然后基于表格结构和内容相似度识别同模板表格, 在此基础上, 结合表头单元格、说明性单元格、表体单元格3类单元格的取值、位置特点, 设置启发式规则进行单元格语义关系的识别, 最后通过实证研究验证本文的方法能够在复杂表格数据化中取得较高的准确率和召回率, 具有可行性。

**关键词:** 复杂表格; 语义关系; 表格数据化; 机器视觉

**中图分类号:** G202 DOI: 10.3772/j.issn.1673-2286.2022.09.004

**引文格式:** 林鑫, 余华娟, 闫奕臻. 复杂表格数据化中的单元格语义关系识别研究[J]. 数字图书馆论坛, 2022(9): 28-35.

因具有简洁规范、信息密度高、数据关系清晰等优点, 表格被广泛应用于各行各业的企事业单位日常管理、政务工作和科学交流中。与非结构化文档相比, 尽管其具有结构化程度高的优势, 但由于常常以独立文档或文档的组成部分存在, 为实现对此类数据的系统管理、共享、复用与挖掘分析, 还需要重新对其进行数据化处理, 即通过表格区域检测、结构识别、单元格语义关系识别, 将表格从文档中提取出来并实现机器可理解<sup>[1-2]</sup>。其中, 表格区域检测是指从包含表格的文档或图片中识别其边界; 结构识别是指识别表格的布局结构、层次结构, 确定各单元格的具体位置及单元格间的位置关系; 单元格语义关系识别首先需要识别单元格的类型, 单元格的类型包括表头、表体、说明性单元格3种, 并在此基础上, 建立具有对应关系的3种单元格间的语义关联, 即将其转换为表头单元格、表体单元格、说明性单元格的元组形式。

根据表头、表体、说明性单元格间的位置关系, 可

以将表格分成简单结构表格与复杂表格, 前者是指表头单元格集中分布在前1行/列或前几行/列的行列表头表格; 后者是指表头与表体单元格混合分布的表格, 其在实践中也广泛应用, 如科研项目申请书、职称评审表、信息登记表、报名表等。当前围绕单元格语义关系识别, 国内外学者主要围绕简单结构表格展开研究, 对复杂表格关注较少。复杂表格不但结构复杂、单元格类型多样, 而且表头类单元格的取值多样, 既难以复用行列表头类表格的处理方法, 也难以通过简单的词典或规则匹配的方法实现单元格语义关系识别, 因此需要开展专门研究。

## 1 相关研究

表格数据化是实现表格数据的系统管理与挖掘分析的重要基础, 因此受到国内外学者的关注, 相关研究主要从表格区域检测、表格结构识别、单元格语义关系

<sup>\*</sup> 本研究得到国家社会科学基金青年项目“社会网络中基于用户认知结构的知识标注研究”(编号: 17CTQ024)资助。

识别这三个方面展开。

## 1.1 表格区域检测与结构识别

表格区域检测与结构识别在早期主要针对的是具有表格形式标记的文档形态,多将其视为两个任务独立进行,但随着时间的推移,研究与实践关注的重点集中到pdf文档、纸质文件扫描件等不具备表格形式标记的情形下,相关研究多基于表格结构实现表格检测,二者逐渐一体化。早期的研究主要是根据表格特征构建启发式规则来完成表格区域检测和结构识别任务。例如:Watanabe等<sup>[3]</sup>和Hirayama<sup>[4]</sup>对于扫描文档,利用文档的水平线和垂直线来定位表格区域;国内学者Fang等<sup>[5]</sup>提出了一种基于表格结构特征和视觉分隔符的表格区域检测方法,该方法通过解析pdf文件的线条和页面分隔符来获得表格的位置;于丰畅等<sup>[6]</sup>针对pdf格式的学术文献,提出了一种几何对象聚类的表格区域检测方法。随着深度学习的发展,为克服启发式规则法受限于表格数据集的特征明显程度以及数据集的规模大小的缺陷,逐渐将深度学习引入表格区域检测和结构识别过程中,如Schreiber等<sup>[7]</sup>采用改进一种循环神经网络Faster R-CNN进行表格区域检测,Hao等<sup>[8]</sup>通过结合松散规则和改进的卷积神经网络确定表格位置。由于深度学习能够适应的场景非常广泛,因此在通过深度学习定位表格区域和结构识别的研究中,文本特征<sup>[9]</sup>、单元格位置特征<sup>[10]</sup>、表格空白特征<sup>[11]</sup>也被加入模型中,大大提升了表格区域检测和结构识别的精确度。

## 1.2 单元格语义关系识别

与表格区域检测和结构识别相似,单元格语义关系识别的方法也可以分为启发式规则法和机器学习法。前者主要基于表格的结构特征、内容特征等构建启发式规则来完成单元格语义关系识别。代表性研究包括:张建东等<sup>[12]</sup>通过词向量计算进行表格内容文本的划分,识别表格表头和表体;赵洪等<sup>[13]</sup>基于规则完成了表头的语义分析和数值信息的抽取,将政府报表信息表示为六元组形式的结构化数据;张元鸣等<sup>[14]</sup>根据字典抽取非结构化表格文档的表头,在此基础上设计了单值区域与多值区域数据抽取的算法,最终将抽取结果组织为六元组DataStr结构化数据模型;Seth等<sup>[15]</sup>根据表格中所有数据单元格均可以索引到行列表头路径的特征,将

单元格分为行表头、列表头、数据、存根表头和额外信息5类。基于机器学习的技术思路,将单元格语义关系识别视为分类问题,进而采用相应的机器学习算法进行实现,在具体的技术模型选用上,既有传统的机器学习模型(如随机森林<sup>[16]</sup>、决策树、支持向量机<sup>[17]</sup>),也包括深度学习<sup>[18]</sup>、预训练模型<sup>[19]</sup>等。

综上所述,国内外围绕表格的数据化进行了多方面研究,针对不同应用场景提出了多种技术解决方案,具有较强的现实性。然而,在单元格语义关系识别方面,相关研究多以行列表头表格为对象,较少关注表头、表体单元格混合分布的复杂表格。针对这一问题,本文拟采用无监督学习思想,综合机器视觉与启发式规则技术,构建一种面向复杂表格数据化的单元格语义关系识别模型,为复杂表格知识抽取提供支持。

## 2 基于无监督学习的单元格语义关系识别模型构建

实践工作中,复杂表格除了因表头与表体单元格混合分布带来的结构复杂外,常常还具有以下特点:①同模板的表格常常不止一份,如教育部每年会收到数万份同模板的教育部人文社科基金申请书;②表格填写过程中可能会对其结构进行微调,如删除一些不需要填写的空行或者增加行、改变行高、单元格宽度等,但表头类单元格一般不会被删除;③单元格类型上,除了表头、表体两类外,还包括辅助性单元格,用于说明对应单元格的取值范围或填写要求等,如2022年国家社会科学基金申请书数据表第3行的第3个单元格内容为“A.重点项目B.一般项目C.青年项目D.一般自选项目E.青年自选项目”,限定了其左侧临近单元格“项目类别”的取值范围;④表头与表体单元格的关系多样,数量上存在1:1、1:n、n:1等不同情形,位置关系上存在左右、上下两种情形。结合上述特点,拟采用无监督学习方法进行复杂表格单元格语义关系识别模型构建:首先利用机器视觉技术将样本数据中的同模板表格聚合到一起;其次利用同模板申请书表头单元格、说明性单元格取值一致的特点,将表体类单元格识别出来;最后结合表头单元格、说明性单元格、表体单元格的取值、位置特点,设置启发式规则进行单元格关联关系的识别。

遵循这一思路,构建了如图1所示的单元格语义关系识别模型,模型输入为未标注单元格类型的复杂表格数据,输出是表格模板及其对应的单元格语义关系;

处理流程上,包括基于机器视觉的表格分割、基于结构与内容相似度的同模板表格识别和基于启发式规则的单元格语义关系识别三个环节。

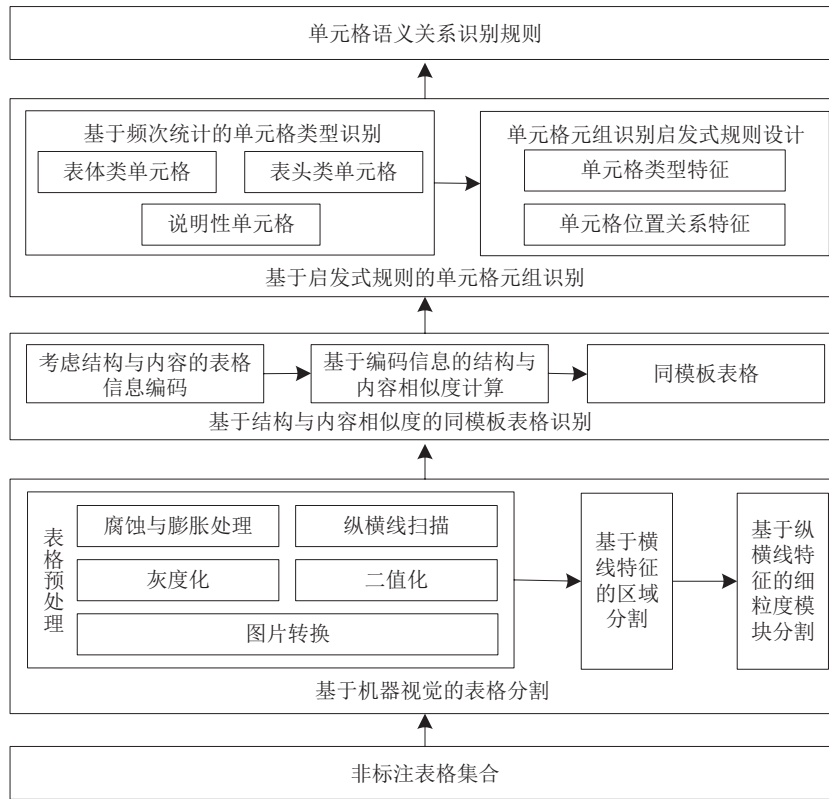


图1 基于无监督学习的单元格语义关系识别模型

## 2.1 基于机器视觉的表格分割

为避免表格填写过程中的行列增删、宽高调整等微调操作对同模板表格识别的干扰,需要对表格进行模块分割与规范化处理,从而实现对其核心视觉与内容特征的提取。鉴于复杂表格以纵向布局为主,增删对象主要是行而非列,因此表格分割中采用横向区域最大化为主进行分割,使得物理上连续且纵线、横线长度均一致的最大区域成为独立模块。如图2所示的复杂表格可以分解为5个模块,规范化处理后,模块3是纵向合并单元格,包括1个单元格;模块1~2、4~5分别包含4个、6个、70个、5个单元格。

鉴于从pdf文档、图片等对象中提取的表格缺乏语义信息,无法准确知晓哪些单元格进行了纵向或横向合并,因此,为提升表格分割算法的通用性,本文拟采用机器视觉技术进行实现。实现步骤为:①表格预处理,将表格统一转换为图片,并进行灰度化、二值化、腐蚀与膨胀处理、纵横线扫描,使表格仅保留纵线与横线特

|        |                |            |                                          |    |                    |              |        |
|--------|----------------|------------|------------------------------------------|----|--------------------|--------------|--------|
| 工作单位   | 中南民族大学公共管理学院 ① |            |                                          |    | 联系电话               | [REDACTED]   |        |
| 身份证件类型 | 身份证            | 身份证件号码     | [REDACTED] ②                             |    | 是否在内(大)陆工作的港澳台研究人员 | 否            |        |
| ③      | 姓名             | 出生年月       | 专业职称                                     | 学位 | 工作单位               | 研究专长         | 本人签字   |
|        | [REDACTED]     | 1961.11.08 | 教授                                       | 硕士 | 中南民族大学公共管理学院       | 农村基层治理与公共服务  | ④      |
|        | [REDACTED]     | 1981.08.04 | 副教授                                      | 博士 | 中南民族大学公共管理学院       | 电子政务与公共服务决策  |        |
|        | [REDACTED]     | 1989.05.01 | 讲师                                       | 博士 | 中南民族大学公共管理学院       | 农村社会保障与公共服务  |        |
|        | [REDACTED]     | 1984.04.02 | 讲师                                       | 硕士 | 中南民族大学现代教育技术中心     | 信息技术管理       |        |
|        | [REDACTED]     | 1991.01.26 | 讲师                                       | 博士 | 武汉软件工程职业学院         | 大数据与人工智能社会治理 |        |
|        | [REDACTED]     | 1987.09.13 | 讲师                                       | 博士 | 湖北经济学院会计学院         | 财务数据库管理      |        |
| 预期成果   | A⑤             |            | A.专著 B.译著 C.论文集 D.研究报告 E.工具书 F.电脑软件 G.其他 |    |                    |              | 字数(千字) |
|        |                |            |                                          |    |                    |              | 230    |

图2 复杂表格分割示例

征;②基于横线特征的区域分割,依据表格中横线的长度、位置及是否相邻(相邻横线长度不一时,将长的横线截断成多条),将复杂表格初步分割成多个区域,每个区域内每条横线起终点的纵坐标均保持一致;③基于纵横线特征的细粒度模块分割,对分割后区域进行

逐行分析,若后一行与前一行内的纵线数量一致且每条纵线的横坐标都一致,则将其视为同一模块,否则将其分割为不同模块。

## 2.2 基于结构与内容相似度的同模板表格识别

经分割处理后,复杂表格的结构有所简化,若仅考虑结构特征,同模板表格识别中存在误判的可能。为此,拟综合考虑结构与内容相似度两个方面进行同模板表格的识别。实现过程中,首先根据表格分割结果,借鉴HTML的表格编码规则对其进行重新编码,然后将结构与内容特征转换为一体化格式,最后根据编码结果的一致性进行同模板表格识别。

(1) 考虑结构与内容的表格信息编码。该环节可以借鉴HTML中有关表格的语法规则进行实现,但为了在突出表格结构与内容特征的同时,又避免对同模板表格识别造成干扰,需要在表格信息编码中注意三个方面。第一,若待编码分割模块包含多行,则只保留前两行,从而在尽量保留表格结构特征的前提下,避免增删行操作对同模板表格识别的影响。同时,在此过程中需要关注对关联模块编码的影响,并作出相应的调整。第二,同一个分割模块编码时,不考虑单元格合并问题,即默认所有的单元格宽度保持一致。第三,内容信息编码主要通过保留每个分割模块左上角单元格取值的方式进行实现。这是因为每个分割区域都可以视为一个独立的表格,无论表头、表体及说明性单元格的分布多复杂,左上角单元格都不可能是表体或说明性单元格,但其他任何一个单元格则无法排除这种可能。

(2) 基于编码信息的结构与内容相似度计算。鉴于编码信息中已经包含结构信息与内容信息,即各个单元格的位置信息与取值信息,因此,表格间的结构与内容相似度计算可以通过编码的相似度进行计算。具体实现方法上,鉴于相似度计算的目标是用于同模板表格识别,因此只需要将相似度的取值范围限定为0和1即可,若待计算表格编码结果字符串完全匹配,则相似度为1;否则,相似度为0。假设表格*i*和表格*j*的编码信息相似度为 $s(con_i, con_j)$ ,则表格*i*与表格*j*的相似度 $s(v_i, v_j)$ 的计算方法如公式(1)所示。

$$s(v_i, v_j) = \begin{cases} 1, & \text{if } s(con_i, con_j) = 1 \\ 0, & \text{else} \end{cases} \quad (1)$$

## 2.3 基于启发式规则的单元格元组识别

识别单元格间语义关系的最终目标是将具有语义关系的表头单元格、表体单元格及说明性单元格关联起来,从而为复杂表格的知识抽取提供支持。基于此,可以将单元格语义关系识别问题转换为元组识别问题,将复杂表格中的所有单元格识别为(表头单元格,表体单元格,说明性单元格)元组,每个元组中表头单元格可以为1个或多个,表体单元格仅有1个,说明性单元格可以为0个或1个。具体实现上,可以进一步分为单元格类型识别与元组识别两个环节。

(1) 基于频次统计的单元格类型识别。只要同模板表格的样本规模不过小,表体单元格的取值不可能完全一致,据此可以设计表体单元格识别策略:对每一组同模板表格,统计每个单元格取值的频次,若单元格取值频次小于表格数量,则将其视为表体类单元格。对于说明性单元格,若是用于明确表体单元格取值范围,则临近表体单元格的取值应是其子集;若是用于解释填写规则,则常常位于同行单元格的最右侧,据此可以设计启发式规则,实现表头单元格与说明性单元格的区分。

(2) 单元格元组识别启发式规则设计。通过对多种类型的复杂表格调研,表头与对应表体单元格的位置分布遵循先右后下的原则,即表头单元格优先与右侧临近单元格构成元组。据此,可以按如下方法进行单元格元组的识别。①若表头单元格的右侧和下方临近单元格均为表体类单元格,则与右侧临近单元格组成元组。②若表头单元格的右侧和下方临近单元格均非表体类单元格,则该元组包含多个表头单元格。③涉及单元格合并时,则将该合并单元格与临近的所有表头类单元格都组成元组。④涉及说明性单元格时,若属于取值范围说明型,则将其与对应单元格组成元组即可;若属于填写规则说明型,则将其与临近的左侧或上面单元格组成元组。

## 2.4 面向增量更新的复杂表格单元格语义关系识别流程

随着时间的推移,待处理的复杂表格类型可能会增加,因此,基于无监督学习的复杂表格单元格语义关系识别在流程上既要支持符合已有模板的表格处理,也需要支持新模板的自动发现与加工,适应复杂表格增量更新的基本现实。为此,设计了如图3所示的面向

增量更新的复杂表格单元格语义关系识别流程。

概括地说,当新增一批待处理的复杂表格时,首先对符合既有模板的表格进行处理;其次,以无法处理的

表格为训练数据,进行新表格模板的学习;最后,对之前未成功处理的表格再次进行处理,直至实现全部表格的单元格语义关系识别。

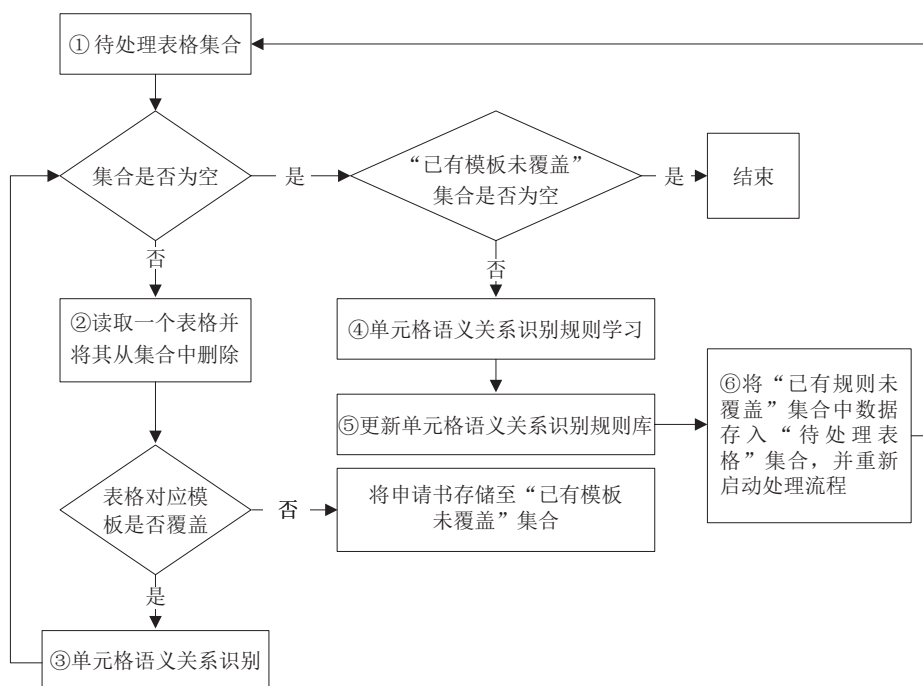


图3 面向增量更新的复杂表格单元格语义关系识别流程

具体而言,面向增量更新的复杂表格单元格语义关系识别流程为:①若待处理表格集合非空,则从中读取一条数据并将其从数据集中删除,之后转步骤②;否则,判断“已有模板未覆盖”数据集是否为空,若为空则结束整个流程,反之转步骤④;②读取各单元格取值,并与模板中表头单元格取值进行比较判断,若模板表格的表头单元格被全部覆盖,而且表头单元格相对位置一致,则将其视为该表格对应的模板并转步骤③,否则,将其存入“已有模板未覆盖”集合;③根据模板对应的单元格语义关联关系对该表格进行处理,之后转步骤①;④若“已有模板未覆盖”数据集非空,随机抽取一定规模的样本,按照前文所述方法进行新模板的学习;⑤将新学习的模板及对应的单元格语义关系映射规则更新至模板库;⑥将“已有模板未覆盖”集合中的数据全部存入待处理表格集合,再次启动处理流程。

### 3 实验

为验证基于无监督学习的复杂表格单元格语义关系识别模型的效果,从互联网上搜集了多种类型的复

杂表格进行了实验,下面对实验设置与结果进行说明。

#### 3.1 实验数据

为保证数据的多样性与代表性,通过互联网搜集了包含复杂表格的180篇文档文件,包括国家社科成果文库申请书、国家重点研发计划项目预申报书、国家自然科学基金项目申请书、科技基础调查基金项目申请书、国家社科基金项目申请书、华中师范大学学位申请与评定书在内的6种模板180篇文档,每类模板均为30个文档。文档的格式包括doc、docx和pdf,每篇文档包含待处理的复杂表格1个,其中前五类文档的复杂表格为基本信息表/数据表,包括项目基本信息、负责人信息、单位信息等内容,后一类文档的复杂表格包括申请人基本信息、学位申请信息等内容。数据集分割上,2/3作为训练数据,用于表格解析模型的无监督学习;1/3作为测试数据。

#### 3.2 评价指标

鉴于各类模板的单元格数量不一,为减少这一因素

的影响,效果评价时首先对每类表格的识别效果进行计算,之后再对其取均值。

(1) 单类表格识别效果评价指标。单类表格识别效果的计算是总体效果评价的基础和核心。假设表格类型 $i$ 中包含的单元格元组有 $m$ 个,算法识别出来的元组有 $n$ 个,其中 $q$ 个正确,则准确率 $p_i$ 和召回率 $r_i$ 计算方法如公式(2)和(3)所示。

$$p_i = \frac{q}{n} \quad (2)$$

$$r_i = \frac{q}{m} \quad (3)$$

(2) 模型总体效果评价指标。假设共有 $s$ 类表格,则模型的准确率 $a_j$ 和召回率 $b_j$ 计算方法如公式(4)和(5)所示。

$$a_j = \frac{\sum_{j=1}^s p_j}{s} \quad (4)$$

$$b_j = \frac{\sum_{j=1}^s r_j}{s} \quad (5)$$

### 3.3 实验过程

实验按照前文所构建的复杂表格单元格语义关系识别模型进行,因此对与模型完全一致的操作步骤不再赘述,仅对表格区域检测及结构识别、表格分割的具体操作方法进行说明。

(1) 表格区域检测及结构识别。鉴于doc和docx文档中的表格存在对应的语义标签信息,因此首先读取两类文档中的表格,之后根据规则筛选出待解析的复杂表格。pdf格式文档的表格区域检测与结构识别通过pdfminer工具完成,并采用同样的规则进行复杂表格筛选。

(2) 表格分割。该环节主要利用OpenCV机器视觉库实现,主要流程包括:①调用word2pdf工具包将doc、docx格式的表格文件转换为pdf文件,并调用pymupdf工具包将pdf表格文件转换为png图片;②调用OpenCV组件对图片进行灰度化、二值化处理,将原图片转换为更加清晰的黑白图片;通过适当的腐蚀与膨胀使横线和竖线更加清晰;同时利用长为100,宽为1的矩形扫描得到 $m$ 条横线;利用长为1,宽为50矩形扫描得到 $n$ 条竖线;③按照前文模型对表格进行分割处理。

### 3.4 实验结果分析

在基于120个表格完成6类复杂表格单元格语义关联关系规则学习后,对剩余的60个测试表格进行处理,效果如表1所示。总体来说,各类表格模板的识别效果均较好,准确率和召回率绝大多数都在90%以上,全部模板的平均召回率92.7%,准确率96.6%,能够较好地满足实践应用需求。

表1 实验结果

| 模板类型           | 召回率/% | 准确率/% |
|----------------|-------|-------|
| 国家社科成果文库申请书    | 90.7  | 95.1  |
| 国家重点研发计划项目预申报书 | 93.5  | 96.7  |
| 国家自然科学基金项目申请书  | 94.0  | 96.9  |
| 科技基础调查基金项目申请书  | 96.4  | 98.1  |
| 华中师范大学学位申请与评定书 | 89.7  | 96.3  |
| 国家社科基金项目申请书    | 91.9  | 96.4  |
| 全部模板           | 92.7  | 96.6  |

实验中同时发现,各类模板都出现了不同程度的误判,其主要原因在于:①部分表体类单元格的取值较为集中,导致各个训练集中对应单元格的取值一致,从而被误判为表头类单元格,如国家社科基金申请书中的“身份证件类型”表头对应的取值绝大多数都是“身份证”,课题负责人的“最后学历”“最后学位”两个表头单元格对应的表体单元格取值在样本中均为“研究生”“博士”;②华中师范大学学位申请与评定书模板中,右上角单元格是照片粘贴处,由于所采集的都是表格电子版,此处并未粘贴申请人的照片,导致系统将其误判为说明性单元格。对于前一个问题,一方面可以根据实践工作建立知识库,收录取值分布易集中的单元格元组,进而指导单元格间语义关系识别;另一方面可以建立预警机制,将包含表头类单元格过多的元组提供给用户进行审核,从而减少单元格间语义关系误识别。对于后一个问题,需要优化单元格元组的概念界定,并探索仅包含单个单元格元组的识别方法。

## 4 结语

数据化是实现表格数据分析挖掘、重用的重要环节,但受结构复杂、规律性弱的特点,申请表、登记表、报名表等复杂表格的单元格间语义关系难以识别,导

致此类表格的知识抽取较为困难。针对这一问题,本文利用待抽取复杂表格常包含多个副本、不同类型单元格间存在一定分布规律的特点,设计了基于无监督学习的复杂表格单元格语义关系识别方法,以便于将具有对应关系的表头、表体及说明性单元格关联起来,从而为表格数据化提供支撑。通过对6类表格进行试验,证实了所构建模型的可行性和有效性。然而,该方法也存在一定局限性,构成未来需要进一步深化研究的问题:一是若训练集中的表体单元格取值一致,容易发生单元格类型误判;二是鉴于表格分割与同模板表格识别时对原始表格进行了简化,因此若待处理表格中存在结构类似的表格时,可能会出现同模板表格识别错误,进而影响单元格语义分析的结果;三是模型默认输入的表格为复杂表格,未考虑复杂表格与行列表头表格混合处理的情况。

#### 参考文献

- [1] 丁培. 学术图表知识发现技术框架及研究进展 [J]. 图书情报工作, 2021, 65 (23): 136-148.
- [2] 高良才, 李一博, 都林, 等. 表格识别技术研究进展 [J]. 中国图象图形学报, 2022, 27 (6): 1898-1917.
- [3] WATANABE T, LUO Q, SUGIE N. Structure recognition methods for various types of documents [J]. Machine Vision and Applications, 1993, 6 (2): 163-176.
- [4] HIRAYAMA Y. A method for table structure analysis using DP matching [C] //Proceedings of 3<sup>rd</sup> international conference on document analysis and recognition. IEEE, 1995, 2: 583-586.
- [5] FANG J, GAO L C, BAI K, et al. A table detection method for multipage pdf documents via visual separators and tabular structures [C] //2011 International Conference on Document Analysis and Recognition. IEEE, 2011: 779-783.
- [6] 于丰畅, 程齐凯, 陆伟. 基于几何对象聚类的学术文献图表定位研究 [J]. 数据分析与知识发现, 2021, 5 (1): 140-149.
- [7] SCHREIBER S, AGNE S, WOLF I, et al. Deepdesrt: Deep learning for detection and structure recognition of tables in document images [C] //2017 14<sup>th</sup> IAPR International Conference on Document Analysis and Recognition. IEEE, 2017, 1: 1162-1167.
- [8] HAO L P, GAO L C, YI X H, et al. A table detection method for pdf documents based on convolutional neural networks [C] //2016 12<sup>th</sup> IAPR Workshop on Document Analysis Systems. IEEE, 2016: 287-292.
- [9] MELINDA L, BHAGVATI C. Parameter-free table detection method [C] //2019 International Conference on Document Analysis and Recognition. IEEE, 2019: 454-460.
- [10] ZHENG X, BURDICK D, POPA L, et al. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context [C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 697-706.
- [11] HUANG Y L, YAN Q Q, LI Y B, et al. A YOLO-based table detection method [C] //2019 International Conference on Document Analysis and Recognition. IEEE, 2019: 813-818.
- [12] 张建东, 陈仕吉, 徐小婷, 等. 基于词向量的PDF表格抽取研究 [J]. 数据分析与知识发现, 2021, 5 (8): 34-44.
- [13] 赵洪, 王芳. 大规模异构的政府统计报表信息抽取与集成融合研究 [J]. 情报学报, 2020, 39 (9): 938-948.
- [14] 张元鸣, 陈苗, 陆佳炜, 等. 非结构化表格文档数据抽取与组织模型研究 [J]. 浙江工业大学学报, 2016, 44 (5): 487-494.
- [15] SETH S, NAGY G. Segmenting tables via indexing of value cells by table headers [C] //2013 12<sup>th</sup> International Conference on Document Analysis and Recognition. IEEE, 2013: 887-891.
- [16] FANG J, MITRA P, TANG Z, et al. Table header detection and classification [C] //Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012: 599-605.
- [17] KOCI E, THIELE M, ROMERO O, et al. A machine learning approach for layout inference in spreadsheets [C] //IC3K 2016: Proceedings of the 8th International Joint Conference on Knowledge Discovery and Information Retrieval. SciTePress, 2016: 77-88.
- [18] GAO L C, HUANG Y L, DÉJEAN H, et al. ICDAR 2019 competition on table detection and recognition (cTDaR) [C] //2019 International Conference on Document Analysis and Recognition. IEEE, 2019: 1510-1515.
- [19] GOL M G, PUJARA J, SZEKELY P. Tabular cell classification using pretrained cell embeddings [C] //Proceedings of 2019 IEEE International Conference on Data Mining. IEEE, 2019: 230-239.

## 作者简介

林鑫, 男, 1987年生, 博士, 副教授, 研究方向: 信息组织与检索, E-mail: xinlin@mail.ccnu.edu.cn。

余华娟, 女, 1997年生, 硕士研究生, 研究方向: 信息组织与检索。

闫奕臻, 女, 2001年生, 本科生, 研究方向: 信息组织与检索。

### Research on Cell Semantic Relation Recognition in Complex Table Digitization

LIN Xin<sup>1,2</sup> YU HuaJuan<sup>1</sup> YAN YiZhen<sup>1</sup>

(1. School of Information Management, Central China Normal University, Wuhan 430079, P. R. China;

2. Research Center for Data Governance and Intelligent Decision Making of Hubei Province, Wuhan 430079, P. R. China )

**Abstract:** Complex tables can describe data in a simple and intuitive way, and are widely used in all walks of life. However, complex tables have problems such as complex structures, diverse cell types, and different forms of table documents. They need to be data processed before they can be shared and reused. Therefore, this paper constructs a cell semantic relationship recognition model based on unsupervised learning to realize the digitization of complex tables. First, it uses machine vision technology to realize the segmentation of complex tables, and then recognizes the same template table based on the similarity of table structure and content. On this basis, heuristic rules are set to identify the semantic relationship of cells in combination with the value and location characteristics of header cells, illustrative cells and table body cells. Finally, the empirical research verifies that the method in this paper can achieve high accuracy and recall rate in complex table digitization, which is feasible.

**Keywords:** Complex Table; Semantic Relationship; Form Digitization; Machine Vision

(收稿日期: 2022-08-27)

## ■ 书 讯 ■

# 《汉语主题词表》

《汉语主题词表》自1980年问世以后, 经1991年进行自然科学版修订, 在我国图书情报界发挥了应有作用, 曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要, 由中国科学技术信息研究所主持, 并联合全国图书情报界相关机构, 自2009年开始进行重新编制工作, 拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表(工程技术卷)》共收录优选词19.6万条, 非优选词16.4万条, 等同率0.84, 在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表(自然科学卷)》共收录专业术语12.4万条, 包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域, 收词系统、完整, 语义关系丰富、严谨, 每条词汇都有相应的学科分类号表现其专业属性, 并与同义英文术语对应。同时, 建立《汉语主题词表》网络服务系统, 提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘, 是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表(工程技术卷)》已于2014年由科学技术文献出版社出版, 分为13个分册, 总定价3 880元。

《汉语主题词表(自然科学卷)》已于2018年5月由科学技术文献出版社出版, 分为5个分册, 总定价1 247元。两卷均可分册购买。