(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109658062 A
(43)申请公布日 2019.04.19

(21)申请号 201811526851.4

(22)申请日 2018.12.13

(71)申请人 广州华资软件技术有限公司
地址 510665 广东省广州市天河区建中路12号首层

(72)发明人 彭本 雷邦宁 张士松 陈辉梓 翁庄明 李自强 余增平

(74)专利代理机构 广州市南锋专利事务所有限公司 44228
代理人 高崇

(51)Int.Cl.
G06Q 10/10(2012.01)
G06Q 50/26(2012.01)
G06F 16/93(2019.01)
G06K 9/00(2006.01)

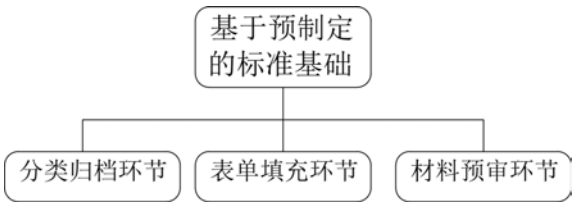
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种基于深度学习的电子档案智能处理方法

(57)摘要

本发明所提供的一种通过人工智能技术进行政务业务材料的自动分类、信息提取自动填表、特征识别自动预审的方法，解决当前依靠人工电子化效率低、易出错、不能及时利用共享材料的技术问题，预先利用高速扫描仪批量的将纸质材料文档扫描成电子材料文档，基于预制定的标准基础作为后续环节的运行规则，对扫描所得的电子材料文档的内容进行识别处理，从而对电子材料文档分别完成分类归档环节、完成表单填充环节以及材料预审环节。



CN 109658062 A

1.一种基于深度学习的电子档案智能处理方法，预先利用高速扫描仪批量的将纸质材料文档扫描成电子材料文档，其特征在于：基于预制定的标准基础作为后续环节的运行规则，对扫描所得的电子材料文档的内容进行识别处理，从而对电子材料文档分别完成分类归档环节、完成表单填充环节以及材料预审环节。

2.根据权利要求1所述的一种基于深度学习的电子档案智能处理方法，其特征在于：在分类归档环节中，根据标准基础识别电子材料文档中预定的特征区域内的文档基本信息，从而将电子材料文档自动分类归档到相应的目录下。

3.根据权利要求2所述的一种基于深度学习的电子档案智能处理方法，其特征在于：所述分类归档环节包括有以下步骤：

步骤A1.针对电子材料文档的标题区域以及相关特征区域进行识别，识别出包含有标题文本及其相关特征文本的文档基本信息；

步骤A2.对文档基本信息进行语义分析，提取出相应的“关键字”，然后根据预先构建的标准分类库进行逐级归类，同时将电子材料文档保存到相应的数据库文件目录下。

4.根据权利要求1所述的一种基于深度学习的电子档案智能处理方法，其特征在于：在表单填充环节中，根据标准基础识别电子材料文档中预定的文本区域内的业务内容信息，从而将所识别的业务内容信息自动填充至预设有的表单模板中。

5.根据权利要求4所述的一种基于深度学习的电子档案智能处理方法，其特征在于：所述表单填充环节包括有以下步骤：

步骤B1.针对电子材料文档的文本区域内通过文本识别出具体的文本内容；

步骤B2.将识别出的文本内容，通过CRF进行“命令实体识别”识别出实体集，从而得出相关的业务内容信息；

步骤B3.将所得到的业务内容信息填充至相对应的表单模板中。

6.根据权利要求1所述的一种基于深度学习的电子档案智能处理方法，其特征在于：在材料预审环节中，根据标准基础识别电子材料文档中预定的要素区域内是否存在有预定的要素图像信息，从而判定电子材料文档是否符合预定的校验规则。

7.根据权利要求6所述的一种基于深度学习的电子档案智能处理方法，其特征在于：所述材料预审环节包括有以下步骤：

步骤C1.通过图像识别针对电子材料文档的要素区域内是否存在有预定的要素信息，并作出相应的识别判断结果；

步骤C2.将识别判断结果一一添加相对应的判断标注，一并反馈至预设有数据库供人工审核。

8.根据权利要求1所述的一种基于深度学习的电子档案智能处理方法，其特征在于：所述标准基础包括有：

-标准分类库：基于业务特征建立，由预设定的“关键字”组成的多级目录类别，用于归档相对应的材料文档；

-识别模板：基于业务内容建立，对不同类型的文档材料相应的内容所在版面位置进行划分定义；

-表单模板：基于业务需求建立，设置不同类型表单模板供相对应的业务内容信息填充；

-实体库:基于业务内容建立,预导入或设定所需词性标注的命名集;

-校验规则:基于业务内容建立,设置有针对不同类型文档材料的预审需求,以便于相对应的电子材料文档进行比对校验。

一种基于深度学习的电子档案智能处理方法

技术领域

[0001] 本发明涉及电子政务文件电子化管理的技术领域，尤其是指一种基于深度学习的电子档案智能处理方法。

背景技术

[0002] 政务办理产生的电子文件是办理业务过程中或签转到其他部门进行业务审核，进行全流程监管的重要文件。

[0003] 政务办理过程中的电子文件包括各个业务办理环节中获取的证件、业务表单、文书、申报书、通知书、证明材料等不同形式的电子材料，需要按照规定，关联对应的业务办理号，分类归档到不同的目录下，如：证件材料、表单材料、证明材料、报告材料等。

[0004] 目前材料电子化的操作，主要依靠人工进行，主要包括扫描和上传两个动作，需要将文件扫描后，分别去修改文件的名称或上传到不同的目录下，由于纸质材料种类繁多，单个业务产生的材料数量多，存在以下明显问题：耗费大量人力，处理效率非常低；人工整理材料不能有效分类，缺乏关联，很难再后面将其有效利用；目前依靠人工不能解决识别要素、自动填表这些需要人工智能才能解决的问题。使得电子材料不能按照统一规范采集、管理、共享。

[0005] 因此政务类相关材料的分类、归档问题需要进一步解决，并引入如要素识别、自动填表等新的方法，进一步优化处理过程。

发明内容

[0006] 本发明的目的在于克服现有技术的不足，提供一种通过人工智能技术进行政务业务材料的自动分类、信息提取自动填表、特征识别自动预审的方法，解决当前依靠人工电子化效率低、易出错、不能及时利用共享材料的技术问题。

[0007] 为了实现上述的目的，本发明所提供的一种基于深度学习的电子档案智能处理方法，预先利用高速扫描仪批量的将纸质材料文档扫描成电子材料文档，基于预制定的标准基础作为后续环节的运行规则，对扫描所得的电子材料文档的内容进行识别处理，从而对电子材料文档分别完成分类归档环节、完成表单填充环节以及材料预审环节。

[0008] 进一步，在分类归档环节中，根据标准基础识别电子材料文档中预定的特征区域内的文档基本信息，从而将电子材料文档自动分类归档到相应的目录下。

[0009] 进一步，所述分类归档环节包括有以下步骤：

步骤A1.针对电子材料文档的标题区域以及相关特征区域进行识别，识别出包含有标题文本及其相关特征文本的文档基本信息；

步骤A2.对文档基本信息进行语义分析，提取出相应的“关键字”，然后根据预先构建的标准分类库进行逐级归类，同时将电子材料文档保存到相应的数据库文件目录下。

[0010] 进一步，在表单填充环节中，根据标准基础识别电子材料文档中预定的文本区域内的业务内容信息，从而将所识别的业务内容信息自动填充至预设有的表单模板中。

[0011] 进一步,所述表单填充环节包括有以下步骤:

步骤B1.针对电子材料文档的文本区域内通过文本识别出具体的文本内容;

步骤B2.将识别出的文本内容,通过CRF进行“命令实体识别”识别出实体集,从而得出相关的业务内容信息;

步骤B3.将所得到的业务内容信息填充至相对应的表单模板中。

[0012] 进一步,在材料预审环节中,根据标准基础识别电子材料文档中预定的要素区域内是否存在有预定的要素图像信息,从而判定电子材料文档是否符合预定的校验规则。

[0013] 进一步,所述材料预审环节包括有以下步骤:

步骤C1.通过图像识别针对电子材料文档的要素区域内是否存在有预定的要素信息,并作出相应的识别判断结果;

步骤C2.将识别判断结果一一添加相对应的判断标注,一并反馈至预设有数据库供人工审核。

[0014] 进一步,所述标准基础包括有:

-标准分类库:基于业务特征建立,由预设定的“关键字”组成的多级目录类别,用于归档相对应的材料文档;

-识别模板:基于业务内容建立,对不同类型的文档材料相应的内容所在版面位置进行划分定义;

-表单模板:基于业务需求建立,设置不同类型表单模板供相对应的业务内容信息填充;

-实体库:基于业务内容建立,预导入或设定所需词性标注的命名集。

[0015] -校验规则:基于业务内容建立,设置有针对不同类型文档材料的预审需求,以便于相对应的电子材料文档进行比对校验。

[0016] 本发明采用上述的方案,其有益效果在于:通过自然语言处理技术、图像识别技术、深度学习算法的提取技术,并深度融合实际的业务情况,可批量的将纸质文档扫描为电子档案,且能进行自动分类,并提取相关的信息数据,完成自动填表和自动预审的功能,不但能高效高质量完成材料电子化处理,并将不断提升材料的利用及共享。这一发明既可以大大减少人工分类、提取信息的工作量,大幅提高准确率,还可以识别提取业务内容信息,自动完成填表;通过对要素信息识别,实现对材料进行预审。

附图说明

[0017] 图1为本实施例的电子档案智能系统的组成示意图。

[0018] 图2为本实施例的分类归档环节的流程示意图。

[0019] 图3为本实施例的表单填充环节的流程示意图。

[0020] 图4为本实施例的分材料预审环节的流程示意图。

具体实施方式

[0021] 下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0022] 参见附图1所示,在本实施例中公开了一种基于深度学习的电子档案智能处理方法,特选用警务相关的业务情况作为实施例进行详细说明,以便于本领域技术人员的便于

理解。本实施例的电子档案智能系统以预先制定的标准基础作为后续工作环节的运行规则，以便于批量扫描的纸质文档在经过一系列工作环节，分别实现了将文档材料完成进行分类归档、识别文本信息进行表单填写以及校验判断进行材料预审的智能化处理。

[0023] 本实施例的标准基础包括有标准分类库、识别模板、表单模板、实体库和校验规则，具体如下：

-标准分类库：基于业务特征建立，由预设定的“关键字”组成的多级目录类别，用于归档相对应的材料文档。所建立的目录类别具有上下级、呈树形结构。即，在本实施例中，一级目录类别分为表单类、证照类、证据材料类等；二级目录类别类别建立在相对应的以及目录类别下，如建立在证照类的一级目录类别下的二级目录类别分为身份证、户口本、驾驶证等；根据实际业务特征相适应的进行目录类别设定，只要采用所需的“关键字”便可。

[0024] -识别模板：基于业务内容建立，对不同类型的文档材料相应的内容所在版面位置进行划分定义，在本实施例中，根据不同的内容主要划分定位为特征区域、文本区域和要素区域，其中，本实施例的特征区域内包括由标题及其相关特征文本组成的文档基本信息；本实施例的文本区域内包括由人员姓名，年龄，户籍等组成的业务内容信息；本实施例的要素区域内包括由公章、手印、签名等组成的要素信息。

[0025] -表单模板：基于业务需求建立，设置不同类型表单模板供相对应的业务内容信息填充，其中，本实施例的表单模板如在审理案件环节，建立表单模板案发情况记录表，分别有作案手段(作案手段、作案特征等)、案发时间(报警时间、案发时间、案发时段等)、案发地点(住宅类型、案发地点、有无监控等)、当事人信息(姓名、性别、年龄等)等信息项。

[0026] -实体库：基于业务内容建立，导入所需词性标注的命名集，可根据不同的行业预设有的命名集进行导入，也可以根据实际需求用户自主添加。

[0027] -校验规则：基于业务内容建立，设置有针对不同类型文档材料的预审需求，以便于相对应的电子材料文档进行比对校验。

[0028] 基于上述预制定的标准基础作为后续环节的运行规则，并且在高速扫描仪批量的将纸质材料文档扫描成电子材料文档后，对扫描所得的电子材料文档的内容进行识别处理，从而对电子材料分别完成分类归档环节、完成表单填充环节以及材料预审环节。

[0029] 如附图2所示，在本实施例中，在分类归档环节中，根据标准基础识别电子材料中预定的特征区域内的文档基本信息(此处的文档基本信息主要包括有标题内容、时间等、地点信息)，从而将电子材料文件自动分类归档到相应的目录下，同时，此环节中所采用的标准基础为标准分类库和识别模板，为了便于理解，该分类归档环节具体包括有以下步骤：

步骤A1.针对电子材料的标题区域以及相关特征区域进行识别，识别出包含有标题文本及其相关特征文本的文档基本信息；

步骤A2.对文档基本信息进行语义分析，提取出相应的“关键字”，然后根据预先构建的标准分类库进行逐级归类，同时将电子材料文档保存到相应的数据库文件目录下。

[0030] 为了便于理解，例如在刑侦案件中，在分类归档环节中对电子材料文件进行识别相关的特征区域内的标题为“XX市公安局**分局 拘留通知书”，并且加以语义分析提出“拘留通知书”作为“关键字”，从而将该电子材料文档归类于“刑事侦查案卷-拘留通知书”的目录类别下，同时，将电子材料文档保存到相应的数据库文件目录以供查阅。

[0031] 如附图3所示，在本实施例中，在表单填充环节中，根据标准基础识别电子材料文

档中预定的文本区域内的业务内容信息(此处的业务内容信息主要包括有人员姓名,年龄,户籍等),从而将所识别的业务内容信息自动填充至预设有的表单模板中,其中,此环节中所采用的标准基础为标准分类库、识别模板、表单模板和实体库。为了便于理解,该分类归档环节具体包括有以下步骤:

步骤B1.针对电子材料的文本区域内通过文本识别出具体的文本内容,其中,此处可根据分类归档环节中的归类情况,选择调用相对应的表单模板以对应相适应的文本区域;

步骤B2.将识别出的文本内容,通过CRF进行"命令实体识别"识别出实体集,再基于实体库的规则,得出相关的业务内容信息,如人员姓名,年龄,户籍等业务内容信息;

步骤B3.将所得到的业务内容信息填充至相对应的表单模板中。

[0032] 为了便于理解,例如针对警务的"简要案情"的材料文档,在表单填充环节中,识别该材料文档的文本区域内所记载"201*年*月*日**时**分,张X利(男,汉族,36岁,户籍地及现住址:广州市天河区XX村X号,居民身份证号码:(110226********3322)拨打110报警称:在广州市天河区XX小区X号楼X单元XXX室家中发现被盗,经工作了解,201*年*月*日*时左右,张X利从广州市天河区建安里小区X号楼X单元XXX室家中离开并将门窗锁好,201*年*月*日**时**分左右张X利回到家中,发现锁被打开且屋内有被翻动的痕迹,经检查发现放在主卧床头柜上边第一个抽屉里的7000元人民币现金、放在主卧衣柜抽屉内的9000元人民币现金和一个翡翠玉坠(具体材质价值不详)被盗,放在次卧电脑桌上面柜子里的一条中华烟(年初买的,时价440元人民币)被盗,后张X利拨打110报警。"的文本内容,从而得出相关的业务内容信息,并且填入预设的"案情简述"表格模板中:

作案手段	
作案手段	技术开锁
技术特征	物品被翻动；锁被打开；屋门闪烁；窗上锁
案发时间	
报警时间	201*-**-** **:**
案发时间	201*-**-** **:**
案发时段	白天
案发地点	
住宅类型	高层
案发地点	广州市天河区XX小区X号楼X单元XXX室家中
有无监控	未描述
丢失物品	
饰品	一个翡翠玉坠（具体材质价值不详）
其他	一条中华烟（年初买的，时价440元人民币）
现金	7000元人民币现金；9000元人民币现金
当事人	
姓名	张X利
性别	男
年龄	36岁
籍贯	广州市天河区XX村X号
身份证号	110226********3322

如附图4所示，在本实施例中，在材料预审环节中，根据标准基础识别电子材料中预定的要素区域内是否存在有预定的要素信息，从而判定电子材料是否符合预定的校验规则，其中，此环节中所采用的标准基础为识别模板和校验规则。为了便于理解，材料预审环节具体包括有以下步骤：

步骤C1.通过图像识别针对电子材料的要素区域内是否存在有预定的要素信息(此处的要素信息包括但不限于公章、手印、签名等信息)，并作出相应的识别判断结果(比如识别要素区域内是否存在公章、手印等要素信息)，并给出相应的标注(例如存在有公章且缺少手印时，标注为：公章(1)、手印(0))；

步骤C2.将识别判断结果一一添加相对应的判断标注，一并反馈至预设有数据库供人工审核。

[0033] 通过材料预审环节预先对材料文档中所存在的一些形式问题进行预审，降低人工审核的工作，并且不符合校验规则的识别判断结果，配合添加上相应的判断标注，能够让审核员直接调取数据库中的材料文档进行人工审核，进一步提升了审核的准确性。

[0034] 综上所述，本实施例的处理方法有效地解决文书材料电子化及归档效率低下问题，节省了人力成本，从而实现办公智能化，降低办公成本；充分利用现有资源，减少重复建设，转化为数字化文档存储；方便共享、检索和传输，使得数据跨越原有的场景，大大提升数据的应用边界和价值，电子数据的应用和开发成为新的价值增长点，且有效提高各政府部

门信息互通,高效业务协同,构建统一的电子化管理平台。

[0035] 以上所述之实施例仅为本发明的较佳实施例,并非对本发明做任何形式上的限制。任何熟悉本领域的技术人员,在不脱离本发明技术方案范围情况下,利用上述揭示的技术内容对本发明技术方案作出更多可能的变动和润饰,或修改均为本发明的等效实施例。故凡未脱离本发明技术方案的内容,依据本发明之思路所作的等同等效变化,均应涵盖于本发明的保护范围内。

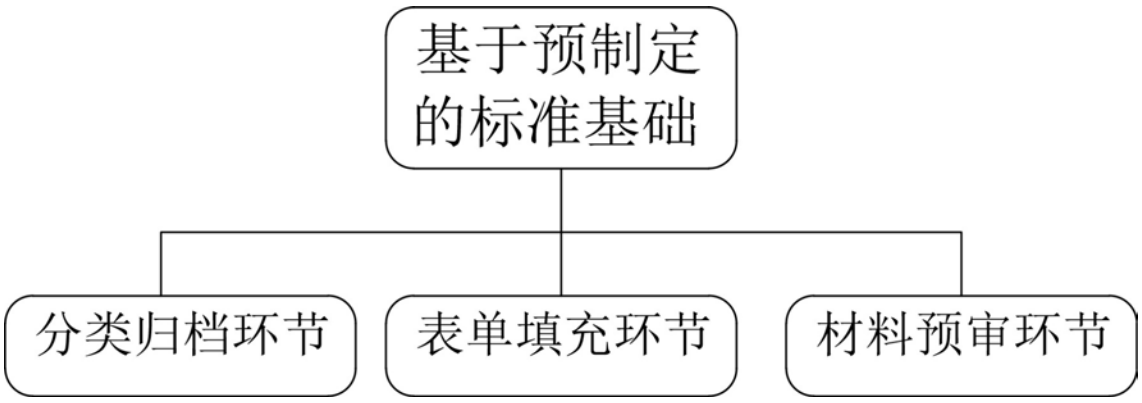

基于预制定的标准基础
分类归档环节
表单填充环节
材料预审环节

图1

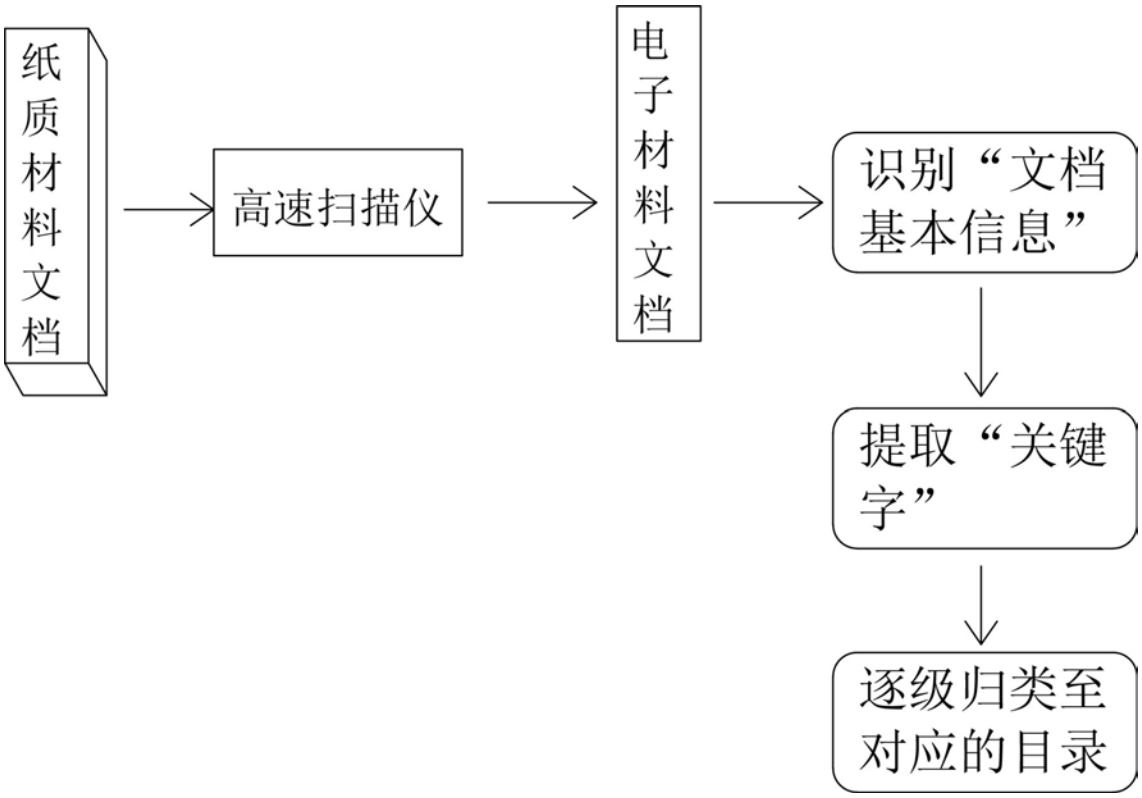

纸质材料文档
高速扫描仪
电子材料文档
识别“文档基本信息”
提取“关键字”
逐级归类至对应的目录

图2

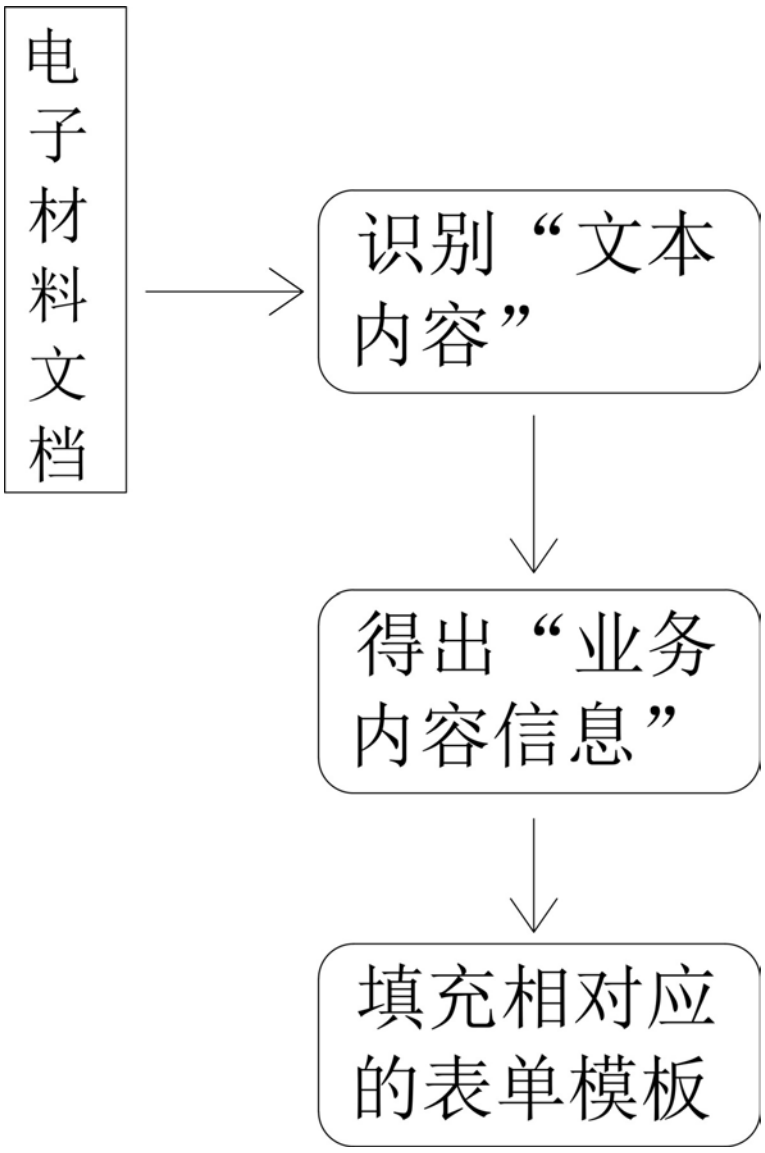


图3

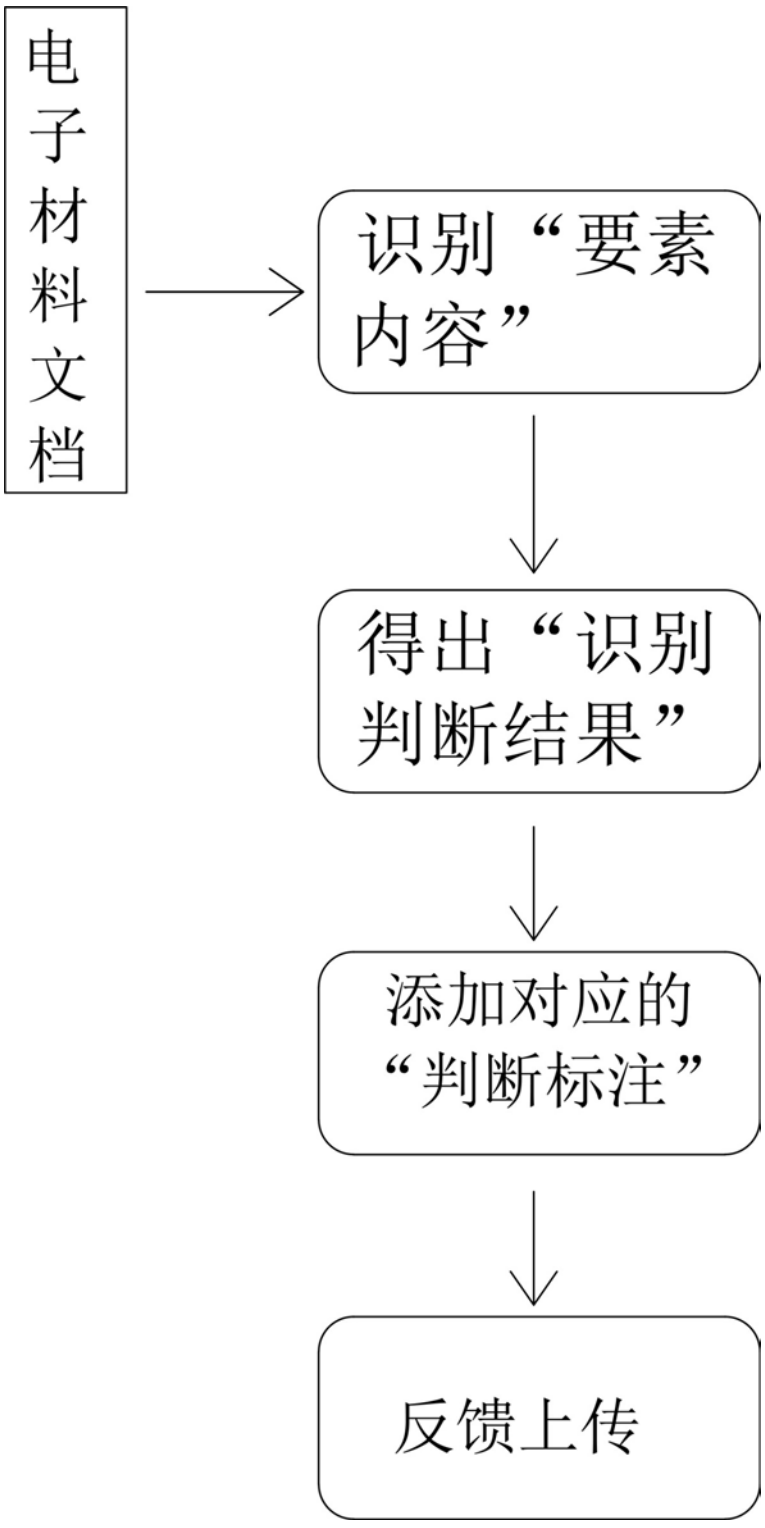


图4