(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114998920 B
(45)授权公告日 2023.04.07

(21)申请号 202210743284.8
(22)申请日 2022.06.27
(65)同一申请的已公布的文献号
申请公布号 CN 114998920 A
(43)申请公布日 2022.09.02
(73)专利权人 北京智慧金源信息科技有限公司
地址 100043 北京市石景山区实兴大街30号院7号楼8层171号(集群注册)
(72)发明人 席国超 徐宝东 张成宏 曾辉 刘建龙 张志刚 常城 李帅
(74)专利代理机构 北京同辉知识产权代理事务所(普通合伙) 11357
专利代理师 于晶晶
(51)Int.Cl.
G06V 30/413(2022.01)
G06V 30/19(2022.01)
G06F 40/30(2020.01)
G06F 40/289(2020.01)
G06F 40/242(2020.01)
(56)对比文件
CN 112597300 A,2021.04.02
CN 113486664 A,2021.10.08
审查员 杨爱林

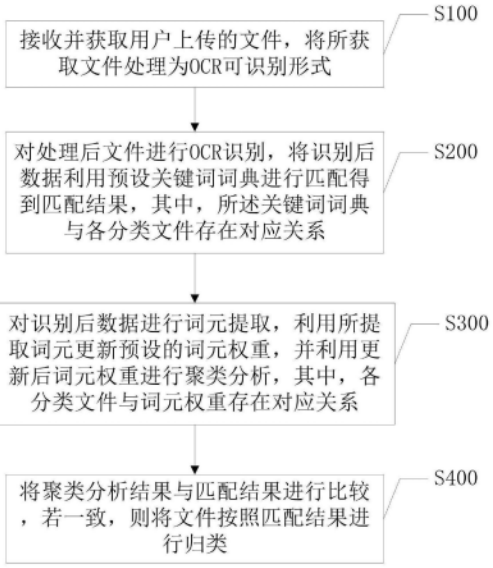
权利要求书2页 说明书9页 附图3页

(54)发明名称

基于NLP语义识别的供应链金融文件管理方法及系统

(57)摘要

本发明公开基于NLP语义识别的供应链金融文件管理方法及系统，该方法包括：获取用户上传的，将其处理为OCR可识别；对处理后文件进行OCR识别，将识别后数据利用预设关键词词典进行匹配，其中，关键词词典与各分类文件存在对应关系；对识别后数据进行词元提取，利用所提取词元更新预设词元权重，并利用更新后词元权重进行聚类分析；比较聚类分析结果与匹配结果，若一致，将文件按照匹配结果进行归类。本发明能够根据语义实现对多类型文件的自动识别归类，尤其适于供应链金融资料管理，在供应链金融场景下，通过文件归类的自动化和智能化，提升了文件管理效率，有利于对各种实际业务的精细化运营，使得金融机构能够更好服务于供应链上客户。



CN 114998920 B

1.一种基于NLP语义识别的供应链金融文件管理方法,其特征在于,包括以下步骤:

接收并获取用户上传的文件,将所获取文件处理为OCR可识别形式;对处理后文件进行OCR识别,将识别后数据利用预设关键词词典进行匹配得到匹配结果,其中,所述关键词词典与各分类文件存在对应关系;

对识别后数据进行词元提取,利用所提取词元更新预设的词元权重,并利用更新后词元权重进行聚类分析,其中,各分类文件与词元权重存在对应关系;

将聚类分析结果与匹配结果进行比较,若一致,则将文件按照匹配结果进行归类;

预先构建分类文件和关键词词典,并建立分类文件与关键词词典间的对应关系;

所述预先构建分类文件和关键词词典,并建立分类文件与关键词词典间的对应关系具体为:

对存量文件OCR识别后进行分词处理,对分词后的词元利用预设权重计算规则计算词元权重;利用计算得到的词元权重进行文本聚类分析,对聚类后的分类结果进行手动标注,得到分类文件及对应的高权重关键词词典;

所述对分词后的词元利用预设权重计算规则计算词元权重具体为:对分词后的词元利用预设公式$W_{td} = TF_{td} \cdot \log\frac{n}{DF_t} + TF_{td} \cdot \frac{AF_t}{n}$计算词元权重,其中,t为词元,$W_{td}$为词元t在文档d中的权重,$TF_{td}$为词元t在文档d中出现的次数,n为文档总数,$DF_t$为包含词元t的文档数,$AF_t$为词元t总共出现的次数;

所述利用计算得到的词元权重进行文本聚类分析具体为:将所得到的的每一存量文件对应的词元权重利用K-Means聚类算法进行聚类分析,其中,使用公式$\theta = \arccos(\frac{\vec{a}\cdot\vec{b}}{||a||\cdot||b||})$替换K-Means算法中的欧式距离进行聚类分析。

2.根据权利要求1所述的基于NLP语义识别的供应链金融文件管理方法,其特征在于,接收并获取用户上传的文件,将所获取文件处理为OCR可识别形式具体为:

接收并获取用户上传的文件,识别所获取文件的格式,并基于预先设定的格式转化规则将文件格式转换为image格式。

3.根据权利要求1所述的基于NLP语义识别的供应链金融文件管理方法,其特征在于,得到分类文件及对应的高权重关键词词典后还包括:通过聚类分析计算得到各分类文件中的最边缘文件,使用关键词词典对最边缘文件进行关键词匹配分析,若关键词匹配结果不一致,则去除相应分类中匹配度最差文件,并重新进行聚类分析。

4.根据权利要求1所述的基于NLP语义识别的供应链金融文件管理方法,其特征在于,若聚类分析结果与匹配结果比较后不一致,则利用聚类分析结果更新关键词词典。

5.一种基于NLP语义识别的供应链金融文件管理系统,其特征在于,包括:

获取模块,用于接收并获取用户上传的文件,将所获取文件处理为OCR可识别形式;

识别模块,用于对处理后文件进行OCR识别;

匹配模块,用于将识别模块识别后数据利用预设关键词词典进行匹配得到匹配结果,其中,所述关键词词典与各分类文件存在对应关系;

分析模块,用于对识别后数据进行词元提取,利用所提取词元更新预设的词元权重,并

利用更新后词元权重进行聚类分析，其中，各分类文件与词元权重存在对应关系；

预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系；

所述预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系具体为：

对存量文件OCR识别后进行分词处理，对分词后的词元利用预设权重计算规则计算词元权重；利用计算得到的词元权重进行文本聚类分析，对聚类后的分类结果进行手动标注，得到分类文件及对应的高权重关键词词典；

所述对分词后的词元利用预设权重计算规则计算词元权重具体为：对分词后的词元利用预设公式 $W_{td} = TF_{td} \cdot \log\frac{n}{DF_t} + TF_{td} \cdot \frac{AF_t}{n}$ 计算词元权重，其中，t为词元，W_{td}为词元t在文档d中的权重，TF_{td}为词元t在文档d中出现的次数，n为文档总数，DF_t为包含词元t的文档数，AF_t为词元t总共出现的次数；

所述利用计算得到的词元权重进行文本聚类分析具体为：将所得到的的每一存量文件对应的词元权重利用K-Means聚类算法进行聚类分析，其中，使用公式 $\theta = \arccos(\frac{\vec{a}\cdot\vec{b}}{\|a\|\cdot\|b\|})$ 替换K-Means算法中的欧式距离进行聚类分析；

比较模块，用于将聚类分析结果与匹配结果进行比较；

执行模块，用于根据比较模块比较结果，对文件进行归类。

6.一种电子设备，包括存储器、至少一个处理器以及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时执行如权利要求1-4中任意一项所述的方法。

基于NLP语义识别的供应链金融文件管理方法及系统

技术领域

[0001] 本发明涉及数据处理技术领域，具体涉及在供应链金融业务中实现的基于NLP语义识别的供应链金融文件管理方法及系统。

背景技术

[0002] 供应链金融是银行围绕核心企业，管理上下游中小企业的资金流、物流和信息流，并把单个企业的不可控风险转变为供应链企业整体的可控风险，通过立体获取各类信息，将风险控制在最低的金融服务。

[0003] 由于供应链金融领域的文件繁杂，例如合同、发票、结算单等，单单合同类型就包含多种，有采购、分包、租赁等多种多样的合同类型，如果涉及工程类的合同，通常还会具有较多的保障性条款。若想对供应链进行精细化管理，就需要确保将链条上的各种文件资料都进行准确的分类。

[0004] 正是由于文件类型多、数据量大，纯粹靠人工进行分类效率低下，出错率也高。而利用现有的文本识别分类技术，即对文件进行OCR(Optical Character Recognition，光学字符识别)识别后得到非结构化的文本，使用常规的聚类分析根本无法做到对多种文件进行分类，更无法做到对各类文件涉及的关键词进行获取等后续操作。

[0005] 因此，现有技术还有待于进一步发展和改进。

发明内容

[0006] 针对现有技术的种种不足，为了解决上述问题，现提出一种基于NLP语义识别的供应链金融文件管理方法及系统。本发明技术方案具体如下：

[0007] 一种基于自然语言处理(NLP)语义识别的供应链金融文件管理方法，其中，包括以下步骤：

[0008] 接收并获取用户上传的文件，将所获取文件处理为OCR可识别形式；对处理后文件进行OCR识别，将识别后数据利用预设关键词词典进行匹配得到匹配结果，其中，所述关键词词典与各分类文件存在对应关系；

[0009] 对识别后数据进行词元提取，利用所提取词元更新预设的词元权重，并利用更新后词元权重进行聚类分析，其中，各分类文件与词元权重存在对应关系；

[0010] 将聚类分析结果与匹配结果进行比较，若一致，则将文件按照匹配结果进行归类；

[0011] 预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系；

[0012] 所述预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系具体为：

[0013] 对存量文件OCR识别后进行分词处理，对分词后的词元利用预设权重计算规则计算词元权重；利用计算得到的词元权重进行文本聚类分析，对聚类后的分类结果进行手动标注，得到分类文件及对应的高权重关键词词典；

[0014] 所述对分词后的词元利用预设权重计算规则计算词元权重具体为：对分词后的词

元利用预设公式$W_{td} = TF_{td} \cdot \log\frac{n}{DF_t} + TF_{td} \cdot \frac{AF_t}{n}$计算词元权重，其中，t为词元，$W_{td}$为词元t在文档d中的权重，$TF_{td}$为词元t在文档d中出现的次数，n为文档总数，$DF_t$为包含词元t的文档数，$AF_t$为词元t总共出现的次数；

[0015] 所述利用计算得到的词元权重进行文本聚类分析具体为：将所得到的的每一存量文件对应的词元权重利用K-Means聚类算法进行聚类分析，其中，使用公式$\theta = \arccos(\frac{\vec{a}\cdot\vec{b}}{||a||\cdot||b||})$替换K-Means算法中的欧式距离进行聚类分析。

[0016] 所述的基于NLP语义识别的供应链金融文件管理方法，其特征在于，接收并获取用户上传的文件，将所获取文件处理为OCR可识别形式具体为：

[0017] 接收并获取用户上传的文件，识别所获取文件的格式，并基于预先设定的格式转化规则将文件格式转换为image格式。

[0018] 所述的基于NLP语义识别的供应链金融文件管理方法，其特征在于，得到分类文件及对应的高权重关键词词典后还包括：通过聚类分析计算得到各分类文件中的最边缘文件，使用关键词词典对最边缘文件进行关键词匹配分析，若关键词匹配结果不一致，则去除相应分类中匹配度最差文件，并重新进行聚类分析。

[0019] 所述的基于NLP语义识别的供应链金融文件管理方法，其特征在于，若聚类分析结果与匹配结果比较后不一致，则利用聚类分析结果更新关键词词典。

[0020] 一种基于NLP语义识别的供应链金融文件管理系统，其中，包括：

[0021] 获取模块，用于接收并获取用户上传的文件，将所获取文件处理为OCR可识别形式；

[0022] 识别模块，用于对处理后文件进行OCR识别；

[0023] 匹配模块，用于将识别模块识别后数据利用预设关键词词典进行匹配得到匹配结果，其中，所述关键词词典与各分类文件存在对应关系；

[0024] 分析模块，用于对识别后数据进行词元提取，利用所提取词元更新预设的词元权重，并利用更新后词元权重进行聚类分析，其中，各分类文件与词元权重存在对应关系；

[0025] 预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系；

[0026] 所述预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系具体为：

[0027] 对存量文件OCR识别后进行分词处理，对分词后的词元利用预设权重计算规则计算词元权重；利用计算得到的词元权重进行文本聚类分析，对聚类后的分类结果进行手动标注，得到分类文件及对应的高权重关键词词典；

[0028] 所述对分词后的词元利用预设权重计算规则计算词元权重具体为：对分词后的词元利用预设公式$W_{td} = TF_{td} \cdot \log\frac{n}{DF_t} + TF_{td} \cdot \frac{AF_t}{n}$计算词元权重，其中，t为词元，$W_{td}$为词元t在文档d中的权重，$TF_{td}$为词元t在文档d中出现的次数，n为文档总数，$DF_t$为包含词元t的文档数，$AF_t$为词元t总共出现的次数；

[0029] 所述利用计算得到的词元权重进行文本聚类分析具体为:将所得到的的每一存量文件对应的词元权重利用K-Means聚类算法进行聚类分析,其中,使用公式

$$\theta = \arccos\left(\frac{\vec{a}\cdot\vec{b}}{||a||\cdot||b||}\right)$$

替换K-Means算法中的欧式距离进行聚类分析;

[0030] 比较模块,用于将聚类分析结果与匹配结果进行比较;

[0031] 执行模块,用于根据比较模块比较结果,对文件进行归类。

[0032] 一种电子设备,包括存储器、至少一个处理器以及存储在所述存储器上并可在所述处理器上运行的计算机程序,其中,所述处理器执行所述程序时执行如上所述的方法。

[0033] 有益效果:

[0034] 本发明基于NLP语义识别的供应链金融文件管理方法能够根据语义实现对多类型文件的自动识别归类,尤其适用于文件类型多,数量多的供应链金融资料的管理,在供应链金融的场景下,通过文件归类的自动化和智能化,提升了文件管理效率,有利于对各种实际业务的精细化运营,使得金融机构能够更好的服务于供应链上的客户。

附图说明

[0035] 图1是本发明具体实施例中基于NLP语义识别的供应链金融文件管理方法流程图;

[0036] 图2是本发明具体实施例中针对存量贸易背景资料实施的本方明方法流程图;

[0037] 图3是本发明具体实施例中针对新增贸易背景资料实施的本方明方法流程图;

[0038] 图4是本发明具体实施例中基于NLP语义识别的供应链金融文件管理系统原理框图。

具体实施方式

[0039] 为了使本领域的人员更好地理解本发明的技术方案,下面结合本发明的实施例,对本发明的技术方案进行清楚、完整的描述,基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动的前提下所获得的其它类同实施例,都应当属于本申请保护的范围。此外,以下实施例中提到的方向用词,例如“上”“下”“左”“右”等仅是参考附图的方向,因此,使用的方向用词是用来说明而非限制本发明创造。

[0040] 本发明技术方案的实现基于OCR识别技术和聚类分析技术的成熟,如图1所示的一种基于NLP语义识别的供应链金融文件管理方法,其中,包括以下步骤:

[0041] S100、接收并获取用户上传的文件,将所获取文件处理为OCR可识别形式。

[0042] 获取用户上传的文件后,识别所获取文件的格式,客户上传的资料可能有多种格式(pdf、png、jpg、doc、docx等),需要对格式进行归一化处理,即基于预先设定的格式转化规则将文件格式转换为image格式。具体的,对于PDF文件,使用FITZ算法进行处理,生成image图片列表;对于DOC、DOCX文件,先使用ExportAsFixedFormat模块进行PDF转化,后将PDF文件进行image转换;对于压缩包形式的资料,则先尝试对压缩包进行解压缩,若解压异常,则进行分卷解压缩处理,后对于解压后的文件进行前述各种形式的image转换。

[0043] S200、对处理后文件进行OCR识别,将识别后数据利用预设关键词词典进行匹配得到匹配结果,其中,所述关键词词典与各分类文件存在对应关系。

[0044] 进一步的,在步骤S100之前还包括:预先构建分类文件和关键词词典,并建立分类

文件与关键词词典间的对应关系。

[0045] 构建分类文件和关键词词典具体发方法为:对存量文件OCR识别后进行分词处理,对分词后的词元利用预设权重计算规则计算词元权重。

[0046] 利用计算得到的词元权重进行文本聚类分析,对聚类后的分类结果进行手动标注,得到分类文件及对应的高权重关键词词典。

[0047] 进一步的,所述对分词后的词元利用预设权重计算规则计算词元权重具体为:对分词后的词元利用预设公式$W_{td} = TF_{td} \cdot \log\frac{n}{DF_t} + TF_{td} \cdot \frac{AF_t}{n}$计算词元权重,其中,t为词元,$W_{td}$为词元t在文档d中的权重,$TF_{td}$为词元t在文档d中出现的次数,n为文档总数,$DF_t$为包含词元t的文档数,$AF_t$为词元t总共出现的次数;

[0048] 所述利用计算得到的词元权重进行文本聚类分析具体为:将所得到的的每一存量文件对应的词元权重利用K-Means聚类算法进行聚类分析,其中,使用公式$\theta = \arccos(\frac{\vec{a}\cdot\vec{b}}{||a||\cdot||b||})$替换K-Means算法中的欧式距离进行聚类分析。

[0049] 得到分类文件及对应的高权重关键词词典后,通过聚类分析计算得到各分类文件中的最边缘文件,使用关键词词典对最边缘文件进行关键词匹配分析,若关键词匹配结果不一致,则去除相应分类中匹配度最差文件,并重新进行聚类分析。

[0050] S300、对识别后数据进行词元提取,利用所提取词元更新预设的词元权重,并利用更新后词元权重进行聚类分析,其中,各分类文件与词元权重存在对应关系。

[0051] S400、将聚类分析结果与匹配结果进行比较,若一致,则将文件按照匹配结果进行归类。若聚类分析结果与匹配结果比较后不一致,则利用聚类分析结果更新关键词词典。

[0052] 本发明通过历史存量数据的分析训练,实现输入一份全新文件(合同),就可以输出其所属类别。在供应链金融的场景下,有利于对各种实际业务的精细化运营,更好的服务于供应链上的每一个客户。

[0053] 本发明方法主要包括两部分,第一部分是利用存量文件建立相对应的文件分类和关键词词典,第二部分则是利用建立好的文件分类和关键词词典,对新增文件的分类过程。下面分别通过实施例1和2进行描述。

[0054] 实施例1:对于存量文件(金融领域称为存量贸易背景资料,包含合同、发票、结算单等资料)的分类以及关键词词典的生成。

[0055] 在此存量文件以合同文件举例,客户上传的合同背景资料进行image转化过后,得到图像数组image_list,使用OCR引擎对图像数组进行图像识别,并对识别结果进行持久化存储。

[0056] 对存量的所有识别后的合同,使用中文分词器进行分词处理,得到每一份合同中包含的词元列表,进一步得到每一份合同中包含那些词元、每一份合同中的词元在该合同中出现的次数、每一个词元共在多少份合同中出现过、每一个词元在所有合同中出现的次数。

[0057] 然后对每一个词元进行权重计算。

[0058] 使用计算出的权重,将词元列表转化为每一份合同的词元权重列表。

[0059] 将每一个词元权重列表看作是一个向量在向量空间中进行计算。

[0060] 使用公式$\theta = \arccos(\frac{\vec{a}\cdot\vec{b}}{\|a\|\cdot\|b\|})$替换K-Means算法中的欧式距离进行聚类分

[0061] 析。

[0062] 进一步地得到分类好的合同，并人工对各个分类进行标注，标注出分类后的分类名称。并根据记录好的词元信息，得到各个分类的高权重关键词，以及各个分类的质心向量。

[0063] 使用质心向量计算出各个分类中最边缘的合同，并使用各个分类中的关键词对分类中的合同进行关键词校验性分析并打分，对上述两个计算中打分较低的合同进行剔除，并重复聚类分析步骤。

[0064] 通过上述步骤，得到合同分类图谱以及对应的关键词词典。

[0065] 实施例2：利用实施例1中得到的合同分类图谱已经对应的关键词词典对新增合同资料进行分类。

[0066] 在供应链金融业务开展的前期阶段，由业务系统异步的将合同资料发送至背景资料处理服务器进行OCR识别及后续处理。

[0067] 进一步地，对OCR识别后得到的文本进行关键词匹配处理，进行合同的初步分类。

[0068] 另外对文本中的词元进行提取，并对词元权重列表进行更新计算，利用新词元权重列表进行变种聚类分析。

[0069] 进一步地，对上述两个步骤的分类进行比较，若相同则直接对业务系统进行分类反馈。若两种识别方式的结果不相同，则对合同进行特殊标注，后续对标注的合同进行处理，若使用关键词匹配进行的分类不准确，则重新进行实施例1的计算步骤，对关键词词典进行更新。

[0070] 本发明技术方案主要是由于供应链金融领域的合同涉及的类型较多导致合同管理困难，无法进行精细化管理所引申出的，由于供应链金融这一问题比较突出，且供应链金融的合同管理对于业务的顺利实施又非常重要，因此，本发明技术方案的顺利实现就非常有必要，下面进一步通过如图2和图3所示的具体实例对本发明方案在供应链金融中的实现进行具体阐述。

[0071] 如图2所示，基本存量贸易背景，本发明方法具体包括以下步骤：

[0072] S1、获取到贸易背景资料信息(物理位置名称、格式、所关联业务编号等)。

[0073] S2、对各种格式的资料进行归一处理，转换为image格式。

[0074] S3、对image_list进行OCR处理。

[0075] S4、通过识别后文本进行分词处理。

[0076] S5、文本聚类分析，得到分类及各类型中的高频词。

[0077] S6、对结果进行可信度和规范性验证。

[0078] 进一步的，如图3所示，当用户上传新增贸易背景资料，本发明方法具体包括以下步骤：

[0079] L1、获取到贸易背景资料信息(物理位置名称、格式、所关联业务编号等)。

[0080] L2、对各种格式的资料进行归一处理，转换为image格式。

[0081] L3、对image_list进行OCR处理。

[0082] L4、使用关键词词典对文本进行匹配性验证并确定贸易背景资料所述类型。

[0083] L5、是否判断正确。若是，则执行步骤L7，若否，则执行步骤L6。

[0084] L6、对关键词词典进行更新计算。

[0085] L7、对业务系统进行反馈。

[0086] 综合图2和图3表述，本发明整体实施步骤具体如下：

[0087] 1、获取到供应链金融服务平台上客户上传的合同之后，由于文件类型多种多样，对于OCR引擎需要传入统一的文件格式，所以在此步骤，将文件类型进行归一化处理。所述归一化处理即建立格式转换规则，首先读取所获取文件的文件格式，在预设的格式转换对应表中查找该文件类型所对应的转换算法，调用该转换算法对所获取文件进行格式转换，本方法中，由于下一步需要对所获取文件进行OCR识别，则需要将不同文件类型转换成统一的image格式。对于PDF文件，使用FITZ算法进行处理，生成image图片列表；对于DOC、DOCX文件，先使用ExportAsFixedFormat模块进行PDF转化，后将PDF文件进行image转换；对于压缩包形式的资料，先尝试对压缩包进行解压缩，若解压异常，则进行分卷解压缩处理，后对于解压后的文件进行前述各种形式的image转换。

[0088] 较佳的，当所获取文件的文件格式无法识别，则将该文件发送给人工处理端进行人工处理，人工处理端系统记录人工对文件的打开路径和打开方式，并将所记录的这些数据返回给系统，系统基于所记录数据自行建立系统对于该类文件格式的识别算法。或者，系统第一次接收到特定格式的且无法识别的文件时，分别将其发送至文件上传端和特定文件处理端，要求上传端和特定文件处理端对该文件进行格式处理以便能够识别，记录上传端和特定文件处理端对该文件的处理过程和处理路径，当其中一方发送该文件可识别版本后，系统自动将该客户端并入系统中，使该客户端赋予系统访问和调用客户端的权限，在之后接收到相同格式的文件后，将该类文件利用该客户端以及所记录的处理路径和处理方法对文件进行转换。

[0089] 或者，人工处理端对于系统发送的未识别文件进行屏幕截图，即文件在人工处理端打开过程中，对所打开文件自动进行屏幕截图，将所截取屏幕保存为image格式，并将其返回给系统，由系统进行下一步的OCR识别，较佳的是，屏幕截图可以以固定频率截取，或者系统记录屏幕发生显示数据变化时(在文件打开状态下)截屏。

[0090] 2、对步骤1得到的image图像列表进行OCR识别，并将识别结果持久化存储。

[0091] 3、对步骤2得到的识别文本，使用JIEBA中文分词器进行分词处理，得到每一份合同中包含的词元列表，进一步得到每一份合同中包含那些词元、每一份合同中的词元在该合同中出现的次数、每一个词元共在多少份合同中出现过、每一个词元在所有合同中出现的次数。使用公式$W_{td} = TF_{td} \cdot \log\frac{n}{DF_t} + TF_{td} \cdot \frac{AF_t}{n}$计算词元权重，其中，t为词元，$W_{td}$为词元t在文档d中的权重，$TF_{td}$为词元t在文档d中出现的次数，n为文档总数，$DF_t$为包含词元t的文档数，$AF_t$为词元t总共出现的次数；利用计算得到的权重能够将词元列表转化为每一份合同的词元权重列表。

[0092] 较佳的是，系统对所识别结果进行初步判断，判断所获取的合同是否为系统中已经存储过的，若查找到系统中存在完全一样的合同，则该合同向上传端发送提示信息，再获得上传端确认信息后，系统再对该合同进行下一步处置。系统对于相同文件的判断需要按

照步骤进行，首先是对于识别后词元列表中的文字词元进行比对，是否完全一致，如果是，接着对词元列表中的数字词元进行比对，当数字词元也完全一致，则说明新上传文件为系统中已经存储的合同，系统向文件上传端发送提示信息，当然，较佳的是，系统对于上传文件(合同)中的盖章进行识别，识别红章中内容和红章的位置信息，当文字词元和数字词元比对完全一致时，利用所识别的红章内容和位置信息进行进一步比对，若红章内容和位置信息完全一致，则证明新上传文件确为系统中已存储文件，发送上传合同已存在于系统，请对合同信息进行确认的提示信息给上传端，若红章内容和位置信息中至少有一项不同，则说明新上传文件可能为新合同，向上传端发送上传合同盖章不同重新确认合同日期的提示信息。进一步的，当文字词元比对完全一致，而数字词元和识别的红章信息比对不完全一致时，则需要系统对数字词元的性质进行判断，区分出是金额、日期等数字词元，具体区分方法为建立所识别的数字词元与文字词元的距离对应表，例如合同日期部分“签订日期:1月2日”其中的数字词元“1”分别与文字词元“日期”和“月”的距离记为0，数字“2”分别与文字词元“月”和“日”距离记为0，利用距离对应表根据距离远近可识别处合同日期和金额的数字词元，数字词元不一致时，具体判断是金额不同还是日期不同，当文字词元比对完全一致时，进一步的比对金额词元和日期词元，金额词元和日期词元均不同时，则说明该上传合同在为系统存在同类型文件，可直接将该上传合同归类到这一类型中，金额词元相同而日期词元不同时，则上传合同应该为续签合同，也是直接将上传合同归类到同一类型中，金额词元不同而日期词元相同时，可能为补充合同，这是需要系统向上传端发送提示信息，由上传端确认。

[0093] 4、根据步骤3得到的权重，对每一份合同生成一个向量。对已有的向量进行聚类分析，本发明使用的是K-Means聚类算法，其中距离计算使用向量的夹角

$$\theta = \arccos\left(\frac{\vec{a}\cdot\vec{b}}{\|a\|\cdot\|b\|}\right)$$

进行替换。使用公式替换K-Means算法中的欧式距离进行

[0094] 聚类分析。

[0095] 5、对聚类分析后的分类结果进行手动标注后，例如标注采购、分包、租赁等标签，得到了分好类的各类合同以及对应的高权重关键词词典，以及各个分类的质心向量信息。为对结果进行校验以及确定是否有噪声样本。使用关键词词典对于各类质心向量夹角最大的合同(分类中最边缘合同)进行关键词匹配分析，若关键词匹配结果不理想，则对合同样本进行特殊标注。对于特殊标注的样本需进行干预后判断是否将样本剔除或重新进行步骤3-5。

[0096] 针对聚类分析后各分类中的最边缘合同利用各分类的关键词词典进行匹配打分，最边缘合同的的分值由匹配上的关键词权重决定，例如，可将匹配上的关键词权重相加即为最边缘合同的分值，将打分中最低的合同从当前分类中剔除，重新进行聚类分析，也可设置打分阈值，在低于相应阈值情况下，相应的合同被剔除当前分类重新聚类分析。

[0097] 6、对于业务发生中的新增合同，在业务的制单前期，通过网络接口方式将合同相关信息(物理位置、名称、格式、所关联业务编号等)传送给系统，系统进行合同分析服务，后续在合同分析服务中，执行步骤1-2。

[0098] 7、使用步骤5得到的关键词词典对步骤6得到的识别文本进行关键词匹配分析，同

时利用词元标识对步骤3中涉及的词元权重进行更新计算，并基于新的词元权重列表进行变种聚类分析，系统可周期性进行变种聚类分析，例如每天进行一次变种聚类分析，若聚类分析结果和关键词匹配结果一致，则新增合同完成分类，若发现聚类分析结果与关键词匹配分析结果有差异，则对新增合同进行特殊标注，后续对标注的合同进行处理，利用新聚类分析结果更新关键词词典，以实现识别算法的自主学习能力。

[0099] 8、根据步骤7得到的分析结果，反馈至业务系统，完成供应链金融合同类型的分类。

[0100] 优选实施例中，利用本发明技术方案中所建立的合同分类机制，例如关键词词典等，进一步建立已分类合同间的关联关系，具体为，对于已分类的合同，利用其关键词词典抓取关于合同方以及合同性质的关键词，利用前述预先建立的词元距离对应表辅助系统进行判断，例如系统获取合同中关键词元“甲方”、“乙方”，进一步获取距离甲方乙方词元距离为0或+1的词元(基于目标词元右侧按照间隔词元数量设置+距离，目标词元左侧按照间隔词元数量设置-距离)，从而识别合同甲乙方的具体单位，进一步通过分类标签或者进一步根据关键词词典中相应关键词和权重或者根据词元距离对应表的进一步分析确认该合同的合同性质，例如为采购、分包等类型，建立关于该合同及甲乙方之间的关联关系(建立关联关系对应表，关联关系对应表中设置该合同的合同方、业务编号、合同类型、金额、日期等)，当接收到新上传文件例如销售发票，识别分类后，利用识别后词元列表并基于上述方法判断出甲乙方、业务编号等根据这些判断信息在系统已存在的关联关系对应表中进行匹配，若匹配成功，则说明该销售发票属于对应合同的，并可进一步通过银行付款回单等文件的匹配进一步验证，之后建立该销售发票与对应合同之间的关联关系，这样方便了文件上传端的上传，可简化其上传工作，无需填写相关信息，通过系统自动识别关联即可，在查看其中某一文件时，系统基于关联关系同时调用其他关联文件以备查看。

[0101] 另一实施例中，文件在上传时，业务人员需要同时填写上传关于该文件的相关信息，包括物理位置、名称、业务编号等

[0102] 如图4所示，本发明还提供一种基于NLP语义识别的供应链金融文件管理系统，其中，包括：

[0103] 获取模块100，用于接收并获取用户上传的文件，将所获取文件处理为OCR可识别形式。

[0104] 识别模块200，用于对处理后文件进行OCR识别。

[0105] 匹配模块300，用于将识别模块识别后数据利用预设关键词词典进行匹配得到匹配结果，其中，所述关键词词典与各分类文件存在对应关系。

[0106] 分析模块400，用于对识别后数据进行词元提取，利用所提取词元更新预设的词元权重，并利用更新后词元权重进行聚类分析，其中，各分类文件与词元权重存在对应关系；

[0107] 预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系；

[0108] 所述预先构建分类文件和关键词词典，并建立分类文件与关键词词典间的对应关系具体为：

[0109] 对存量文件OCR识别后进行分词处理，对分词后的词元利用预设权重计算规则计算词元权重；利用计算得到的词元权重进行文本聚类分析，对聚类后的分类结果进行手动标注，得到分类文件及对应的高权重关键词词典；

[0110] 所述对分词后的词元利用预设权重计算规则计算词元权重具体为:对分词后的词元利用预设公式$W_{td} = TF_{td} \cdot \log\frac{n}{DF_t} + TF_{td} \cdot \frac{AF_t}{n}$计算词元权重,其中,t为词元,$W_{td}$为词元t在文档d中的权重,$TF_{td}$为词元t在文档d中出现的次数,n为文档总数,$DF_t$为包含词元t的文档数,$AF_t$为词元t总共出现的次数;

[0111] 所述利用计算得到的词元权重进行文本聚类分析具体为:将所得到的的每一存量文件对应的词元权重利用K-Means聚类算法进行聚类分析,其中,使用公式$\theta = \arccos(\frac{\vec{a}\cdot\vec{b}}{||a||\cdot||b||})$替换K-Means算法中的欧式距离进行聚类分析;

[0112] 比较模块500,用于将聚类分析结果与匹配结果进行比较。

[0113] 执行模块600,用于根据比较模块比较结果,对文件进行归类。

[0114] 本发明提供一种电子设备,包括存储器、至少一个处理器以及存储在所述存储器上并可在所述处理器上运行的计算机程序,其中,所述处理器执行所述程序时执行如上所述的方法。

[0115] 本发明基于NLP语义识别的供应链金融文件管理方法能够根据语义实现对多类型文件的自动识别归类,尤其适用于文件类型多,数量多的供应链金融资料的管理,在供应链金融的场景下,通过文件归类的自动化和智能化,提升了文件管理效率,有利于对各种实际业务的精细化运营,使得金融机构能够更好的服务于供应链上的客户。

[0116] 以上已将本发明做一详细说明,以上所述,仅为本发明之较佳实施例而已,当不能限定本发明实施范围,即凡依本申请范围所作均等变化与修饰,皆应仍属本发明涵盖范围内。

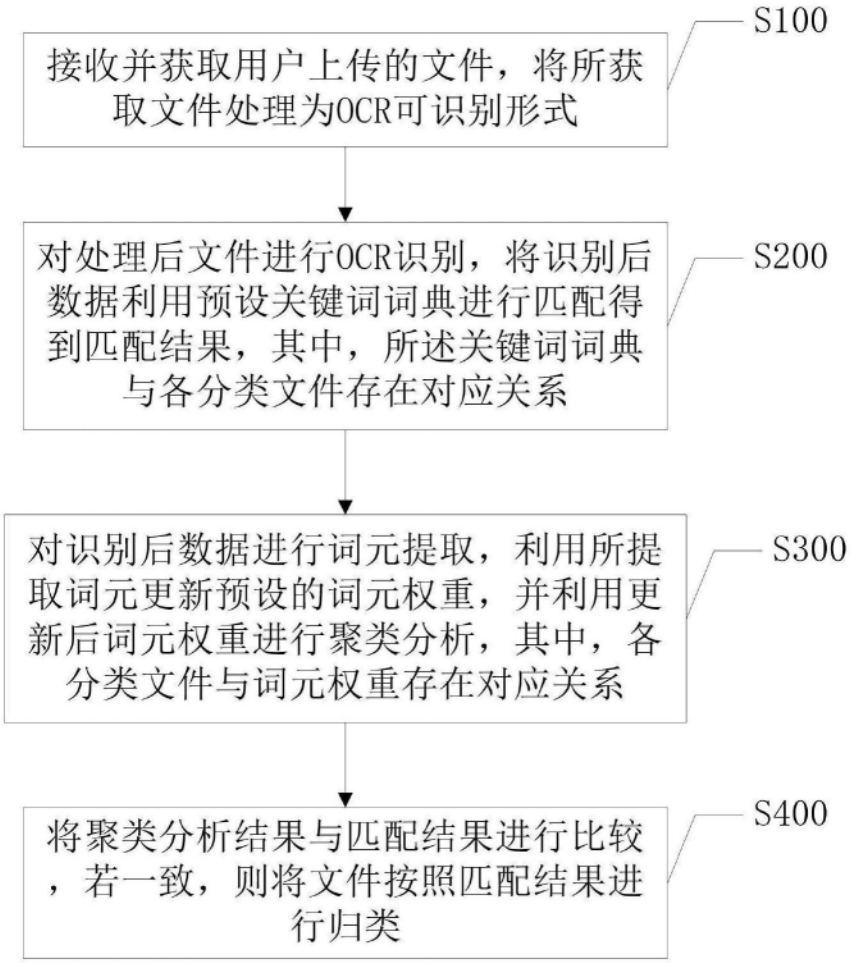

S100
接收并获取用户上传的文件，将所获取文件处理为OCR可识别形式
S200
对处理后文件进行OCR识别，将识别后数据利用预设关键词词典进行匹配得到匹配结果，其中，所述关键词词典与各分类文件存在对应关系
S300
对识别后数据进行词元提取，利用所提取词元更新预设的词元权重，并利用更新后词元权重进行聚类分析，其中，各分类文件与词元权重存在对应关系
S400
将聚类分析结果与匹配结果进行比较，若一致，则将文件按照匹配结果进行归类

图1

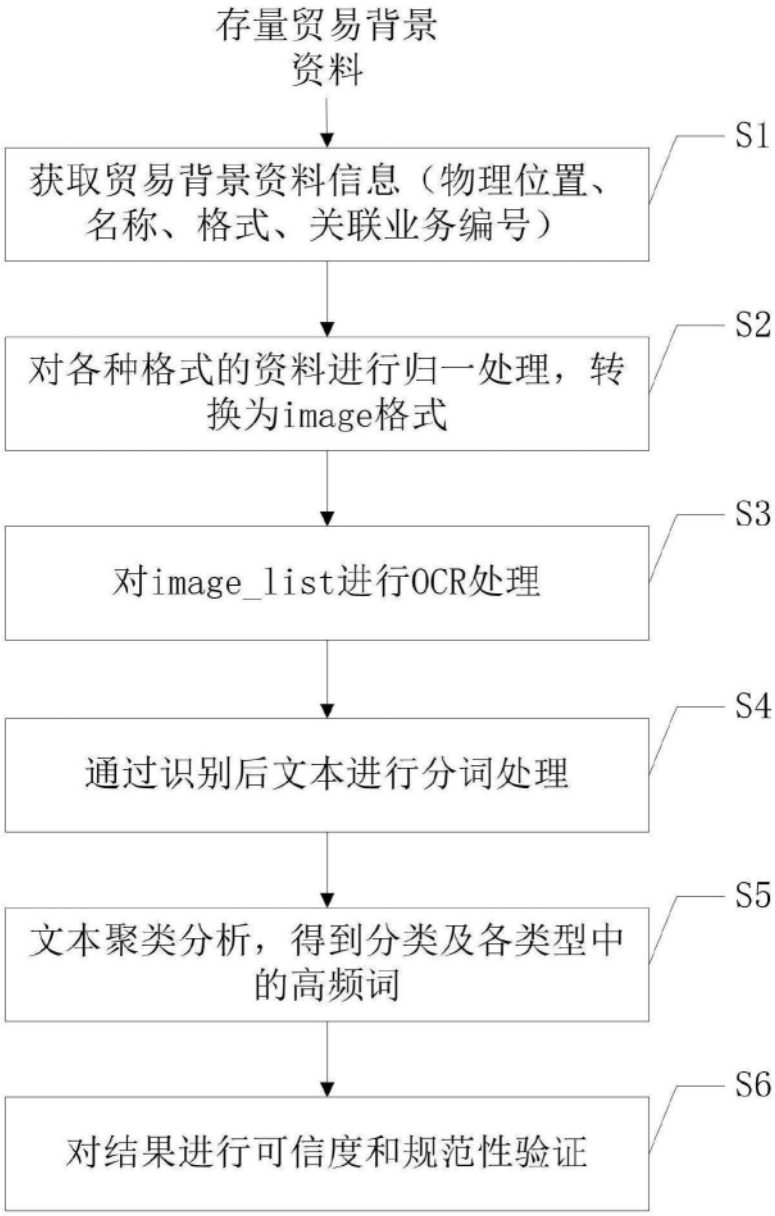

存量贸易背景资料
获取贸易背景资料信息（物理位置、名称、格式、关联业务编号）
S1
对各种格式的资料进行归一处理，转换为image格式
S2
对image_list进行OCR处理
S3
通过识别后文本进行分词处理
S4
文本聚类分析，得到分类及各类型中的高频词
S5
对结果进行可信度和规范性验证
S6

图2

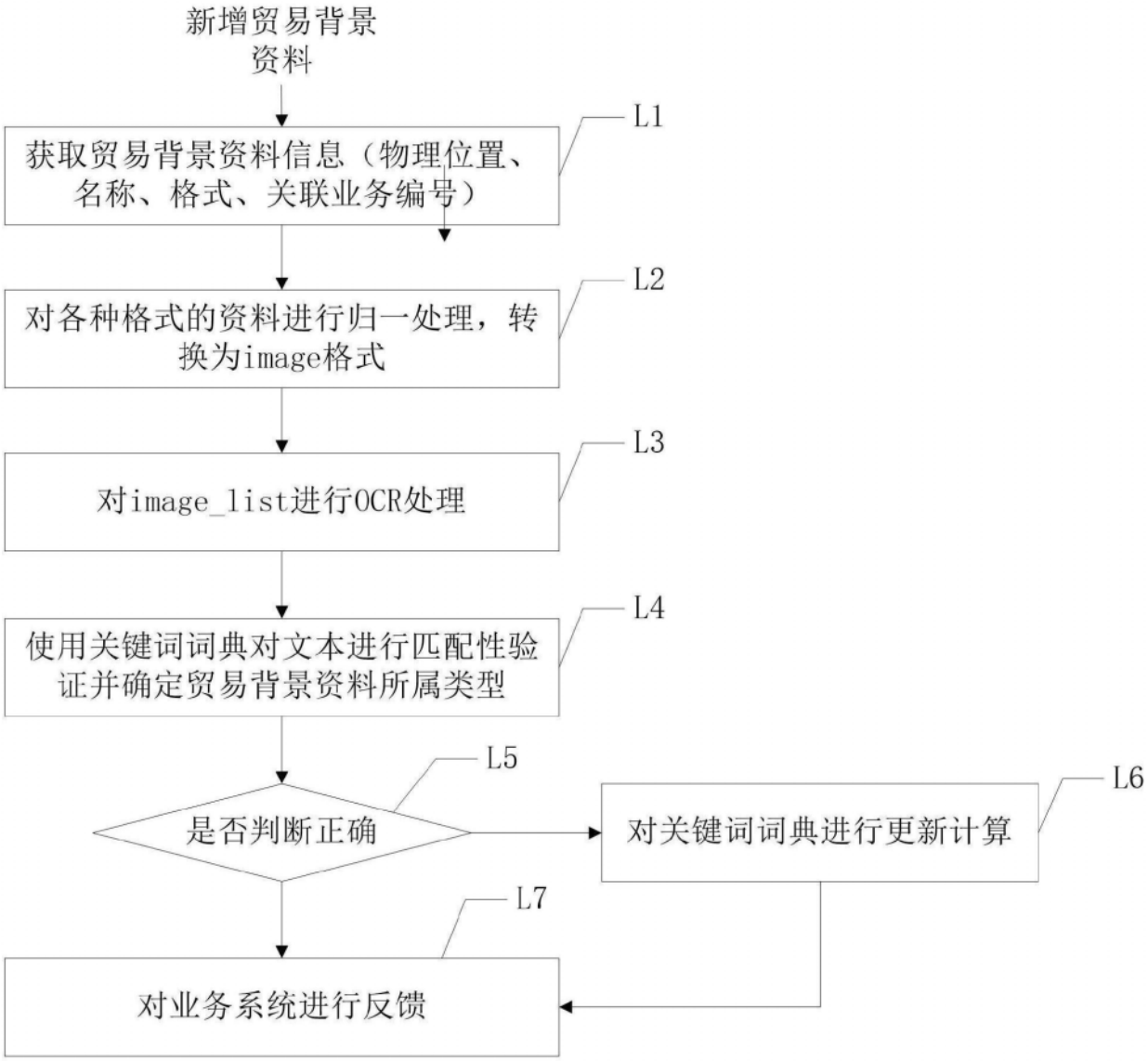

新增贸易背景资料
获取贸易背景资料信息（物理位置、名称、格式、关联业务编号）
L1
对各种格式的资料进行归一处理，转换为image格式
L2
对image_list进行OCR处理
L3
使用关键词词典对文本进行匹配性验证并确定贸易背景资料所属类型
L4
是否判断正确
L5
对关键词词典进行更新计算
L6
对业务系统进行反馈
L7

图3

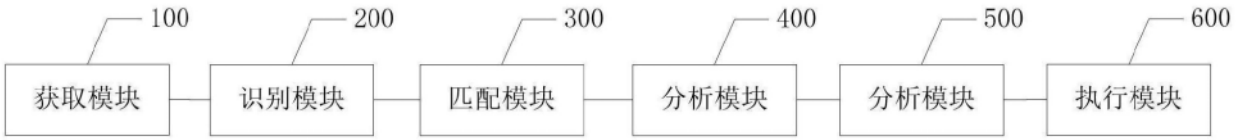

100
获取模块
200
识别模块
300
匹配模块
400
分析模块
500
分析模块
600
执行模块

图4