(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117574009 A
(43)申请公布日 2024.02.20

(21)申请号 202311431850.2

(22)申请日 2023.10.31

(71)申请人 灵犀科技有限公司
地址 266426 山东省青岛市中国(山东)自由贸易试验区青岛片区龙门山路136号803室

(72)发明人 姜琳杰 吴楠 赛哲锋

(74)专利代理机构 北京万思博知识产权代理有限公司 11694
专利代理师 徐敏

(51)Int.Cl.
G06F 16/958(2019.01)
G06F 16/951(2019.01)
G06F 40/30(2020.01)
G06F 40/289(2020.01)
G06Q 50/26(2024.01)

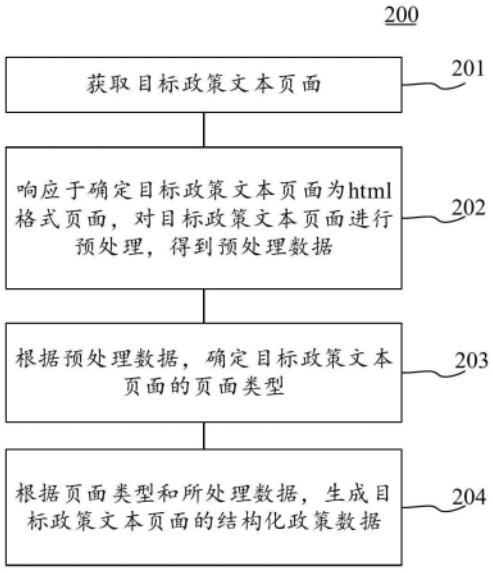
权利要求书2页 说明书9页 附图3页

(54)发明名称

结构化政策数据生成方法、装置、电子设备和可读介质

(57)摘要

本公开的实施例公开了结构化政策数据生成方法、装置、电子设备和可读介质。该方法的一具体实施方式包括:获取目标政策文本页面;响应于确定目标政策文本页面为html格式页面,对目标政策文本页面进行预处理,得到预处理数据;根据预处理数据,确定目标政策文本页面的页面类型;根据页面类型和预处理数据,生成目标政策文本页面的结构化政策数据。该实施方式针对政策类网站进行复杂字段和多样化数据提取,可以精准提取政策内容、字段,还保存html页面正文内容中的图片、表格等内容以及相应位置,使重新解析渲染后的页面图文并茂,更加便于阅读。



CN 117574009 A

1.一种结构化政策数据生成方法,其特征在于,包括:

获取目标政策文本页面;

响应于确定所述目标政策文本页面为html格式页面,对所述目标政策文本页面进行预处理,得到预处理数据;

根据所述预处理数据,确定所述目标政策文本页面的页面类型;

根据所述页面类型和所述预处理数据,生成目标政策文本页面的结构化政策数据。

2.根据权利要求1所述的方法,其特征在于,所述根据所述预处理数据,确定所述目标政策文本页面的页面类型,包括:

将所述预处理数据输入至预先训练的页面分类模型中,得到所述目标政策文本页面的页面类型,其中,所述页面分类模型是以样本预处理数据为输入,样本预处理数据对应的样本页面类型为期望输出,对SVM模型进行训练得到的。

3.根据权利要求1所述的方法,其特征在于,所述根据所述页面类型和所述预处理数据,生成结构化政策数据,包括:

响应于确定所述页面类型为列表页类型,从所述预处理数据中确定节点集;

从所述节点集中确定相似节点集,并将所述相似节点集合并为组节点;

提取所述组节点的标题和超链接地址,并根据所述标题和所述超链接地址生成结构化政策数据。

4.根据权利要求1所述的方法,其特征在于,所述根据所述页面类型和所述预处理数据,生成结构化政策数据,包括:

响应于确定所述页面类型为正文页类型,对所述预处理数据进行特征提取,得到所述目标政策文本页面的文本特征;

根据所述文本特征将所述目标政策文本页面分为正文分区和标题分区;

将所述正文分区和所述标题分区作为结构化政策数据。

5.根据权利要求4所述的方法,其特征在于,所述根据所述文本特征将所述目标政策文本页面分为正文分区和标题分区,包括:

将所述文本特征输入至预先训练的页面分区模型中,得到所述目标政策文本页面的分区结果,其中,所述页面分区模型是以样本文本特征为输入,样本文本特征对应的样本分区结果为期望输出,将结合GNE文本及标点符号密度提取算法、历史权重算法以及newspaper nlp算法的SVM模型作为初始模型进行训练得到的。

6.根据权利要求1所述的方法,其特征在于,所述方法还包括:

从所述结构化政策数据中确定正文数据;

将所述正文数据输入至预先训练的摘要生成模型,得到所述正文数据对应的文本摘要,其中,所述摘要生成模型是以样本正文数据为输入,样本正文数据对应的样本文本摘要为期望输出,以交叉熵为损失函数,对LSTM模型进行训练得到的。

7.一种结构化政策数据生成装置,其特征在于,包括:

获取单元,被配置成获取目标政策文本页面;

预处理单元,被配置成响应于确定所述目标政策文本页面为html格式页面,对所述目标政策文本页面进行预处理,得到预处理数据;

确定单元,被配置成根据所述预处理数据,确定所述目标政策文本页面的页面类型;

生成单元,被配置成根据所述页面类型和所述预处理数据,生成目标政策文本页面的结构化政策数据。

8.根据权利要求7所述的装置,其特征在于,所述确定单元被进一步配置成:

将所述预处理数据输入至预先训练的页面分类模型中,得到所述目标政策文本页面的页面类型,其中,所述页面分类模型是以样本预处理数据为输入,样本预处理数据对应的样本页面类型为期望输出,对SVM模型进行训练得到的。

9.一种电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,其上存储有一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-6中任一所述的方法。

10.一种计算机可读介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现如权利要求1-6中任一所述的方法。

结构化政策数据生成方法、装置、电子设备和可读介质

技术领域

[0001] 本公开的实施例涉及计算机技术领域,具体涉及结构化政策数据生成方法、装置、电子设备和计算机可读介质。

背景技术

[0002] 目前的网页解析,主要针对场景为新闻舆情类,提取字段为发布时间、作者以及新闻主要内容。但政策类内容的结构与新闻类结构不同,新闻类结构主要标题、内容两大部分组成;政策类为文号、时间等组成一部分,政策内容部分、下方各市县网站部分、页面附件、视频以及表格等部分。

[0003] 因此,对于政策类网页进行网页解析时,往往会出现多种问题:页面组成内容较多比较杂乱,算法提取会造成一定的干扰,另外,采集程序的开发主要集中在网页解析上,文本类型相关网站非常多,不同网站的网页布局、排版、风格、内容各不相同,每个网站、网页都需要单独编写解析逻辑,实际操作过程中,程序维护非常频繁,维护成本高,维护效率差,而且未维护的程序会导致,后续数据混乱,更新不及时,影响线上数据,因此,为了降低人工编写成本,提高线上数据实时性,也就需要编写或少量编写解析逻辑就可以自动提取网页数据的方法。

发明内容

[0004] 本公开的内容部分用于以简要的形式介绍构思,这些构思将在后面的具体实施方式部分被详细描述。本公开的内容部分并不旨在标识要求保护的技术方案的关键特征或必要特征,也不旨在用于限制所要求的保护的技术方案的范围。

[0005] 本公开的一些实施例提出了结构化政策数据生成方法、装置、电子设备和计算机可读介质,来解决以上背景技术部分提到的技术问题。

[0006] 第一方面,本公开的一些实施例提供了一种结构化政策数据生成方法,该方法包括:获取目标政策文本页面;响应于确定上述目标政策文本页面为html格式页面,对上述目标政策文本页面进行预处理,得到预处理数据;根据上述预处理数据,确定上述目标政策文本页面的页面类型;根据上述页面类型和上述预处理数据,生成目标政策文本页面的结构化政策数据。

[0007] 第二方面,本公开的一些实施例提供了一种结构化政策数据生成装置,装置包括:获取单元,被配置成获取目标政策文本页面;预处理单元,被配置成响应于确定上述目标政策文本页面为html格式页面,对上述目标政策文本页面进行预处理,得到预处理数据;确定单元,被配置成根据上述预处理数据,确定上述目标政策文本页面的页面类型;生成单元,被配置成根据上述页面类型和上述预处理数据,生成目标政策文本页面的结构化政策数据。

[0008] 第三方面,本申请实施例提供了一种电子设备,该网络设备包括:一个或多个处理器;存储装置,用于存储一个或多个程序;当一个或多个程序被一个或多个处理器执行,使

得一个或多个处理器实现如第一方面中任一实现方式描述的方法。

[0009] 第四方面,本申请实施例提供了一种计算机可读介质,其上存储有计算机程序,该计算机程序被处理器执行时实现如第一方面中任一实现方式描述的方法。

[0010] 本公开的上述各个实施例中的一个实施例具有如下有益效果:针对政策类网站,实现复杂字段和多样化数据的提取,针对表格、附件以及政策相关字段做了相应的处理,从而可以精准提取政策内容以及字段,另外,本申请为了增加解析完整性,还保存html页面正文内容中的图片、表格等内容以及相应位置,使重新解析渲染后的页面图文并茂,更加便于阅读。

附图说明

[0011] 结合附图并参考以下具体实施方式,本公开各实施例的上述和其他特征、优点及方面将变得更加明显。贯穿附图中,相同或相似的附图标记表示相同或相似的元素。应当理解附图是示意性的,元件和元素不一定按照比例绘制。

[0012] 图1是根据本公开的一些实施例的结构化政策数据生成方法的一个应用场景的示意图;

图2是根据本公开的结构化政策数据生成方法的一些实施例的流程图;

图3是根据本公开的结构化政策数据生成装置的一些实施例的结构示意图;

图4是适于用来实现本公开的一些实施例的电子设备的结构示意图。

具体实施方式

[0013] 下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施例,然而应当理解的是,本公开可以通过各种形式来实现,而且不应该被解释为限于这里阐述的实施例。相反,提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是,本公开的附图及实施例仅用于示例性作用,并非用于限制本公开的保护范围。

[0014] 另外还需要说明的是,为了便于描述,附图中仅示出了与有关发明相关的部分。在不冲突的情况下,本公开中的实施例及实施例中的特征可以相互组合。

[0015] 需要注意,本公开中提及的“第一”、“第二”等概念仅用于对不同的装置、模块或单元进行区分,并非用于限定这些装置、模块或单元所执行的功能的顺序或者相互依存关系。

[0016] 需要注意,本公开中提及的“一个”、“多个”的修饰是示意性而非限制性的,本领域技术人员应当理解,除非在上下文另有明确指出,否则应该理解为“一个或多个”。

[0017] 本公开实施方式中的多个装置之间所交互的消息或者信息的名称仅用于说明性的目的,而并不是用于对这些消息或信息的范围进行限制。

[0018] 下面将参考附图并结合实施例来详细说明本公开。

[0019] 图1是根据本公开一些实施例的结构化政策数据生成方法的一个应用场景的示意图。

[0020] 如图1所示,结构化政策数据生成方法得执行主体服务器101可以获取目标政策文本页面102,之后,响应于确定目标政策文本页面102为html格式页面,对目标政策文本页面102进行预处理,得到预处理数据103,再根据预处理数据103,确定目标政策文本页面102的页面类型104,最后,服务器101根据页面类型104和预处理数据103,生成目标政策文本页面

102的结构化政策数据105。

[0021] 可以理解的是,结构化政策数据生成方法可以是由终端设备来执行,或者也可以是由服务器101来执行,上述方法的执行主体还可以包括上述终端设备与上述服务器101通过网络相集成所构成的设备,或者还可以是各种软件程序来执行。其中,终端设备可以是具有信息处理能力的各种电子设备,包括但不限于智能手机、平板电脑、电子书阅读器、膝上型便携计算机和台式计算机等等。执行主体也可以体现为服务器101、软件等。当执行主体为软件时,可以安装在上述所列举的电子设备中。其可以实现成例如用来提供分布式服务的多个软件或软件模块,也可以实现成单个软件或软件模块。在此不做具体限定。

[0022] 应该理解,图1中的服务器数目仅仅是示意性的。根据实现需要,可以具有任意数目的服务器。

[0023] 继续参考图2,示出了根据本公开的结构化政策数据生成方法的一些实施例的流程200。该结构化政策数据生成方法,包括以下步骤:

步骤201,获取目标政策文本页面。

[0024] 在一些实施例中,结构化政策数据生成方法的执行主体(例如图1所示的服务器)可以获取目标政策文本页面。

[0025] 步骤202,响应于确定上述目标政策文本页面为html格式页面,对上述目标政策文本页面进行预处理,得到预处理数据。

[0026] 在一些实施例中,响应于确定上述目标政策文本页面为html格式页面,上述执行主体(例如图1所示的服务器)可以对上述目标政策文本页面进行预处理,得到预处理数据。具体的,上述执行主体可以对目标政策文本页面中option、embed、media、style大多广告、评论、样式等干扰标签以及对应的数据进行清除。

[0027] 步骤203,根据上述预处理数据,确定上述目标政策文本页面的页面类型。

[0028] 在一些实施例中,上述执行主体可以根据上述预处理数据,确定上述目标政策文本页面的页面类型。

[0029] 在一些实施例的一些可选的实现方式中,上述执行主体可以将上述预处理数据输入至预先训练的页面分类模型中,得到上述目标政策文本页面的页面类型,其中,上述页面分类模型是以样本预处理数据为输入,样本预处理数据对应的样本页面类型为期望输出,对SVM模型进行训练得到的。

[0030] 在这里,上述SVM模型的基本公式:

$f(x)=sign(\sum_{i=1}^{n} a_i y_i K(x,x_i)+b)$,其中,x 是输入样本,$a_i$是对应样本的拉格朗日乘子,$y_i$是样本的类别,K 是核函数。

[0031] 具体的,上述执行主体可以从预处理数据提取文本密集度、超链接节点的数量等特征,选取页面分类模型,进行分区识别,并根据分类识别结果的置信度score值,给定相应页面类型。

[0032] 步骤204,根据上述页面类型和上述预处理数据,生成目标政策文本页面的结构化政策数据。

[0033] 在一些实施例中,上述执行主体可以根据上述页面类型和上述预处理数据,生成目标政策文本页面的结构化政策数据。

[0034] 在一些实施例的一些可选的实现方式中,响应于确定上述页面类型为列表页类

型,上述执行主体可以从上述预处理数据中确定节点集;从上述节点集中确定相似节点集,并将上述相似节点集合并为组节点;提取上述组节点的标题和超链接地址,并根据上述标题和上述超链接地址生成结构化政策数据。

[0035] 具体的,列表页类型页面处理逻辑如下:使用lxml库,转为Element对象,选取相似组节点,例如列表页大部分由
、<talbe><td>等标签实现,下方包裹超链接<a>标签,且有多个同类型相似标签组合而成。根据这些特征,可以提取出多个相似标签,然后选取其中相似度最高组节点,进行合并,提取标题、超链接地址,通过二维列表组合后返回结构化政策数据。

[0036] 在一些实施例的一些可选的实现方式中,响应于确定上述页面类型为正文页类型,上述执行主体可以对上述预处理数据进行特征提取,得到上述目标政策文本页面的文本特征;根据上述文本特征将上述目标政策文本页面分为正文分区和标题分区;将上述正文分区和上述标题分区作为结构化政策数据。

[0037] 具体的,详情页分区处理逻辑如下:使用lxml库,转为Element对象,进行分区,筛选出类似正文部分的node节点,单独提取出来。筛选条件:提取node节点中的文本内容,获取该内容中包含的指定的政策文件相关停词数量;获取文本中包含超链接的数量,通过判断停词数量以及超链接的数据,来筛选该节点;对筛选出的节点进行数据处理,删除顶部节点;清洗 把a标签节点中的文本内容取出,并替换该节点;把br标签节点也就是正文中的换行符改为\n;ul、li标签取出文本内容,使用\n来保留样式;替换b、strong、i、sup等标签这些都是正文中用于设置样式相关标签,去除后可以保证正文内容的格式规范;删除没有文本内容的标签;删除node中的最后一个顶级节点,还有DOM深度太深的节点,因为大多为媒体资源、加载库、相关网站等数据;对节点中的图片、表格、附件等进行定位,页面位置替换为MD5加密参数,从而保留页面图片、表格、附件等位置;进行节点文本提取,并选取文字密集度较高的,保留格式以字符串形式返回。

[0038] 另外,在将上述正文分区和上述标题分区作为结构化政策数据的步骤之前,上述执行主体还会对上述正文分区和上述标题分区的数据进行以下处理:

对于页面分区中的正文分区:提取正文分区的文本特征(大多为表格样式,例如标签:table、ul、等类似标签),利用文本特征将提取正文分区的相关节点;提取上述相关节点中的文本内容,使用编写的正则规则(文本类型对应关系表)对文本内容和文本类型进行匹配,从而获取索引号、发文字号、主题分类、发布机构等字段内容。

[0039] 对于页面分区中的标题分区的标题字段:提取meta标签、根据contains定位的标签、h1、h2标题常用标签、title标签的文本内容,用两个字符串(两个标签的文本内容)的交集字符数量除以两个字符串的并集字符数量,选取置信度最高结果,进行干扰数据清洗,如br等标签。

[0040] 对于页面分区中的标题分区的时间字段:时间字段是大部分网站存在字段且格式有迹可循,所以提取规则如下:使用正则匹配url中可能包含的时间信息;根据通用时间信息所使用的标签,编写xpath正则,对其内容进行提取;根据所有时间格式,编写正则表达式来提取。

[0041] 对进行三种处理后的处理结果进行时间类型判断,并统一时间格式后进行整合,返回结构化政策数据。

[0042] 通过上述方式针对表格、附件以及政策相关字段做了相应的处理,从而可以精准提取政策内容和相应字段,同时为增加解析的完整性,定位保存html页面正文内容中的图片、表格等位置,使重新解析渲染后的页面图文并茂,更加便于阅读,对近百个字段名附近标签的对比,生成结构化数据,提高了网页数据提取的准确性和自动化程度,采用分布式采集并经过多轮调参,提高了该方式的稳定性和可靠性,同时减少了人工编写解析逻辑的成本和频率,提高了线上数据的实时性。

[0043] 在一些实施例的一些可选的实现方式中,上述执行主体可以将上述文本特征输入至预先训练的页面分区模型中,得到上述目标政策文本页面的分区结果,其中,上述页面分区模型是以样本文本特征为输入,样本文本特征对应的样本分区结果为期望输出,将结合GNE文本及标点符号密度提取算法、历史权重算法以及newspaper nlp算法的SVM模型作为初始模型进行训练得到的。

[0044] 具体的,文本密度提取算法:

$\mathrm{TF-IDF(t,d)=TF(t,d)*}log\left(\frac{N}{DF(t)}\right)$,其中,t 是单词,d 是文档,N 是文档总数,TF(t,d) 是单词 t 在文档 d 中的出现次数, DF(t) 是包含单词 t 的文档数量。

[0045] 分区识别与权重标记:使用 SVM 模型训练一个分类器,将 HTML 内容分为标题、正文等分区。SVM 的训练过程会涉及特征工程和训练集的标注。根据 GerapyAutoExtractor 和 GNE 等算法,对各个分区进行权重标记,考虑历史权重、NLP 分析权重以及文本密集度等因素。

[0046] 借鉴 newspaper 中相关节点上下文连接密度算法,对近百个字段名附近标签进行对比,生成结构化数据,涉及对 HTML 树的解析,以及一些基于规则或者启发式的方法来提取结构化信息。

[0047] 1、分区识别与权重标记表示

svm_model = train_svm_model(training_data);

partition_labels = svm_model.predict(html_content);

weighted_partitions=mark_partitions(partition_labels,history_weights,nlp_weights,text_density_weights),

其中,raining_data表示用于 SVM 模型训练的数据集,包括标记好的 HTML 内容以及它们对应的分区标签;html_content表示输入值html页面;svm_model表示训练好的 SVM 模型,用于对新的 HTML 内容进行分区识别;partition_labels表示从 SVM 模型中得到的 HTML 分区标签,指示那个部分是字段、正文等;history_weights表示历史权重;nlp_weights表示nlp权重;text_density_weights表示文本密集度权重;weighted_partitions表示对分区进行的权重标记,包括历史权重、NLP 分析权重、文本密集度权重等信息。

[0048] 结构化数据生成表示

structured_data=generate_structured_data(html_content,field_names),

其中,html_content表示待处理的 HTML 内容;field_names:预设的字段名称;structured_data表示生成的结构化数据,一个字典、JSON 对象,包含从 HTML 中提取的结构信息。

[0049] 历史权重、nlp权重、文本密集度权重为预先设定好的权重值,用于对各个分区进

行不同程度加权。通过权重标记,可以更加准确的确定那部分是关键内容,以及更好的适应不同类型的网页。svm在样本数据集偏小的情况下也有好的效果,泛化能力和鲁棒性较好,用于分区识别。文本及标点符号密度提取算法可以用于提取关键信息;nlp相关算法可以用于生成摘要;svm模型可以提供分区的初始识别结果,然后通过gen算法和newspaper nlp算法进行进一步的分析和提取,从而得到更准确的结果,通过该算法对网页进行分区识别和权重标记,得到列表页分区和正文分区的HTML,从而解决了解析过程中的干扰问题。

[0050] 在一些实施例的一些可选的实现方式中,上述执行主体可以从上述结构化政策数据中确定正文数据;将上述正文数据输入至预先训练的摘要生成模型,得到上述正文数据对应的文本摘要,其中,上述摘要生成模型是以样本正文数据为输入,样本正文数据对应的样本文本摘要为期望输出,以交叉熵为损失函数,对LSTM模型进行训练得到的。

[0051] 具体的,基于深度学习的生成式摘要:

lstm_model = train_lstm_model(text_data, summary_data);

generated_summary = generate_summary(lstm_model, text_data),其中,text_data表示文本数据集;text_data表示摘要数据集;lstm_model表示训练好的LSTM模型,用于生成新闻文章的摘要;lstm_model表示训练好的LSTM模型,用于生成新闻文章的摘要;text_data表示文本输入值;generated_summary表示由LSTM模型生成的摘要。

[0052] 基于深度学习的生成式摘要算法,通过该算法生成与原始文本相关的简短摘要,提高了数据处理效率和准确性。

[0053] 本公开的上述各个实施例中的一个实施例具有如下有益效果:针对政策类网站,实现复杂字段和多样化数据的提取,针对表格、附件以及政策相关字段做了相应的处理,从而可以精准提取政策内容以及字段,另外,本申请为了增加解析完整性,还保存html页面正文内容中的图片、表格等内容以及相应位置,使重新解析渲染后的页面图文并茂,更加便于阅读。

[0054] 进一步参考图3,作为对上述各图所示方法的实现,本公开提供了一种结构化政策数据生成装置的一些实施例,这些装置实施例与图2所示的那些方法实施例相对应,该装置具体可以应用于各种电子设备中。

[0055] 如图3所示,一些实施例的结构化政策数据生成装置300包括:获取单元301、预处理单元302、确定单元303和生成单元304。其中,获取单元,被配置成获取目标政策文本页面;预处理单元,被配置成响应于确定上述目标政策文本页面为html格式页面,对上述目标政策文本页面进行预处理,得到预处理数据;确定单元,被配置成根据上述预处理数据,确定上述目标政策文本页面的页面类型;生成单元,被配置成根据上述页面类型和上述预处理数据,生成目标政策文本页面的结构化政策数据。

[0056] 在一些实施例的可选实现方式中,上述确定单元被进一步配置成:将上述预处理数据输入至预先训练的页面分类模型中,得到上述目标政策文本页面的页面类型,其中,上述页面分类模型是以样本预处理数据为输入,样本预处理数据对应的样本页面类型为期望输出,对SVM模型进行训练得到的。

[0057] 在一些实施例的可选实现方式中,上述生成单元被进一步配置成:响应于确定上述页面类型为列表页类型,从上述预处理数据中确定节点集;从上述节点集中确定相似节点集,并将上述相似节点集合并为组节点;提取上述组节点的标题和超链接地址,并根据上

述标题和上述超链接地址生成结构化政策数据。

[0058] 在一些实施例的可选实现方式中,上述生成单元被进一步配置成:响应于确定上述页面类型为正文页类型,对上述预处理数据进行特征提取,得到上述目标政策文本页面的文本特征;根据上述文本特征将上述目标政策文本页面分为正文分区和标题分区;将上述正文分区和上述标题分区作为结构化政策数据。

[0059] 在一些实施例的可选实现方式中,上述生成单元被进一步配置成:将上述文本特征输入至预先训练的页面分区模型中,得到上述目标政策文本页面的分区结果,其中,上述页面分区模型是以样本文本特征为输入,样本文本特征对应的样本分区结果为期望输出,将结合GNE文本及标点符号密度提取算法、历史权重算法以及newspaper nlp算法的SVM模型作为初始模型进行训练得到的。

[0060] 在一些实施例的可选实现方式中,上述装置还包括摘要生成单元,被配置成:从上述结构化政策数据中确定正文数据;将上述正文数据输入至预先训练的摘要生成模型,得到上述正文数据对应的文本摘要,其中,上述摘要生成模型是以样本正文数据为输入,样本正文数据对应的样本文本摘要为期望输出,以交叉熵为损失函数,对LSTM模型进行训练得到的。

[0061] 可以理解的是,该装置300中记载的诸单元与参考图2描述的方法中的各个步骤相对应。由此,上文针对方法描述的操作、特征以及产生的有益效果同样适用于装置300及其中包含的单元,在此不再赘述。

[0062] 本公开的上述各个实施例中的一个实施例具有如下有益效果:针对政策类网站,实现复杂字段和多样化数据的提取,针对表格、附件以及政策相关字段做了相应的处理,从而可以精准提取政策内容以及字段,另外,本申请为了增加解析完整性,还保存html页面正文内容中的图片、表格等内容以及相应位置,使重新解析渲染后的页面图文并茂,更加便于阅读。

[0063] 下面参考图4,其示出了适于用来实现本公开的一些实施例的电子设备(例如图1中的服务器)400的结构示意图。图4示出的电子设备仅仅是一个示例,不应对本公开的实施例的功能和使用范围带来任何限制。

[0064] 如图4所示,电子设备400可以包括处理装置(例如中央处理器、图形处理器等)401,其可以根据存储在只读存储器(ROM)402中的程序或者从存储装置408加载到随机访问存储器(RAM)403中的程序而执行各种适当的动作和处理。在RAM 403中,还存储有电子设备400操作所需的各种程序和数据。处理装置401、ROM 402以及RAM 403通过总线404彼此相连。输入/输出(I/O)接口405也连接至总线404。

[0065] 通常,以下装置可以连接至I/O接口405:包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置406;包括例如液晶显示器(LCD)、扬声器、振动器等的输出装置407;包括例如磁带、硬盘等的存储装置408;以及通信装置409。通信装置409可以允许电子设备400与其他设备进行无线或有线通信以交换数据。虽然图4示出了具有各种装置的电子设备400,但是应理解的是,并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图4中示出的每个方框可以代表一个装置,也可以根据需要代表多个装置。

[0066] 特别地,根据本公开的一些实施例,上文参考流程图描述的过程可以被实现为计

算机软件程序。例如,本公开的一些实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的一些实施例中,该计算机程序可以通过通信装置409从网络上被下载和安装,或者从存储装置408被安装,或者从ROM 402被安装。在该计算机程序被处理装置401执行时,执行本公开的一些实施例的方法中限定的上述功能。

[0067] 需要说明的是,本公开的一些实施例上述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开的一些实施例中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开的一些实施例中,计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:电线、光缆、RF(射频)等等,或者上述的任意合适的组合。

[0068] 在一些实施方式中,客户端、服务器可以利用诸如HTTP(HyperText Transfer Protocol,超文本传输协议)之类的任何当前已知或未来研发的网络协议进行通信,并且可以与任意形式或介质的数字数据通信(例如,通信网络)互连。通信网络的示例包括局域网(“LAN”),广域网(“WAN”),网际网(例如,互联网)以及端对端网络(例如,ad hoc端对端网络),以及任何当前已知或未来研发的网络。

[0069] 上述计算机可读介质可以是上述电子设备中所包含的;也可以是单独存在,而未装配入该电子设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被该电子设备执行时,使得该电子设备:获取目标政策文本页面;响应于确定上述目标政策文本页面为html格式页面,对上述目标政策文本页面进行预处理,得到预处理数据;根据上述预处理数据,确定上述目标政策文本页面的页面类型;根据上述页面类型和上述预处理数据,生成目标政策文本页面的结构化政策数据。

[0070] 可以以一种或多种程序设计语言或其组合来编写用于执行本公开的一些实施例的操作的计算机程序代码,上述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++,还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)——连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0071] 附图中的流程图和框图,图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0072] 描述于本公开的一些实施例中的单元可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中,例如,可以描述为:一种处理器包括获取单元、预处理单元、确定单元和生成单元。其中,这些单元的名称在某种情况下并不构成对该单元本身的限定,例如,获取单元还可以被描述为“获取目标政策文本页面的单元”。

[0073] 本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如,非限制性地,可以使用的示范类型的硬件逻辑部件包括:现场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、片上系统(SOC)、复杂可编程逻辑设备(CPLD)等等。

[0074] 以上描述仅为本公开的一些较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解,本公开的实施例中所涉及的发明范围,并不限于上述技术特征的特定组合而成的技术方案,同时也应涵盖在不脱离上述发明构思的情况下,由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开的实施例中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

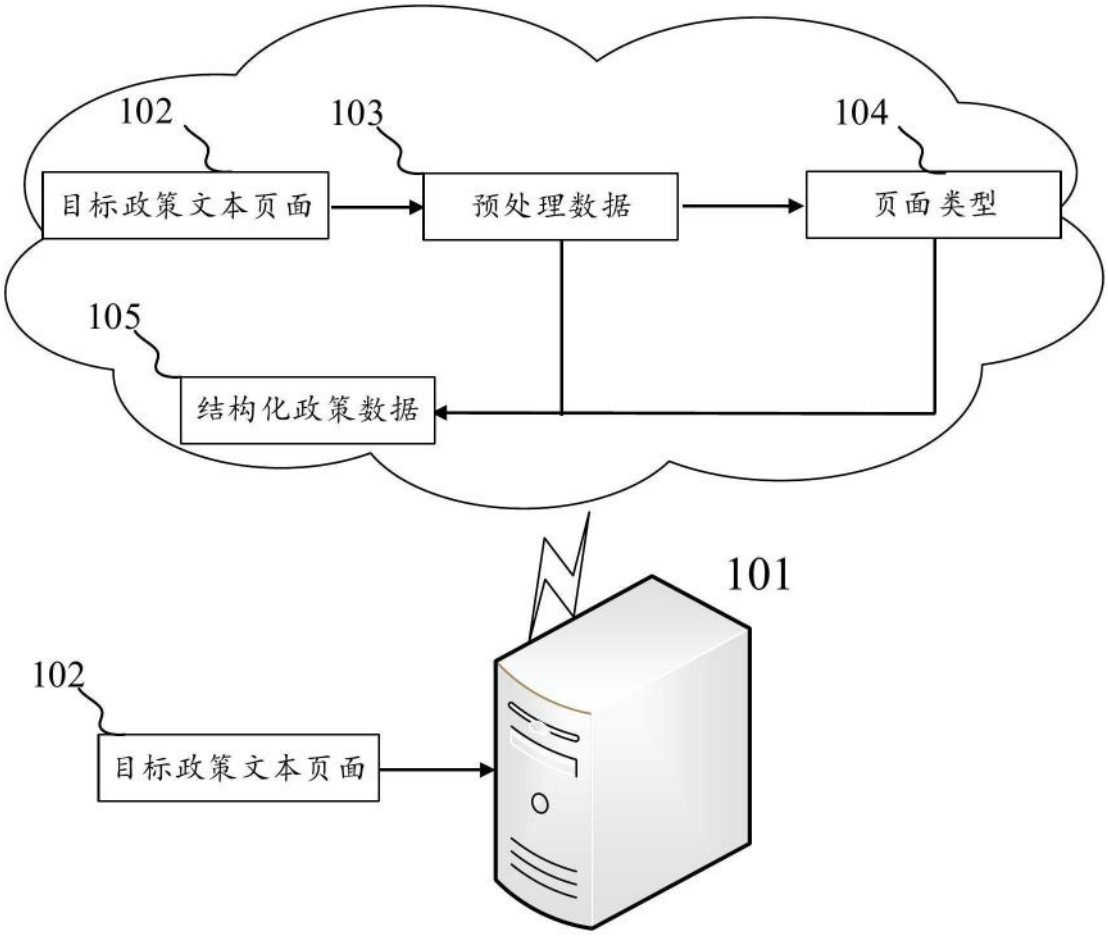

102
目标政策文本页面
103
预处理数据
104
页面类型
105
结构化政策数据
101
102
目标政策文本页面

图1

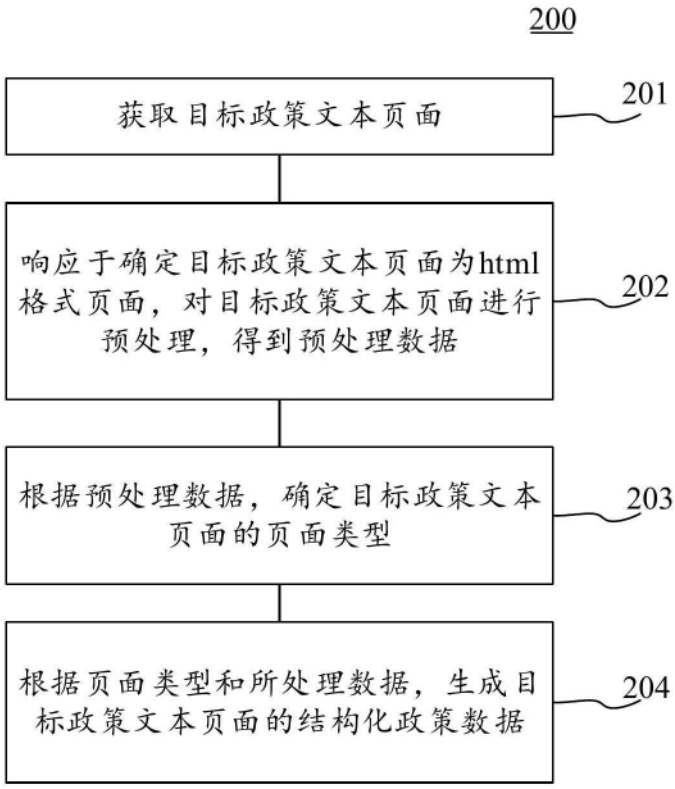

200
获取目标政策文本页面
201
响应于确定目标政策文本页面为html格式页面，对目标政策文本页面进行预处理，得到预处理数据
202
根据预处理数据，确定目标政策文本页面的页面类型
203
根据页面类型和所处理数据，生成目标政策文本页面的结构化政策数据
204

图2

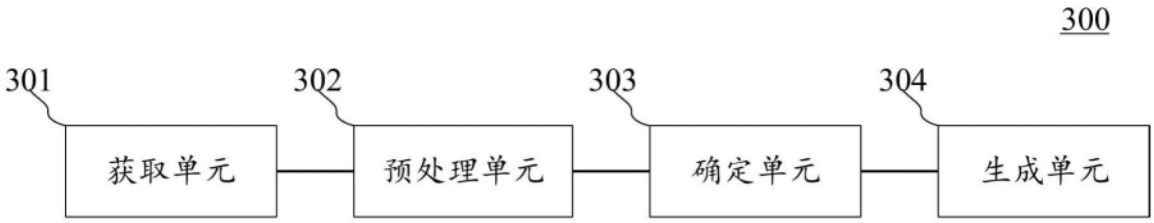

300
301
获取单元
302
预处理单元
303
确定单元
304
生成单元

图3

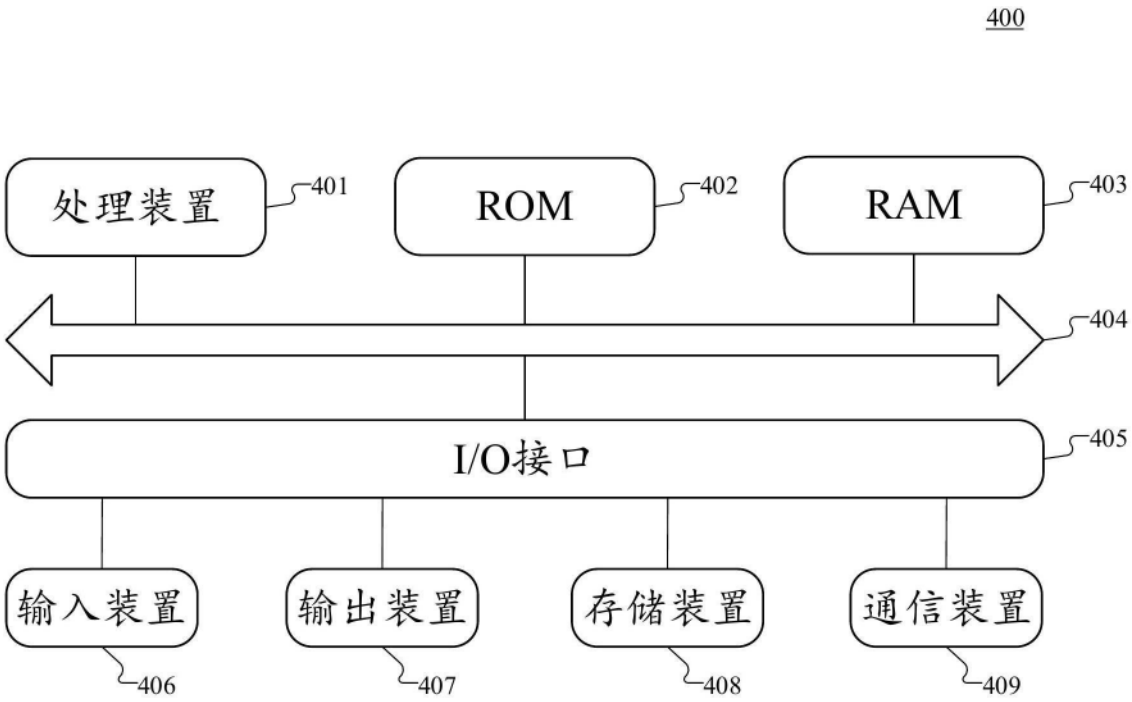

400
处理装置
401
ROM
402
RAM
403
404
I/O接口
405
输入装置
406
输出装置
407
存储装置
408
通信装置
409

图4