发明名称
一种纳税人资料图像采集云端处理方法

摘要
本发明提供一种纳税人资料图像采集云端处理方法，包括以下步骤：A. PC上的图像采集云端软件集成 TWAIN 兼容驱动实现；当扫描仪扫描图像后，图像采集云端软件通过内置的 OCR 组件将图像识别图像文本，将文本传递到物联网服务软件；B. 物联网服务软件收到文本信息后，分析出该图像属于何种资料，并根据资料的种类将图像处理的算法，包括加密、压缩、去噪、格式、大小精度，传递到云端软件；C. 云端软件接收到资料算法后，通过内置的图像处理组件，依照下载的算法进行图像处理；D. 云端软件向云服务软件发起请求，通过 HTTP 方式将图像向云服务软件传递；E. 云端软件由 DHT 网络组成，可实现对云服务软件的离散下载，并对通讯量进行排序和优化调整，减少对网络带宽的消耗。
1. 一种纳税人资料图像采集云端处理方法，其特征在于包括以下处理步骤：
   A. 客户端 PC 上的图像采集云端软件集成 TWAIN 兼容驱动，当扫描仪扫描图像后，图像采集云端软件通过内置的 OCR 组件将图像识别图像文本，将文本传递到物联网服务软件；
   B. 物联网服务软件收到文本信息后，分析出该图像属于何种资料，并根据资料的种类将图像处理的算法，包括加密、压缩、去噪、格式、大小精度，传递到云端软件；
   C. 云端软件接收到资料算法后，通过内置的图像处理组件，依照下载的算法进行图像处理；
   D. 云端软件向云服务软件发起请求，通过 HTTP 方式将图像向云服务软件传递；
   E. 云端软件由 DHT 网络组成，可实现对云服务软件的离散下载，并对通讯量进行排序和优化调整，减少对网络带宽的消耗。

该方法的实现环境为客户端 PC 通过 USB 接口或高速 SCSI 接口连接扫描仪，并通过以太网或互联网连接物联网处理服务器和云计算处理服务器。客户端 PC 安装图像采集云端软件，物联网处理服务器安装物联网服务软件，云计算处理服务器装云服务软件，其中图像采集云端软件，包括：1) TWAIN 兼容驱动软件；2) 图像 OCR 处理软件；3) DHT 网络上传下载及反向代理软件；4) 传输过程的通讯量调整软件；5) 算法装卸软件，其中：

1) TWAIN 兼容驱动软件
   TWAIN 兼容驱动采用 USB 接口适配，处理过程包括：
   1) 加载 Twain Source Manager，获得 DSM_Entry 区域；
   2) 启动 Twain Source Manager；
   3) 加载 Twain the Source；
   4) 启动 Twain the Source；
   5) 配 Twain the Source；
   6) 获得调整信号数据；
   7) 识别传输信道命令格式；
   8) 启动传输；
   9) 完成传输；
   10) 关闭 TWAIN 会话；

2) 图像 OCR 处理软件
   采用业界常用的 OCR 软件，包括基于 DSP 版面分析或文字特征分析算法，最终实现对图像文字的提取；

3) DHT 网络上传下载及其反向代理软件
   云端软件需要对扫描的图像文件和 OCR 文本进行上传，以加载到云计算服务器上，采用 DHT 反向代理技术以提高传输速度，以最大化的利用 ADSL 网络传输的带宽耗用；

   云端软件采用 DHT 反向代理软件，提供向不同云服务端的分块上传，处理过程包括：
   (1) 加载云服务 server table；
   (2) 测试连接速度，并排序；
   (3) 读取图像和文本的分块节点位置和节点信息；
   (4) Upload 各节点信息到 servers table 各云服务上；
   (5) 更新 server table；
（6）制定云服务 Upload 主节点；
（7）启动 UDP 连接，按照更新后的 server table 进行上传；
（8）完成传输；
（9）关闭会话；

云端软件浏览图像时，需要从云服务端下载图像或相关文本信息，此时采用 DHT 网络直接下载；

4）传输排序的通讯量调整软件
在云端软件通过 DHT 网络进行传输时，可能对网络更宽造成资源的冲突和浪费，此时应采用通讯量的调整软件，即 Traffic Shaping，处理方法如下：
（1）在内存建立 ACK tables，采用 hashtable 进行存储；
（2）当创建 DHT 网络传输包，封包 ACK 时，将该包的文件名、块位置和大小保存入 ACK tables；
（3）对 ACK tables 按照文件名进行排序；
（4）每次启动块传输时，查询 ACK tables 进行传输；
（5）算法装配软件
算法装配软件，在一定程度上类似沙盘软件，但当前的沙盘软件主要提供虚拟存储区域，而不能提供一个应用的自动计算环境；而算法装配软件提供了一种适用于应用自动下载自动执行的环境，可根据服务端定制的算法要求，提供对应的算法服务，其处理过程包括：
（1）在内存创建三个区域，算法加载区、算法析构区、算法执行区；
（2）算法加载区，将网络下载的算法进行加载、分析语法结构、进行文本检查；
（3）算法析构区，将算法文本进行拆解分析，形成指令集合；
（4）算法执行区，利用常用的编译执行环境，包括 VC、加载析构后的指令集和执行算法。
一种纳税人资料图像采集云端处理方法

[0001]

技术领域
[0002] 本发明涉及一种计算机应用技术领域，具体地说是一种通过云计算、物联网技术对税务常用的资料扫描采集，实现扫描图像的云端处理及分布计算。

背景技术
[0003] 随着信息科技的发展，影像文件逐渐成为信息处理的主要需求之一，尤其是税务部门，纳税人数量激增，其经济活动也趋向于多元化，需要采集的纳税人的纸质资料越来越多，手工管理工作量越来越大。
[0004] 税务机关在日常的税收征管过程中，需要对纳税人的营业执照、办税证明、身份证件和合同等资料进行采集和归档，当前的主要管理方式是手工收集方式，不利于资料的管理，同时给纳税人带来了不方便，办理不同的业务需要反复提交相关资料，也增加了办税成本；随着信息技术的发展，部分税务机关逐渐使用扫描仪、数码相机等图像采集设备采集图像，并利用相关的影像存储管理软件进行管理，但随之又带来了以下问题：

（1）图像质量问题，由于前台操作人员的操作水平和计算机技术能力，对扫描后的图像质量很难做到统一，图像格式的规范性不便于管理。
[0005] （2）工作效率问题，采用扫描仪、数码相机采集大量的资料耗费了较多的时间，反而降低了办税效率。
[0006] （3）电子影像文件的安全可靠存储问题，纳税人的资料涉及商业秘密，只有具有税务调查、稽查权限的税务干部才能查询和调阅，同时电子文件在服务器的存储要避免应设备故障而导致的电子文件损坏问题。
[0007] 而使用电子图像技术近来也发展迅猛，扫描设备方面，馈纸式、平板式扫描仪已经成熟应用，鉴于其体积和重量较大，近年出现了便携式扫描仪，其以便携性和易用性也逐渐被大众所认可，提高了扫描仪的普及率；另外，数码相机技术快速发展，图像清晰度也能够逐渐达到商业影像管理的要求，也成为影像采集的重要工具。云计算和物联网则是最新的技术潮流和趋势，通过云计算平台，可将图像文件存储在云平台上，减少云端的文件存储的压力和安全性问题；而物联网可提供一个联网、智能的设备协同管理平台。

发明内容
[0008] 本发明的目的是提供一种纳税人资料图像采集云端处理及分布计算方法，可以实现和现有的云计算和物联网数据中心进行连接整合，实现图像采集设备上层软件的联网以及网络化的安全存储。
[0009] 本发明的目的方法是按以下方式实现的，包括以下步骤：

A．PC上的图像采集云端软件集成 TWAIN 兼容驱动实现，当扫描仪扫描图像后，图像采集云端软件通过内置的 OCR 组件将图像识别图像文本，将文本传递到物联网服务软件
B. 物联网服务软件收到文本信息后，分析出该图像属于何种资料，并根据资料的种类
将图像处理的算法，包括加密，压缩，去噪，格式，大小精度，传递到云端软件
C. 云端软件接收到资料算法后，通过内置的图像处理组件，依照下载的算法进行图像
处理
D. 云端软件向云服务软件发起请求，通过 HTTP 方式将图像向云服务软件传递
E. 云端软件由 DHT 网络组成，可实现对云服务软件的离散下载，并对通讯量进行排序
和优化调整，减少对网络带宽的消耗
实现本发明方法的环境为客户端 PC 通过 USB 接口或高速 SCSI 接口连接扫描仪，并
通过以太网或互联网连接物联网处理服务器和云计算处理服务器，客户端 PC 安装图像采
集云端软件，物联网处理服务器安装物联网服务软件，云计算处理服务器安装云服务软件，
其中，图像采集云端软件，包括 1) TWAIN 兼容驱动软件 ;2) 图像 OCR 处理软件 ;3) DHT 网络
上传下载及反向代理软件 ;4) 传输排序的通讯量调整软件 ;5) 算法装载软件，其中
1) TWAIN 兼容驱动软件
TWAIN 兼容驱动采用 USB 接口适配，处理过程包括：
1) 加载 Twain Source Manager，获得 DSM_Entry 区域；
2) 启动 Twain Source Manager；
3) 加载 Twain the Source；
4) 启动 Twain the Source；
5) 适配 Twain the Source；
6) 获得调整信号数据；
7) 识别传输信道命令格式；
8) 启动传输；
9) 完成传输；
10) 关闭 TWAIN 会话；
2) 图像 OCR 处理软件
采用业界常用的 OCR 软件，包括基于 DSP 版面分析或文字特征分析算法，最终实现对图
像文字的提取；
3) DHT 网络上传下载及其反向代理软件
云端软件需要对扫描的图像文件和 OCR 文本进行上传，以加载到云计算服务器上，采
用 DHT 反向代理技术以提高传输速度，以最大化的利用 ADSL 网络传输的带宽耗用；
云端软件采用 DHT 反向代理技术，提供向不同云服务端的分块上传，处理过程包括：
1) 加载云服务 server table；
2) 测试连接速度，并排序；
3) 读取图像和文本的分块节点位置和节点信息；
4) Upload 各节点信息到 servers table 各云服务上；
5) 更新 server table；
6) 制定云服务 Upload 主节点；
7) 启动 UDP 连接，按照更新后的 server table 进行上传；
8) 完成传输；
（9）关闭会话；
云端软件浏览图像时，需要从云服务端下载图像或相关文本信息，此时采用 DHT 网络直接下载；
4）传输排序的通讯量调整软件
在云端软件通过 DHT 网络进行传输时，可能对网络带宽造成资源的冲突和浪费，此时应采用通讯量的调整技术，即 Traffic Shaping，处理方法如下：
（1）内存建立 ACK tables，采用 hashtable 进行存储；
（2）当创建 DHT 网络传输包，封包 ACK 时，将该包的文件名、块位置和大小保存入 ACK tables；
（3）对 ACK tables 按照文件名进行排序；
（4）每次启动块传输时，查询 ACK tables 进行传输；
5）算法装载软件
算法装载软件，在一定程度上类似沙盘软件，但当前的沙盘软件主要提供虚拟存储区域，而不能提供一个应用的自动计算机环境；而算法装载技术提供了一种适用于应用自动装载执行的环境，可根据服务端定制的算法要求，提供对应的算法服务，其处理过程包括：
（1）在内存创建三个区域，算法加载区、算法析构区、算法执行区
（2）算法加载区，将网络下载的算法进行加载，分析语义结构，进行文本检验；
（3）算法析构区，将算法文本进行拆解分析，形成指令集合；
（4）算法执行区，利用常用的编译执行环境，包括 VC、加载析构后的指令集和执行算法。

[0010] 本发明的优异效果：本发明主要应用于税务或其他部门，需要采集影像信息的场景，可结合物联网和云计算提供整体的影像处理技术，提高工作效率，而云端技术方法属于直接和操作人员打交道的系统，尤为重要，主要实施的方式包括：
1、建立云计算中心或物联网处理中心；
2、客户开发包含有云端技术方法的云端软件；
3、通过云端软件的配置文件连接云计算中心或物联网中心。

[0011] 其中，云计算中心和物联网处理中心建设需要考虑图像的分布式存储技术、税务业务支持词汇的分词技术、图像常用的优化算法技术、DHT 网络接收技术、文件备份与恢复技术等，但目前这些技术均已经在有类似的成熟案例，因此在操作上实际可行。

[0012] 该方法已经应用于浪潮品牌的税务电子影像管理软件产品上，并在济南、北京进行实际使用和验证。采用本方法，实现了对税务常用纸质档案资料的电子化存储，提供了一种行之有效的电子档案化操作工具。

附图说明
[0013] 图 1 为纳税人资料图像采集云端处理及分布计算方法的整体结构示意图；图 2 为纳税人资料图像采集云端处理及分布计算方法的整体流程图。

具体实施方式
[0014] 参照说明书附图对本发明的方法作以下详细地说明。
本发明的纳税入资料图像采集云端处理方法，包括以下步骤：

A. PC 上的图像采集云端软件集成 TWAIN 兼容驱动实现，当扫描仪扫描图像后，图像采集云端软件通过内置的 OCR 组件将图像识别图像文本，将文本传递到物联网服务软件；

B. 物联网服务软件收到文本信息后，分析出该图像属于何种资料，并根据资料的种类将图像处理的算法，包括加密、压缩、去噪、格式、大小精度，传递到云端软件；

C. 云端软件接收到资料算法后，通过内置的图像处理组件，依照下载的算法进行图像处理；

D. 云端软件向云服务软件发起请求，通过 HTTP 方式将图像向云服务软件传递；

E. 云端软件由 DHT 网络组成，可实现对云服务软件的离散下载，并对通讯量进行排序和优化调整，减少对网络带宽的消耗；

实现本发明方法的环境为客户端 PC 通过 USB 接口或高速 SCSI 接口连接扫描仪，并通过以太网或互联网连接物联网处理服务器和云计算处理服务器，客户端 PC 安装图像采集云端软件，物联网处理服务器安装物联网服务软件，云计算处理服务器安装云服务软件，其中：图像采集云端软件，包括：1) TWAIN 兼容驱动软件；2) 图像 OCR 处理软件；3) DHT 网络上传下载及反向代理软件；4) 传输排序的通讯量调整软件；5) 算法装载软件，其中：

1) TWAIN 兼容驱动软件

TWAIN 兼容驱动采用 USB 接口适配，处理过程包括：

- 加载 Twain Source Manager，获得 DSM_Entry 区域；
- 启动 Twain Source Manager；
- 加载 Twain the Source；
- 启动 Twain the Source；
- 适配 Twain the Source；
- 获得调整信号数据；
- 识别传输信道命令格式；
- 启动传输；
- 完成传输；
- 关闭 TWAIN 会话。

2) 图像 OCR 处理软件

采用业界常用的 OCR 算法即可，包括基于 DSP 版面分析或文字特征分析算法即可最终实现对图像文字的提取；

3) DHT 网络上传下载及其反向代理软件

云端软件需要对扫描的图像文件和 OCR 文本进行上传，以加载到云计算服务器上，采用 DHT 反向代理技术可以在提高传输速度，以最大化的利用 ADSL 等网络传输的带宽耗用。

传统的 DHT 提供一种分布式存储方法。在不需要服务器的情况下，每个客户端负责一个或几个小范围的路由，并负责存储一小部分数据，从而实现整个 DHT 网络的寻址和存储，而云端软件采用 DHT 反向代理技术，提供向不同云服务端的分块上传技术，其主要处理过程包括：
加载云服务 server table；
测试连接速度，并排序；
读取图像和文本的分块节点位置和节点信息；
Upload 各节点信息到 servers table 各云服务上；
更新 server table；
制定云服务 Upload 主节点；
启动 UDP 连接，按照更新后的 server table 进行上传；
完成传输；
关闭会话。

[0018] 而云端软件浏览图像时，需要从云服务端下载图像或相关文本信息，此时采用 DHT 网络直接下载。

[0019] 4) 传输排序的通讯量调整软件

在云端软件通过 DHT 网络进行传输时，可能对网络带宽造成资源的冲突和浪费，此时应采用通讯量的调整技术，即 Traffic Shaping，主要处理方法：

- 在内存建立 ACK tables，采用 hashtable 进行存储；
- 当创建 DHT 网络传输包，封包 ACK 时，将该包的文件名、块位置和大小保存入 ACK tables；
- 对 ACK tables 按照文件名进行排序；
- 每次启动块传输时，查询 ACK tables 进行传输。

[0020] 5) 算法装载软件

算法装载技术在一定程度上类似沙盘技术，但当前的沙盘技术主要提供虚拟存储区域，而不能提供一个应用的自动计算环境；而算法装载技术提供了一种适用于应用自动下载自动执行的环境，可根据服务端定制的算法要求，提供对应的算法服务，其处理过程包括：

（1）在内存创建三个区域，算法加载区、算法析构区、算法执行区
（2）算法加载区，将网络下载的算法进行加载，分析语法结构，进行文本检验；
（3）算法析构区，将算法文本进行拆解分析，形成指令主体；
（4）算法执行区，利用常用的编译执行环境，如 VC，加载析构后的指令集，执行算法。

实施例

[0021] 本发明主要应用于税务或者其他部门，需要采集影像资料的场景，可结合物联网和云计算提供整体的影像处理技术，提高工作效率，而云端技术方法属于直接和操作人员打交道的系统，尤为重要，主要实施的方式包括：

1) 建立云计算中心或物联网处理中心；
2) 客户端开发包含有云端技术方法的云端软件；
3) 通过云端软件的配置文件连接云计算中心或物联网中心。

[0022] 其中，云计算中心和物联网处理中心建设需要考虑图像的分布式存储技术、税务业务支持词汇的分词技术、图像常用的优化算法技术、DHT 网络接收技术，文件的备份与恢复技术等，但目前这些技术均已经在有类似的成熟案例，因此在操作上实际可行。
【0023】本发明的方法已经应用于浪潮品牌的税务电子影像管理软件产品上，并在济南、北京进行实际使用和验证。采用本方法，实现了对税务常用纸质档案资料的电子化存储，提供了一种行之有效的电子档案化的操作工具。
【0024】除说明书所述的技术特征外，均为本专业技术人员的已知技术。
图 1
图 2