(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114564586 A
(43)申请公布日 2022.05.31

(21)申请号 202210211955.6

(22)申请日 2022.03.04

(71)申请人 中信银行股份有限公司
地址 100020 北京市朝阳区光华路10号院1号楼6-30层、32-42层

(72)发明人 刁培金 王湛 赵宾 孙航 赵翠 张兆海 周佟

(74)专利代理机构 北京市兰台律师事务所 11354
专利代理师 于越 张峰

(51)Int.Cl.
G06F 16/35(2019.01)
G06V 10/764(2022.01)

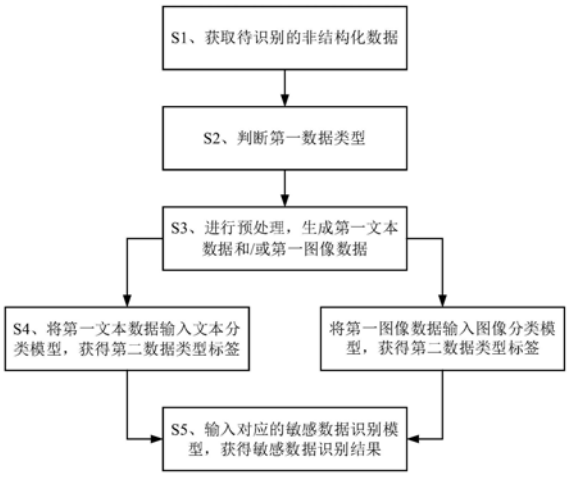
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种非结构化敏感数据识别方法及系统

(57)摘要

本发明涉及一种非结构化敏感数据识别方法及系统，使用相互独立模型分别处理不同类型的非结构化数据敏感识别，用于处理非结构化图像数据和非结构化文本数据，对于非结构化敏感图像数据，能够准确对图像类的非结构化敏感数据进行识别；对于非结构化敏感文本数据，获取上下文的语义结构关系同时能够加大对关键信息的关注度，降低与主题内容关联性较低的冗余信息对分类结果的干扰；解决对特定图像这类非结构化非文本敏感数据的识别问题。



CN 114564586 A

1.一种非结构化敏感数据识别方法,其特征在于,包括:

S1、获取待识别的非结构化数据;

S2、判断非结构化数据的第一数据类型,所述第一数据类型包括非结构化文本数据和非结构化图像数据;

S3、分别对非结构化文本数据和/或非结构化图像数据进行预处理,生成第一文本数据和/或第一图像数据;

S4、将第一文本数据输入文本分类模型,获得对应第一文本数据的第二数据类型标签;

将第一图像数据输入图像分类模型,获得对应第一图像数据的第二数据类型标签;

S5、根据第二数据类型标签,将第一文本数据和/或第一图像数据输入对应的敏感数据识别模型,获得待识别的非结构化数据的敏感数据识别结果。

2.如权利要求1所述的方法,其特征在于,所述第二数据类型标签选自通讯录、好友列表、指纹、网络拓扑结构图和产品设计图纸中的任意一种。

3.如权利要求1所述的方法,其特征在于,所述步骤S2包括:

根据待识别的非结构化数据的文件后缀名判断非结构化数据的第一数据类型。

4.如权利要求1所述的方法,其特征在于,所述预处理包括下列操作中的任意一种或多种组合:

转换文件格式;

更改文件编码;

修改文件名;

文件解密脱壳;

清洗不可用文件;

文件只读属性修改;

色彩空间调整。

5.如权利要求1所述的方法,其特征在于,所述文本分类模型和图像分类模型的生成过程包括:

获取历史数据,对历史数据添加对应的第二数据类型标签生成第一训练数据集;

使用第一训练数据集训练获得文本分类模型和图像分类模型。

6.如权利要求1所述的方法,其特征在于,所述敏感数据识别模型的生成过程包括:

获取历史数据,对历史数据添加对应的第二数据类型标签和敏感数据识别结果生成第二训练数据集;

使用第二训练数据集训练获得针对各第二数据类型标签的敏感数据识别模型。

7.一种非结构化敏感数据识别系统,其特征在于,包括:

数据交互模块,用于获取待识别的非结构化数据并反馈敏感数据识别结果;

第一数据类型分类模块,用于判断非结构化数据的第一数据类型;

数据预处理模块,用于对非结构化文本数据和/或非结构化图像数据进行预处理,生成第一文本数据和/或第一图像数据;

第二数据类型分类模块,用于获得对应第一文本数据和/或第一图像数据的第二数据类型标签;

敏感数据识别模块,用于获得待识别的非结构化数据的敏感数据识别结果。

8.一种计算机可读存储介质，其特征在于，所述存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现权利要求1至6中任一项所述的方法。

9.一种电子设备，其特征在于，包括处理器和存储器；

所述存储器，用于存储第一文本数据和第一图像数据；

所述处理器，用于通过调用第一文本数据和第一图像数据，执行权利要求1至6中任一项所述的方法。

10.一种计算机程序产品，包括计算机程序和/或指令，其特征在于，该计算机程序和/或指令被处理器执行时实现权利要求1至6中任一项所述方法的步骤。

一种非结构化敏感数据识别方法及系统

技术领域

[0001] 本发明涉及数据安全保护技术领域,尤其涉及一种非结构化敏感数据识别方法及系统。

背景技术

[0002] 在数据安全技术领域,现有的敏感数据识别主要以面向结构化数据的方法为主,例如通过对结构化数据的敏感属性识别,实现将结构化数据表中的敏感属性实现自动化识别。此类方法一般采用聚类分析的学习方式,在未知样本集的分类情况下,根据属性间相似度自动实现数据属性的分类;在敏感属性的最终识别阶段,考虑敏感属性与疑似敏感属性之间的关联关系,解决链接攻击的问题,从而进一步挖掘与敏感属性有关联的属性,减少隐私的泄露程度。

[0003] 另一类现有技术将自然语言处理等AI方法应用于该领域。此类技术通过基于人工智能的敏感数据自动识别方法及系统将人工智能技术应用于敏感数据和关联关系识别阶段,有效解决传统正则方式性能与准确性不可兼得的痛点,并且也可省去专业人员对于正则识别规则的维护。该技术使用二叉树和支持向量机(SVM)训练分类模型,使用双向长短期记忆网络(Bi-LSTM)和条件随机场(CRF)结合训练识别模型。

[0004] 但是,现有结构化数据的敏感属性识别方法缺乏对非结构化数据的处理能力;基于人工智能的敏感数据自动分类识别方法及系统在自动分类部分一般采用决策树和SVM的传统方法,没有使用深度学习方法导致效果不理想;敏感数据识别采用Bi-LSTM+CRF的方式,一方面,在为长序列文本建模时会对模型的训练过程产生严重的负担,另一方面,由于长时间的步骤迭代,模型在处理比较靠后的信息时可能已经丢失前文序列中较为重要的局部关键信息;无法兼顾文本信息的空间分布以及时序特性,往往只能得到单一维度的信息,在为序列建模的迭代过程中容易忽略局部细节信息,简单的将各类神经网络进行层次叠加会导致模型过于复杂且泛化能力较差。同时,越来越多的图像等非文本数据出现导致现有技术不能很好地应对处理。

发明内容

[0005] 为解决现有技术的不足,本发明提出一种非结构化敏感数据识别方法及系统,用于处理非结构化图像数据和非结构化文本数据,对于非结构化敏感图像数据,能够准确对图像类的非结构化敏感数据进行识别;对于非结构化敏感文本数据,获取上下文的语义结构关系同时能够加大对关键信息的关注度,降低与主题内容关联性较低的冗余信息对分类结果的干扰;解决对特定图像这类非结构化非文本敏感数据的识别问题。

[0006] 为实现以上目的,本发明所采用的技术方案包括:

[0007] 一种非结构化敏感数据识别方法,其特征在于,包括:

[0008] S1、获取待识别的非结构化数据;

[0009] S2、判断非结构化数据的第一数据类型,所述第一数据类型包括非结构化文本数

据和非结构化图像数据；

[0010] S3、分别对非结构化文本数据和/或非结构化图像数据进行预处理，生成第一文本数据和/或第一图像数据；

[0011] S4、将第一文本数据输入文本分类模型，获得对应第一文本数据的第二数据类型标签；

[0012] 将第一图像数据输入图像分类模型，获得对应第一图像数据的第二数据类型标签；

[0013] S5、根据第二数据类型标签，将第一文本数据和/或第一图像数据输入对应的敏感数据识别模型，获得待识别的非结构化数据的敏感数据识别结果。

[0014] 进一步地，所述第二数据类型标签选自通讯录、好友列表、指纹、网络拓扑结构图和产品设计图纸中的任意一种。

[0015] 进一步地，所述步骤S2包括：

[0016] 根据待识别的非结构化数据的文件后缀名判断非结构化数据的第一数据类型。

[0017] 进一步地，所述预处理包括下列操作中的任意一种或多种组合：

[0018] 转换文件格式；

[0019] 更改文件编码；

[0020] 修改文件名；

[0021] 文件解密脱壳；

[0022] 清洗不可用文件；

[0023] 文件只读属性修改；

[0024] 色彩空间调整。

[0025] 进一步地，所述文本分类模型和图像分类模型的生成过程包括：

[0026] 获取历史数据，对历史数据添加对应的第二数据类型标签生成第一训练数据集；

[0027] 使用第一训练数据集训练获得文本分类模型和图像分类模型。

[0028] 进一步地，所述敏感数据识别模型的生成过程包括：

[0029] 获取历史数据，对历史数据添加对应的第二数据类型标签和敏感数据识别结果生成第二训练数据集；

[0030] 使用第二训练数据集训练获得针对各第二数据类型标签的敏感数据识别模型。

[0031] 本发明还涉及一种非结构化敏感数据识别系统，其特征在于，包括：

[0032] 数据交互模块，用于获取待识别的非结构化数据并反馈敏感数据识别结果；

[0033] 第一数据类型分类模块，用于判断非结构化数据的第一数据类型；

[0034] 数据预处理模块，用于对非结构化文本数据和/或非结构化图像数据进行预处理，生成第一文本数据和/或第一图像数据；

[0035] 第二数据类型分类模块，用于获得对应第一文本数据和/或第一图像数据的第二数据类型标签；

[0036] 敏感数据识别模块，用于获得待识别的非结构化数据的敏感数据识别结果。

[0037] 本发明还涉及一种计算机可读存储介质，其特征在于，所述存储介质上存储有计算机程序，所述计算机程序被处理器执行时实现上述的方法。

[0038] 本发明还涉及一种电子设备，其特征在于，包括处理器和存储器；

[0039] 所述存储器，用于存储第一文本数据和第一图像数据；

[0040] 所述处理器，用于通过调用第一文本数据和第一图像数据，执行上述的方法。

[0041] 本发明还涉及一种计算机程序产品，包括计算机程序和/或指令，其特征在于，该计算机程序和/或指令被处理器执行时实现上述方法的步骤。

[0042] 本发明的有益效果为：

[0043] 采用本发明所述非结构化敏感数据识别方法及系统，用于处理非结构化图像数据和非结构化文本数据，对于非结构化敏感图像数据；对于非结构化敏感文本数据，获取上下文的语义结构关系同时能够加大对关键信息的关注度，降低与主题内容关联性较低的冗余信息对分类结果的干扰；解决对特定图像这类非结构化非文本敏感数据的识别问题。

附图说明

[0044] 图1为本发明非结构化敏感数据识别方法流程示意图。

[0045] 图2为本发明非结构化敏感数据识别系统结构示意图。

具体实施方式

[0046] 为了更清楚的理解本发明的内容，将结合附图和实施例详细说明。

[0047] 本发明第一方面涉及一种步骤流程如图1所示的方法，包括：

[0048] S1、获取待识别的非结构化数据。

[0049] S2、判断非结构化数据的第一数据类型，所述第一数据类型包括非结构化文本数据和非结构化图像数据。特别是，可以根据待识别的非结构化数据的文件后缀名判断非结构化数据的第一数据类型。

[0050] S3、分别对非结构化文本数据和/或非结构化图像数据进行预处理，生成第一文本数据和/或第一图像数据。

[0051] 所述的预处理目的是为了将非结构化文本数据和/或非结构化图像数据转化为能够适用文本分类模型、图像分类模型的格式，常用的处理方式包括转换文件格式、更改文件编码、修改文件名、文件解密脱壳、清洗不可用文件、文件只读属性修改、色彩空间调整，当然其他的常规操作方式在需要的情况下也是适用的，经过预处理后的第一文本数据、第一图像数据应根据文本分类模型、图像分类模型的需要形成统一的格式，以确保分类的准确性。

[0052] S4、将第一文本数据输入文本分类模型，获得对应第一文本数据的第二数据类型标签；将第一图像数据输入图像分类模型，获得对应第一图像数据的第二数据类型标签。

[0053] 第二数据类型标签选自通讯录、好友列表、指纹、网络拓扑结构图和产品设计图纸中的任意一种，其中非结构化文本数据分为通讯录和好友列表两类，非结构化图像数据分为指纹、网络拓扑结构图、产品设计图纸。

[0054] 文本分类模型和图像分类模型的生成通过使用如历史数据等现有数据作为训练集获得，特别是可以对历史数据添加对应的第二数据类型标签生成第一训练数据集，并使用第一训练数据集训练获得文本分类模型和图像分类模型。针对性的，可以设置非结构化敏感文本数据分类模型结构为D-BGRU-SA(引入中断信息流和Attention机制的BGRU)，既能提取上下文远距离依赖关系又具有类似卷积核的位置不变性，从而兼顾到文本的时间特征

及空间特征。在此基础上融合了自注意力机制，进一步学习特征之间的依赖关系，为重要特征分配较大权值以降低噪声冗余，强化模型对关键信息的提取能力，实现文本特征的优化操作；设置非结构化敏感图像数据分类模型结构为Oriented R-CNN(引入RPN和Rols的R-CNN)，兼顾效率和准确性，且对旋转目标的检测也有很好的效果。

[0055] 当然，对应不同的敏感数据识别需求，也可以选择使用ALBERT模型和半监督学习的方法，引入有监督学习和无监督学习的一致性训练方法，在有标签数据较少的情况下实现文本识别。还可采用训练信号退火(training signal annealing，TSA)收敛策略，以降低模型过拟合的可能性。还可以针对敏感图像数据分类模型和敏感图像数据识别模型使用YOLO-level Feature方法和半监督学习的方法，引入有监督学习和无监督学习的一致性训练方法，在有标签数据较少的情况下实现图像分类识别。

[0056] S5、根据第二数据类型标签，将第一文本数据和/或第一图像数据输入对应的敏感数据识别模型，获得待识别的非结构化数据的敏感数据识别结果。

[0057] 类似的，敏感数据识别模型的生成也可以通过现有数据作为训练集获得，例如在第一训练数据集基础上对数据额外增加敏感数据识别结果生成第二训练数据集，并使用第二训练数据集训练获得针对各第二数据类型标签的敏感数据识别模型。

[0058] 由于针对不同分类的数据采用各自相对独立的敏感数据识别模型，因此能够获得更加具有针对性的敏感数据分类识别结果。例如，对于非结构化敏感文本数据，可以选择将中断信息流的思想引入双向门控循环单元(BGRU)中，既能提取上下文远距离依赖关系又具有类似卷积核的位置不变性，从而兼顾到文本的时间特征及空间特征，在此基础上融合了自注意力机制，进一步学习特征之间的依赖关系，为重要特征分配较大权值以降低噪声冗余，强化模型对关键信息的提取能力，实现文本特征的优化操作，最终使得敏感数据识别结果更准确。而对于非结构化敏感图像数据，将区域生成网络(RPN)和Rols(Regions of Interest)的思想引入基于R-CNN的图像识别单元中，能够准确对图像类的非结构化敏感数据进行识别。

[0059] 本发明另一方面还涉及一种系统，其结构如图2所示，包括：

[0060] 数据交互模块，用于获取待识别的非结构化数据并反馈敏感数据识别结果；

[0061] 第一数据类型分类模块，用于判断非结构化数据的第一数据类型；

[0062] 数据预处理模块，用于对非结构化文本数据和/或非结构化图像数据进行预处理，生成第一文本数据和/或第一图像数据；

[0063] 第二数据类型分类模块，用于获得对应第一文本数据和/或第一图像数据的第二数据类型标签；

[0064] 敏感数据识别模块，用于获得待识别的非结构化数据的敏感数据识别结果。

[0065] 通过使用该系统，能够执行上述的运算处理方法并实现对应的技术效果。

[0066] 优选的，还可以包括模型训练模块，用于生成和更新文本分类模型、图像分类模型和敏感数据识别模型

[0067] 本发明的实施例还提供能够实现上述实施例中的方法中全部步骤的一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，该计算机程序被处理器执行时实现上述实施例中的方法的全部步骤。

[0068] 本发明的实施例还提供一种用于执行上述方法的电子设备，作为该方法的实现装

置，所述电子设备至少具备有处理器和存储器，特别是该存储器上存储有执行方法所需的数据和相关的计算机程序，例如第一文本数据和第一图像数据等，并通过由处理器调用存储器中的数据、程序执行实现方法的全部步骤，并获得对应的技术效果。

[0069] 优选的，该电子设备可以包含有总线架构，总线可以包括任意数量的互联的总线和桥，总线将包括由一个或多个处理器和存储器的各种电路链接在一起。总线还可以将诸如外围设备、稳压器和功率管理电路等之类的各种其他电路链接在一起，这些都是本领域所公知的，因此，本文不再对其进行进一步描述。总线接口在总线和接收器和发送器之间提供接口。接收器和发送器可以是同一个元件，即收发机，提供用于在传输介质上与各种其他系统通信的单元。处理器负责管理总线和通常的处理，而存储器可以被用于存储处理器在执行操作时所使用的数据。

[0070] 额外的，所述电子设备还可以进一步包括通信模块、输入单元、音频处理器、显示器、电源等部件。其所采用的处理器(或称为控制器、操作控件)可以包括微处理器或其他处理器装置和/或逻辑装置，该处理器接收输入并控制电子设备的各个部件的操作；存储器可以是缓存器、闪存、硬驱、可移动介质、易失性存储器、非易失性存储器或其它合适装置中的一种或更多种，可储存上述有关的数据信息，此外还可存储执行有关信息的程序，并且处理器可执行该存储器存储的该程序，以实现信息存储或处理等；输入单元用于向处理器提供输入，例如可以为按键或触摸输入装置；电源用于向电子设备提供电力；显示器用于进行图像和文字等显示对象的显示，例如可为LCD显示器。通信模块即为经由天线发送和接收信号的发送机/接收机。通信模块(发送机/接收机)耦合到处理器，以提供输入信号和接收输出信号，这可以和常规移动通信终端的情况相同。基于不同的通信技术，在同一电子设备中，可以设置有多个通信模块，如蜂窝网络模块、蓝牙模块和/或无线局域网模块等。通信模块(发送机/接收机)还经由音频处理器耦合到扬声器和麦克风，以经由扬声器提供音频输出，并接收来自麦克风的音频输入，从而实现通常的电信功能。音频处理器可以包括任何合适的缓冲器、解码器、放大器等。另外，音频处理器还耦合到中央处理器，从而使得可以通过麦克风能够在本机上录音，且使得可以通过扬声器来播放本机上存储的声音。

[0071] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0072] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的系统。

[0073] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令系统的制造品，该指令系统实现在流程图一个流程或多个流程和/或方框图一个方框或

多个方框中指定的功能。

[0074] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0075] 以上所述仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换等都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求书的保护范围为准。

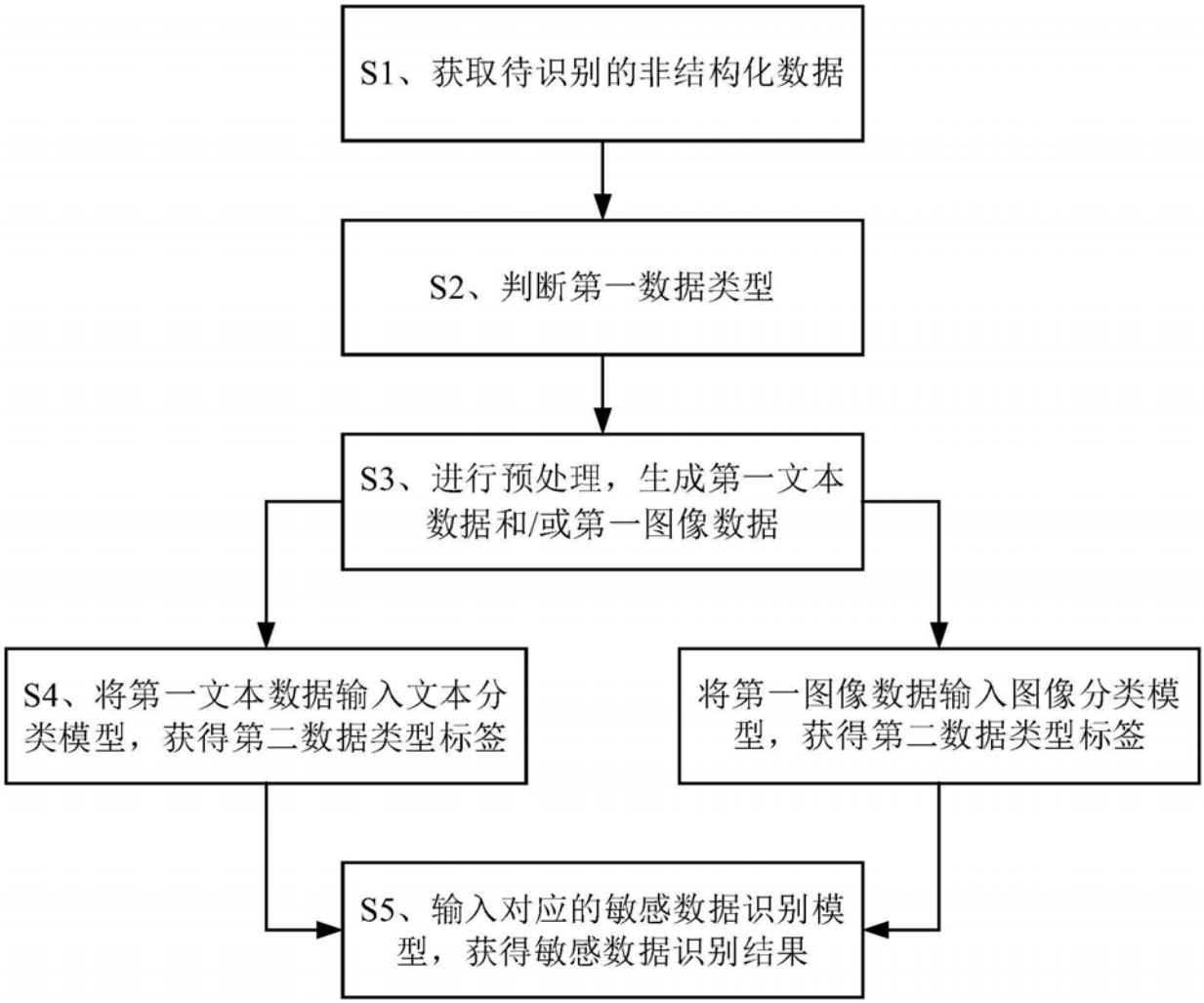

S1、获取待识别的非结构化数据
S2、判断第一数据类型
S3、进行预处理，生成第一文本数据和/或第一图像数据
S4、将第一文本数据输入文本分类模型，获得第二数据类型标签
将第一图像数据输入图像分类模型，获得第二数据类型标签
S5、输入对应的敏感数据识别模型，获得敏感数据识别结果

图1

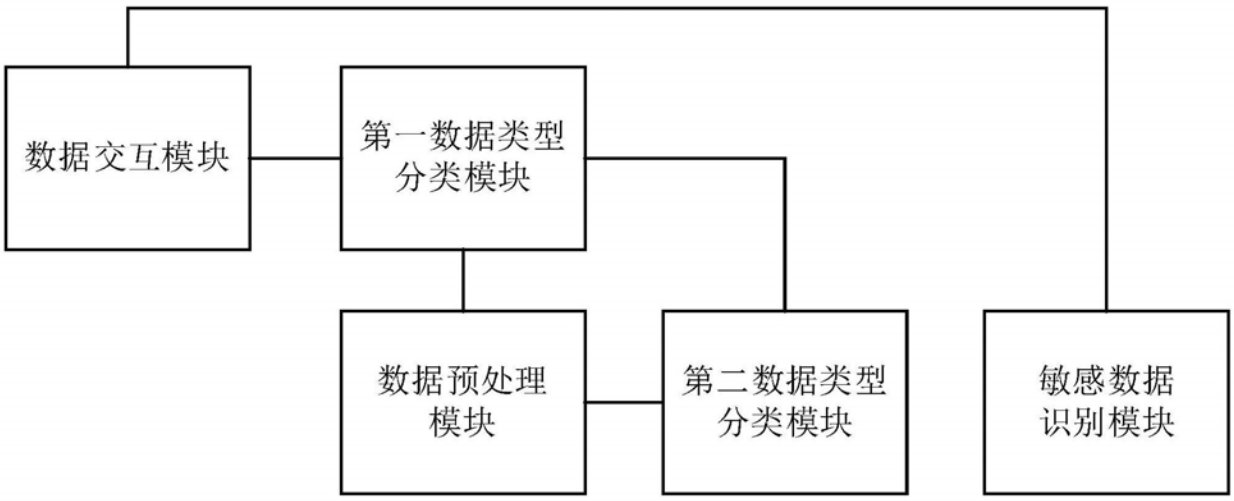

数据交互模块
第一数据类型分类模块
数据预处理模块
第二数据类型分类模块
敏感数据识别模块

图2