(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 113673889 A
(43)申请公布日 2021.11.19

(21)申请号 202110985791.8

(22)申请日 2021.08.26

(71)申请人 上海罗盘信息科技有限公司
地址 200030 上海市徐汇区零陵路583号9楼503室

(72)发明人 林松 郝艳丰 陆鸿强 马力 徐渊博 李刚华 姚东鸿 林永东

(74)专利代理机构 深圳市创富知识产权代理有限公司 44367
代理人 朱易顺

(51)Int.Cl.
G06Q 10/06(2012.01)
G06F 16/35(2019.01)
G06F 40/279(2020.01)

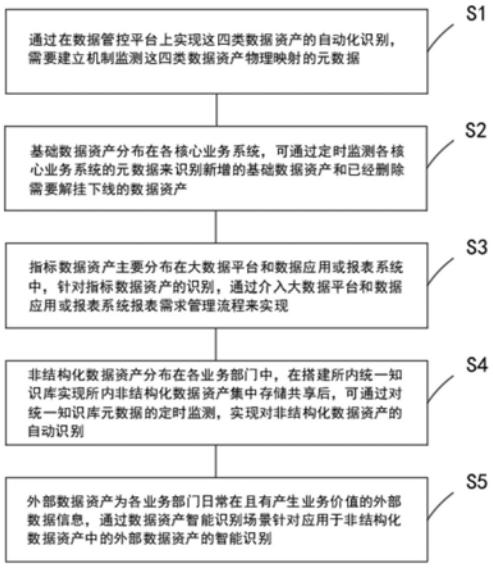
权利要求书3页 说明书6页 附图1页

(54)发明名称

一种智能化数据资产识别的方法

(57)摘要

本发明属于信息技术领域，尤其是涉及一种智能化数据资产识别的方法，包括以下步骤：通过在数据管控平台上实现这四类数据资产的自动化识别，需要建立机制监测这四类数据资产物理映射的元数据；基础数据资产分布在各核心业务系统，可通过定时监测各核心业务系统的元数据来识别新增的基础数据资产和已经删除需要解挂下线的数据资产；指标数据资产主要分布在大数据平台和数据应用或报表系统中，针对指标数据资产的识别，通过介入大数据平台和数据应用或报表系统报表需求管理流程来实现。本发明根据现有元数据设计出元模型，然后将大数据平台中的元数据按元模型集中汇总并关联到一起，达到企业对大数据平台的数据统一管理与应用的目的。



CN 113673889 A

1.一种智能化数据资产识别的方法,其特征在于,所述数据资产智能识别包括基础数据资产识别、指标数据资产识别、非结构化数据资产和外部数据资产识别,所述数据资产智能识别的方法包括以下步骤:

S1、通过在数据管控平台上实现这四类数据资产的自动化识别,需要建立机制监测这四类数据资产物理映射的元数据;

S2、基础数据资产分布在各核心业务系统,可通过定时监测各核心业务系统的元数据来识别新增的基础数据资产和已经删除需要解挂下线的数据资产;

S3、指标数据资产主要分布在大数据平台和数据应用或报表系统中,针对指标数据资产的识别,通过介入大数据平台和数据应用或报表系统报表需求管理流程来实现;

S4、非结构化数据资产分布在各业务部门中,在搭建所内统一知识库实现所内非结构化数据资产集中存储共享后,可通过对统一知识库元数据的定时监测,实现对非结构化数据资产的自动识别;

S5、外部数据资产为各业务部门日常在且有产生业务价值的外部数据信息,通过数据资产智能识别场景针对应用于非结构化数据资产中的外部数据资产的智能识别。

2.根据权利要求1所述的一种智能化数据资产识别的方法,其特征在于,所述步骤S1中元数据包括技术元数据和业务元数据,技术元数据的采集,根据现有元数据设计出元模型,然后将大数据平台中的元数据按元模型集中汇总并关联到一起,达到企业对大数据平台的数据统一管理与应用的目的,并且对于元数据管理工具支持的格式可直接进行导入,而对于一些自定义的规则,需要进行格式转换并导入。

3.根据权利要求2所述的一种智能化数据资产识别的方法,其特征在于,所述数据管控平台应具备CSV适配器、XML适配器、DB适配器和API接入适配器,以支持大数据平台、统一知识库元数据的顺利接入,且数据管控平台应具有数据资产识别引擎,该引擎可根据基础数据资产、指标数据资产、非结构化数据资产准入规则,识别新增的各类数据资产。

4.根据权利要求1所述的一种智能化数据资产识别的方法,其特征在于,所述步骤S2中基础数据资产通过大数据平台ODS层和DW层的元数据比对,识别出可能新增和变更的基础数据资产信息,对基础数据资产目录和资产项进行补充和完善;建立大数据平台常用系统表名关键字和系统控制字段名关键字的“过滤库”,用于对新增和变更的元数据进行识别筛选。

5.根据权利要求1所述的一种智能化数据资产识别的方法,其特征在于,所述步骤S3中指标数据资产来自于大数据平台报表元数据,根据比对可以识别新增和变更的元数据信息,进而根据指标数据资产准入规则对新增和变更的元数据信息进行判别,识别新的指标数据资产。

6.根据权利要求1所述的一种智能化数据资产识别的方法,其特征在于,所述步骤S5中外部数据资产分为两类包括:被指标数据资产所引用的外部数据信息项和被内部非结构化数据资产所引用的外部数据资产项目,针对这两类外部数据资产,其主要满足的规则如下:对于指标因子,判断该外部数据信息项有没有被内部指标所引用;对于重要标签,判断该外部数据信息项有没有被内部非结构化数据资产所引用。

7.根据权利要求1-6任一所述的一种智能化数据资产识别的方法,其特征在于,针对各类数据资产,通过设计适配于各类数据资产智能管理模型,并且该模型应可根据数据资产

名称、定义、来源等属性，与资产分类树进行智能匹配，推荐合适的数据资产挂载点，以提高工作效率，降低人工出错几率。

8.根据权利要求7所述的一种智能化数据资产识别的方法，其特征在于，所述数据资产智能挂载的核心是对文本的自动化分类，建立、选择适当的分类规则从而进行正确分类的这一过程，其建立分类规则的基本过程是：先从已分类结果中倒推寻找分类规则，即先从已分类的训练文本中根据不同类别的文本所具有的不同特征；进而搜寻提取到一定准确、适当的分类规则；再将待分类文本按照以上规则进行归类；最终使得分类结果与目标结果相一致；

所述文本分类用计算公式如式(1-1)所示可定义为如下：

F(D,C)＝{True,False}　　(1-1)

上述公式(4-1)中，集合D＝{d1,d2,…,di,…dn}是指待分类的文本集合，其中，di表示其中的第i个待分类文本，而n是指待分类文本集合D中包含待分类文本的数量大小，集合C＝{c1,c2,…,cj,…cm}则是指我们预先定义的类别集合，其中cj表示其中的第j个类别，而m是指类别集合C中所预先定义的类别数量；而F函数，在这代表的是一种映射关系，若F(di,cj)＝True，则代表数据集合中第i个待分类文本di它的分类结果是第j个类别cj；反之，若F(di,cj)＝False，则是指数据集合中第i个待分类文本di的分类结果并不是第j个类别cj，数学集合中的映射概念存在有一对一、一对多这两种映射关系，同理在文本分类中也可以分为单标签分类和多标签分类。

9.根据权利要求8所述的一种智能化数据资产识别的方法，其特征在于，在拿到数据集后的第一个处理步骤是对文本数据进行数据预处理操作，这一过程中，按照处理顺序我们需要对文本数据进行如下操作：文本标记、分词以及去除停用词处理，且经过文本预处理环节之后再对文本数据进行文本表示环节，具体的利用VSM模型对文本D进行文本表示，词项以及词项的权重值将成为文本表示这个模型的组成部分，文本D就能被n个词项以及他们的权重值所组成的特征向量代表，表示形式如下：D＝{(t1,w1),(t2,w2),…,(ti,wj),…(tn,wn)}，其中ti,wj分别是指对应的第i个特征词以及第i个特征词的权重值，wj∈(0-1)。

10.根据权利要求9所述的一种智能化数据资产识别的方法，其特征在于，在对数据进行预处理操作以及文本表示后需对文本数据进行特征选取，具体的步骤：根据文本数据集特点，通过选定流程选取适合的特征计算函数，对数据集中每一条文本中的每个词项分别进行特征计算得到量化结果，将结果按照由大到小进行顺序排列，根据提前设定的阈值情况，从中选出一定数量的特征项作为原始文本数据的代表；具体的算法采用卡方统计算法即CHI算法，需要设定最小阈值和最大阈值，假设词项t_i与类别c_j满足一阶自由度的卡方(x2)分布，通过函数计算出词项与类别之间的相关度，来提供选取标准，利用卡方统计算法可计算得到每条文本的所有词项的相关度，再根据相关度大小，对词项进行选取，其计算公式1-2所示：

$$CHI(t_i,c_j)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1-2)$$

式(1-2)中，A是指类别结果为c_j的文档中存在词项t_i的文本数量，C是指类别结果为c_j的文档中不存在词项t_i的文本数量，B是指训练文本数据集中类别结果为非c_j的文档中存在词项t_i的文本数量，N是指整个训练文本数据集中所包含的文本数量，D是指训练文本数据

集中类别结果为非c_j的文档中不存在词项t_i的文本数量，且N＝A+B+C+D；

在公式(1-2)中，CHI(t_i,c_j)表示的是词项t_i与类别c_j的卡方统计值，这是处理单个词项的单分类问题时的计算方法，但在处理多类分类问题时，卡方统计结果需要再进行一步处理，可以使用加权平均或者求和两种计算法，两种计算公式分别如式(1-3)、式(1-4)所示：

$$CHI_{avg}(t_i)=\sum_{j=1}^{m}p(c_j)CHI(t_i,c_j) \qquad (1-3)$$

$$CHI_{\max(t_i)}=\max_{j=1,\ldots,m}\left(CHI(t_i,c_j)\right) \qquad (1-4)$$

采用上述算法以完成对数据进行预处理操作以及文本表示后需对文本数据进行特征选取。

一种智能化数据资产识别的方法

技术领域

[0001] 本发明涉及信息技术领域，尤其涉及一种智能化数据资产识别的方法。

背景技术

[0002] “数据资产运营”作为重要的建设内容之一，以实现业务价值为导向，以数据资产使用部门为中心，为企业不同层面数据资产使用部门提供数据便利，设计数据资产评价体系，建立数据资产内部共享和运营流通等机制，从而进一步推动某某企业数据使用、数据共享，降低数据资产成本，促进数据价值发挥，目前市面上对于资产的管理方式主要还是以手工记账的管理方式为主，由于管理资产众多、盘点工作繁重、物品属性复杂，需占用大量的人力物力，而且管理者对固定资产的历史操作和资产统计工作异常困难，此外资产随着使用年限的增加，残存值也在不断下降，这就很可能导致资产统计不准确、资产流失和资产重复购买等多种问题。

[0003] 为更好的提升企业数据资产运营效率，计划应用AI技术进行数据资产运营自动化领域进行探索，确定相关智能化场景和落地方式，以便集成至相关系统平台，以提高数据资产运营的相关工作效率，降低人工出错几率，我们提出一种智能化数据资产识别的方法来改善上述问题。

发明内容

[0004] 本发明的目的是为了解决现有技术中存在的缺点，而提出的一种智能化数据资产识别的方法。

[0005] 为了实现上述目的，本发明采用了如下技术方案：

[0006] 一种智能化数据资产识别的方法，所述数据资产智能识别包括基础数据资产识别、指标数据资产识别、非结构化数据资产和外部数据资产识别，所述数据资产智能识别的方法包括以下步骤：

[0007] S1、通过在数据管控平台上实现这四类数据资产的自动化识别，需要建立机制监测这四类数据资产物理映射的元数据；

[0008] S2、基础数据资产分布在各核心业务系统，可通过定时监测各核心业务系统的元数据来识别新增的基础数据资产和已经删除需要解挂下线的数据资产；

[0009] S3、指标数据资产主要分布在大数据平台和数据应用或报表系统中，针对指标数据资产的识别，通过介入大数据平台和数据应用或报表系统报表需求管理流程来实现；

[0010] S4、非结构化数据资产分布在各业务部门中，在搭建所内统一知识库实现所内非结构化数据资产集中存储共享后，可通过对统一知识库元数据的定时监测，实现对非结构化数据资产的自动识别；

[0011] S5、外部数据资产为各业务部门日常在且有产生业务价值的外部数据信息，通过数据资产智能识别场景针对应用于非结构化数据资产中的外部数据资产的智能识别。

[0012] 在上述的智能化数据资产识别的方法中，所述步骤S1中元数据包括技术元数据和

业务元数据,技术元数据的采集,根据现有元数据设计出元模型,然后将大数据平台中的元数据按元模型集中汇总并关联到一起,达到企业对大数据平台的数据统一管理与应用的目的,并且对于元数据管理工具支持的格式可直接进行导入,而对于一些自定义的规则,需要进行格式转换并导入。

[0013] 在上述的智能化数据资产识别的方法中,所述数据管控平台应具备CSV适配器、XML适配器、DB适配器和API接入适配器,以支持大数据平台、统一知识库元数据的顺利接入,且数据管控平台应具有数据资产识别引擎,该引擎可根据基础数据资产、指标数据资产、非结构化数据资产准入规则,识别新增的各类数据资产。

[0014] 在上述的智能化数据资产识别的方法中,所述步骤S2中基础数据资产通过大数据平台ODS层和DW层的元数据比对,识别出可能新增和变更的基础数据资产信息,对基础数据资产目录和资产项进行补充和完善;建立大数据平台常用系统表名关键字和系统控制字段名关键字的“过滤库”,用于对新增和变更的元数据进行识别筛选。

[0015] 在上述的智能化数据资产识别的方法中,所述步骤S3中指标数据资产来自于大数据平台报表元数据,根据比对可以识别新增和变更的元数据信息,进而根据指标数据资产准入规则对新增和变更的元数据信息进行判别,识别新的指标数据资产。

[0016] 在上述的智能化数据资产识别的方法中,所述步骤S5中外部数据资产分为两类包括:被指标数据资产所引用的外部数据信息项和被内部非结构化数据资产所引用的外部数据资产项目,针对这两类外部数据资产,其主要满足的规则如下:对于指标因子,判断该外部数据信息项有没有被内部指标所引用;对于重要标签,判断该外部数据信息项有没有被内部非结构化数据资产所引用。

[0017] 在上述的智能化数据资产识别的方法中,针对各类数据资产,通过设计适配于各类数据资产智能管理模型,并且该模型应可根据数据资产名称、定义、来源等属性,与资产分类树进行智能匹配,推荐合适的数据资产挂载点,以提高工作效率,降低人工出错几率。

[0018] 在上述的智能化数据资产识别的方法中,所述数据资产智能挂载的核心是对文本的自动化分类,建立、选择适当的分类规则从而进行正确分类的这一过程,其建立分类规则的基本过程是:先从已分类结果中倒推寻找分类规则,即先从已分类的训练文本中根据不同类别的文本所具有的不同特征;进而搜寻提取到一定准确、适当的分类规则;再将待分类文本按照以上规则进行归类;最终使得分类结果与目标结果相一致;

[0019] 所述文本分类用计算公式如式(1-1)所示可定义为如下:

[0020] $F(D,C)=\{True,False\}$ (1-1)

[0021] 上述公式(4-1)中,集合$D=\{d1,d2,\cdots,di,\cdots dn\}$是指待分类的文本集合,其中,di表示其中的第i个待分类文本,而n是指待分类文本集合D中包含待分类文本的数量大小,集合$C=\{c1,c2,\cdots,cj,\cdots cm\}$ 则是指我们预先定义的类别集合,其中cj表示其中的第j个类别,而 m是指类别集合C中所预先定义的类别数量;而F函数,在这代表的是一种映射关系,若$F(di,cj)=True$,则代表数据集合中第i个待分类文本di它的分类结果是第j个类别cj;反之,若$F(di,cj)=False$,则是指数据集合中第i个待分类文本di的分类结果并不是第j个类别cj,数学集合中的映射概念存在有一对一、一对多这两种映射关系,同理在文本分类中也可以分为单标签分类和多标签分类。

[0022] 在上述的智能化数据资产识别的方法中,在拿到数据集后的第一个处理步骤是对

文本数据进行数据预处理操作，这一过程中，按照处理顺序我们需要对文本数据进行如下操作：文本标记、分词以及去除停用词处理，且经过文本预处理环节之后再对文本数据进行文本表示环节，具体的利用VSM模型对文本D进行文本表示，词项以及词项的权重值将成为文本表示这个模型的组成部分，文本D就能被n个词项以及他们的权重值所组成的特征向量代表，表示形式如下：D＝{(t1,w1),(t2,w2),…,(ti,wj),…(tn,wn)}，其中ti,wj分别是指对应的第i个特征词以及第i个特征词的权重值，wj∈(0-1)。

[0023] 在上述的智能化数据资产识别的方法中，在对数据进行预处理操作以及文本表示后需对文本数据进行特征选取，具体的步骤：根据文本数据集特点，通过选定流程选取适合的特征计算函数，对数据集中每一条文本中的每个词项分别进行特征计算得到量化结果，将结果按照由大到小进行顺序排列，根据提前设定的阈值情况，从中选出一定数量的特征项作为原始文本数据的代表；具体的算法采用卡方统计算法即CHI算法，需要设定最小阈值和最大阈值，假设词项t_i与类别c_j满足一阶自由度的卡方(x2)分布，通过函数计算出词项与类别之间的相关度，来提供选取标准，利用卡方统计算法可计算得到每条文本的所有词项的相关度，再根据相关度大小，对词项进行选取，其计算公式1-2所示：

[0024] $$CHI(t_i,c_j)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1-2)$$

[0025] 式(1-2)中，A是指类别结果为c_j的文档中存在词项t_i的文本数量，C是指类别结果为c_j的文档中不存在词项t_i的文本数量，B 是指训练文本数据集中类别结果为非c_j的文档中存在词项t_i的文本数量，N是指整个训练文本数据集中所包含的文本数量，D是指训练文本数据集中类别结果为非c_j的文档中不存在词项t_i的文本数量，且N＝A+B+C+D；

[0026] 在公式(1-2)中，CHI(t_i,c_j)表示的是词项t_i与类别c_j的卡方统计值，这是处理单个词项的单分类问题时的计算方法，但在处理多类分类问题时，卡方统计结果需要再进行一步处理，可以使用加权平均或者求和两种计算法，两种计算公式分别如式(1-3)、式(1-4)所示：

[0027] $$CHI_{avg}(t_i)=\sum_{j=1}^{m}p(c_j)CHI(t_i,c_j) \qquad (1-3)$$

[0028] $$CHI_{\max(t_i)}=\max_{j=1,\ldots,m}\left(CHI(t_i,c_j)\right) \qquad (1-4)$$

[0029] 采用上述算法以完成对数据进行预处理操作以及文本表示后需对文本数据进行特征选取。

[0030] 与现有技术相比，本一种智能化数据资产识别的方法的优点在于：

[0031] 1、本发明根据现有元数据设计出元模型，然后将大数据平台中的元数据按元模型集中汇总并关联到一起，达到企业对大数据平台的数据统一管理与应用的目的；

[0032] 2、本发明通过VSM模型表示方法，可以将文本的相似度计算问题转化为对文本对应的特征向量进行夹角余弦的计算问题，如此就使得文本计算的复杂度得到了明显的简化。

附图说明

[0033] 图1为本发明提出的一种智能化数据资产识别的方法的方法步骤图；

[0034] 图2为本发明提出的一种智能化数据资产识别的方法的逻辑架构图。

具体实施方式

[0035] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。

[0036] 实施例

[0037] 参照图1-2，一种智能化数据资产识别的方法，数据资产智能识别包括基础数据资产识别、指标数据资产识别、非结构化数据资产和外部数据资产识别，数据资产智能识别的方法包括以下步骤：

[0038] S1、通过在数据管控平台上实现这四类数据资产的自动化识别，需要建立机制监测这四类数据资产物理映射的元数据；

[0039] S2、基础数据资产分布在各核心业务系统，可通过定时监测各核心业务系统的元数据来识别新增的基础数据资产和已经删除需要解挂下线的数据资产；

[0040] S3、指标数据资产主要分布在大数据平台和数据应用或报表系统中，针对指标数据资产的识别，通过介入大数据平台和数据应用或报表系统报表需求管理流程来实现；

[0041] S4、非结构化数据资产分布在各业务部门中，在搭建所内统一知识库实现所内非结构化数据资产集中存储共享后，可通过对统一知识库元数据的定时监测，实现对非结构化数据资产的自动识别；

[0042] S5、外部数据资产为各业务部门日常在且有产生业务价值的外部数据信息，通过数据资产智能识别场景针对应用于非结构化数据资产中的外部数据资产的智能识别。

[0043] 指标数据资产主要分布在大数据平台和数据应用或报表系统中，针对指标数据资产的识别，可通过介入大数据平台和数据应用或报表系统报表需求管理流程来实现，非结构化数据资产当前主要分布在各业务部门中，并未集中存储，后期搭建所内统一知识库实现所内非结构化数据资产集中存储共享后，可通过对统一知识库元数据的定时监测，实现对非结构化数据资产的自动识别；外部数据资产主要为各业务部门日常在且有产生业务价值的外部数据信息，现阶段由于外部数据资产涉及的范围广，并无统一物理落地之处，较难通过技术手段对外部数据资产进行自动化识别，故数据资产智能识别场景只针对应用于非结构化数据资产中的外部数据资产的智能识别。

[0044] 其中，步骤S1中元数据包括技术元数据和业务元数据，技术元数据的采集，根据现有元数据设计出元模型，然后将大数据平台中的元数据按元模型集中汇总并关联到一起，达到企业对大数据平台的数据统一管理与应用的目的，并且对于元数据管理工具支持的格式可直接进行导入，而对于一些自定义的规则，需要进行格式转换并导入，进一步的，数据管控平台应具备CSV适配器、XML适配器、DB适配器和API接入适配器，以支持大数据平台、统一知识库元数据的顺利接入，且数据管控平台应具有数据资产识别引擎，该引擎可根据基础数据资产、指标数据资产、非结构化数据资产准入规则，识别新增的各类数据资产。

[0045] 其中，步骤S2中基础数据资产通过大数据平台ODS层和DW层的元数据比对，识别出可能新增和变更的基础数据资产信息，对基础数据资产目录和资产项进行补充和完善；建立大数据平台常用系统表名关键字和系统控制字段名关键字的“过滤库”，用于对新增和变更的元数据进行识别筛选，基础数据资产应满足下面三条规则：1、因新业务、新功能模块产

生的元数据信息;2、因业务调整而发生改变的元数据信息;3、不在“过滤库”里的元数据信息。

[0046] 其中,步骤S3中指标数据资产来自于大数据平台报表元数据,根据比对可以识别新增和变更的元数据信息,进而根据指标数据资产准入规则对新增和变更的元数据信息进行判别,识别新的指标数据资产,指标数据资产准入规则如下:1、元数据信息是否是度量值;2、和已有指标名称比对,判别是否是新的指标;3、具有重要业务价值。

[0047] 步骤S4中的内部非结构数据资产主要为当前各业务部门的手工制作的统计报告和研究报告,各部门的非结构化信息,若要纳入到非结构化数据资产目录,应满足如下规则:由某某企业内部产生,拥有独立的知识产权;有明确的归属部门和作者;材料描述信息可对所内公开。

[0048] 其中,步骤S5中外部数据资产分为两类包括:被指标数据资产所引用的外部数据信息项和被内部非结构化数据资产所引用的外部数据资产项目,针对这两类外部数据资产,其主要满足的规则如下:对于指标因子,判断该外部数据信息项有没有被内部指标所引用;对于重要标签,判断该外部数据信息项有没有被内部非结构化数据资产所引用。

[0049] 进一步的,针对各类数据资产,通过设计适配于各类数据资产智能管理模型,并且该模型应可根据数据资产名称、定义、来源等属性,与资产分类树进行智能匹配,推荐合适的数据资产挂载点,以提高工作效率,降低人工出错几率,具体的,数据资产智能挂载的核心是对文本的自动化分类,建立、选择适当的分类规则从而进行正确分类的这一过程,其建立分类规则的基本过程是:先从已分类结果中倒推寻找分类规则,即先从已分类的训练文本中根据不同类别的文本所具有的不同特征;进而搜寻提取到一定准确、适当的分类规则;再将待分类文本按照以上规则进行归类;最终使得分类结果与目标结果相一致;

[0050] 文本分类用计算公式如式(1-1)所示可定义为如下:

[0051] $F(D,C)=\{True,False\}$ (1-1)

[0052] 上述公式(4-1)中,集合$D=\{d1,d2,\cdots,di,\cdots dn\}$是指待分类的文本集合,其中,di表示其中的第i个待分类文本,而n是指待分类文本集合D中包含待分类文本的数量大小,集合$C=\{c1,c2,\cdots,cj,\cdots cm\}$ 则是指我们预先定义的类别集合,其中cj表示其中的第j个类别,而 m是指类别集合C中所预先定义的类别数量;而F函数,在这代表的是一种映射关系,若$F(di,cj)=True$,则代表数据集合中第i个待分类文本di它的分类结果是第j个类别cj;反之,若$F(di,cj)=False$,则是指数据集合中第i个待分类文本di的分类结果并不是第j个类别 cj,数学集合中的映射概念存在有一对一、一对多这两种映射关系,同理在文本分类中也可以分为单标签分类和多标签分类,其中,单标签分类是指待分类的文本只能被划分到一个类别中,数据资产的分类属于单标签分类,本方案对多标签分类不做相关解释。

[0053] 更进一步的,在拿到数据集后的第一个处理步骤是对文本数据进行数据预处理操作,这一过程中,按照处理顺序我们需要对文本数据进行如下操作:文本标记、分词以及去除停用词处理,且经过文本预处理环节之后再对文本数据进行文本表示环节,具体的利用VSM模型对文本D进行文本表示,词项以及词项的权重值将成为文本表示这个模型的组成部分,文本D就能被n个词项以及他们的权重值所组成的特征向量代表,表示形式如下:$D=\{(t1,w1),(t2,w2),\cdots,(ti,wj),\cdots (tn,wn)\}$,其中ti,wj分别是指对应的第i个特征词以及第i个特征词的权重值,$wj\in(0-1)$,通过VSM模型表示方法,可以将文本的相似度计算问

题转化为对文本对应的特征向量进行夹角余弦的计算问题，如此就使得文本计算的复杂度得到了明显的简化。

[0054] 其中，在对数据进行预处理操作以及文本表示后需对文本数据进行特征选取，具体的步骤：根据文本数据集特点，通过选定流程选取适合的特征计算函数，对数据集中每一条文本中的每个词项分别进行特征计算得到量化结果，将结果按照由大到小进行顺序排列，根据提前设定的阈值情况，从中选出一定数量的特征项作为原始文本数据的代表；具体的算法采用卡方统计算法即CHI算法，需要设定最小阈值和最大阈值，假设词项t_i与类别c_j满足一阶自由度的卡方(x2)分布，通过函数计算出词项与类别之间的相关度，来提供选取标准，利用卡方统计算法可计算得到每条文本的所有词项的相关度，再根据相关度大小，对词项进行选取，其计算公式1-2所示：

[0055] $$CHI(t_i,c_j)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1-2)$$

[0056] 式(1-2)中，A是指类别结果为c_j的文档中存在词项t_i的文本数量，C是指类别结果为c_j的文档中不存在词项t_i的文本数量，B 是指训练文本数据集中类别结果为非c_j的文档中存在词项t_i的文本数量，N是指整个训练文本数据集中所包含的文本数量，D是指训练文本数据集中类别结果为非c_j的文档中不存在词项t_i的文本数量，且N＝A+B+C+D；

[0057] 在公式(1-2)中，CHI(t_i,c_j)表示的是词项t_i与类别c_j的卡方统计值，这是处理单个词项的单分类问题时的计算方法，但在处理多类分类问题时，卡方统计结果需要再进行一步处理，可以使用加权平均或者求和两种计算法，两种计算公式分别如式(1-3)、式(1-4)所示：

[0058] $$CHI_{avg}(t_i)=\sum_{j=1}^{m}p(c_j)CHI(t_i,c_j) \qquad (1-3)$$

[0059] $$CHI_{\max(t_i)}=\max_{j=1,\ldots,m}(CHI(t_i,c_j)) \qquad (1-4)，$$

[0060] 采用上述算法以完成对数据进行预处理操作以及文本表示后需对文本数据进行特征选取，卡方统计算法是在假设词项与类别间存在卡方分布的前提下展开运算的，它考虑了词项在不同类别之间的分布情况。

[0061] 需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。

[0062] 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，根据本发明的技术方案及其发明构思加以等同替换或改变，都应涵盖在本发明的保护范围之内。

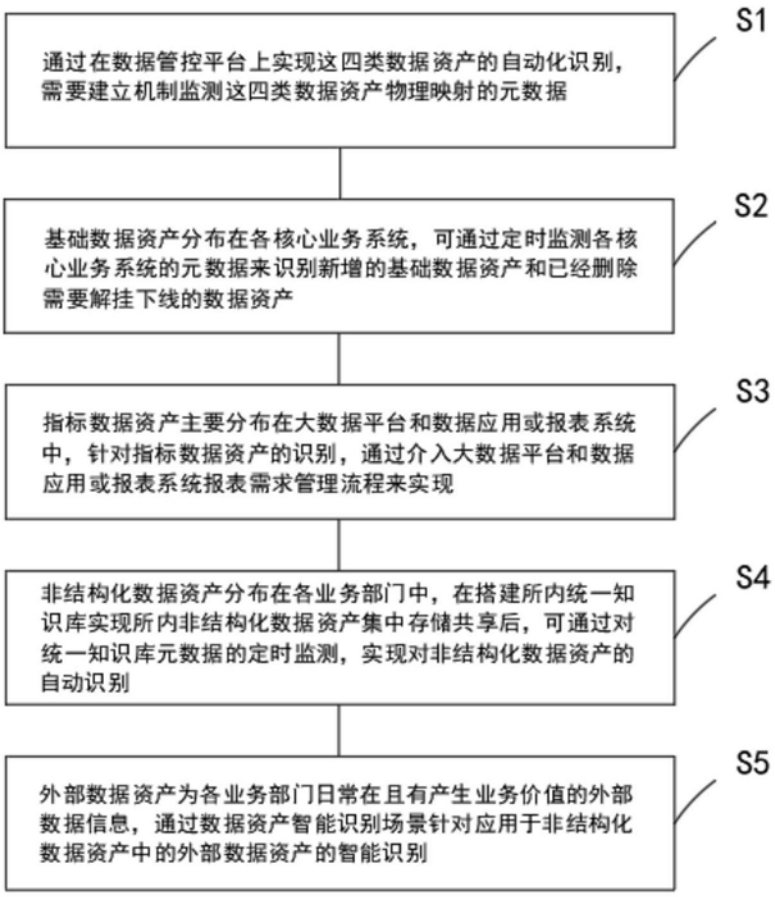

S1
通过在数据管控平台上实现这四类数据资产的自动化识别，需要建立机制监测这四类数据资产物理映射的元数据
S2
基础数据资产分布在各核心业务系统，可通过定时监测各核心业务系统的元数据来识别新增的基础数据资产和已经删除需要解挂下线的数据资产
S3
指标数据资产主要分布在大数据平台和数据应用或报表系统中，针对指标数据资产的识别，通过介入大数据平台和数据应用或报表系统报表需求管理流程来实现
S4
非结构化数据资产分布在各业务部门中，在搭建所内统一知识库实现所内非结构化数据资产集中存储共享后，可通过对统一知识库元数据的定时监测，实现对非结构化数据资产的自动识别
S5
外部数据资产为各业务部门日常在且有产生业务价值的外部数据信息，通过数据资产智能识别场景针对应用于非结构化数据资产中的外部数据资产的智能识别

图1

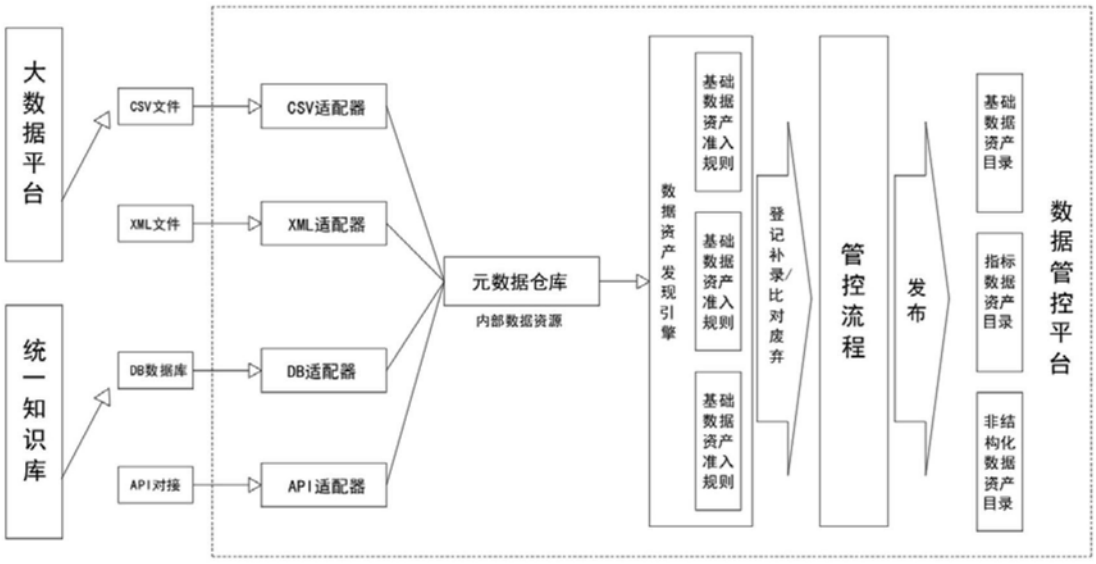

大数据平台
统一知识库
CSV文件
XML文件
DB数据库
API对接
CSV适配器
XML适配器
DB适配器
API适配器
元数据仓库
内部数据资源
数据资产发现引擎
基础数据资产准入规则
基础数据资产准入规则
基础数据资产准入规则
登记补录/比对废弃
管控流程
发布
基础数据资产目录
指标数据资产目录
非结构化数据资产目录
数据管控平台

图2