(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114443783 B
(45)授权公告日 2022.06.24

(21)申请号 202210374815.0
(22)申请日 2022.04.11
(65)同一申请的已公布的文献号
申请公布号 CN 114443783 A
(43)申请公布日 2022.05.06
(73)专利权人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号
专利权人 物产中大数字科技有限公司
(72)发明人 朱海洋 陈为 季永炜 周俊 金慧颖 应石磊 孙元园 朱建龙
(74)专利代理机构 北京亿腾知识产权代理事务所(普通合伙) 11309
专利代理师 张明
(51)Int.Cl.
G06F 16/28(2019.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
(56)对比文件
CN 102521386 A,2012.06.27
CN 111723292 A,2020.09.29
CN 112434188 A,2021.03.02
CN 113536155 A,2021.10.22
CN 109597855 A,2019.04.09
US 2016328406 A1,2016.11.10
蔡威林等.基于影响度的标签传播算法.《佳木斯大学学报》.2022,第40卷(第1期),
Maria D.Chikina等.An effective statistical evaluation of ChIPseq dataset similarity.《Bioinformatics》.2012,
审查员 单娟

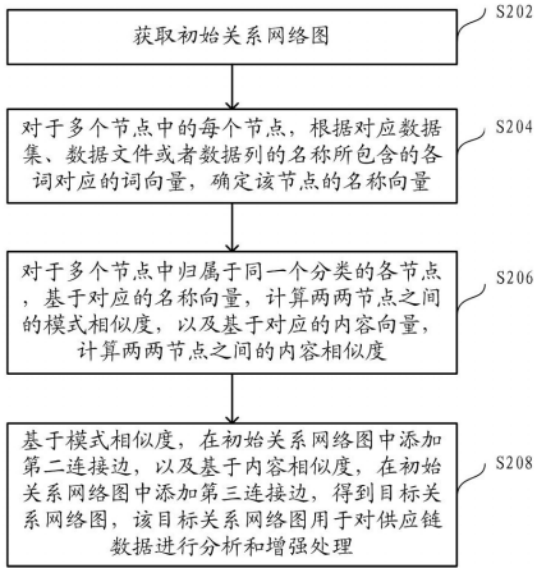
权利要求书3页 说明书10页 附图3页

(54)发明名称

一种供应链数据分析和增强处理方法及装置

(57)摘要

本说明书实施例提供一种供应链数据分析和增强处理方法及装置,获取初始关系网络图,其中包括分别归属于三个分类的多个节点,其中,第一类节点与数据集相对应,第二类节点与数据文件相对应,第三类节点与数据列相对应,在具有隶属关系的两个分类的节点之间通过第一连接边连接。对于每个节点,根据对应数据对象的名称包含的各词对应的词向量,确定对应的名称向量。对于归属于同一个分类的各节点,分别基于对应的名称向量和内容向量,计算两两节点之间的模式相似度以及内容相似度,并基于模式相似度和内容相似度,在初始关系网络图中添加第二连接边和第三连接边,得到目标关系网络图。基于目标关系网络图针对供应链数据进行分析和增强处理。



CN 114443783 B

1.一种供应链数据分析处理方法,包括:

获取初始关系网络图;所述初始关系网络图包括分别归属于三个分类的多个节点,其中,归属于第一个分类的节点为第一类节点,该第一类节点与数据仓库中的数据集相对应;归属于第二个分类的节点为第二类节点,该第二类节点与数据集中的数据文件相对应;归属于第三个分类的节点为第三类节点,该第三类节点与从数据文件中抽取的数据列相对应;在具有隶属关系的两个分类的节点之间通过第一连接边连接;

对于所述多个节点中的每个节点,根据对应数据集、数据文件或者数据列的名称包含的各词对应的词向量,确定该节点的名称向量;

对于所述多个节点中归属于同一个分类的各节点,基于对应的名称向量,计算两两节点之间的模式相似度,以及基于对应的内容向量,计算两两节点之间的内容相似度;其中,任一节点的内容向量,根据对应数据集、数据文件或者数据列的内容包含的各词对应的词向量而确定;

基于所述模式相似度,在所述初始关系网络图中添加第二连接边,以及基于所述内容相似度,在所述初始关系网络图中添加第三连接边,得到目标关系网络图,所述目标关系网络图用于对供应链数据进行分析处理;

对于所述目标关系网络图中的任一节点,在所述目标关系网络图中,确定出经过预定数量K1以内的第二连接边到达的、与所述任一节点归属于同一个分类的第一目标节点;将所述第一目标节点对应的数据集、数据文件或者数据列,作为针对所述任一节点进行模式相似性分析的分析结果;和/或,

对于所述目标关系网络图中的任一节点,在所述目标关系网络图中,确定出经过预定数量K2以内的第三连接边到达的、与所述任一节点归属于同一个分类的第二目标节点;将所述第二目标节点对应的数据集、数据文件或者数据列,作为针对所述任一节点进行内容相似性分析的分析结果。

2.根据权利要求1所述的方法,还包括:

对于所述目标关系网络图,判断所述第三类节点对应的数据列是否为所述第二类节点对应的数据文件的主键或者外键,若是,则在所述目标关系网络图中添加第四连接边。

3.根据权利要求1所述的方法,其中,所述确定该节点的名称向量,包括:

基于word2vec网络,确定该节点对应的数据集、数据文件或者数据列的名称包含的各词对应的第一词向量,以及基于WordNet,确定该节点对应的数据集、数据文件或者数据列的名称包含的各词对应的第二词向量;

对所述各词对应的第一词向量和第二词向量求平均,得到所述各词的向量表示;

对所述各词的向量表示进行融合,得到该节点的名称向量。

4.根据权利要求1所述的方法,其中,所述归属于同一个分类的各节点包括第一节点和第二节点;

所述基于所述模式相似度,在所述初始关系网络图中添加第二连接边,包括:判断所述第一节点与第二节点之间的模式相似度是否大于第一阈值,若是,则在所述第一节点与第二节点之间构建第二连接边,且将所述模式相似度作为所述第二连接边的权重;

所述基于所述内容相似度,在所述初始关系网络图中添加第三连接边,包括:判断所述第一节点与第二节点之间的内容相似度是否大于第二阈值,若是,则在所述第一节点与第

二节点之间构建第三连接边，且将所述内容相似度作为所述第三连接边的权重。

5.根据权利要求1所述的方法，还包括：

利用所述目标关系网络图对图神经网络进行训练，得到所述目标关系网络图中每个节点的节点向量；

基于所述节点向量，计算两两节点之间的打分，所述打分指示两个节点之间存在连接边的概率；所述打分用于在所述目标关系网络图中添加新连接边。

6.根据权利要求5所述的方法，还包括：

获取目标数据集，并针对所述目标数据集构建对应的目标子图；

利用所述图神经网络对所述目标子图进行图嵌入处理，得到所述目标子图中每个节点的节点向量；

将所述目标子图整合到所述目标关系网络图；所述将所述目标子图整合到所述目标关系网络图至少包括：基于所述目标子图以及目标关系网络图中各个节点的节点向量，依次计算所述目标子图中各个节点与所述目标关系网络图中各个节点之间的链接相似度，并基于该链接相似度，在所述目标关系网络图中添加新连接边。

7.根据权利要求1所述的方法，还包括：

对于所述目标关系网络图中归属于同一个分类的两个节点，基于所述目标关系网络图，确定分别以该两个节点为起始节点和终止节点的目标路径；

基于所述目标路径包含的各连接边，确定该两个节点对应的数据集、数据文件或者数据列之间的关系类型。

8.一种供应链数据分析处理装置，包括：

获取单元，用于获取初始关系网络图；所述初始关系网络图包括分别归属于三个分类的多个节点，其中，归属于第一个分类的节点为第一类节点，该第一类节点与数据仓库中的数据集相对应；归属于第二个分类的节点为第二类节点，该第二类节点与数据集中的数据文件相对应；归属于第三个分类的节点为第三类节点，该第三类节点与从数据文件中抽取的数据列相对应；在具有隶属关系的两个分类的节点之间通过第一连接边连接；

确定单元，用于对于所述多个节点中的每个节点，根据对应数据集、数据文件或者数据列的名称包含的各词对应的词向量，确定该节点的名称向量；

计算单元，用于对于所述多个节点中归属于同一个分类的各节点，基于对应的名称向量，计算两两节点之间的模式相似度，以及基于对应的内容向量，计算两两节点之间的内容相似度；其中，任一节点的内容向量，根据对应数据集、数据文件或者数据列的内容包含的各词对应的词向量而确定；

添加单元，用于基于所述模式相似度，在所述初始关系网络图中添加第二连接边，以及基于所述内容相似度，在所述初始关系网络图中添加第三连接边，得到目标关系网络图，所述目标关系网络图用于对供应链数据进行分析处理；

分析单元，用于对于所述目标关系网络图中的任一节点，在所述目标关系网络图中，确定出经过预定数量K1以内的第二连接边到达的、与所述任一节点归属于同一个分类的第一目标节点；将所述第一目标节点对应的数据集、数据文件或者数据列，作为针对所述任一节点进行模式相似性分析的分析结果；和/或，

对于所述目标关系网络图中的任一节点，在所述目标关系网络图中，确定出经过预定

数量K2以内的第三连接边到达的、与所述任一节点归属于同一个分类的第二目标节点;将所述第二目标节点对应的数据集、数据文件或者数据列,作为针对所述任一节点进行内容相似性分析的分析结果。

一种供应链数据分析和增强处理方法及装置

技术领域

[0001] 本说明书一个或多个实施例涉及计算机技术领域，尤其涉及一种供应链数据分析和增强处理方法及装置。

背景技术

[0002] 大型供应链集成服务集团公司的超大规模供应链数据主要涉及主数据、行为数据、业务数据、财务数据及第三方数据等诸多方面，普遍存在数据冗余、数据缺失、数据格式不一致及数据分布不平衡等问题，迫切需要研究一种新型数据分析和增强的技术解决方案，以便更高效、准确地改善数据质量，为大型供应链集成服务集团公司数字化转型奠定基础。数据分析和增强是数据处理过程中的一项基本任务，用于确定并优化与数据处理及数据应用项目相关的超大规模数据集。大宗商品供应链集成服务集团公司汇聚了超大规模的机器可读和结构化数据集。这些数据一般收集在一个被称为数据湖或者数据仓库的数据存储中。数据所有者一般通过一个数据中台系统提供这些数据集或构建新的数据集，如，国内的阿里云MaxCompute系统、网易有数系统、袋鼠云数栈系统、数澜科技数栖系统和国外的Talend系统等，这些大数据平台一般是通过数据资产目录或数据标签的形式支持检索目标数据，使得用于寻找、下载、准备和整合相关数据的时间和精力都比较多。随着数据应用的大量增加和数据分类、分级的管制，相似数据集大量存在，目标数据检索难度增大，数据处理效率大大降低。因此，迫切需要提供一种解决方案，用于数据模型相似度评估、分析、处理、优化，以及基于数据安全相关法律法规要求下进行数据服务和应用的数据处理工作，提高数据处理工作效率和数据使用合规性，以实现高效和可扩展的数据分析和增强。

发明内容

[0003] 本说明书一个或多个实施例描述了一种供应链数据分析和增强处理方法及装置，可以更高效更准确地对供应链数据进行分析和增强。

[0004] 第一方面，提供了一种供应链数据分析和增强处理方法，包括：

[0005] 获取初始关系网络图；所述初始关系网络图包括分别归属于三个分类的多个节点，其中，归属于第一个分类的节点为第一类节点，该第一类节点与数据仓库中的数据集相对应；归属于第二个分类的节点为第二类节点，该第二类节点与数据集中的数据文件相对应；归属于第三个分类的节点为第三类节点，该第三类节点与从数据文件中抽取的数据列相对应；在具有隶属关系的两个分类的节点之间通过第一连接边连接；

[0006] 对于所述多个节点中的每个节点，根据对应数据集、数据文件或者数据列的名称包含的各词对应的词向量，确定该节点的名称向量；

[0007] 对于所述多个节点中归属于同一个分类的各节点，基于对应的名称向量，计算两两节点之间的模式相似度，以及基于对应的内容向量，计算两两节点之间的内容相似度；其中，任一节点的内容向量，根据对应数据集、数据文件或者数据列的内容包含的各词对应的词向量而确定；

[0008] 基于所述模式相似度，在所述初始关系网络图中添加第二连接边，以及基于所述内容相似度，在所述初始关系网络图中添加第三连接边，得到目标关系网络图，所述目标关系网络图用于对供应链数据进行分析和增强处理。

[0009] 第二方面，提供了一种供应链数据分析和增强处理装置，包括：

[0010] 获取单元，用于获取初始关系网络图；所述初始关系网络图包括分别归属于三个分类的多个节点，其中，归属于第一个分类的节点为第一类节点，该第一类节点与数据仓库中的数据集相对应；归属于第二个分类的节点为第二类节点，该第二类节点与数据集中的数据文件相对应；归属于第三个分类的节点为第三类节点，该第三类节点与从数据文件中抽取的数据列相对应；在具有隶属关系的两个分类的节点之间通过第一连接边连接；

[0011] 确定单元，用于对于所述多个节点中的每个节点，根据对应数据集、数据文件或者数据列的名称包含的各词对应的词向量，确定该节点的名称向量；

[0012] 计算单元，用于对于所述多个节点中归属于同一个分类的各节点，基于对应的名称向量，计算两两节点之间的模式相似度，以及基于对应的内容向量，计算两两节点之间的内容相似度；其中，任一节点的内容向量，根据对应数据集、数据文件或者数据列的内容包含的各词对应的词向量而确定；a

[0013] 添加单元，用于基于所述模式相似度，在所述初始关系网络图中添加第二连接边，以及基于所述内容相似度，在所述初始关系网络图中添加第三连接边，得到目标关系网络图，所述目标关系网络图用于对供应链数据进行分析和增强处理。

[0014] 本说明书一个或多个实施例提供的供应链数据分析和增强处理方法及装置，从不同类节点之间的隶属关系、同类节点之间的内容相似性和模式相似性等方面来构建目标关系网络图，由此使得所构建的目标关系网络图可以用于多维度的数据分析和增强，进而可以提升数据分析和增强效率。

附图说明

[0015] 为了更清楚地说明本说明书实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本说明书的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

[0016] 图1示出根据一个实施例的供应链数据分析和增强处理系统示意图；

[0017] 图2示出根据一个实施例的关系网络图的构建方法流程图；

[0018] 图3示出根据一个实施例的目标关系网络图的示意图；

[0019] 图4示出根据一个实施例的供应链数据分析方法流程图；

[0020] 图5示出根据一个实施例的供应链数据分析和增强处理装置示意图。

具体实施方式

[0021] 下面结合附图，对本说明书提供的方案进行描述。

[0022] 图1示出根据一个实施例的供应链数据分析和增强处理系统示意图。图1中，该系统可以包括剖析装置、构建装置和链接器。

[0023] 具体地，在将数据湖或数据仓库中的供应链数据(以下简称数据)通过数据接口的

方式汇聚到系统之后，在剖析装置，可以梳理出数据湖或者数据仓库中数据列、数据文件和数据集之间的关系，同时计算出数据列、数据文件和数据集各自的内容向量并保存，以便于在后续计算内容相似度时使用。在构建装置，可以构建出若干种核心关系，形成一张完整的关系网络图。这里的若干种核心关系包括但不限于模式相似性关系、内容相似性关系以及主外键关系等等。针对关系网络图，还可以基于业务人员核心知识进行整合和构建。在链接器，基于已有核心关系，对图神经网络进行训练，得到每个节点的节点向量。之后，基于每个节点的节点向量，可以计算任意两个节点之间的打分，从而对关系网络图进行补全。同时对新增的数据集，可以推断出新增的数据集对应的子图中的节点与已有的关系网络图中的节点的关系，完成对新增数据集关系的推断，增强数据发现和增强已有的关系网络图。

[0024] 需要说明，上述系统可以对外提供服务接口。应理解，通过该服务接口，可以额外提供多种数据分析方法，从而相比于传统的大数据平台，能够增强数据中的语义、模式等信息，提升数据挖掘分析的效率和准确率，帮助准确发现和补全新的数据，提升企业的数据赋能价值创新能力。

[0025] 以下对上述关系网络图的构建方法进行说明。

[0026] 图2示出根据一个实施例的关系网络图的构建方法流程图。可以理解，该方法可以通过任何具有计算、处理能力的装置、设备、平台、设备集群来执行。如图2所示，该构建方法至少可以包括如下步骤。

[0027] 步骤202，获取初始关系网络图。

[0028] 该初始关系网络图包括分别归属于三个分类的多个节点，其中，归属于第一个分类的各节点统称为第一类节点，该第一类节点与数据仓库(或者数据湖)中的数据集(dataset)相对应(即每个第一类节点对应于一个数据集)。归属于第二个分类的各节点统称为第二类节点，该第二类节点与数据集中的数据文件相对应(即每个第二类节点对应于一个数据文件)。归属于第三个分类的各节点统称为第三类节点，该第三类节点与从数据文件中抽取的数据列(column)相对应(即每个第三类节点对应于一个数据列)。

[0029] 上述多个节点中，在具有隶属关系(即isPartof关系)的两个分类的节点之间通过第一连接边连接。从而基于该第一连接边，可以查找具有隶属关系的数据列与数据文件，以及数据文件与数据集，进而执行数据去重处理。比如，对于同属于一个数据集的两个数据文件，可以进一步分析该两个数据文件的内容是否一致，并在相一致的情况下，去除一个数据文件等等。

[0030] 以下对上述数据集、数据文件和数据列等概念一一进行说明。

[0031] 首先，一个数据集通常由若干数据文件构成(也即一个数据文件可以为一个数据集的一部分)。这里的数据文件可以为以下中的任一种：表格文件、CSV文件、JSON文件以及分布在大数据环境中的各种分布式数据(比如，以HDFS存储的大数据文件(big table))。本说明书中，数据文件所存储的数据可以为多模态数据，比如，既可以为文本数据，也可以为图像数据，以下描述中以文本数据为例进行说明。此外，一个数据文件可以包含若干数据列(也即一个数据列可以为一个数据文件的一部分)。这里的数据列可以由字段以及对应的字段值构成，这里的字段即为数据列的名称，字段值即为数据列的内容。

[0032] 需要说明，对于初始关系网络图中的每个节点，可以确定对应的内容向量。具体地，对于每个第三类节点，可以基于word2vec网络，确定该节点对应数据列的内容(即字段

值)包含的各词对应的词向量。之后,可以对各词对应的词向量进行求平均,得到第三类节点的固定长度的内容向量,也即得到第三类节点对应数据列的内容向量。

[0033] 对于每个第二类节点,可以对对应数据文件包含的各数据列的内容向量进行求平均,得到该第二类节点的内容向量,也即得到第二类节点对应数据文件的内容向量。

[0034] 对于每个第一类节点,可以对对应数据集包含的各数据文件的内容向量进行求平均,得到该第一类节点的内容向量,也即得到第一类节点对应数据集的内容向量。

[0035] 在得到目标关系网络图中多个节点各自的内容向量之后可以对其进行保存,以便后续使用。需要说明,由于这里只存储了每个节点的内容向量,而并没有存储原始数据,从而可以减少隐私风险。此外,由于在本说明书实施例中,针对数据集、数据文件以及数据列均确定了对应的内容向量,从而为后续从不同粒度进行内容相似性分析奠定了基础。

[0036] 步骤204,对于多个节点中的每个节点,根据对应数据集、数据文件或者数据列的名称包含的各词对应的词向量,确定该节点的名称向量。

[0037] 上述确定该节点的名称向量具体可以包括:基于word2vec网络,确定该节点对应的数据集、数据文件或者数据列的名称包含的各词对应的第一词向量,以及基于WordNet(由普林斯顿大学认识科学实验室建立和维护的英语字典),确定该节点对应的数据集、数据文件或者数据列的名称包含的各词对应的第二词向量。对各词对应的第一词向量和第二词向量求平均,得到各词的向量表示。对各词的向量表示进行融合(比如,求平均),得到该节点的名称向量。如此,就可以得到目标关系网络图中多个节点各自的名称向量。针对该名称向量也可以保存,以便后续使用。

[0038] 步骤206,对于多个节点中归属于同一个分类的各节点,基于对应的名称向量,计算两两节点之间的模式相似度,以及基于对应的内容向量,计算两两节点之间的内容相似度。

[0039] 这里的模式相似度或内容相似度可以包括但不限于余弦相似度或者Ochiai系数等。当然,在实际应用中,也可以基于欧氏距离、曼哈顿距离或者皮尔逊相关系数等,确定上述模式相似度或者内容相似度。

[0040] 步骤206具体可以为:基于各第一类节点的名称向量,计算两两第一类节点之间的模式相似度;基于各第二类节点的名称向量,计算两两第二类节点之间的模式相似度;以及基于各第三类节点的名称向量,计算两两第三类节点之间的模式相似度。同理,基于预先存储的各第一类节点的内容向量,计算两两第一类节点之间的内容相似度;基于预先存储的各第二类节点的内容向量,计算两两第二类节点之间的内容相似度;以及基于预先存储的各第三类节点的内容向量,计算两两第三类节点之间的内容相似度。

[0041] 需要说明,上述可以是针对每一类节点中的所有节点计算两两节点之间的模式相似度和内容相似度,也可以只针对部分节点计算模式相似度和内容相似度,以减少计算量。比如,只针对可能存在关联关系的节点计算模式相似度和内容相似度。这里的可能存在关联关系的节点例如可以为对应数据对象(包括数据集、数据文件或者数据列)选自同一个部门等等。

[0042] 步骤208,基于模式相似度,在初始关系网络图中添加第二连接边,以及基于内容相似度,在初始关系网络图中添加第三连接边,得到目标关系网络图,该目标关系网络图用于对供应链数据进行分析和增强处理。

[0043] 以上述多个分类中任一分类为例来说，假设归属于该分类的两个节点分别为第一节点和第二节点，那么上述在初始关系网络图中添加第二连接边可以包括：判断第一节点与第二节点之间的模式相似度是否大于第一阈值，若是，则在第一节点与第二节点之间构建第二连接边，且将该两者的模式相似度作为第二连接边的权重。类似地，可以在归属于每个分类的节点之间构建出第二连接边。

[0044] 在一个示例中，上述第二连接边可以表示为：<第一节点，第二节点，模式相似度：xx>。

[0045] 需要说明，上述模式相似度大于第一阈值，也可以理解为是第一节点和第二节点之间具有模式相似性(schemaSimilarity)关系。从而，基于该第二连接边，可以查找具有模式相似性关系的数据集、数据文件或者数据列。

[0046] 还以上述第一节点和第二节点为例来说，上述在初始关系网络图中添加第三连接边可以包括：判断第一节点与第二节点之间的内容相似度是否大于第二阈值，若是，则在第一节点与第二节点之间构建第三连接边，且将内容相似度作为第三连接边的权重。

[0047] 在一个示例中，上述第三连接边可以表示为：<第一节点，第二节点，内容相似度：yy>。

[0048] 需要说明，上述内容相似度大于第二阈值，也可以理解为是第一节点和第二节点之间具有内容相似性(contentSimilarity)关系。从而，基于该第三连接边，可以查找具有内容相似性关系的数据集、数据文件或者数据列。

[0049] 当然，在实际应用中，还可以在上述目标关系网络图中添加新连接边，以指示新的关系类型。

[0050] 在一个示例中，上述添加新连接边可以包括：对于目标关系网络图，判断第三类节点对应的数据列是否为第二类节点对应的数据文件的主键(PrimaryKey)或者外键(ForeignKey)，若是，则在目标关系网络图中添加第四连接边。

[0051] 也就是说，基于该第四连接边，可以查找具有主外键(prikeyForkey)关系的数据列和数据文件。

[0052] 图3示出根据一个实施例的目标关系网络图的示意图。图3中，目标关系网络图包括归属于三个分类的多个节点，其中，第一类节点通过点虚线框示出，其代表数据集；第二类节点通过横线虚线框示出，其代表数据文件；第三类节点通过实线框示出，其代表数据列。此外，目标关系图还包括四种类型的连接边。其中，第一连接边连接具有隶属关系的节点，第二连接边连接具有模型相似性关系的节点，第三连接边连接具有内容相似性关系的节点，第四连接边连接具有主外键关系的节点。最后，对应于每个连接边的数字，代表对应的权重。比如：“主外键关系:0.95”代表节点“数据列1”与节点“数据列2”之间的第四连接边的权重为0.95。

[0053] 当然，在实际应用中，业务人员也可以对上述目标关系网络图进行标注或者编辑，加入领域知识。比如图3中的“数据集3”和“数据集4”两个节点之间，可能不具有上述四种关系，但业务人员基于领域知识会发现，这两个数据集实际上存在contentSimilarity关系，那么可以对该目标关系网络图的连接边进行完善，以加入更多领域知识，由此可以形成更加全面的关系网络图，为后续数据的丰富和增强做好准备。

[0054] 在另一个示例中，上述添加新连接边可以包括：利用目标关系网络图对图神经网

络(Graph Neural Network,GNN)进行训练,得到目标关系网络图中每个节点的节点向量。基于各节点的节点向量,计算两两节点之间的打分(比如,余弦相似度)。该打分指示两个节点之间存在连接边的概率。输出打分大于阈值分数的节点对。之后,由业务人员确定是否在该节点对之间构建新连接边。

[0055] 同上所述,这里可以是针对所有节点计算两两打分,也可以只针对部分节点计算打分,本说明书对此不作限定。

[0056] 需要说明,在对GNN进行训练之前,可以先基于目标关系网络图中节点之间的连接边的类型,确定节点的若干关系标签。这里的若干关系标签可以包括isPartof关系标签、schemaSimilarity关系标签、contentSimilarity关系标签以及prikeyForkey关系标签中的至少一项。举例来说,假设归属于第二个分类的节点1通过第一连接边与归属于第二个分类的节点2连接,且节点1与其它节点不存在任何的连接边,那么节点1的isPartof关系标签为1,其它关系标签均为0。类似地,可以确定出每个节点的若干关系标签。

[0057] 之后,可以将目标关系网络图输入GNN,得到预测结果。这里的预测结果可以包括两两节点之间的若干相似度。该若干相似度与上述若干关系标签相对应。以及根据预测结果与若干关系标签,确定GNN中参数的训练梯度,基于训练梯度,更新GNN中的参数,得到训练后的GNN。

[0058] 需要说明,在完成针对GNN的训练后,可以同时获得目标关系网络图中每个节点的节点向量。

[0059] 应理解,上述目标关系网络图是基于数据湖或数据仓库中的已有数据集而建立的。当数据湖或数据仓库中新增数据集时,还可以对目标关系网络图执行如下的图更新操作。

[0060] 该图更新操作具体可以包括:获取目标数据集,并针对该目标数据集构建对应的目标子图。这里的目标子图的构建方法可以参见上述步骤202-步骤208。利用预先训练的图神经网络对目标子图进行图嵌入处理,得到目标子图中每个节点的节点向量。将目标子图整合到目标关系网络图,得到更新的目标关系网络图。

[0061] 上述将目标子图整合到目标关系网络图至少可以包括:基于目标子图以及目标关系网络图中各个节点的节点向量,依次计算目标子图中各个节点与目标关系网络图中各个节点之间的链接相似度,并基于该链接相似度,在目标关系网络图中添加新连接边。这里的链接相似度例如可以为余弦相似度等。

[0062] 具体地,如果任意的两个节点之间的链接相似度大于预定阈值,那么在该两个节点之间构建第五连接边。并将该链接相似度作为第五连接边的权重。

[0063] 当然,在实际应用中,为提升整合效率,可以只针对目标子图和目标关系网络图中对应于数据集和数据文件的节点计算链接相似度,本说明书对此不作限定。

[0064] 总之,本说明书实施例提供的方案,可以利用图神经网络,逐步加强图中不同数据之间的关系。

[0065] 图4示出根据一个实施例的供应链数据分析方法流程图。可以理解,该方法可以通过任何具有计算、处理能力的装置、设备、平台、设备集群来执行。如图4所示,该构建方法至少可以包括如下步骤。

[0066] 步骤402,获取目标关系网络图。

[0067] 该目标关系网络图可以是基于图2示出的各方法步骤构建得到。

[0068] 在一个示例中，该目标关系网络图可以如图3所示。

[0069] 步骤404，基于目标关系网络图进行数据分析。

[0070] 在一个示例中，上述基于目标关系网络图进行数据分析可以包括：对于目标关系网络图中的任一节点，在目标关系网络图中，确定出经过预定数量K1以内的第二连接边到达的、与该任一节点归属于同一个分类的第一目标节点。将该第一目标节点对应的数据集、数据文件或者数据列，作为针对该任一节点进行模式相似性分析的分析结果；和/或，

[0071] 对于目标关系网络图中的任一节点，在目标关系网络图中，确定出经过预定数量K2以内的第三连接边到达的、与该任一节点归属于同一个分类的第二目标节点。将第二目标节点对应的数据集、数据文件或者数据列，作为针对该任一节点进行内容相似性分析的分析结果。

[0072] 也就是说，基于本说明书实施例构建的目标关系网络图，可以查找具有模式相似性(或内容相似性)的数据集、数据文件或者数据列。

[0073] 在另一个示例中，上述基于目标关系网络图进行数据分析还可以包括：对于目标关系网络图中的任一节点，基于该节点的内容向量以及其它节点的内容向量，查找该任一节点的相似节点。比如，将其它节点中与该任一节点的内容相似度大于阈值相似度的节点作为相似节点。

[0074] 类似地，也可以基于名称向量，查找相似节点。

[0075] 在又一个示例中，上述基于目标关系网络图进行数据分析还可以包括：对于目标关系网络图中归属于同一个分类的两个节点，基于该两个节点各自的内容向量，计算内容相似度；或者，基于该两个节点各自的名称向量，计算模式相似度；或者，基于该两个节点各自的节点向量，计算综合相似度。

[0076] 在还一个示例中，上述基于目标关系网络图进行数据分析还可以包括：对于目标关系网络图中归属于同一个分类的两个节点，基于目标关系网络图，确定分别以该两个节点为起始节点和终止节点的目标路径。基于目标路径包含的各连接边，确定该两个节点所对应的数据集、数据文件或者数据列之间的关系类型。

[0077] 应理解，通过上述数据分析方法，能够便于数据工程师更有效发现数据，并对现有的数据找到相似性(包括内容相似性、模式相似性以及综合相似性等)，来实现数据丰富和增强，提升数据分析的效果和效率。

[0078] 综上，本说明书实施例提供的方案，可以对大型供应链集成服务集团公司在数据湖或数据仓库中拥有的超大规模数据集，进行归档并创建一个关系网络图来进行数据整合。基于该关系网络图，可以让不同的团队在不接触原始数据的情况下，根据嵌入向量等检查不同部门的数据。基于嵌入向量的相似性的数据发现，允许企业在不牺牲隐私的情况下挖掘数据价值和创新潜力，且可容易地集成到现有的大数据平台中。此外，通过该方案，可以支持进行高效的数据发现、数据整合、数据探索和数据增强。

[0079] 最后，通过本说明书实施例提供的方案，可以帮助进行数据发现和增强，并丰富现有数据湖或者数据仓库中的数据信息。通过这些可扩展的有效发现操作，支持大型供应链集成服务集团公司超大规模数据集中找到相关的数据，以便更好地利用它们，最大限度地发挥挖掘大数据平台的价值和潜力。

[0080] 本说明书一个实施例还提供的一种供应链数据分析和增强处理装置，如图5所示，该装置可以包括：

[0081] 获取单元502，用于获取初始关系网络图，该初始关系网络图包括分别归属于三个分类的多个节点，其中，归属于第一个分类的节点为第一类节点，该第一类节点与数据仓库中的数据集相对应。归属于第二个分类的节点为第二类节点，该第二类节点与数据集中的数据文件相对应。归属于第三个分类的节点为第三类节点，该第三类节点与从数据文件中抽取的数据列相对应，在具有隶属关系的两个分类的节点之间通过第一连接边连接。

[0082] 确定单元504，用于对于多个节点中的每个节点，根据对应数据集、数据文件或者数据列的名称包含的各词对应的词向量，确定该节点的名称向量。

[0083] 确定单元504具体用于：

[0084] 基于word2vec网络，确定该节点对应的数据集、数据文件或者数据列的名称包含的各词对应的第一词向量，以及基于WordNet，确定该节点对应的数据集、数据文件或者数据列的名称包含的各词对应的第二词向量；

[0085] 对各词对应的第一词向量和第二词向量求平均，得到各词的向量表示；

[0086] 对各词的向量表示进行融合，得到该节点的名称向量。

[0087] 计算单元506，用于对于多个节点中归属于同一个分类的各节点，基于对应的名称向量，计算两两节点之间的模式相似度，以及基于对应的内容向量，计算两两节点之间的内容相似度。其中，任一节点的内容向量，根据对应数据集、数据文件或者数据列的内容包含的各词对应的词向量而确定。

[0088] 添加单元508，用于基于模式相似度，在初始关系网络图中添加第二连接边，以及基于内容相似度，在初始关系网络图中添加第三连接边，得到目标关系网络图，该目标关系网络图用于对供应链数据进行分析和增强处理。

[0089] 其中，归属于同一个分类的各节点包括第一节点和第二节点；

[0090] 添加单元508具体用于：判断第一节点与第二节点之间的模式相似度是否大于第一阈值，若是，则在第一节点与第二节点之间构建第二连接边，且将模式相似度作为第二连接边的权重；

[0091] 添加单元508具体还用于：判断第一节点与第二节点之间的内容相似度是否大于第二阈值，若是，则在第一节点与第二节点之间构建第三连接边，且将内容相似度作为第三连接边的权重。

[0092] 可选地，该装置还可以包括：

[0093] 判断单元510，用于对于目标关系网络图，判断第三类节点对应的数据列是否为第二类节点对应的数据文件的主键或者外键，若是，则在目标关系网络图中添加第四连接边。

[0094] 可选地，该装置还可以包括：

[0095] 训练单元512，用于利用目标关系网络图对图神经网络进行训练，得到目标关系网络图中每个节点的节点向量；

[0096] 计算单元506，还用于基于节点向量，计算两两节点之间的打分，该打分指示两个节点之间存在连接边的概率，该打分用于在目标关系网络图中添加新连接边。

[0097] 可选地，该装置还可以包括：

[0098] 构建单元514，用于获取目标数据集，并针对所述目标数据集构建对应的目标子

图；

[0099] 处理单元516，用于利用图神经网络对目标子图进行图嵌入处理，得到目标子图中每个节点的节点向量；

[0100] 整合单元518，用于将目标子图整合到目标关系网络图，将目标子图整合到目标关系网络图至少包括：基于目标子图以及目标关系网络图中各个节点的节点向量，依次计算目标子图中各个节点与目标关系网络图中各个节点之间的链接相似度，并基于该链接相似度，在目标关系网络图中添加新连接边。

[0101] 可选地，该装置还可以包括：

[0102] 分析单元520，用于基于目标关系网络图进行数据分析。

[0103] 分析单元520具体用于：

[0104] 对于目标关系网络图中的任一节点，在目标关系网络图中，确定出经过预定数量K1以内的第二连接边到达的、与任一节点归属于同一个分类的第一目标节点；将第一目标节点对应的数据集、数据文件或者数据列，作为针对任一节点进行模式相似性分析的分析结果；和/或，

[0105] 对于目标关系网络图中的任一节点，在目标关系网络图中，确定出经过预定数量K2以内的第三连接边到达的、与所述任一节点归属于同一个分类的第二目标节点；将第二目标节点对应的数据集、数据文件或者数据列，作为针对任一节点进行内容相似性分析的分析结果。

[0106] 分析单元520还具体用于：

[0107] 对于目标关系网络图中归属于同一个分类的两个节点，基于目标关系网络图，确定分别以该两个节点为起始节点和终止节点的目标路径；

[0108] 基于目标路径包含的各连接边，确定该两个节点所对应的数据集、数据文件或者数据列之间的关系类型。

[0109] 本说明书上述实施例装置的各功能模块的功能，可以通过上述方法实施例的各步骤来实现，因此，本说明书一个实施例提供的装置的具体工作过程，在此不复赘述。

[0110] 本说明书一个实施例提供的供应链数据分析和增强处理装置，可以提升数据分析和增强效率。

[0111] 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于装置实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0112] 结合本说明书公开内容所描述的方法或者算法的步骤可以硬件的方式来实现，也可以是由处理器执行软件指令的方式来实现。软件指令可以由相应的软件模块组成，软件模块可以被存放于RAM存储器、闪存、ROM存储器、EPROM存储器、EEPROM存储器、寄存器、硬盘、移动硬盘、CD-ROM或者本领域熟知的任何其它形式的存储介质中。一种示例性的存储介质耦合至处理器，从而使处理器能够从该存储介质读取信息，且可向该存储介质写入信息。当然，存储介质也可以是处理器的组成部分。处理器和存储介质可以位于ASIC中。另外，该ASIC可以位于服务器中。当然，处理器和存储介质也可以作为分立组件存在于服务器中。

[0113] 本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功

能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。计算机可读介质包括计算机存储介质和通信介质，其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是通用或专用计算机能够存取的任何可用介质。

[0114] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

[0115] 以上所述的具体实施方式，对本说明书的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本说明书的具体实施方式而已，并不用于限定本说明书的保护范围，凡在本说明书的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本说明书的保护范围之内。

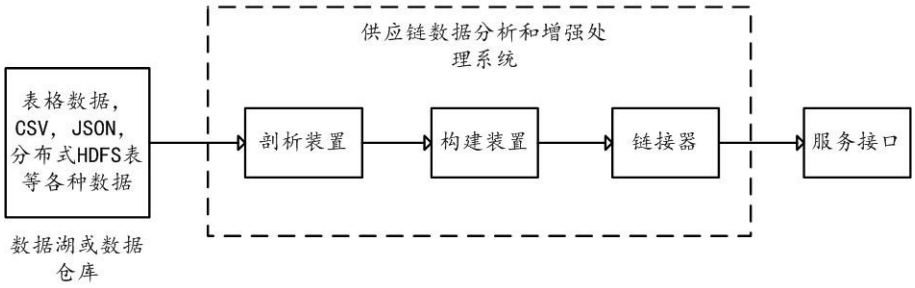

供应链数据分析和增强处理系统
表格数据，CSV，JSON，分布式HDFS表等各种数据
剖析装置
构建装置
链接器
服务接口
数据湖或数据仓库

图1

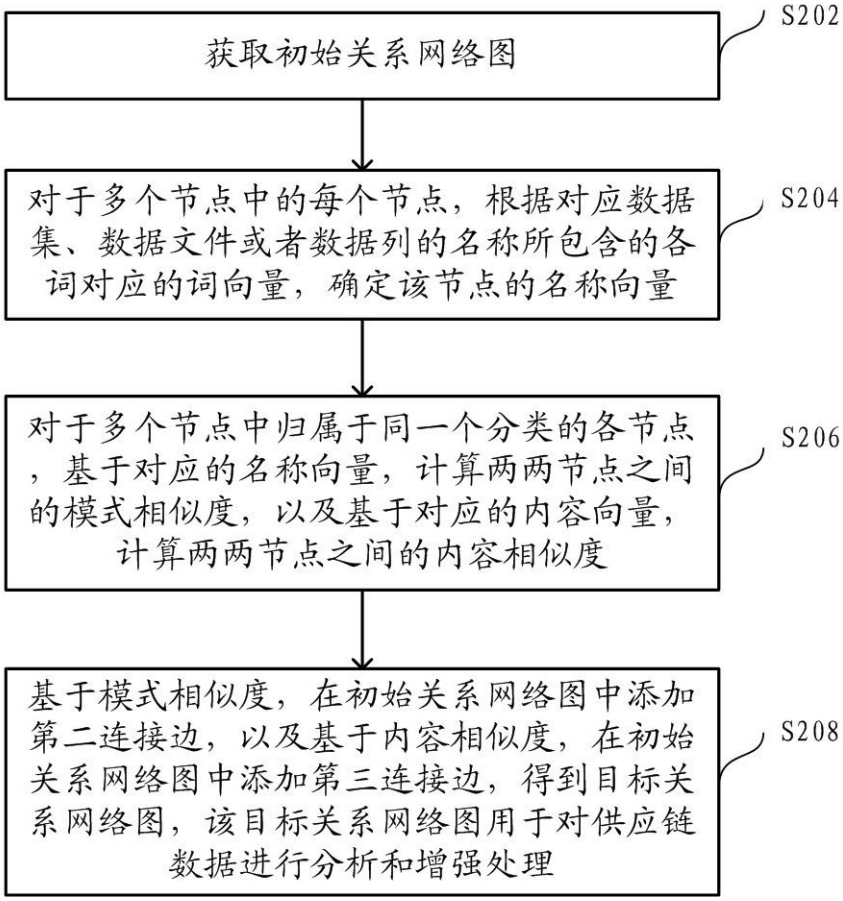

获取初始关系网络图
S202
对于多个节点中的每个节点，根据对应数据集、数据文件或者数据列的名称所包含的各词对应的词向量，确定该节点的名称向量
S204
对于多个节点中归属于同一个分类的各节点，基于对应的名称向量，计算两两节点之间的模式相似度，以及基于对应的内容向量，计算两两节点之间的内容相似度
S206
基于模式相似度，在初始关系网络图中添加第二连接边，以及基于内容相似度，在初始关系网络图中添加第三连接边，得到目标关系网络图，该目标关系网络图用于对供应链数据进行分析和增强处理
S208

图2

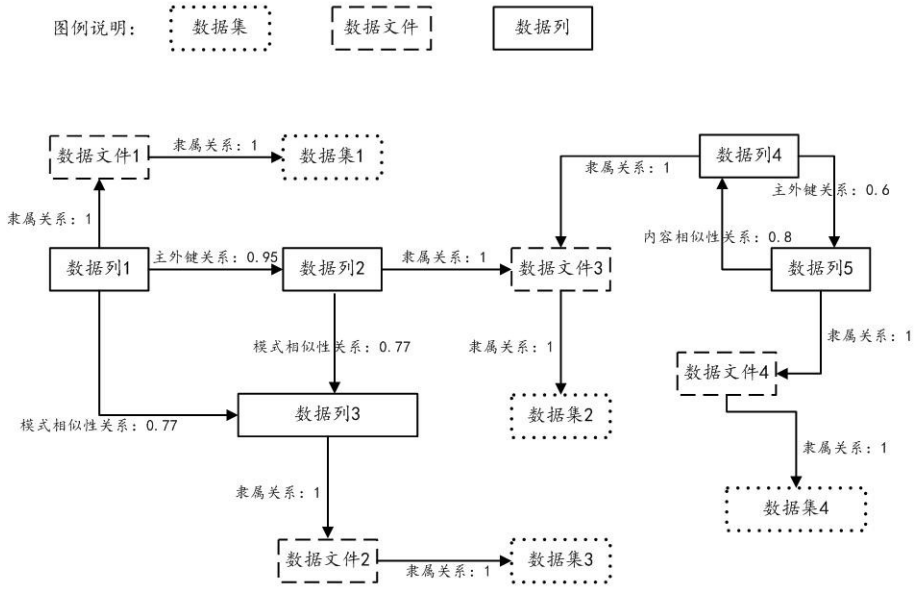

图例说明：
数据集
数据文件
数据列
数据文件1
隶属关系：1
数据集1
隶属关系：1
数据列1
主外键关系：0.95
数据列2
隶属关系：1
数据文件3
隶属关系：1
数据列4
主外键关系：0.6
内容相似性关系：0.8
数据列5
隶属关系：1
数据文件4
隶属关系：1
数据集4
模式相似性关系：0.77
数据列3
模式相似性关系：0.77
隶属关系：1
数据集2
隶属关系：1
数据文件2
隶属关系：1
数据集3

图3

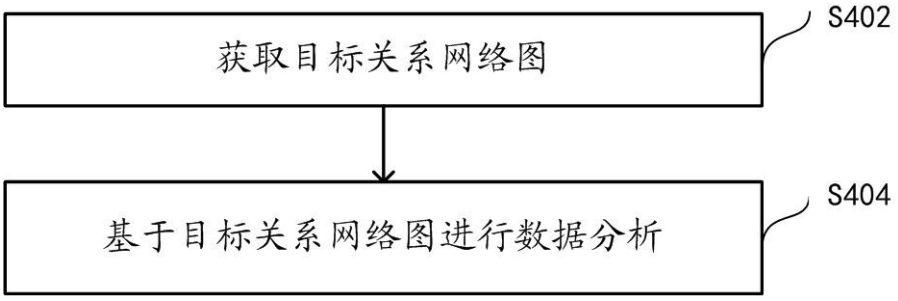

获取目标关系网络图
S402
基于目标关系网络图进行数据分析
S404

图4

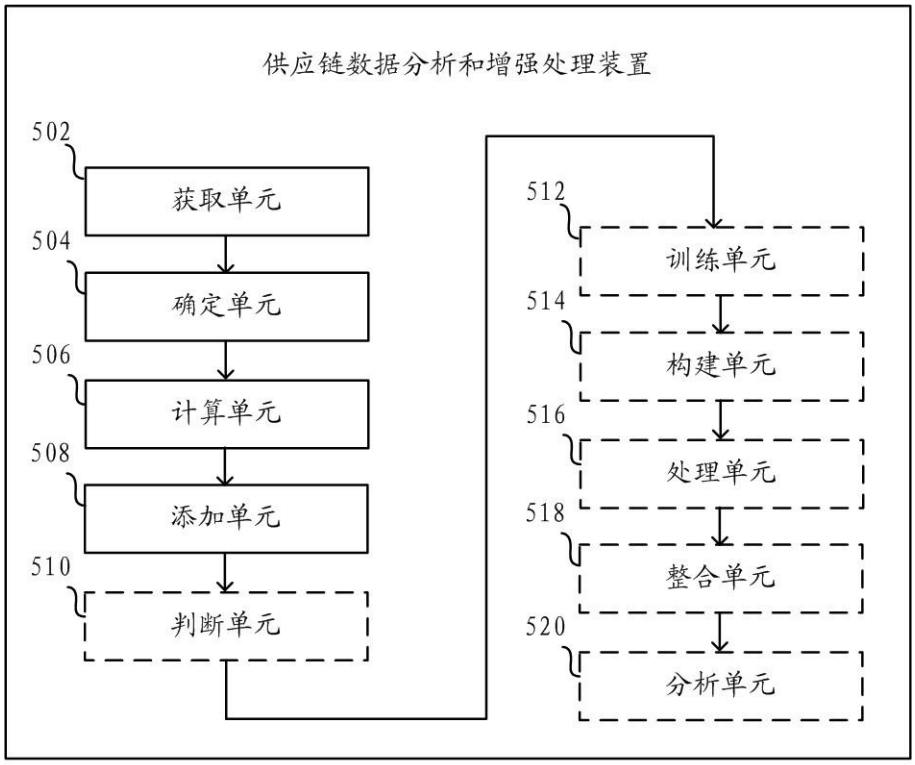

供应链数据分析和增强处理装置
502
获取单元
504
确定单元
506
计算单元
508
添加单元
510
判断单元
512
训练单元
514
构建单元
516
处理单元
518
整合单元
520
分析单元

图5