(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112948469 B
(45)授权公告日 2023.10.13

(21)申请号 202110410056.4
(22)申请日 2021.04.16
(65)同一申请的已公布的文献号
申请公布号 CN 112948469 A
(43)申请公布日 2021.06.11
(73)专利权人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼
(72)发明人 任霁野 王媛 汪伟
(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 熊永强
(51)Int.Cl.
G06F 16/2458(2019.01)
G06F 16/28(2019.01)

(56)对比文件
CN 110647522 A,2020.01.03
CN 105740381 A,2016.07.06
CN 102693317 A,2012.09.26
CN 107784111 A,2018.03.09
CN 109299090 A,2019.02.01
CN 110135890 A,2019.08.16
CN 111309776 A,2020.06.19
CN 111858720 A,2020.10.30
CN 112001649 A,2020.11.27
CN 112070402 A,2020.12.11
CN 112231350 A,2021.01.15
EP 1675060 A1,2006.06.28
郑吉;周莲英.基于层次性过滤的社交网络关键节点挖掘算法研究.数据通信.2018,(第04期),第30-35页.

审查员 孙士博

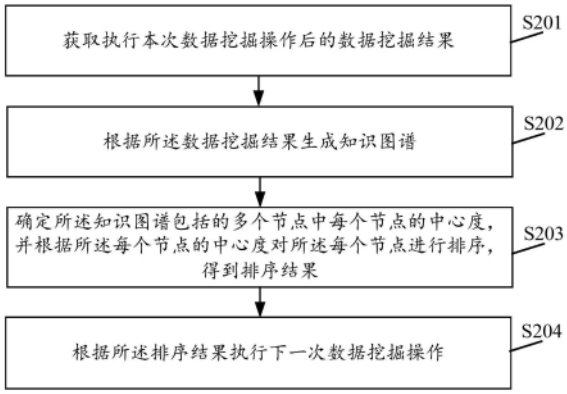
权利要求书2页 说明书13页 附图5页

(54)发明名称

数据挖掘方法、装置、计算机设备及存储介质

(57)摘要

本申请实施例提供了一种数据挖掘方法、装置、计算机设备及存储介质，该方法应用于大数据技术领域，该方法包括：获取执行本次数据挖掘操作后的数据挖掘结果；根据所述数据挖掘结果生成知识图谱；确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果；根据所述排序结果执行下一次数据挖掘操作。采用本申请，可以基于不同知识体系有效地进行数据挖掘。本申请涉及区块链技术，如可将数据挖掘结果写入区块链中。



CN 112948469 B

1.一种数据挖掘方法,其特征在于,包括:

获取执行本次数据挖掘操作后的数据挖掘结果;

根据所述数据挖掘结果生成知识图谱;

确定所述知识图谱包括的多个节点中每个节点的中心度,并根据所述每个节点的中心度对所述每个节点进行排序,得到排序结果;

确定已执行数据挖掘操作的次数;

在已执行数据挖掘操作的次数小于或等于预设次数时,根据所述排序结果确定针对所述每个节点的挖掘优先级;

根据所述每个节点的挖掘优先级执行下一次数据挖掘操作。

2.根据权利要求1所述的方法,其特征在于,所述确定所述知识图谱包括的多个节点中每个节点的中心度,包括:

确定所述知识图谱包括的多个节点中每个节点所在的最短路径的数量,并确定所述多个节点之间的最短路径的数量;

根据所述每个节点所在的最短路径的数量以及所述多个节点之间的最短路径的数量,确定所述每个节点的间接中心度以作为所述每个节点的中心度。

3.根据权利要求1所述的方法,其特征在于,所述确定所述知识图谱包括的多个节点中每个节点的中心度,包括:

确定所述多个节点中每个节点所连接的目标属性的节点的数量;

根据所述每个节点所连接的目标属性的节点的数量,确定所述每个节点的度中心度以作为所述每个节点的中心度。

4.根据权利要求1所述的方法,其特征在于,所述确定所述知识图谱包括的多个节点中每个节点的中心度,包括:

确定所述多个节点中每个节点所连接的目标属性的节点的数量,并确定所述多个节点之间的目标路径的数量;

根据所述每个节点所连接的目标属性的节点的数量以及所述多个节点之间的目标路径的数量,确定所述多个节点的度中心度以作为所述每个节点的中心度。

5.根据权利要求4所述的方法,其特征在于,所述根据所述排序结果执行下一次数据挖掘操作,还包括:

在已执行数据挖掘操作的次数大于预设次数时,计算所述每个节点的度增益;

根据所述每个节点的度增益,从所述多个节点中确定出度增益大于或等于预设数值的目标节点;

在所述目标节点为多个时,根据所述排序结果以及每个所述目标节点的度增益确定针对每个目标节点的挖掘优先级;

根据所述每个目标节点的挖掘优先级执行数据挖掘操作。

6.根据权利要求5所述的方法,其特征在于,所述计算所述每个节点的度增益,包括:

获取执行上一次数据挖掘操作后得到的所述每个节点的中心度;

根据所述执行上一次数据挖掘操作后得到的所述每个节点的中心度,以及执行本次数据挖掘操作后得到的所述每个节点的中心度,计算所述每个节点的度增益。

7.一种数据挖掘装置,其特征在于,包括:

数据挖掘模块，用于获取执行本次数据挖掘操作后的数据挖掘结果；

生成模块，用于根据所述数据挖掘结果生成知识图谱；

排序模块，用于确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果；

所述数据挖掘模块，还用于确定已执行数据挖掘操作的次数；在已执行数据挖掘操作的次数小于或等于预设次数时，根据所述排序结果确定针对所述每个节点的挖掘优先级；根据所述每个节点的挖掘优先级执行下一次数据挖掘操作。

8.一种计算机设备，其特征在于，包括处理器和存储器，所述处理器和所述存储器相互连接，其中，所述存储器用于存储计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，执行如权利要求1-6任一项所述的方法。

9.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行以实现如权利要求1-6任一项所述的方法。

数据挖掘方法、装置、计算机设备及存储介质

技术领域

[0001] 本申请涉及计算机技术领域，尤其涉及大数据技术领域，涉及一种数据挖掘方法、装置、计算机设备及存储介质。

背景技术

[0002] 知识图谱是目前大数据和人工智能领域的热门研究方向，因为它不光能以可视化的形式展现事物之间的联系，同时它包含了许多技术的应用，例如图论、数据库技术、可视化、数据挖掘、深度学习等。

[0003] 知识图谱在企业或机构的应用一般是以集合了数据挖掘、实体识别、实体关联等技术的系统形式展现的。当知识图谱技术需要应用在一个需要主动挖掘数据的场景时，整个流程的自动化程度和信息的准确度将会成为该系统表现的一个重要考量；针对自动化程度这一议题，不同的行业企业或团队针对其业务都有自身的解决方案，例如社交领域的知识图谱有持续的流数据输入，数据采集本身是自动化的，业务分析模型主要负责标识实体的属性和实体间的关系。

[0004] 但是对于需要主动向外挖掘数据进行扩张的，知识上有一定专业壁垒，对于大众陌生的知识图谱应用领域，例如政治，或是金融、生物学等纵深领域的知识梳理场景，往往需要有一定专业背景的人员参与机型识别效果的验证以及挖掘策略的制定，然而，这些过程因为知识体系的不同导致数据挖掘过程十分困难。因此，如何基于不同知识体系有效地进行数据挖掘成为亟待解决的问题。

发明内容

[0005] 本申请实施了提供了一种数据挖掘方法、装置、计算机设备及存储介质，可以基于不同知识体系有效地进行数据挖掘。

[0006] 第一方面，本申请实施了提供了一种数据挖掘方法，包括：

[0007] 获取执行本次数据挖掘操作后的数据挖掘结果；

[0008] 根据所述数据挖掘结果生成知识图谱；

[0009] 确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果；

[0010] 根据所述排序结果执行下一次数据挖掘操作。

[0011] 可选的，所述确定所述知识图谱包括的多个节点中每个节点的中心度，包括：

[0012] 确定所述知识图谱包括的多个节点中每个节点所在的最短路径的数量，并确定所述多个节点之间的最短路径的数量；

[0013] 根据所述每个节点所在的最短路径的数量以及所述多个节点之间的最短路径的数量，确定所述每个节点的间接中心度以作为所述每个节点的中心度。

[0014] 可选的，所述确定所述知识图谱包括的多个节点中每个节点的中心度，包括：

[0015] 确定所述多个节点中每个节点所连接的目标属性的节点的数量；

[0016] 根据所述每个节点所连接的目标属性的节点的数量，确定所述每个节点的度中心度以作为所述每个节点的中心度。

[0017] 可选的，所述确定所述知识图谱包括的多个节点中每个节点的中心度，包括：

[0018] 确定所述多个节点中每个节点所连接的目标属性的节点的数量，并确定所述多个节点之间的目标路径的数量；

[0019] 根据所述每个节点所连接的目标属性的节点的数量以及所述多个节点之间的目标路径的数量，确定所述多个节点的度中心度以作为所述每个节点的中心度。

[0020] 可选的，所述根据所述排序结果执行下一次数据挖掘操作，包括：

[0021] 确定已执行数据挖掘操作的次数；

[0022] 在已执行数据挖掘操作的次数小于或等于预设次数时，根据所述排序结果确定针对所述每个节点的挖掘优先级；

[0023] 根据所述每个节点的挖掘优先级执行下一次数据挖掘操作。

[0024] 可选的，所述根据所述排序结果执行下一次数据挖掘操作，还包括：

[0025] 在已执行数据挖掘操作的次数大于预设次数时，计算所述每个节点的度增益；

[0026] 根据所述每个节点的度增益，从所述多个节点中确定出度增益大于或等于预设数值的目标节点；

[0027] 在所述目标节点为多个时，根据所述排序结果以及每个所述目标节点的度增益确定针对每个目标节点的挖掘优先级；

[0028] 根据所述每个目标节点的挖掘优先级执行数据挖掘操作。

[0029] 可选的，所述计算所述每个节点的度增益，包括：

[0030] 获取执行上一次数据挖掘操作后得到的所述每个节点的中心度；

[0031] 根据所述执行上一次数据挖掘操作后得到的所述每个节点的中心度，以及执行本次数据挖掘操作后得到的所述每个节点的中心度，计算所述每个节点的度增益。

[0032] 第二方面，本申请实施例提供了一种数据挖掘装置，包括：

[0033] 数据挖掘模块，用于获取执行本次数据挖掘操作后的数据挖掘结果；

[0034] 生成模块，用于根据所述数据挖掘结果生成知识图谱；

[0035] 排序模块，用于确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果；

[0036] 所述数据挖掘模块，还用于根据所述排序结果执行下一次数据挖掘操作。

[0037] 第三方面，本申请实施例提供了一种计算机设备，包括处理器和存储器，所述处理器和所述存储器相互连接，其中，所述存储器用于存储计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，执行如第一方面所述的方法。

[0038] 第四方面，本申请实施例提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行以实现如第一方面所述的方法。

[0039] 综上所述，计算机设备可以获取执行本次数据挖掘操作后的数据挖掘结果，并根据数据挖掘结果生成知识图谱以确定知识图谱包括的多个节点中每个节点的中心度，并根据每个节点的中心度对每个节点进行排序，得到排序结果，从而根据排序结果执行下一次数据挖掘操作，本申请能够基于不同知识体系有效地进行数据挖掘，在面对大众陌生的知识图谱应用领域等领域时，本申请无需如现有技术般需要有一定专业背景的人员参与机型

识别效果的验证以及挖掘策略的制定，因此在数据挖掘效率有较大的提升，另外本申请通过分析节点的中心度，然后根据节点的中心度来进行数据挖掘的方式，也使得数据挖掘质量有较大的保障。

附图说明

[0040] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术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申请的一些实施例，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他的附图。

[0041] 图1a是本申请实施例提供的一种数据挖掘流程示意图；

[0042] 图1b是本申请实施例提供的另一种数据挖掘流程的示意图；

[0043] 图1c是本申请实施例提供的一种数据挖掘情景示意图；

[0044] 图2是本申请实施例提供的一种数据挖掘方法的流程示意图；

[0045] 图3是本申请实施例提供的一种知识图谱的示意图；

[0046] 图4是本申请实施例提供的另一种数据挖掘方法的流程示意图；

[0047] 图5是本申请实施例提供的一种数据挖掘装置的结构示意图；

[0048] 图6是本申请实施例提供的一种计算机设备的结构示意图。

具体实施方式

[0049] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行描述。

[0050] 数据挖掘，也可以叫数据开采，数据采掘等，是按照既定的业务目标从海量数据中提取出潜在、有效的信息的处理过程.在浅层次上，它利用现有数据库管理系统等数据源管理系统的查询、检索及报表功能，与多维分析、统计分析方法相结合，进行联机分析处理，从而得出可供决策参考的统计分析数据.在深层次上，它可以从数据库等数据源中发现隐含的、先前未知的具有潜在价值的信息。

[0051] 数据挖掘是一个多学科领域，它融合了数据库技术、人工智能、机器学习、模式识别、模糊数学和数理统计等最新技术的研究成果，可以用来支持商业智能应用和决策分析，目前广泛应用于金融、医疗等行业。数据挖掘技术的发展对于各行各业来说，都具有重要的现实意义。

[0052] 其中，一个结合知识图谱的简单的数据挖掘流程可以参见图1a。图1a所示的数据挖掘流程包括如下步骤。

[0053] 1、数据挖掘。该过程为由业务驱动或是知识驱动的数据收集步骤，通常需要对业务或知识领域熟悉的专业人士来制定数据挖掘策略。

[0054] 2、实体识别。该过程通过自然语言处理、图像识别、声纹识别的算法对文本、图像或声音形式的数据进行分析，挖掘其中的目标实体，通常实体识别模型需要丰富的语料等训练数据以及后期的调优才能达到良好的识别效果。

[0055] 3、生成知识图谱。将实体和关系以节点和连接线的形式展现。在一个实施例中，生成知识图谱的过程可以包括实体识别的过程。

[0056] 4、模型效果检视。该过程根据知识图谱的效果对步骤2的实体识别模型和实体识

别策略的效果进行检验。该过程通常需要有在该领域有一定专业积累的人士进行判断。

[0057] 5、模型调优。根据上一步骤的模型效果制定实体识别模型的优化措施。以便后续利用优化的实体识别模型来实体识别，以便后续再获取数据挖掘结果后可以根据优化的实体识别模型获取更为准确的知识图谱，从而根据更为准确的知识图谱来进行数据挖掘。

[0058] 6、数据挖掘。在模型调优后，根据知识所涉及的领域，可能需要对业务或知识领域熟悉的专业人士再制定新一轮的数据挖掘方向或策略。

[0059] 上述过程在针对陌生或纵深的研究领域内，有时难免会需要人工介入，从专业角度出发判断节点集群的丰富程度并制定下一次的挖掘任务，若是知识图谱构建者缺乏该领域的专业知识则会因为难以判断每轮挖掘任务的丰富程度而制定新一轮的数据挖掘方向。为此，本申请提出了一种数据挖掘策略，可以使用图论中衡量节点在网络中的重要性的“中心性”概念，通过中心性概念自动测算表示应用领域中的实体的节点的重要性，并根据节点的重要性进行排序，从而排序结果来开展数据挖掘工作，在一个实施了中，可以将排序结果(通常是排序了的节点的名称或节点的图像)传送至数据挖掘程序，以便数据挖掘程序根据排序结果来开展数据挖掘工作。在一个实施例中，参见图1b所示的数据挖掘流程，在图1b所示的数据挖掘流程中，中心性计算和度数增益计算的过程衔接在“模型调优”之后。在中心度计算步骤中，可自行根据业务场景选择使用度中心性、间接中心性或两者结合的方式分析已有节点的中心性。度中心性相关公式适用于定位主题性最高或影响度最广的节点并推进挖掘工作，间接中心性相关公式适用于定位路径流量最高的节点；或者，还可以根据实际应用场景自行设定两种中心性计算结果的权重结合使用。

[0060] 在一个实施例中，结合图1c来阐述所述的数据挖掘策略，计算机设备可以遍历知识图谱中多个节点，如所有节点来计算每个节点的中心度，而后根据每个节点的中心度对每个节点进行排序，得到排序结果，此处的排序结果可以为节点列表。在第一次执行数据挖掘操作时，可以根据排序结果确定每个节点的挖掘优先级，根据每个节点的挖掘优先级执行数据挖掘工作。由于在第一次进行数据挖掘时，度增益是无法计算的，因此可以通过数据挖掘程序直接读取排序结果启动数据挖掘任务。在第N(大于1)次执行数据挖掘操作时，可以计算每个节点的度增益，并根据每个节点的度增益执行数据挖掘工作，在这个过程中，具体可以确定度增益大于0的各节点的挖掘优先级，然后根据度增益大于0的各节点的挖掘优先级执行数据挖掘工作。

[0061] 请参阅图2，为本申请实施例提供的一种数据挖掘方法的流程示意图。该方法可以应用于计算机设备。计算机设备可以为服务器等设备。服务器可以一个服务器或服务器集群。具体地，该方法可以包括以下步骤：

[0062] S201、获取执行本次数据挖掘操作后的数据挖掘结果。

[0063] 本申请实施例中，计算机设备可以执行本次数据挖掘操作，得到执行本次数据挖掘操作后的数据挖掘结果。本次数据挖掘操作可以是第一次的数据挖掘操作，也可以是非第一次的数据挖掘操作。本次数据挖掘操作为非第一次的数据挖掘操作，表明在本次数据挖掘之前已经执行过数据挖掘操作。数据挖掘结果指经数据挖掘得到的数据。

[0064] S202、根据所述数据挖掘结果生成知识图谱。

[0065] 本申请实施例中，计算机设备可以对数据挖掘结果进行知识抽取，得到多个三元组，并对多个三元组进行知识融合，得到知识融合结果。计算机设备在设备知识融合结果

后，可以对知识融合结果进行知识加工，得到知识图谱。其中，知识抽取的过程包括实体抽取、关系抽取和属性抽取。知识融合的过程包括本体匹配和实体对齐。知识加工包括知识推理、知识发现和质量评估。其中，所述的实体抽取的过程可以包括上述实体识别的过程。在一个实施例中，该实体识别的过程可以是经由模型检视和模型调优后得到的优化后的实体识别模型来实现的。

[0066] S203、确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果。

[0067] 其中，所述的多个节点可以是知识图谱包括的所有节点，也可以是知识图谱包括的部分节点。在一个实施例中，知识图谱包括的各节点可以被划分至各节点集群。节点集群可以根据业务目标划分，在此不做赘述。相应地，所述的多个节点可以是目标节点集群包括的所有节点，也可以是目标节点集群包括的部分节点。排序结果指示了排序后的每个节点。其中，排序方式可以为按照中心度由大到小将每个节点由前到后排序，或按照中心度由小到大将每个节点由前到后排序，等等。

[0068] 本申请实施例中，计算机设备可以调取中心性算法来确定知识图谱包括的多个节点中每个节点的中心度，并根据每个节点的中心度对每个节点进行排序，得到排序结果。其中，所述的中心性算法可以包括间接中心性算法和/或度中心性算法等中心性算法。下面依次对两种算法进行讲解。

[0069] 其中，间接中心性，也可以叫中介中心性。经由间接中心性算法确定出的中心度可以称之为间接中心度。间接中心度可以用于表征节点的间接中心性。某节点的间接中间性高，说明该节点在已探索的网络结构中的“中介属性”强，该节点的持续挖掘价值在于找出其它使用到其“中介能力”的节点。

[0070] 在一个实施例中，计算机设备可以调用间接中心性算法确定知识图谱包括的多个节点中每个节点的中心度。具体地，计算机设备可以确定知识图谱包括的多个节点中每个节点所在的最短路径的数量，并确定多个节点之间的最短路径的数量，然后根据每个节点所在的最短路径的数量以及多个节点之间的最短路径的数量，确定每个节点的间接中心度作为每个节点的中心度。其中，所述的间接中心性算法如下：

[0071] $$BC(v)=\sum\frac{dst(v)}{dst}$$ 公式1.1；

[0072] 其中，BC为间接中心度。dst()是从其它节点s到标的节点t的最短路径中经过节点v的数量。dst表示从其它节点s到标的节点t的最短路径的数量。其中，标的节点t为所述的多个节点中的节点，其它节点s为所述多个节点中除标的节点t之外的节点。公式1.1中的节点v为除标的节点t和其它节点s之外的节点。

[0073] 举例来说，在企业风控和信息公示的场景下，知识图谱往往需要呈现企业工商信息、高管信息、企业或高管的持股情况。参见图3所示的知识图谱，图3所示的知识图谱包括多个节点，节点包括公司节点或高管节点，节点的属性为公司名称或人物名称，边的属性为公司与公司之间的关系、人物与公司之间的关系或人物与人物之间的关系。图3所示的知识图谱可以被划分为公司1节点所在节点集群1以及公司5节点所在节点集群2。其中，公司2、公司3和公司4均为公司1的子公司。公司6、公司7、公司8均为公司5的子公司。在此示例中，中介性高可被定义为持有股份多，反之，中介性低可被定义为持有股份少。下面说明计算机

设备如何调用间接中心算法确定节点集群1中各节点的间接中心度。

[0074] 计算机设备将节点集群1中各节点作为标的节点t，并将节点集群1中除标的节点之外的节点作为其它节点s。为了计算这些节点的间接中心度，需要统计其它节点s通往标的节点t的最短路径的数量(即dst)，在本示例中具体可统计其它节点s通往标的节点t的边属性包括“持股”或“股东”的最短路径的数量，并且还需要统计节点集群1中各节点间的最短路径的数量(即dst())，在本示例中具体可统计节点集群1中各节点间通过某中介节点的边的属性包括“持股”或“股东”的最短路径的数量。基于上述步骤，可以梳理出如下两个统计表：

[0075] 表1:最短路径穷举

[0076]

节点对	最短路径	最短路径的中介节点（最短路径经过的节点）
公司4-公司1	公司4-公司2-公司1	公司2
公司3-公司1	公司3-公司2-公司1	公司2
公司4-高管1	公司4-公司2-公司1-高管1	公司2、公司1
公司4-高管2	公司4-公司2-公司1-高管2	公司2、公司1
公司2-高管1	公司2-公司1-高管1	公司1
公司2-高管2	公司2-公司1-高管2	公司1
公司3-高管1	公司3-公司2-公司1-高管1	公司2、公司1
公司3-高管2	公司3-公司2-公司1-高管2	公司2、公司1

[0077] 表2:间接中心度统计

[0078]

公司节点	dst()	dst	BC
公司1	6	8	0.75

公司2	6	8	0.75
公司3	0	8	0
公司4	0	8	0
高管1	0	8	0
高管2	0	8	0

[0079] 其中,表1穷举了由标的节点t和其它节点s构成的节点对,以及每个节点对的最短路径,以及每个节点对的最短路径的中介节点。表2罗列了各节点在其它节点s通往标的节点t的最短路径的数量,以及各节点间的最短路径的数量,简单将这些数值代入公式1.1即可算出节点集群1中各节点的间接中心度,参见表2的最后一列数据。公司2由于持股情况复杂,所以中介性最强,即间接中心度最高。以公司2节点为例,将它的dst()的数值以及dst的数值代入公式1.1,可计算得到公司2节点的间接中心度,如下。

[0080] $$BC(公司2)=\sum\frac{dst(公司2)}{dst}=\frac{6}{8}=0.75$$

[0081] 其中,经由度中心性算法确定出的中心度可以为度中心度。度中心度可以用于表征节点的度中心性。在一个实施例中,关于度中心度用于表征节点的度中心性可以有以下两个层面的含义:一种是度中心度用于表征节点自身的度中心性,另一种是度中心度用于表征节点所在多个节点的度中心性。某节点的度中心性高,则说明该节点在已探索的网络结构中的“关系繁荣性“强,该节点的持续挖掘价值在于从多个方向或多个维度发散性的找出于其有关联性的其它节点。

[0082] 在一个实施例中,计算机设备可以调用度中心性算法确定知识图谱包括的多个节点中每个节点的中心度,其中一种方式为,计算机设备可以确定多个节点中每个节点所连接的目标属性的节点的数量,并根据每个节点所连接的目标属性的节点的数量,确定每个节点的度中心度作为每个节点的中心度。其中,目标属性可以为多个节点属性中的任一节点属性或指定节点属性。其中,所述的度中心性算法如下:

[0083] DC(v)=deg(v) 公式1.2;

[0084] 其中,DC为度中心度。节点v在公式1.2中可表示多个节点中的任一节点。公式1.2中的deg()表示与节点v连接的满足指定条件的节点的数量。

[0085] 举例来说,在金融分析的过程中,有时需要快速定位资源广、规模大的企业。针对某个市场或行业,这种通过关系网络判断企业资源是否广泛或关系规模是否庞大的场景,可以通过计算企业在知识图谱中的度中心度实现。在本示例中,图3所示的知识图谱还可以包括行业A节点。在需要确定行业A中资源广、关系规模大的企业时,可以计算图3所示的知识图谱中属于行业A的各公司节点的度中心度,并依据属于行业A的各公司节点的度中心度确定属于行业A的各公司的资源广泛程度、关系规模庞大程度。具体地,计算机设备可以确定多个节点中每个节点连接的满足指定条件的节点的数量,并根据每个节点连接的满足指定条件的节点的数量,确定每个节点的度中心度。例如,计算机设备可以将每个节点连接的为目标属性(如公司或高管)的节点确定为每个节点连接的满足指定条件的节点,再如,计算机设备可以确定每个节点连接的为指定属性(如就职或持股)的边,然后将每个节点连接的为指定属性的边所连接的节点确定为每个节点连接的满足指定条件的节点。计算机设备可以调取度中心性算法计算出公司1节点的度中心度为3,并计算出公司5节点的度中心度

为5。公司5节点的度中心度高于公司1节点的度中心度，说明公司5的资源要比公司1广、公司5的关系规模要比公司2大。

[0086] 在一个实施例中，计算机设备调用度中心性算法确定知识图谱包括的多个节点中每个节点的中心度，另一种方式为计算机设备确定多个节点中每个节点所连接的目标属性的节点的数量，并确定多个节点之间的目标路径的数量，并根据每个节点所连接的目标属性的节点的数量以及多个节点之间的目标路径的数量，确定多个节点的度中心度以作为每个节点的中心度。在一个实施例中，计算机设备可以根据多个节点中每个节点所连接的目标属性的节点的数量，确定每个节点的度中心度，然后根据每个节点的度中心度以及多个节点之间的最短路径的数量，确定多个节点的度中心度以作为每个节点的度中心度。其中，此处所述的度中心度算法如下所示：

[0087] $$DC = \frac{\sum_{i=1}^{V}\left(DC(n^*)\right) - DC(v_i))}{(V-1)(V-2)}$$

[0088] 其中DC表示多个节点的度中心度。n*表示多个节点中度中心度最高的节点，而DC(n*)为n*的度中心度。DC(v_i)为多个节点中其它节点v_i的度中心度。(V-1)(V-2)表示最大可能相连情况，V可以为多个节点间的目标路径的数量。在一个实施例中，此处的目标路径可以包括n*与n*连接的节点之间的路径。在一个实施例中，V可以理解为n*的最大连接数。在一个实施例中，V可以为多个节点的数量。即，计算机设备确定多个节点中每个节点所连接的目标属性的节点的数量，并确定多个节点的数量，根据每个节点所连接的目标属性的节点的数量以及多个节点的数量，确定多个节点的度中心度以作为每个节点的中心度。

[0089] 在企业分析中的一些特定场景，例如规模测算或行业集中性测算，会需要统计某个行业中各个头部企业占有市场规模的比重。在此场景下，可以通过集群度中心性的视角，统计由某些节点构成的节点集群在整个关系网络中占有的中心性，并结合节点所映射的实体属性来判断其对网络的影响性。以图3为例，假设需要判断公司5节点对行业A节点的规模性影响，则需要计算节点集群2的度中心度。具体地，计算机设备可以确定节点集群2中的n*为公司5节点，并分别统计公司5节点的度中心度以及节点集群2中的除公司5节点之外的其它公司节点的度中心度。在本示例中，若需要分析行业下各集团的公司总数规模及其在该行业中的突出性，则可以不用考虑高管节点的作用，只考虑公司节点间的联系。同时，在此有向图中，可以将行业A节点连至公司5节点的路径考虑在内。至此，公司5节点的度中心度DC(n*)为3、公司6节点、公司7节点、公司8节点的度中心度DC(v_i)均为1，公司5节点的最大连接数V，为公司5节点的出度数以及来自行业节点的入度数之和，即为4。将这些数值代入公式1.3，可算得节点集群2的度中心度，如下。

[0090] $$DC = \frac{\sum_{i=1}^{v}\left(DC(n^*)\right) - DC(v_i))}{(V-1)(V-2)} = \frac{(3-1)+(3-1)+(3-1)}{(4-1)(4-2)} = \frac{6}{6} = 1$$

[0091] 节点集群2的度中心度，可以作为节点集群2中各公司节点的中心度。

[0092] 在一个实施例中，计算机设备可以得到多个节点中每个节点的度中心度，以及多个节点中每个节点的间接中心度。在得到每个节点的度中心度以及每个节点的间接中心度后，计算机设备可以将每个节点的度中心度分别乘以第一权重，得到每个节点对应的第一权重结果；并将每个节点的间接中心度分别乘以第二权重，得到每个节点对应的第二权重

结果;计算机设备将每个节点对应的第一权重结果分别与每个节点对应的第二权重结果相加,得到每个节点对应的中心度。

[0093] S204、根据所述排序结果执行下一次数据挖掘操作。

[0094] 计算机设备可以根据排序结果确定每个节点的挖掘优先级,按照每个节点的挖掘优先级执行下一次数据挖掘操作。例如,计算机设备可以对优先级高的节点优先执行数据挖掘操作,对优先级低的节点靠后执行数据挖掘操作。在每个节点的中心度为多个节点的度中心度时,每个节点具有相同的挖掘优先级,计算机设备可以同时对每个节点执行下一次数据挖掘操作。

[0095] 可见,图2所示的实施例中,计算机设备可以获取执行本次数据挖掘操作后的数据挖掘结果,并根据数据挖掘结果生成知识图谱以确定知识图谱包括的多个节点中每个节点的中心度,并根据每个节点的中心度对每个节点进行排序,得到排序结果,从而根据排序结果执行下一次数据挖掘操作,该过程能够给予不同知识体系有效地进行数据挖掘,不仅可以提升数据挖掘效率还可以保证数据挖掘质量。

[0096] 请参阅图4,为本申请实施例提供的另一种数据挖掘方法的流程示意图。该方法可以应用于计算机设备中。计算机设备可以为服务器等设备。服务器可以一个服务器或服务器集群。具体地,该方法可以包括以下步骤:

[0097] S401、获取执行本次数据挖掘操作后的数据挖掘结果。

[0098] S402、根据所述数据挖掘结果生成知识图谱。

[0099] S403、确定所述知识图谱包括的多个节点中每个节点的中心度,并根据所述每个节点的中心度对所述每个节点进行排序,得到排序结果。

[0100] 其中,步骤S401-步骤S403可以参见图2实施例中的步骤S201-步骤S203,在此不做赘述。

[0101] S404、确定已执行数据挖掘操作的次数。

[0102] 其中,计算机设备在执行步骤S404后,根据已执行数据挖掘操作的次数判断执行步骤S405还是步骤S407。

[0103] S405、在已执行数据挖掘操作的次数小于或等于预设次数时,根据所述排序结果确定针对所述每个节点的挖掘优先级。

[0104] S406、根据所述每个节点的挖掘优先级执行下一次数据挖掘操作。

[0105] 在步骤S405-步骤S406中,计算机设备在已执行数据挖掘操作的次数小于或等于预设次数时,根据所述排序结果确定针对所述每个节点的挖掘优先级,并根据所述每个节点的挖掘优先级执行下一次数据挖掘操作。

[0106] S407、在已执行数据挖掘操作的次数大于预设次数时,计算所述每个节点的度增益。

[0107] S408、根据所述每个节点的度增益,从所述多个节点中确定出度增益大于或等于预设数值的目标节点。

[0108] S409、在所述目标节点为多个时,根据所述排序结果以及每个所述目标节点的度增益确定针对每个目标节点的挖掘优先级。

[0109] S410、根据所述每个目标节点的挖掘优先级执行数据挖掘操作。

[0110] 在步骤S406-步骤S410中,计算机设备可以在已执行数据挖掘操作的次数大于预

设次数时，计算该每个节点的度增益，并根据该排序结果和该每个节点的度增益执行下一次数据挖掘操作。在一个实施例中，计算机设备可以从多个节点中确定出度增益大于或等于预设数值(如0)的目标节点，并可以在该目标节点为多个时，根据每个该目标节点的度增益确定针对每个目标节点的挖掘优先级，从而根据该每个目标节点的挖掘优先级执行数据挖掘操作。在一个实施例中，计算机设备可以根据排序结果以及该目标节点的度增益确定针对每个目标节点的挖掘优先级。其中，所述预设次数可以设置为1等次数。在一个实施例中，度增益高的节点的挖掘优先级高，度增益低的节点的挖掘优先级低。

[0111] 为了保持数据挖掘工作的效率，需要一套标准来判断每一轮的数据挖掘相较于上一轮是否有更好的拓展以及某些节点是否已经停止了生长，本申请实施例可以将这套标准成为“度增益”。例如，计算机设备可以在确定已执行数据挖掘操作的次数大于1时，计算每个节点的度增益。计算机设备根据每个节点的度增益从多个节点中确定出度增益大于或等于预设数值的目标节点，并在目标节点为多个时，根据排序结果以及每个目标节点的度增益确定每个目标节点的优先级，从而根据每个目标节点的优先级执行数据挖掘操作。

[0112] 在一个实施例中，计算机设备计算所述每个节点的度增益的方式具体为计算机设备获取执行上一次数据挖掘操作后得到的所述每个节点的中心度，并根据所述执行上一次数据挖掘操作后得到的所述每个节点的中心度，以及执行本次数据挖掘操作后得到的所述每个节点的中心度，计算所述每个节点的度增益。

[0113] 其中，度增益的计算方式可以如下：

[0114] $$D = \frac{v_i - v_{i-1}}{v_{i-1}} \times 100\%$$

[0115] 其中，D表示度增益。i表示本次执行数据挖掘操作后，已执行数据挖掘操作的次数。v_i为节点在本次执行数据挖掘操作后得到的中心度，V_{i-1}为节点在上一次执行数据挖掘后得到的中心度。随着挖掘次数的增加，度增益通常是呈先上升后下降趋近于零的趋势，当增益为零时可以停止对该节点的挖掘。

[0116] 可见，图4所示的实施例中，数据挖掘装置可以获取执行本次数据挖掘操作后的数据挖掘结果，并根据数据挖掘结果生成知识图谱以确定知识图谱包括的多个节点中每个节点的中心度，并根据每个节点的中心度对每个节点进行排序，得到排序结果，从而根据排序结果执行下一次数据挖掘操作，该过程能够提升数据挖掘效率并保障数据挖掘质量。

[0117] 本申请涉及区块链技术，如可将数据挖掘结果写入区块链中，或基于区块链存储的数据来执行不同轮次的数据挖掘操作。

[0118] 请参阅图5，为本申请实施例提供的一种数据挖掘装置的结构示意图。具体地，该装置可以应用于计算机设备，具体地，该装置可以包括：

[0119] 数据挖掘模块501，用于获取执行本次数据挖掘操作后的数据挖掘结果。

[0120] 生成模块502，用于根据所述数据挖掘结果生成知识图谱。

[0121] 排序模块503，用于确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果。

[0122] 数据挖掘模块501，还用于根据所述排序结果执行下一次数据挖掘操作。

[0123] 在一种可选的实施方式中，排序模块503确定所述知识图谱包括的多个节点中每个节点的中心度，具体为确定所述知识图谱包括的多个节点中每个节点所在的最短路径的

数量,并确定所述多个节点之间的最短路径的数量;根据所述每个节点所在的最短路径的数量以及所述多个节点之间的最短路径的数量,确定所述每个节点的间接中心度以作为所述每个节点的中心度。

[0124] 在一种可选的实施方式中,排序模块503确定所述知识图谱包括的多个节点中每个节点的中心度,具体为确定所述多个节点中每个节点所连接的目标属性的节点的数量;根据所述每个节点所连接的目标属性的节点的数量,确定所述每个节点的度中心度以作为所述每个节点的中心度。

[0125] 在一种可选的实施方式中,排序模块503计算所述知识图谱中各个节点的中心度,具体为确定所述多个节点中每个节点所连接的目标属性的节点的数量,并确定所述多个节点之间的目标路径的数量;根据所述每个节点所连接的目标属性的节点的数量以及所述多个节点之间的目标路径的数量,确定所述多个节点的度中心度以作为所述每个节点的中心度。

[0126] 在一种可选的实施方式中,数据挖掘模块501根据所述排序结果执行下一次数据挖掘操作,具体为确定已执行数据挖掘操作的次数;在已执行数据挖掘操作的次数小于或等于预设次数时,根据所述排序结果确定针对所述每个节点的挖掘优先级;根据所述每个节点的挖掘优先级执行下一次数据挖掘操作。

[0127] 在一种可选的实施方式中,数据挖掘模块501根据所述排序结果执行下一次数据挖掘操作,还具体为在已执行数据挖掘操作的次数大于预设次数时,计算所述每个节点的度增益;根据所述每个节点的度增益,从所述多个节点中确定出度增益大于或等于预设数值的目标节点;在所述目标节点为多个时,根据所述排序结果以及每个所述目标节点的度增益确定针对每个目标节点的挖掘优先级;根据所述每个目标节点的挖掘优先级执行数据挖掘操作。

[0128] 在一种可选的实施方式中,数据挖掘模块501计算所述每个节点的度增益,具体为获取执行上一次数据挖掘操作后得到的所述每个节点的中心度;根据所述执行上一次数据挖掘操作后得到的所述每个节点的中心度,以及执行本次数据挖掘操作后得到的所述每个节点的中心度,计算所述每个节点的度增益。

[0129] 可见,图5所示的实施例中,数据挖掘装置可以获取执行本次数据挖掘操作后的数据挖掘结果,并根据数据挖掘结果生成知识图谱以确定知识图谱包括的多个节点中每个节点的中心度,并根据每个节点的中心度对每个节点进行排序,得到排序结果,从而根据排序结果执行下一次数据挖掘操作,该过程能够给予不同知识体系有效地进行数据挖掘,不仅可以提升数据挖掘效率还可以保证数据挖掘质量。

[0130] 请参阅图6,为本申请实施例提供的一种计算机设备的结构示意图。本实施例中所描述的计算机设备可以包括:一个或多个处理器1000和存储器2000。处理器1000和存储器2000可以通过总线等方式连接。

[0131] 处理器1000可以是中央处理模块(Central Processing Unit,CPU),该处理器还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器

等。

[0132] 存储器2000可以是高速RAM存储器，也可为非不稳定的存储器(non-volatile memory)，例如磁盘存储器。存储器2000用于存储一组程序代码，处理器1000可以调用存储器2000中存储的程序代码。具体地：

[0133] 处理器1000，用于获取执行本次数据挖掘操作后的数据挖掘结果；根据所述数据挖掘结果生成知识图谱；确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果；根据所述排序结果执行下一次数据挖掘操作。

[0134] 在一个实施例中，处理器1000确定所述知识图谱包括的多个节点中每个节点的中心度，具体为确定所述知识图谱包括的多个节点中每个节点所在的最短路径的数量，并确定所述多个节点之间的最短路径的数量；根据所述每个节点所在的最短路径的数量以及所述多个节点之间的最短路径的数量，确定所述每个节点的间接中心度以作为所述每个节点的中心度。

[0135] 在一个实施例中，处理器1000确定所述知识图谱包括的多个节点中每个节点的中心度，具体为确定所述多个节点中每个节点所连接的目标属性的节点的数量；根据所述每个节点所连接的目标属性的节点的数量，确定所述每个节点的度中心度以作为所述每个节点的中心度。

[0136] 在一个实施例中，处理器1000确定所述知识图谱包括的多个节点中每个节点的中心度，具体为确定所述多个节点中每个节点所连接的目标属性的节点的数量，并确定所述多个节点之间的目标路径的数量；根据所述每个节点所连接的目标属性的节点的数量以及所述多个节点之间的目标路径的数量，确定所述多个节点的度中心度以作为所述每个节点的中心度。

[0137] 在一个实施例中，处理器1000根据所述排序结果执行下一次数据挖掘操作，具体为确定已执行数据挖掘操作的次数；在已执行数据挖掘操作的次数小于或等于预设次数时，根据所述排序结果确定针对所述每个节点的挖掘优先级；根据所述每个节点的挖掘优先级执行下一次数据挖掘操作。

[0138] 在一个实施例中，处理器1000根据所述排序结果执行下一次数据挖掘操作，还具体为在已执行数据挖掘操作的次数大于预设次数时，计算所述每个节点的度增益；根据所述每个节点的度增益，从所述多个节点中确定出度增益大于或等于预设数值的目标节点；在所述目标节点为多个时，根据所述排序结果以及每个所述目标节点的度增益确定针对每个目标节点的挖掘优先级；根据所述每个目标节点的挖掘优先级执行数据挖掘操作。

[0139] 在一个实施例中，处理器1000计算所述每个节点的度增益，具体为获取执行上一次数据挖掘操作后得到的所述每个节点的中心度；根据所述执行上一次数据挖掘操作后得到的所述每个节点的中心度，以及执行本次数据挖掘操作后得到的所述每个节点的中心度，计算所述每个节点的度增益。

[0140] 具体实现中，本申请实施例中所描述的处理器1000可执行图2实施例、图4实施例所描述的实现方式，也可执行本申请实施例所描述的实现方式，在此不再赘述。

[0141] 在本申请各个实施例中的各功能模块可以集成在一个处理模块中，也可以是各个模块单独物理存在，也可以是两个或两个以上模块集成在一个模块中。上述集成的模块既

可以采样硬件的形式实现，也可以采样软件功能模块的形式实现。

[0142] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的计算机可读存储介质可为易失性的或非易失性的。例如，该计算机存储介质可以为磁碟、光盘、只读存储记忆体(Read-Only Memory，ROM)或随机存储记忆体(Random Access Memory，RAM)等。所述的计算机可读存储介质可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序等；存储数据区可存储根据区块链节点的使用所创建的数据等。

[0143] 其中，本申请所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain)，本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0144] 以上所揭露的仅为本申请一种较佳实施例而已，当然不能以此来限定本申请之权利范围，本领域普通技术人员可以理解实现上述实施例的全部或部分流程，并依本申请权利要求所作的等同变化，仍属于本申请所涵盖的范围。

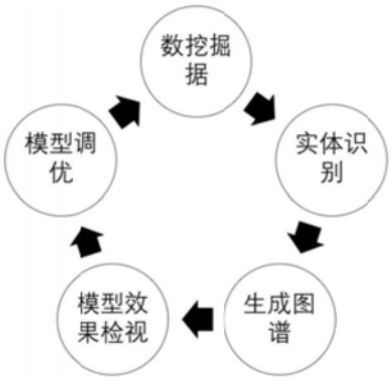

数挖掘据
实体识别
生成图谱
模型效果检视
模型调优

图1a

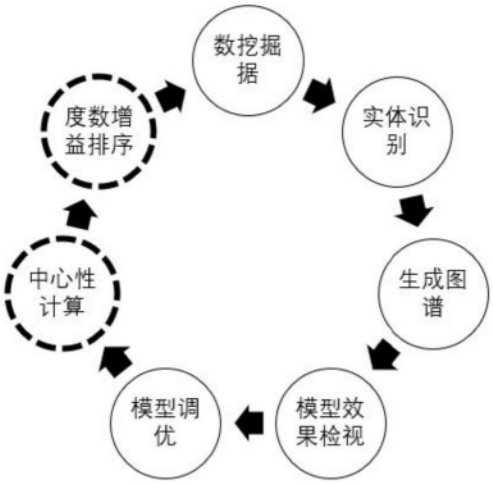

数挖掘据
实体识别
生成图谱
模型效果检视
模型调优
中心性计算
度数增益排序

图1b

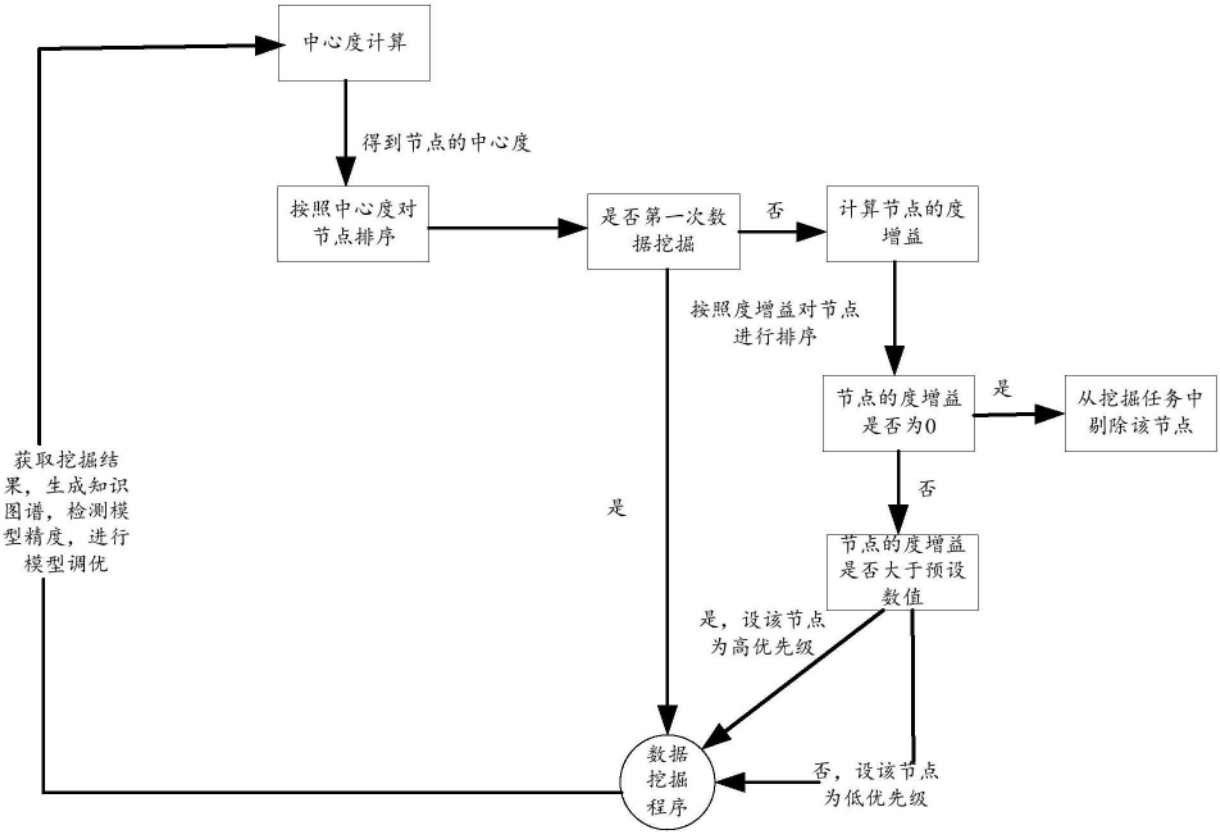

中心度计算
得到节点的中心度
按照中心度对节点排序
是否第一次数据挖掘
否
计算节点的度增益
按照度增益对节点进行排序
节点的度增益是否为0
是
从挖掘任务中剔除该节点
否
节点的度增益是否大于预设数值
是，设该节点为高优先级
否，设该节点为低优先级
是
数据挖掘程序
获取挖掘结果，生成知识图谱，检测模型精度，进行模型调优

图1c

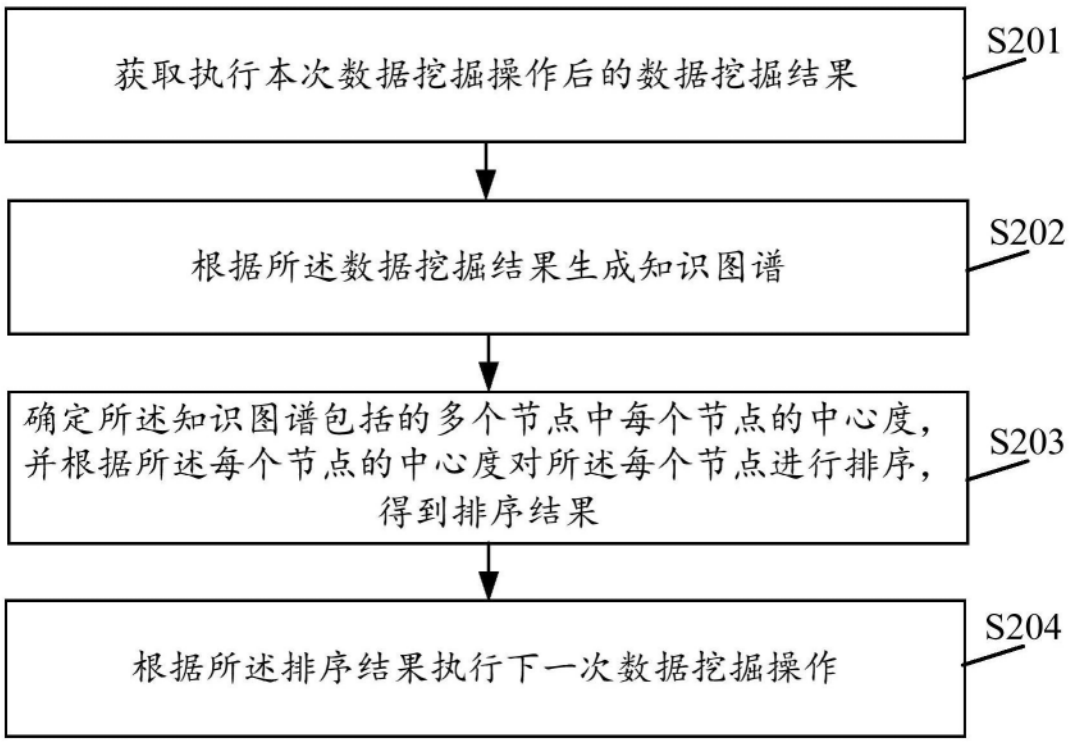

获取执行本次数据挖掘操作后的数据挖掘结果
S201
根据所述数据挖掘结果生成知识图谱
S202
确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果
S203
根据所述排序结果执行下一次数据挖掘操作
S204

图2

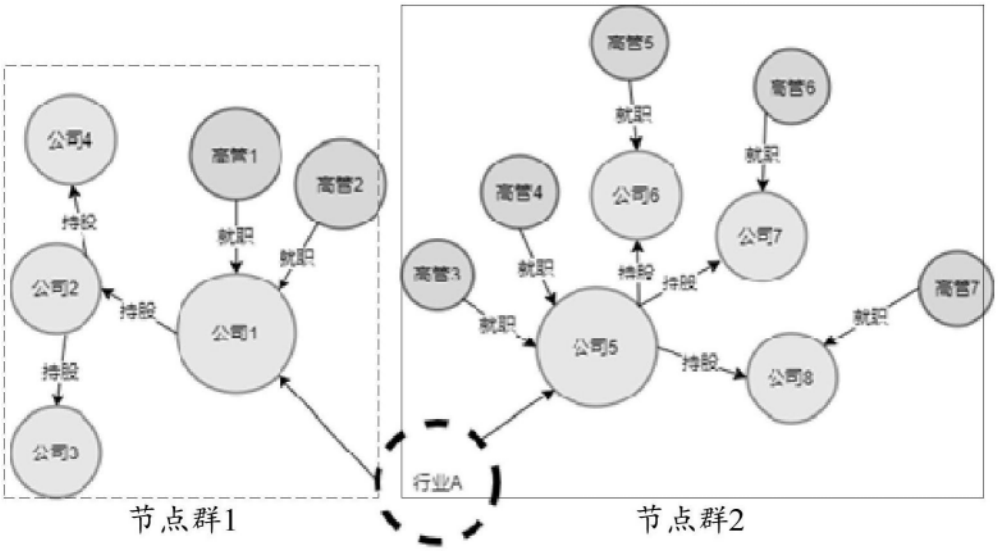

公司4
高管1
高管2
持股
就职
就职
公司2
持股
公司1
持股
公司3
节点群1
高管5
高管6
就职
就职
高管4
公司6
公司7
高管3
就职
持股
持股
高管7
就职
公司5
持股
公司8
行业A
节点群2

图3

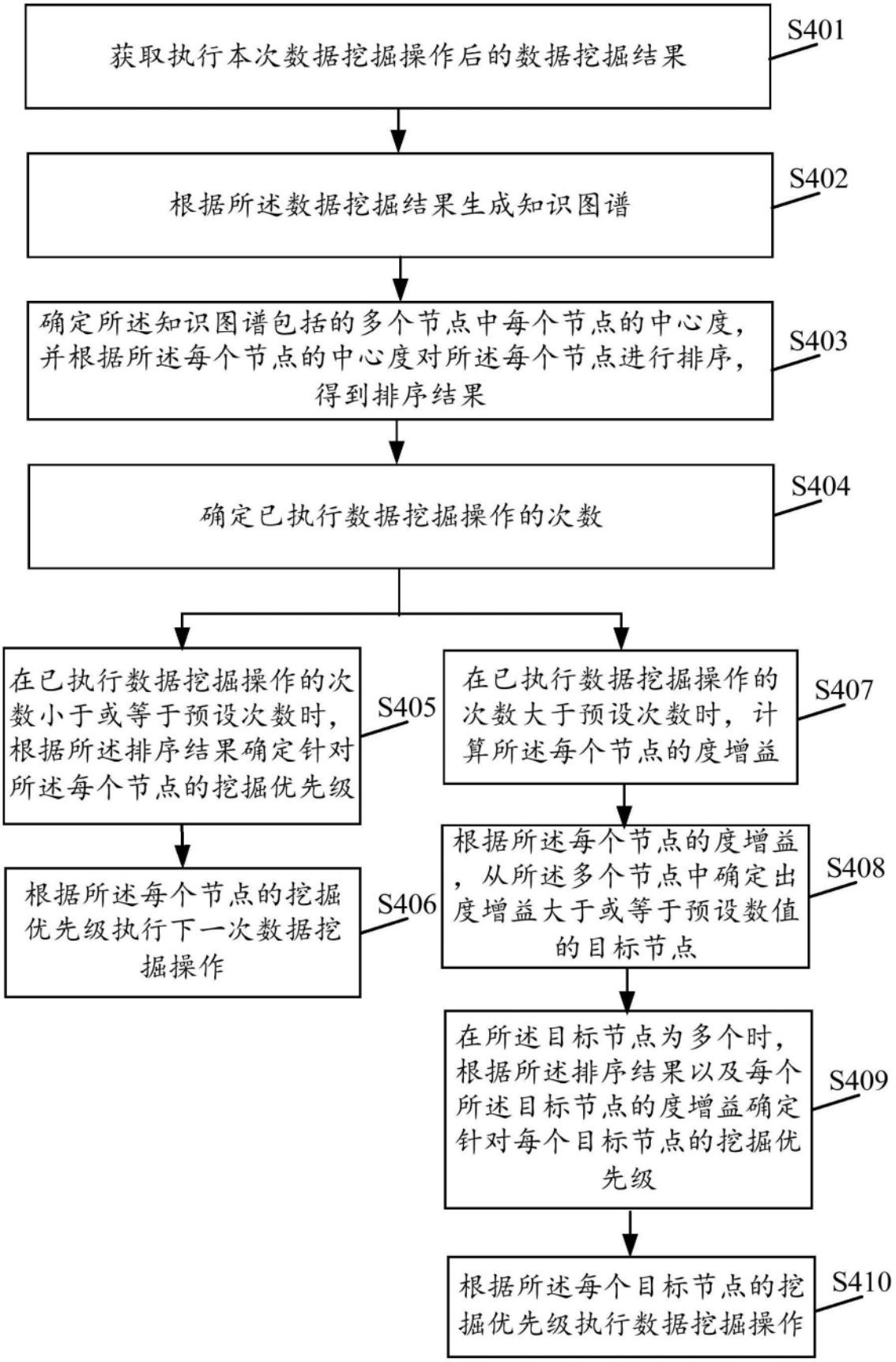

获取执行本次数据挖掘操作后的数据挖掘结果
S401
根据所述数据挖掘结果生成知识图谱
S402
确定所述知识图谱包括的多个节点中每个节点的中心度，并根据所述每个节点的中心度对所述每个节点进行排序，得到排序结果
S403
确定已执行数据挖掘操作的次数
S404
在已执行数据挖掘操作的次数小于或等于预设次数时，根据所述排序结果确定针对所述每个节点的挖掘优先级
S405
根据所述每个节点的挖掘优先级执行下一次数据挖掘操作
S406
在已执行数据挖掘操作的次数大于预设次数时，计算所述每个节点的度增益
S407
根据所述每个节点的度增益，从所述多个节点中确定出度增益大于或等于预设数值的目标节点
S408
在所述目标节点为多个时，根据所述排序结果以及每个所述目标节点的度增益确定针对每个目标节点的挖掘优先级
S409
根据所述每个目标节点的挖掘优先级执行数据挖掘操作
S410

图4

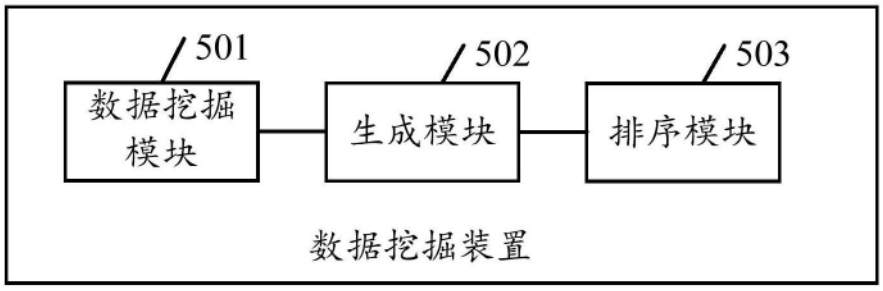


图5

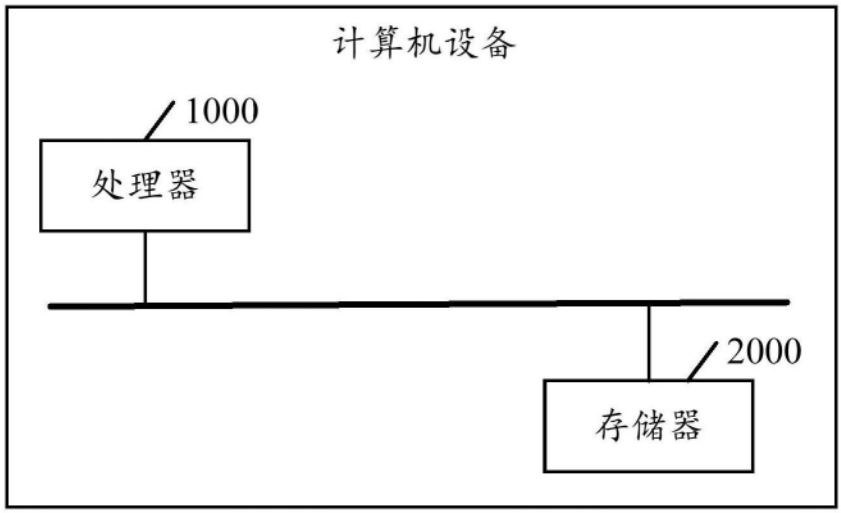


图6