(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 111177150 A
(43)申请公布日 2020.05.19

(21)申请号 201911299296.0

(22)申请日 2019.12.17

(71)申请人 北京明略软件系统有限公司
地址 100084 北京市海淀区中关村东路1号院1号楼10层A1002

(72)发明人 刘鹏飞 耿少华

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 王素燕 栗若木

(51)Int.Cl.
*G06F 16/22*(2019.01)
*G06F 16/245*(2019.01)
*G06F 16/901*(2019.01)
*G06Q 30/00*(2012.01)

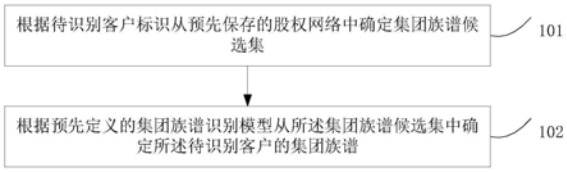
权利要求书2页 说明书9页 附图4页

(54)发明名称

一种识别集团族谱的方法及系统

(57)摘要

本发明实施例公开了一种识别集团族谱的方法及系统，其中该方法包括：根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集；根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。如此，能够从海量股权关系中识别出客户的集团族谱，提升了集团族谱的识别效率。




CN 111177150 A

1.一种识别集团族谱的方法,其特征在于,包括:

根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集;

根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。

2.根据权利要求1所述的方法,其特征在于,

所述股权网络是由点与点之间有向连接构成的点边关系图,其中点关系存储对应客户的属性,点与点之间连接的边关系存储对应关联客户属性及关联关系属性;

所述集团族谱识别模型中的集团类型包括以下至少之一:

两个或两个以上的客户共同被第三方客户所控制共同组成的集团;在股权上或者经营决策上直接或间接控制其他客户或被其他客户控制的客户共同组成的集团;由主要投资者个人、关键管理人员或与其近亲属共同直接控制或间接控制的客户共同组成的集团。

3.根据权利要求1所述的方法,其特征在于,所述根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集,包括:

利用图计算引擎加载预先保存的股权网络;

利用最大连通图算法从所述股份网络中识别出所述待识别客户标识关联的股权子网络,作为所述待识别客户的集团族谱候选集。

4.根据权利要求1所述的方法,其特征在于,所述根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱,包括:

利用图计算算法根据所述集团族谱识别模型遍历所述集团族谱候选集,从中识别出所述待识别客户对应的全部集团族谱。

5.根据权利要求1所述的方法,其特征在于,在根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集之前,该方法还包括:

利用图谱抽取工具从控股数据及亲属关系数据中抽取事先已梳理好的点边关系,形成所述股权网络,并保存到数据库中。

6.根据权利要求3至5任一项所述的方法,其特征在于,

所述图计算引擎为spark graphx图计算引擎,所述最大连通图算法为深度优先图搜索算法,所述图计算算法为基于spark graphx的Pregel,所述图谱抽取工具为hive sql图谱抽取工具,所述数据库为hive数据库。

7.根据权利要求1所述的方法,其特征在于,该方法还包括:

通过图展示工具展示所述待识别实体的集团族谱。

8.一种识别集团族谱的系统,其特征在于,包括:

第一确定单元,用于根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集;

第二确定单元,用于根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。

9.一种识别集团族谱的系统,其特征在于,包括:存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,所述计算机程序被所述处理器执行时实现如权利要求1至7中任一项所述识别集团族谱的方法。

10.一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上存储有信息处

理程序,所述信息处理程序被处理器执行时实现如权利要求1至7中任一项所述识别集团族谱的方法的步骤。

# 一种识别集团族谱的方法及系统

## 技术领域

[0001] 本发明实施例涉及数据挖掘技术，尤指一种识别集团族谱的方法及系统。

## 背景技术

[0002] 各种跨国企业、跨行业企业、跨地区企业越来越多，集团性客户在商业银行中的地位和比重越来越高。相对于单个的企业而言，集团客户的经济实力更为雄厚，诚信度也比单个企业的好些，而且他们的需求更为多样化，可以给银行带来很大的利益。但是集团客户内部关联交易日益复杂，这些跨行业、跨地区经营带来的银企信息不对称等增加了银行授信资产的潜在风险。一旦这些风险暴露，将产生多米诺骨牌效应，会牵涉很多的债权银行，这不仅对银行企业会产生影响，对整个国家的金融系统都会带来巨大的冲击。因此，商业银行必须有效识别集团客户，并以此控制和防范集团客户的信贷风险，加强对集团客户授信业务的风险管理，才能促进各项业务健康和稳健发展。

[0003] 目前为了识别集团族谱，银行基于现有系统及数据采用了以结构化数据库为核心，采用存储过程sql(结构化查询语言，Structured Query Language)实现集团族谱识别方案。该方案的主要思路是采用深度优先搜索算法(Depth First Search，DFS)对股权关系生成的图(Graph)进行遍历，对所有股权关系进行穷尽搜索，直至所有的满足规则的节点都被触达，即形成最终的控股路径，即集团图谱，并为每户企业打上相应的集团标识。例如，依据上述遍历过程，集团图谱识别的具体步骤如下：第一步，提取股权关系并生成图。从数据库中导出所有股权关系(去除重复的股权关系)，以(x，y)表示一条股权关系，x为控制结点(控制人)，y为被控制结点(被控制人)。以控股关系为边，以企业(自然人或法人)为结点，构成一个图。由于控股关系是有向的，所以将其表示为有向图。第二步，将图表示为邻接矩阵(X，Y)。邻接矩阵是一个二维数组，其中每一维度均为图中的所有结点(即企业名，或者自然人或法人)。当结点i和结点j之间存在一条边时(即存在控股关系)，第i行、第j列对应的元素的值为1，否则为0。邻接矩阵将复杂的股权关系表示为清晰的二维矩阵，有利于DFS快速查找图中任意结点的所有相邻结点，确保搜索的高效和准确。第三步，采用DFS算法进行集团客户识别。利用DFS算法，对邻接矩阵(X，Y)进行搜索遍历，得出集团族谱识别结果。

[0004] 银行企业客户数量快速增长，伴随着大量不断变化的股权关系，以此构建的股权关系数量急剧增加，基于现有集团族谱识别方案一般难以满足复杂股权关系挖掘的性能要求。如此，基于上述集团族谱识别方案，由于股权关系网络的复杂性，导致目前的集团图谱识别策略难以满足海量股权关系数据挖掘的要求。

## 发明内容

[0005] 有鉴于此，本发明实施例提供了一种识别集团族谱的方法，包括：

[0006] 根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集；

[0007] 根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。

[0008] 本发明实施例还提供了一种识别集团族谱的系统，包括：

[0009] 第一确定单元，用于根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集；

[0010] 第二确定单元，用于根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。

[0011] 本发明实施例还提供了一种识别集团族谱的系统，包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述计算机程序被所述处理器执行时实现上述识别集团族谱的方法。

[0012] 本发明实施例还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上存储有信息处理程序，所述信息处理程序被处理器执行时实现上述识别集团族谱的方法的步骤。

[0013] 本发明实施例提供的技术方案，能够从海量股权关系中识别出客户的集团族谱，提升了集团族谱的识别效率。

[0014] 本申请的其它特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本申请而了解。本申请的其他优点可通过在说明书以及附图中所描述的方案来实现和获得。

## 附图说明

[0015] 附图用来提供对本申请技术方案的理解，并且构成说明书的一部分，与本申请的实施例一起用于解释本申请的技术方案，并不构成对本申请技术方案的限制。

[0016] 图1为本发明一实施例提供的一种识别集团族谱的方法的流程示意图；

[0017] 图2为本发明另一实施例提供的一种识别集团族谱的方法的流程示意图；

[0018] 图3a为本发明实施例中集团族谱识别模型中一种集团类型的示意图；

[0019] 图3b为本发明实施例中集团族谱识别模型中一种集团类型的示意图；

[0020] 图3c为本发明实施例中集团族谱识别模型中一种集团类型的示意图；

[0021] 图4为本发明另一实施例提供的一种识别集团族谱的方法的流程示意图；

[0022] 图5为本发明一实施例提供的一种识别集团族谱的系统的结构示意图；

[0023] 图6为本发明另一实施例提供的一种识别集团族谱的系统的结构示意图；

[0024] 图7为本发明一实施例中识别出的集团族谱的展示示意图；

[0025] 图8为本发明另一实施例提供的一种识别集团族谱的系统的结构示意图。

## 具体实施方式

[0026] 本申请描述了多个实施例，但是该描述是示例性的，而不是限制性的，并且对于本领域的普通技术人员来说显而易见的是，在本申请所描述的实施例包含的范围内可以有更多的实施例和实现方案。尽管在附图中示出了许多可能的特征组合，并在具体实施方式中进行了讨论，但是所公开的特征的许多其它组合方式也是可能的。除非特意加以限制的情况以外，任何实施例的任何特征或元件可以与任何其它实施例中的任何其他特征或元件结合使用，或可以替代任何其它实施例中的任何其他特征或元件。

[0027] 本申请包括并设想了与本领域普通技术人员已知的特征和元件的组合。本申请已

经公开的实施例、特征和元件也可以与任何常规特征或元件组合，以形成由权利要求限定的独特的发明方案。任何实施例的任何特征或元件也可以与来自其它发明方案的特征或元件组合，以形成另一个由权利要求限定的独特的发明方案。因此，应当理解，在本申请中示出和/或讨论的任何特征可以单独地或以任何适当的组合来实现。因此，除了根据所附权利要求及其等同替换所做的限制以外，实施例不受其它限制。此外，可以在所附权利要求的保护范围内进行各种修改和改变。

[0028] 此外，在描述具有代表性的实施例时，说明书可能已经将方法和/或过程呈现为特定的步骤序列。然而，在该方法或过程不依赖于本文所述步骤的特定顺序的程度上，该方法或过程不应限于所述的特定顺序的步骤。如本领域普通技术人员将理解的，其它的步骤顺序也是可能的。因此，说明书中阐述的步骤的特定顺序不应被解释为对权利要求的限制。此外，针对该方法和/或过程的权利要求不应限于按照所写顺序执行它们的步骤，本领域技术人员可以容易地理解，这些顺序可以变化，并且仍然保持在本申请实施例的精神和范围内。

[0029] 图1为本发明一实施例提供的一种识别集团族谱的方法的流程示意图，如图1所示，该方法包括：

[0030] 步骤101，根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集；

[0031] 步骤102，根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。

[0032] 可选地，所述股权网络是由点与点之间有向连接构成的点边关系图，其中点关系存储对应客户的属性，点与点之间连接的边关系存储对应关联客户属性及关联关系属性；

[0033] 所述集团族谱识别模型中的集团类型包括以下至少之一：

[0034] 两个或两个以上的客户共同被第三方客户所控制共同组成的集团；在股权上或者经营决策上直接或间接控制其他客户或被其他客户控制的客户共同组成的集团；由主要投资者个人、关键管理人员或与其近亲属共同直接控制或间接控制的客户共同组成的集团。

[0035] 可选地，所述根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集，包括：

[0036] 利用图计算引擎加载预先保存的股权网络；

[0037] 利用最大连通图算法从所述股份网络中识别出所述待识别客户标识关联的股权子网络，作为所述待识别客户的集团族谱候选集。

[0038] 可选地，所述根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱，包括：

[0039] 利用图计算算法根据所述集团族谱识别模型遍历所述集团族谱候选集，从中识别出所述待识别客户对应的全部集团族谱。

[0040] 可选地，在根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集之前，该方法还包括：

[0041] 利用图谱抽取工具从控股数据及亲属关系数据中抽取事先已梳理好的点边关系，形成所述股权网络，并保存到数据库中。

[0042] 可选地，所述图计算引擎为spark graphx图计算引擎，所述最大连通图算法为深度优先图搜索算法，所述图计算算法为基于spark graphx的Pregel，所述图谱抽取工具为hive sql图谱抽取工具，所述数据库为hive数据库。

[0043] 可选地，该方法还包括：

[0044] 通过图展示工具展示所述待识别客户的集团族谱。

[0045] 本发明实施例提供的技术方案，能够从海量股权关系中识别出客户的集团族谱，提升了集团族谱的识别效率。

[0046] 图2为本发明另一实施例提供的一种识别集团族谱的方法的流程示意图，如图2所示，该方法包括：

[0047] 步骤201，利用图谱抽取工具从控股数据及亲属关系数据中抽取事先已梳理好的点边关系，形成股权网络，并保存到数据库中；

[0048] 其中，所述股权网络是由点与点之间有向连接构成的点边关系图，其中点关系存储对应客户的属性，点与点之间连接的边关系存储对应关联客户属性及关联关系属性。

[0049] 可选地，所述图谱抽取工具为现有技术中任一种图谱抽取工具，例如hive sql图谱抽取工具，所述数据库为现有技术中任一种数据库，例如hive数据库。例如，对控股数据及亲属关系数据，可以基于事先已梳理好的点边关系，通过hive sql图谱抽取工具，将点边关系抽取出来，形成股权图谱存储到hive中。

[0050] 步骤202，利用图计算引擎加载预先保存的股权网络；

[0051] 可选地，所述图计算引擎为现有技术中任一种图计算引擎，例如spark graphx图计算引擎。例如，以spark graphx为图计算引擎，加载hive中存储的点边关系数据。

[0052] 步骤203，利用最大连通图算法从所述股权网络中识别出所述待识别客户标识关联的股权子网络，作为所述待识别客户的集团族谱候选集；

[0053] 可选地，所述最大连通图算法为现有技术中任一种最大连通图算法，例如深度优先图搜索算法，例如，通过深度优先图搜索算法等最大连通图算法，识别股权最小图谱中的股权关系子图，并以节点id作为子图标识，存储在节点属性当中，以此将无关企业及相关关系剔除，获取集团族谱候选集。

[0054] 其中，所述待识别客户是指待识别的实体，例如请求贷款的客户。所述待识别客户标识是指待识别的实体标识，例如客户id或者名称等表示客户身份的标识。该实体可以为自然人或者法人或者其他组织。

[0055] 步骤204，利用图计算算法根据集团族谱识别模型遍历所述集团族谱候选集，从中识别出所述待识别客户对应的全部集团族谱；

[0056] 可选地，所述集团族谱识别模型中的集团类型包括以下至少之一：

[0057] 两个或两个以上的客户共同被第三方客户所控制共同组成的集团；在股权上或者经营决策上直接或间接控制其他客户或被其他客户控制的客户共同组成的集团；由主要投资者个人、关键管理人员或与其近亲属共同直接控制或间接控制的客户共同组成的集团。

[0058] 其中，两个或两个以上的客户共同被第三方客户所控制共同组成的集团：例如图3a所示，某两个客户共同被第三方企事业法人所控制，其中A客户作为控制方，从股权上控制B和C客户，客户A、B、C组成的控股路径即为一种集团族谱；另外，图3a中的控股路径也可以扩展为多个客户或多条控股路径的情况；

[0059] 其中，在股权上或者经营决策上直接或间接控制其他客户或被其他客户控制的客户共同组成的集团：例如图3b所示，在股权上或者经营决策上直接或间接控制其他企事业法人或被其他企事业法人控制的，其中A客户作为控制方，从股权上控制B客户，同时B客户

又作为控制方控制C客户，客户A、B、C组成的控股路径即为一种集团族谱；另外，图3b中的控股路径也可以扩展为多个客户或多条控股路径的情况；

[0060] 其中，由主要投资者个人、关键管理人员或与其近亲属共同直接控制或间接控制的客户共同组成的集团：例如图3c所示，主要投资者个人、关键管理人员或与其近亲属(包括三代以内直系亲属关系和二代以内旁系亲属关系)共同直接控制或间接控制的集团，其中，自然人A、B都对客户C具有控股关系，自然人A与自然人B具有亲属关系，A、B、C组成的控股路径即为一种集团族谱；另外，图3c中的控股路径也可以扩展为多个客户或多条控股路径的情况。

[0061] 可选地，所述图计算算法为现有技术中任一种图计算算法，例如基于spark graphx的Pregel。例如，基于步骤203中获得的集团族谱候选集，结合上述集团族谱识别模型，通过pregel实现候选集遍历，对候选集进行筛选，获得所述待识别客户的全部集团族谱。

[0062] 可选地，通过pregel遍历集团族谱候选集的具体实施步骤包括：

[0063] 步骤1，遍历集团族谱候选集中所有的节点，为目标节点(即待识别客户对应的节点)赋予初始链圈标识id，其他的所有节点设置为‘NULL’；并为目标节点关联的边赋予属性，标识是否遍历；

[0064] 其中，股权关系图中的节点的属性(即点关系属性)包括以下至少之一：标识id、对应客户名称、是否为“NULL”等。股权关系图中点与点之间连接的边关系存储对应关联客户属性及关联关系属性，关联关系属性例如为控股人、亲属关系，占股权比例等，另外边关系是具有方向性的，例如节点A与节点B的边关系为从节点A指向节点B，表示节点A与节点B的关系为节点A对应客户是节点B对应客户的股权控制人。

[0065] 步骤2，如果目的节点属性为‘NULL’，则源节点向目的节点发送消息；如果源节点属性为‘NULL’，则目的节点向源节点发送消息；如果两端节点都为’NULL’，则不发送消息；如果两端节点都出现id，并且边属性为未遍历，则出现集团族谱，基于集团族谱识别模型为该边做集团族谱标识；

[0066] 以此类推，基于上述迭代，可获得目标节点的全部集团族谱。

[0067] 步骤205，通过图展示工具展示所述待识别客户的集团族谱。

[0068] 可选地，所述图展示工具可以为现有技术中任一个图展示工具，例如echarts等。

[0069] 本发明实施例提供的技术方案，能够从海量股权关系中识别出客户的集团族谱，提升了集团族谱的识别效率。

[0070] 图4为本发明另一实施例提供的一种识别集团族谱的方法的流程示意图，如图4所示，该方法包括：

[0071] 步骤401，对控股数据及亲属关系数据，基于事先已梳理好的点边关系，通过hive sql图谱抽取工具，将点边关系抽取出来，形成股权图谱存储到hive中；

[0072] 其中，所述股权图谱即是指上一实施例中的股权网络。

[0073] 具体而言，对股权及亲属关系数据进行梳理，提取数据中的相关实体，属性，关联关系。对所涉及股权关系进行统一表示，以企业为实体，股权关系为边构建股权图谱。

[0074] 步骤402，以spark graphx为图计算引擎加载hive中存储的股权图谱，通过最大连通图算法识别出待识别客户的股权最小图谱；

[0075] 其中，所述股权最小图谱即是指上一实施例中的股权子网络，作为集团族谱候选集。

[0076] 具体而言，以spark graphx为图计算引擎，加载股权图谱。通过实现的连通图算法，识别股权图谱中的股权关系子图。以此将无关企业及相关关系剔除，获取集团族谱候选集。

[0077] 本步骤中，将无关企业及相关关系从股权网络中剔除，获取集团族谱候选集。

[0078] 步骤403，通过pregel根据集团族谱识别模型遍历所述股权最小图谱，从中识别出所述待识别客户对应的全部集团族谱；

[0079] 具体而言，针对已获得集团族谱候选集，结合集团族谱识别模型，以待识别客户为起始点，通过pregel实现的深度优先算法，遍历候选集，对候选集进行筛选，获得集团族谱。

[0080] 步骤404，通过图展示工具展示所述待识别客户对应的全部集团族谱。

[0081] 可选地，所述图展示工具为现有任一种图展示工具，例如echarts等。

[0082] 本发明实施例提供的技术方案，通过hive数据库的使用解决了海量图数据存储表示问题，并通过spark graphx图计算引擎，解决了传统sql遍历复杂网络存在的性能问题，如此能够从海量图数据中识别出集团族谱。集团族谱的识别集团族谱的识别，有利于银行对集团客户的关系管理、日常业务管理、风险管理、效益分析等功能，达到动态掌握客户信息，实施有效监控，为客户提供差异化服务的目的，同事有助于银行集团客户管理部门提高风险预警和防范水平，促进集团客户业务的精细化、规范化管理。

[0083] 图5为本发明一实施例提供的一种识别集团族谱的系统的结构示意图，如图5所示，该系统包括：

[0084] 第一确定单元，用于根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集；

[0085] 第二确定单元，用于根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。

[0086] 可选地，所述股权网络是由点与点之间有向连接构成的点边关系图，其中点关系存储对应客户的属性，点与点之间连接的边关系存储对应关联客户属性及关联关系属性；

[0087] 所述集团族谱识别模型中的集团类型包括以下至少之一：

[0088] 两个或两个以上的客户共同被第三方客户所控制共同组成的集团；在股权上或者经营决策上直接或间接控制其他客户或被其他客户控制的客户共同组成的集团；由主要投资者个人、关键管理人员或与其近亲属共同直接控制或间接控制的客户共同组成的集团。

[0089] 可选地，第一确定单元，具体用于利用图计算引擎加载预先保存的股权网络；

[0090] 利用最大连通图算法从所述股份网络中识别出所述待识别客户标识关联的股权子网络，作为所述待识别客户的集团族谱候选集。

[0091] 可选地，第二确定单元，具体用于利用图计算算法根据所述集团族谱识别模型遍历所述集团族谱候选集，从中识别出所述待识别客户对应的全部集团族谱。

[0092] 可选地，该系统还包括：

[0093] 第三确定单元，用于在根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集之前，利用图谱抽取工具从控股数据及亲属关系数据中抽取事先已梳理好的点边关系，形成所述股权网络，并保存到数据库中。

[0094] 可选地，所述图计算引擎为spark graphx图计算引擎，所述最大连通图算法为深度优先图搜索算法，所述图计算算法为基于spark graphx的Pregel，所述图谱抽取工具为hive sql图谱抽取工具，所述数据库为hive数据库。

[0095] 可选地，该系统还包括：展示单元，用于通过图展示工具展示所述待识别客户的集团族谱。

[0096] 本发明实施例提供的技术方案，能够从海量股权关系中识别出客户的集团族谱，提升了集团族谱的识别效率。

[0097] 图6为本发明另一实施例提供的一种识别集团族谱的系统的结构示意图，如图6所示，该系统包括：

[0098] 连通图API(Application Programming Interface，应用程序接口)和集团族谱过滤API；

[0099] 其中，连通图API对应于上述实施例中的第一确定单元，集团族谱过滤API对应于上述实施例中的第二确定单元。

[0100] 其中，连通图API，用于根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集；

[0101] 可选地，所述股权网络是由点与点之间有向连接构成的点边关系图，其中点关系存储对应客户的属性，点与点之间连接的边关系存储对应关联客户属性及关联关系属性；

[0102] 所述集团族谱识别模型中的集团类型包括以下至少之一：

[0103] 两个或两个以上的客户共同被第三方客户所控制共同组成的集团；在股权上或者经营决策上直接或间接控制其他客户或被其他客户控制的客户共同组成的集团；由主要投资者个人、关键管理人员或与其近亲属共同直接控制或间接控制的客户共同组成的集团。

[0104] 可选地，连通图API，具体用于利用图计算引擎加载预先保存的股权网络；

[0105] 利用最大连通图算法从所述股权网络中识别出所述待识别客户标识关联的股权子网络，作为所述待识别客户的集团族谱候选集。

[0106] 可选地，所述图计算引擎为现有任一种图计算引擎，例如spark graphx图计算引擎，所述最大连通图算法为现有任一种最大连通图算法，例如深度优先图搜索算法。例如，将上述股权网络作为连通图API的输入，计算股权网络中的集团族谱候选集。

[0107] 其中，该系统还包括：

[0108] 第三确定单元，用于利用图谱抽取工具从控股数据及亲属关系数据中抽取事先已梳理好的点边关系，形成所述股权网络，并保存到数据库中。

[0109] 可选地，所述图谱抽取工具为现有任一种图谱抽取工具，例如hive sql图谱抽取工具，所述数据库为现有任一种数据库，例如hive数据库。

[0110] 例如，本实施例中，基于某商业银行客户的控股数据及亲属关系数据进行说明，从控股数据及亲属关系数据数据中按事先定义的点边定义，通过hive sql工具抽取上述数据中的点边关系，构造股权网络，分别存储点和边关系。点关系存储客户及其相关属性，边关系存储以关联客户id为主体的信息以及关联关系属性。然后，将hive数据库中的股权网络作为连通图API的输入，计算股权网络中的集团族谱候选集。

[0111] 其中，集团族谱过滤API，用于根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱。

[0112] 可选地，集团族谱过滤API，具体用于利用图计算算法根据所述集团族谱识别模型遍历所述集团族谱候选集，从中识别出所述待识别客户对应的全部集团族谱。

[0113] 可选地，所述图计算算法为现有任一种图计算算法，例如基于spark graphx的Pregel。

[0114] 例如，本实施例中，以待识别客户(即待识别实体)和集团族谱候选集为输入，调用集团族谱过滤API，筛选待识别客户所在的全部集团族谱。

[0115] 其中，该系统还包括：显示单元，

[0116] 所述显示单元，用于通过图展示工具展示识别出的全部集团族谱。

[0117] 可选地，所述图展示工具为现有任一种图展示工具，例如echarts等。例如图7所示，即为图展示工具展示的一种集团族谱示例图，其中，A、B、C、D、E、F、G分别为图中各个节点可以分别代表不同的实体(客户)，每一个实体可以代表不同的客户，其中每两个节点之间的边关系代表了这两个节点的关联关系，例如A(自然人)、B(企业)之间的边关系为企业股东与企业的关系，股东A占企业B股权的60％。

[0118] 本发明实施例提供的技术方案，能够从海量股权关系中识别出客户的集团族谱，提升了集团族谱的识别效率。

[0119] 图8为本发明另一实施例提供的一种识别集团族谱的系统的结构示意图，如图8所示，该系统包括：

[0120] 股权图谱API、连通图API、集团族谱过滤API和显示单元；

[0121] 其中，股权图谱API对应于上述实施例中的第三确定单元。

[0122] 其中，股权图谱API，用于对控股数据及亲属关系数据，基于事先已梳理好的点边关系，通过hive sql图谱抽取工具，将点边关系抽取出来，形成股权图谱存储到hive中；

[0123] 其中，所述股权图谱即是指上述实施例中的股权网络。

[0124] 其中，连通图API，用于以spark graphx为图计算引擎加载hive中存储的股权图谱，通过最大连通图算法识别出待识别客户的股权最小图谱；

[0125] 其中，所述股权最小图谱即是指上一实施例中的股权子网络，作为集团族谱候选集。

[0126] 本步骤中，将无关企业及相关关系从股权网络中剔除，获取集团族谱候选集。

[0127] 其中，集团族谱过滤API，用于通过pregel根据集团族谱识别模型遍历所述股权最小图谱，从中识别出所述待识别客户对应的全部集团族谱；

[0128] 其中，显示单元，用于通过图展示工具展示所述待识别客户对应的全部集团族谱。

[0129] 可选地，所述图展示工具为现有任一种图展示工具，例如echarts等。

[0130] 本发明实施例提供的技术方案，对股权控股数据进行治理提取出业务相关的实体及相关属性，以及梳理和完善客户的控股关系，构建股权图谱；并采用hive作为图谱存储介质，在构建的股权图谱的基础上，基于spark graphx的Pregel实现最大连通子图、深度优先等图搜索算法，实现对股权图谱的穷尽搜索；并结业务规则设计了集团族谱识别模型，以此完成对集团族谱的识别。

[0131] 本发明实施例还提供了一种识别集团族谱的系统，包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述计算机程序被所述处理器执行时实现上述任一项所述识别集团族谱的方法。

[0132] 本发明实施例还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上存储有信息处理程序，所述信息处理程序被处理器执行时实现上述任一项所述识别集团族谱的方法的步骤。

[0133] 本领域普通技术人员可以理解，上文中所公开方法中的全部或某些步骤、系统、装置中的功能模块/单元可以被实施为软件、固件、硬件及其适当的组合。在硬件实施方式中，在以上描述中提及的功能模块/单元之间的划分不一定对应于物理组件的划分；例如，一个物理组件可以具有多个功能，或者一个功能或步骤可以由若干物理组件合作执行。某些组件或所有组件可以被实施为由处理器，如数字信号处理器或微处理器执行的软件，或者被实施为硬件，或者被实施为集成电路，如专用集成电路。这样的软件可以分布在计算机可读介质上，计算机可读介质可以包括计算机存储介质(或非暂时性介质)和通信介质(或暂时性介质)。如本领域普通技术人员公知的，术语计算机存储介质包括在用于存储信息(诸如计算机可读指令、数据结构、程序模块或其他数据)的任何方法或技术中实施的易失性和非易失性、可移除和不可移除介质。计算机存储介质包括但不限于RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光盘存储、磁盒、磁带、磁盘存储或其他磁存储装置、或者可以用于存储期望的信息并且可以被计算机访问的任何其他的介质。此外，本领域普通技术人员公知的是，通信介质通常包含计算机可读指令、数据结构、程序模块或者诸如载波或其他传输机制之类的调制数据信号中的其他数据，并且可包括任何信息递送介质。

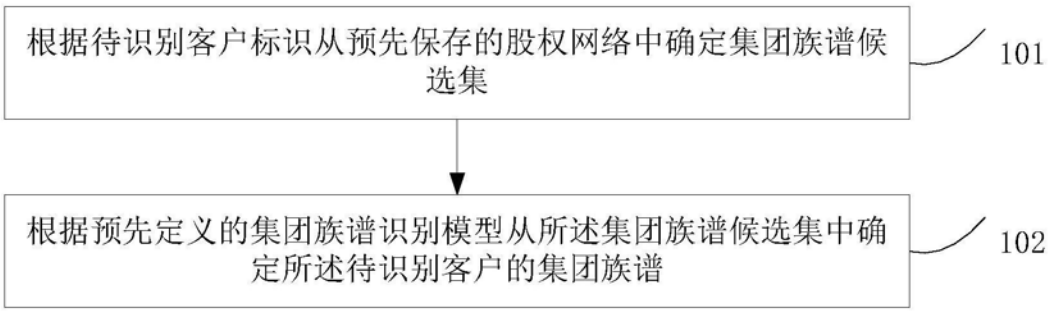

根据待识别客户标识从预先保存的股权网络中确定集团族谱候选集
101
根据预先定义的集团族谱识别模型从所述集团族谱候选集中确定所述待识别客户的集团族谱
102

图1

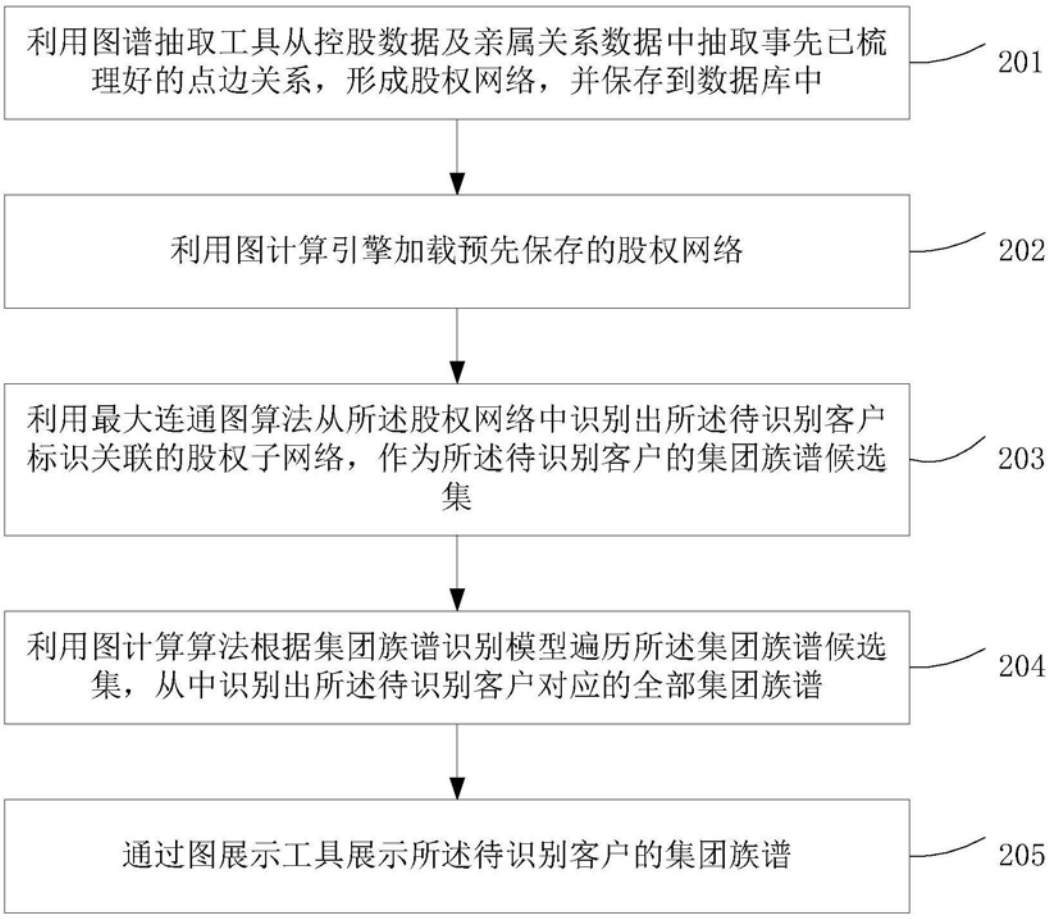

利用图谱抽取工具从控股数据及亲属关系数据中抽取事先已梳理好的点边关系，形成股权网络，并保存到数据库中
201
利用图计算引擎加载预先保存的股权网络
202
利用最大连通图算法从所述股权网络中识别出所述待识别客户标识关联的股权子网络，作为所述待识别客户的集团族谱候选集
203
利用图计算算法根据集团族谱识别模型遍历所述集团族谱候选集，从中识别出所述待识别客户对应的全部集团族谱
204
通过图展示工具展示所述待识别客户的集团族谱
205

图2

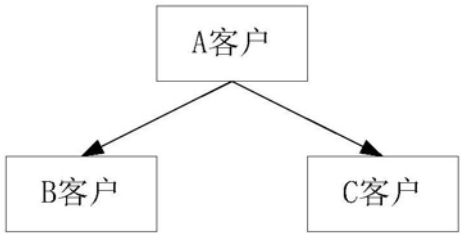



图3a

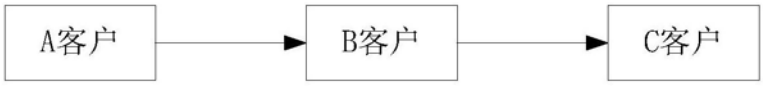



图3b

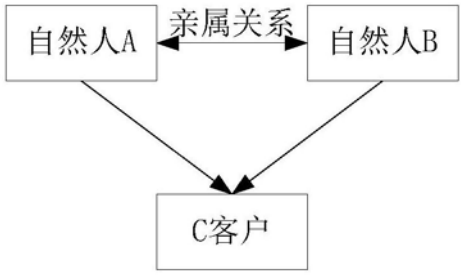



图3c

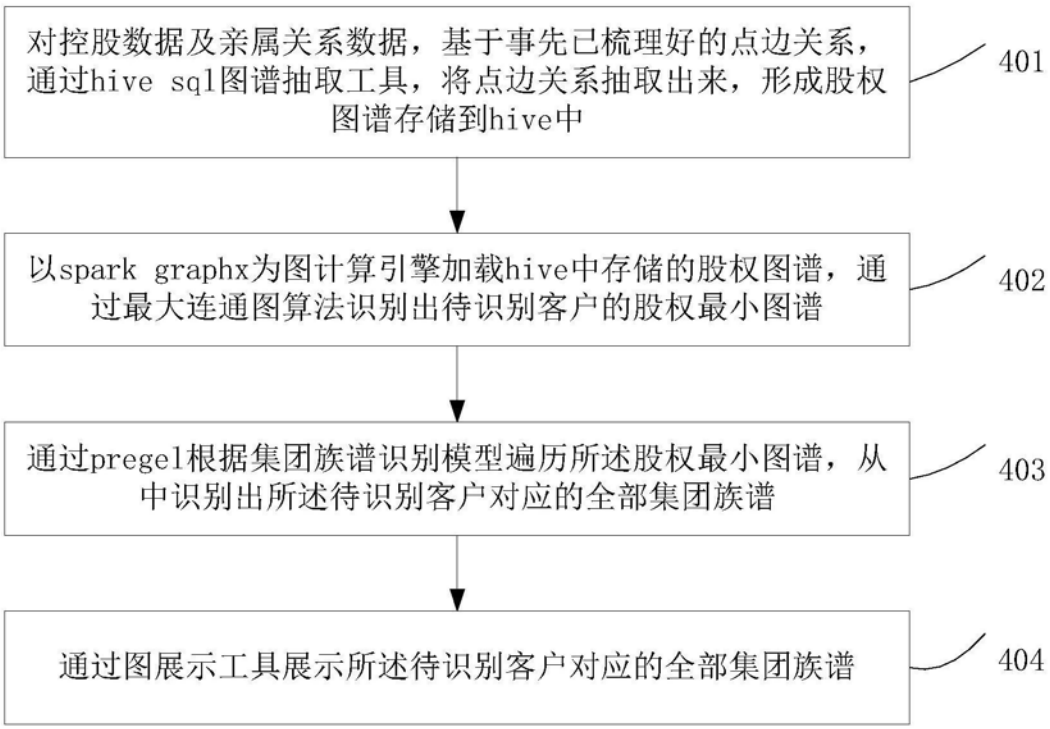

对控股数据及亲属关系数据，基于事先已梳理好的点边关系，通过hive sql图谱抽取工具，将点边关系抽取出来，形成股权图谱存储到hive中
401
以spark graphx为图计算引擎加载hive中存储的股权图谱，通过最大连通图算法识别出待识别客户的股权最小图谱
402
通过pregel根据集团族谱识别模型遍历所述股权最小图谱，从中识别出所述待识别客户对应的全部集团族谱
403
通过图展示工具展示所述待识别客户对应的全部集团族谱
404

图4

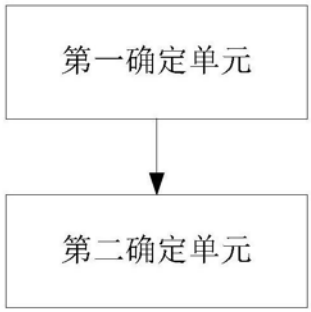

第一确定单元
第二确定单元

图5


连通图API
集团族谱过滤API

图6

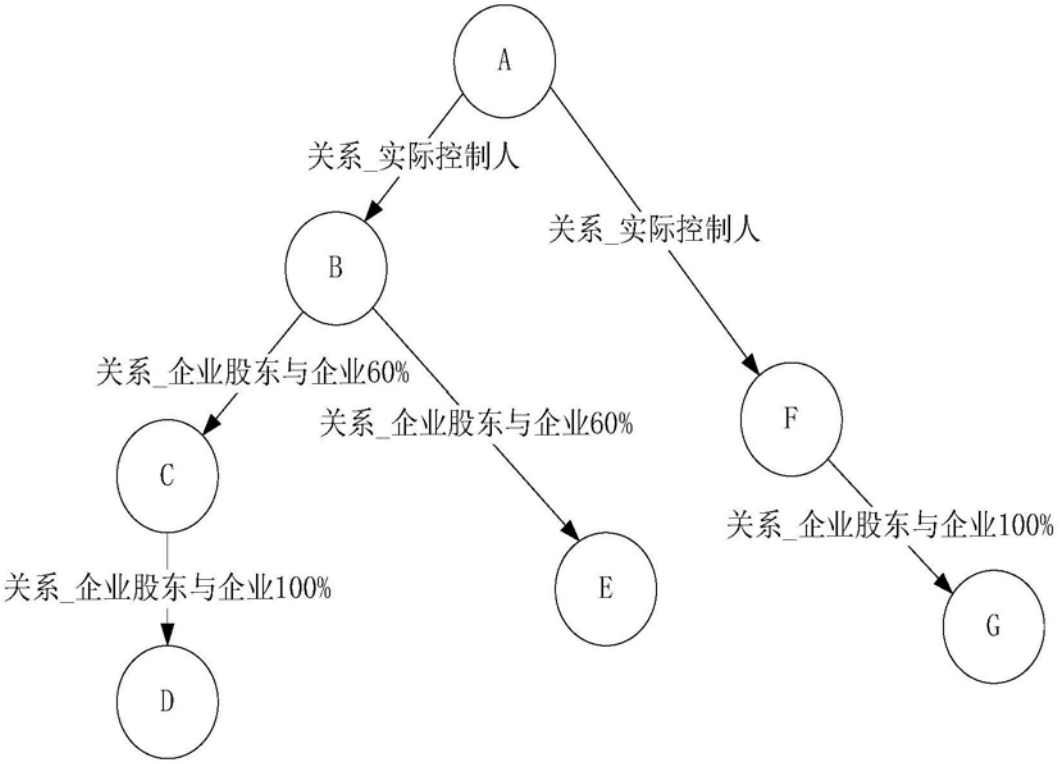

A
关系_实际控制人
关系_实际控制人
B
关系_企业股东与企业60%
关系_企业股东与企业60%
F
C
关系_企业股东与企业100%
关系_企业股东与企业100%
E
G
D

图7

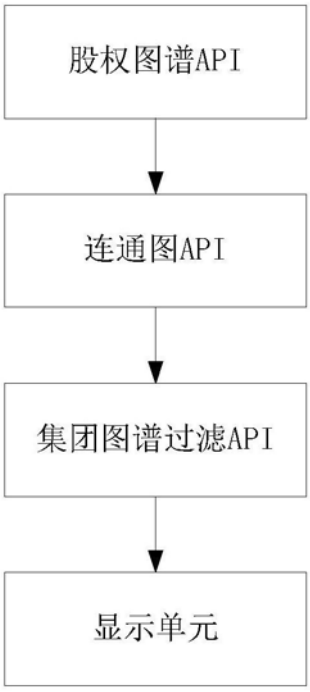

股权图谱API
连通图API
集团图谱过滤API
显示单元

图8