(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108427956 B
(45)授权公告日 2019.08.06

(21)申请号 201710078997.6

(22)申请日 2017.02.14

(65)同一申请的已公布的文献号
申请公布号 CN 108427956 A

(43)申请公布日 2018.08.21

(73)专利权人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 李霖 陈培炫 陈谦

(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
代理人 王仲凯

(51)Int.Cl.
G06K 9/62(2006.01)
G06F 16/95(2019.01)

(56)对比文件
CN 103020163 A,2013.04.03,
CN 103136303 A,2013.06.05,
CN 102063458 A,2011.05.18,
US 2014089324 A1,2014.03.27,

审查员 程琼

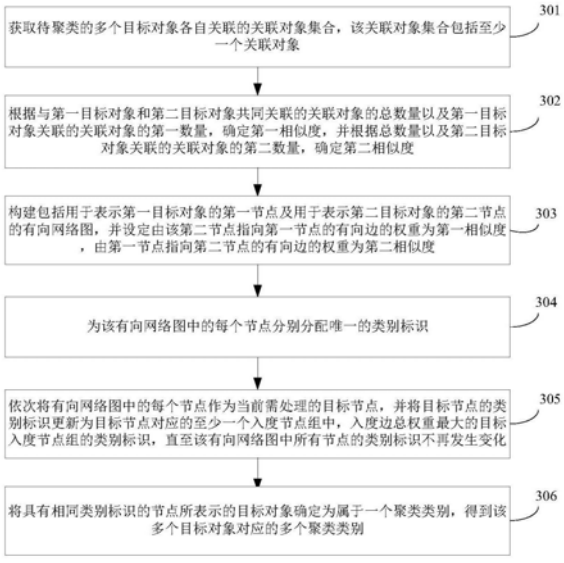
权利要求书3页 说明书15页 附图6页

(54)发明名称

一种对象聚类方法和装置

(57)摘要

本申请提供了一种对象聚类方法和装置,该方案中,针对多个待聚类的目标对象构建有向网络图,并基于任意两个目标对象之间关联的关联对象的相似程度,确定出有向网络图表征所述两个目标对象的两个节点之间有向边的权重;且为每个节点分配唯一的类别标识;依次将每个节点作为目标节点,将目标节点的类别标识更新为该目标节点对应的多个入度节点组中,指向该目标节点的有向边的总权重最大的目标入度节点组的类别标识,直至所有节点的类别标识均不再发生变化,得到多个类别标识表示的多个聚类类别。本申请的方案可以提高对象聚类的精准度。



CN 108427956 B

1.一种对象聚类方法,其特征在于,包括:

获取待聚类的多个目标对象各自关联的关联对象集合,所述关联对象集合包括至少一个关联对象;

基于任意两个目标对象之间关联的关联对象的相似程度,确定待构建的有向网络图中表征所述两个目标对象的两个节点之间有向边的权重,并构建所述有向网络图;

为所述有向网络图中的每个节点分别分配唯一的类别标识;

依次将每个所述节点作为当前需处理的目标节点,从所述目标节点对应的至少一个入度节点组中,确定出指向该目标节点的有向边的总权重最大的目标入度节点组,并将所述节点的类别标识更新为所述目标入度节点组的目标标识,直至所述有向网络图中所有节点的类别标识不再发生变化,其中,所述入度节点组包括有向边指向所述目标节点且具有相同类别标识的至少一个入度节点;

将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别,得到该多个目标对象对应的多个聚类类别。

2.根据权利要求1所述的对象聚类方法,其特征在于,所述基于任意两个目标对象之间关联的关联对象的相似程度,确定待构建的有向网络图中表征所述两个目标对象的两个节点之间有向边的权重,包括:

对于所述多个目标对象中任意的第一目标对象以及第二目标对象,根据与所述第一目标对象和所述第二目标对象都关联的关联对象的总数量,以及所述第一目标对象关联的关联对象的第一数量,确定出待构建的有向网络图中,从表示所述第二目标对象的第二节点指向表示所述第一目标对象的第一节点的有向边的权重;

根据所述总数量,以及所述第二目标对象关联的关联对象的第二数量,确定出从所述第一节点指向所述第二节点的有向边的权重。

3.根据权利要求2所述的对象聚类方法,其特征在于,所述根据与所述第一目标对象和所述第二目标对象都关联的关联对象的总数量,以及所述第一目标对象关联的关联对象的第一数量,确定出从表示所述第二目标对象的第二节点指向表示所述第一目标对象的第一节点的有向边的权重,包括:

将与所述第一目标对象以及所述第二目标对象都关联的关联对象的总数量,与所述第一目标对象关联的关联对象的第一数量的比值,确定为从表示所述第二目标对象的第二节点指向表示所述第一目标对象的第一节点的有向边的权重;

所述根据所述总数量,以及所述第二目标对象关联的关联对象的第二数量,确定出从所述第一节点指向所述第二节点的有向边的权重,包括:

将所述总数量,与所述第二目标对象关联的关联对象的第二数量的比值,确定为从所述第一节点指向所述第二节点的有向边的权重。

4.根据权利要求1所述的对象聚类方法,其特征在于,获取待聚类的多个目标对象各自关联的关联对象集合,包括:

获取待聚类的多个目标对象各自对应的至少一个数据关系,所述数据关系包括所述目标对象的标识与所述目标对象关联的关联对象之间的对应关系;

所述为所述有向网络图中的每个节点分配一个唯一的类别标识,包括:

将有向网络图中每个节点表示的目标对象的标识作为所述节点的类别标识。

5.根据权利要求1所述的对象聚类方法，其特征在于，所述依次将每个所述节点作为当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所述有向网络图中所有节点的类别标识不再发生变化，包括：

将所述有向网络图中的所有节点均作为待处理节点；

如果存在未处理的待处理节点，从未处理的所述待处理节点中选取当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所有待处理节点均作为目标节点被处理为止；

如果存在更新后的类别标识与更新前的类别标识不同的节点，则将更新前的类别标识与更新后的类别标识不同的节点确定为待处理节点，并返回执行所述如果存在未处理的待处理节点，从未处理的所述待处理节点中选取当前需处理的目标节点的操作；

如果不存在更新后的类别标识与更新前的类别标识不同的节点，则执行所述将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别的操作。

6.根据权利要求1所述的对象聚类方法，其特征在于，所述获取待聚类的多个目标对象各自关联的关联对象集合，所述关联对象集合包括至少一个关联对象，包括：

获取待聚类的多个用户设备各自关联的用户标识集合，所述用户标识集合包括多个用户标识，其中，用户设备关联的用户标识为通过所述用户设备访问预设网络的用户的标识。

7.根据权利要求1所述的对象聚类方法，其特征在于，所述获取待聚类的多个目标对象各自关联的关联对象集合，所述关联对象集合包括至少一个关联对象，包括：

获取待聚类的多篇文档各自关联的社群标识集合，所述社群标识集合包括至少一个社群标识；

所述将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，具体为：

将具有相同类别标识的节点所表示的文档确定为属于同一个主题。

8.一种对象聚类装置，其特征在于，包括：

数据获取单元，用于获取待聚类的多个目标对象各自关联的关联对象集合，所述关联对象集合包括至少一个关联对象；

有向图构建单元，用于基于任意两个目标对象之间关联的关联对象的相似程度，确定待构建的有向网络图中表征所述两个目标对象的两个节点之间有向边的权重，并构建所述有向网络图；

类别初始化单元，用于为所述有向网络图中的每个节点分别分配唯一的类别标识；

聚类分析单元，用于依次将每个所述节点作为当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并将所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所述有向网络图中所有节点的类别标识不再发生变化，其中，所述入度节点组包括有向边指向所述目标节点且具有相同类别标识的至少一个入度节点；

类别提取单元，用于将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别。

9.根据权利要求8所述的对象聚类装置，其特征在于，所述有向图构建单元，包括：

第一权重确定单元，用于对于所述多个目标对象中任意的第一目标对象以及第二目标对象，根据与所述第一目标对象和所述第二目标对象都关联的关联对象的总数量，以及所述第一目标对象关联的关联对象的第一数量，确定出待构建的有向网络图中，从表示所述第二目标对象的第二节点指向表示所述第一目标对象的第一节点的有向边的权重；

第二权重确定单元，用于根据所述总数量，以及所述第二目标对象关联的关联对象的第二数量，确定出从所述第一节点指向所述第二节点的有向边的权重；

有向图构建子单元，用于构建所述有向网络图。

10.根据权利要求9所述的对象聚类装置，其特征在于，所述第一权重确定单元，具体用于将与所述第一目标对象以及所述第二目标对象都关联的关联对象的总数量，与所述第一目标对象关联的关联对象的第一数量的比值，确定为从表示所述第二目标对象的第二节点指向表示所述第一目标对象的第一节点的有向边的权重；

所述第二权重确定单元，具体用于将所述总数量，与所述第二目标对象关联的关联对象的第二数量的比值，确定为从所述第一节点指向所述第二节点的有向边的权重。

11.根据权利要求8所述的对象聚类装置，其特征在于，所述数据获取单元，具体为，用于获取待聚类的多个目标对象各自对应的至少一个数据关系，所述数据关系包括所述目标对象的标识与所述目标对象关联的关联对象之间的对应关系；

所述类别初始化单元，包括：

类别初始化子单元，用于将有向网络图中每个节点表示的目标对象的标识作为所述节点的类别标识。

12.根据权利要求8所述的对象聚类装置，其特征在于，所述聚类分析单元，包括：

节点初始处理单元，用于将所述有向网络图中的所有节点均作为待处理节点；

聚类分析子单元，用于如果存在未处理的待处理节点，从未处理的所述待处理节点中选取当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所有待处理节点均作为目标节点被处理为止；

循环控制单元，用于如果存在更新后的类别标识与更新前的类别标识不同的节点，则将更新前的类别标识与更新后的类别标识不同的节点确定为待处理节点，并触发返回执行所述聚类分析子单元的操作；

所述类别提取单元具体为，用于如果不存在更新后的类别标识与更新前的类别标识不同的节点，则将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别。

13.根据权利要求8所述的对象聚类装置，其特征在于，所述数据获取单元具体为，用于获取待聚类的多个用户设备各自关联的用户标识集合，所述用户标识集合包括多个用户标识，其中，用户设备关联的用户标识为通过所述用户设备访问预设网络的用户的标识。

一种对象聚类方法和装置

技术领域

[0001] 本申请涉及数据处理技术领域，尤其涉及一种对象聚类方法和装置。

背景技术

[0002] 随着互联网技术的不断发展，网络中的信息量日益增多。为了能够有效利用网络信息，很多情况下，需要对网络信息中所包含同种对象进行聚类。如，同一网络用户可能会使用多个不同的用户设备来进行网络访问，例如，用户可以利用自己或家人的手机或者其他终端设备来登录即时通讯系统或者论坛等等，而为了防御恶意访问或者是有针对性的对用户提供服务，就可能需要确定出哪些用户设备是由同一个用户经常使用的，从而需要对用户设备进行聚类。

[0003] 然而，针对网络中的一种目标对象进行聚类时，仅仅是根据目标对象关联的关联对象的标识，将关联有相同关联对象的标识的目标对象聚类到一起，如，如果多个用户设备访问网络系统时，所采用的用户账号相同，则认为该多个用户设备为同一个用户经常使用的，将这多个用户设备聚类到一起。根据目标对象关联的关联对象的标识，对目标对象进行聚类，使得目标对象聚类出的类别较多，不能将关联性较强的目标对象聚类到一起，导致聚类的精准度低。

发明内容

[0004] 有鉴于此，本申请提供了一种对象聚类方法和装置，以最大程度的挖掘待聚类的目标对象之间的关联度，提高聚类的精准度。

[0005] 为实现上述目的，一方面，本申请提供了一种对象聚类方法，包括：

[0006] 获取待聚类的多个目标对象各自关联的关联对象集合，所述关联对象集合包括至少一个关联对象；

[0007] 基于任意两个目标对象之间关联的关联对象的相似程度，确定待构建的有向网络图中表征所述两个目标对象的两个节点之间有向边的权重，并构建所述有向网络图；

[0008] 为所述有向网络图中的每个节点分别分配唯一的类别标识；

[0009] 依次将每个所述节点作为当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所述有向网络图中所有节点的类别标识不再发生变化，其中，所述入度节点组包括有向边指向所述目标节点且具有相同类别标识的至少一个入度节点；

[0010] 将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别。

[0011] 另一方面，本申请实施例还提供了一种对象聚类装置，包括：

[0012] 数据获取单元，用于获取待聚类的多个目标对象各自关联的关联对象集合，所述关联对象集合包括至少一个关联对象；

[0013] 有向图构建单元，用于基于任意两个目标对象之间关联的关联对象的相似程度，确定待构建的有向网络图中表征所述两个目标对象的两个节点之间有向边的权重，并构建所述有向网络图；

[0014] 类别初始化单元，用于为所述有向网络图中的每个节点分别分配唯一的类别标识；

[0015] 聚类分析单元，用于依次将每个所述节点作为当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所述有向网络图中所有节点的类别标识不再发生变化，其中，所述入度节点组包括有向边指向所述目标节点且具有相同类别标识的至少一个入度节点；

[0016] 类别提取单元，用于将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别。

[0017] 由以上内容可知，在本申请实施例中，根据待聚类的任意两个目标对象之间的所关联的关联对象的相似程度，确定出待构建的有向网络图中表示这两个目标对象的目标节点之间有向边的权重，并构建出有向网络图之后，基于有向网络图中各个节点之间有向边的权重，对有向网络图中不同节点进行类别聚类，从而有利于从全局角度挖掘目标对象之间的相似程度，并对目标对象进行聚类，进而可以有效提高目标对象聚类的精准度。

附图说明

[0018] 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，还可以根据提供的附图获得其他的附图。

[0019] 图1为本申请实施例公开一种对象聚类方法所适用的一种计算机设备的一种可能的组成架构示意图；

[0020] 图2为本申请实施例公开的一种对象聚类方法所适合的一种系统组成架构示意图；

[0021] 图3为本申请公开的一种对象聚类方法一个实施例的流程示意图；

[0022] 图4为本申请公开的一种对象聚类方法又一个实施例的流程示意图；

[0023] 图5a示出了本申请实施例构建出的一种有向网络图的部分组成结构示意图；

[0024] 图5b示出了利用权重最大的入度边对应的入度节点的社区标识对图5a所示的有向网络图中一个节点的社区标识进行更新之后的有向网络图的示意图；

[0025] 图6示出了本申请公开的一种对象聚类方法又一个实施例的流程示意图；

[0026] 图7示出了本申请公开的一种对象聚类装置一个实施例的组成结构示意图。

具体实施方式

[0027] 本申请实施例提供了的对象聚类方法和装置适用于对登录社交网络的多台用户设备进行聚类，或者是多篇文章进行主题发现。

[0028] 本实施例的方法和装置适用于单台计算机设备，也可以是分布式计算系统。

[0029] 如图1，其示出了本申请实施例的对象聚类方法和装置所适用的计算机设备的一种组成结构示意图。在图1中，该计算机设备可以包括：存储器101、处理器102、通信模块103、显示器104、输入单元105以及通信总线106等部件。其中，处理器101、存储器102、通信接口103、显示器104以及输入单元105均通过通信总线106完成相互间的通信。

[0030] 其中，存储器101可用于存储软件程序以及模块。存储器120可存储操作系统、至少一个功能(比如，图像播放功能)所需的应用程序等；还可以存储根据终端的使用所创建的数据等。其中，存储器101可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0031] 处理器102是终端的控制中心，利用各种接口和线路连接整个终端的各个部分，通过运行或执行存储在存储器101内的软件程序和/或模块，以及调用存储在存储器101内的数据，执行终端的各种功能和处理数据，从而对终端进行整体监控。可选的，处理器102可包括一个或多个处理单元。

[0032] 该通信模块103可以用于收发信息，或者数据处理过程中信号的接收与发送；或者通过网络与其他设备进行通信等。

[0033] 该显示器104可用于窗口界面，并在窗口界面中显示所处理的数据、图形，有向网络图等等；还可以显示由用户输入的信息，或者提供给用户的信息，以及计算机设备的各种图形用户接口，这些图形用户接口可以由图形、文本、图片等任意组合来构成。该显示器可以包括显示面板，如，可以为采用液晶显示器、有机发光二极管等形式来配置的显示面板。进一步的，该显示器可以包括具备采集触摸事件的触摸显示面板。

[0034] 输入单元105可用于接收输入的用户输入的字符、数字等信息，以及产生与用户设置以及功能控制有关的信号输入。该输入单元可以包括但不限于物理键盘、鼠标、操作杆等中的一种或多种。

[0035] 可以理解的是，无论在何种场景中，该终端均可以为任意能够实现访问服务平台的设备，如，该终端可以为手机、平板电脑、台式电脑等等。

[0036] 当然，为了提高数据处理能力，本申请实施例的对象聚类方法也可以适用于分布式计算系统，如图2所示，其示出了本申请的对象聚类方法所适用的一种分布式计算系统的组成结构示意图。

[0037] 由图2可知，该分布式计算系统可以包括多台计算机设备201，这多台计算机设备201之间可以通过网络相连，这多台计算机设备201之间可以相互配合以完成本申请实施例的对象聚类方法和装置中所涉及到的数据处理。

[0038] 基于以上共性，在本申请的对象聚类方法中，获取到待聚类的多个目标对象各自关联至少一个关联对象之后，基于任意两个目标对象之间关联的关联对象的相似程度，确定出有向网络图中表征所述两个目标对象的两个节点之间有向边的权重，并构建表示有该多个关联对象的多个节点，以及节点之间具有有向边的有向网络图；同时，为有向网络图中的每个节点分别分配唯一的类别标识；然后依次将每个所述节点作为当前需处理的目标节点，从目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所述有向网络图中所有节点的类别标识不再发生变化，其中，所述入度节点组包括指向所述目标节点且具有相同类别标识的至少一个入度节点；将具有相同类别标识的节点所

表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别，从而实现了依据目标对象所关联的关联对象构建有向网络图，并基于有向网络图对目标对象进行精准聚类。

[0039] 基于图1和图2，下面结合不同实施例对本申请实施例的对象聚类方法进行相似介绍。

[0040] 参见图3，其示出了本申请一种对象聚类方法一个实施例的流程示意图，本实施例的方法可以应用于如上所示的计算机设备或者计算机系统，本实施例的方法可以包括：

[0041] S301，获取待聚类的多个目标对象各自关联的关联对象集合，该关联对象集合包括至少一个关联对象。

[0042] 其中，待聚类的目标对象可以根据需要选取，相应的，针对不同的目标对象，对该目标对象进行聚类所需获取的该目标对象关联的关联对象也会有所差异。如，当需要通过聚类，将同一个用户经常使用的用户设备聚类到一起时，则目标对象可以为用户设备，而目标对象关联的关联对象可以为用户的账号、用户名等用户标识。

[0043] 可以理解的是，为了确定目标对象关联的关联对象集合，可以获取是获取到各个目标对象关联的所有关联对象的信息，也可以是获取多个数据关系，每个数据关系中包括一个目标对象的标识与该目标对象关联的关联对象的对应关系，根据目标对象的标识，可以确定出每个目标对象关联有哪些关联对象。

[0044] S302，对于该多个目标对象中任意的第一目标对象以及第二目标对象，根据与第一目标对象以及第二目标对象都关联的关联对象的总数量以及与第一目标对象关联的关联对象的第一数量，确定出待构建的有向网络图中，第二目标对象指向第一目标对象的第一相似度，并根据该总数量以及与第二目标对象关联的关联对象的第二数量，确定待构建的有向网络图中第二目标对象指向第一目标对象的第二相似度。

[0045] 可以理解的是，为了便于描述任意两个目标对象之间的相似度，对于任意两个不同的目标对象，可以将其中一个目标对象称为第一目标对象，并将第二目标对象称为第二目标对象。其中，该第一目标对象与第二目标对象不同，需要说明的是，在本申请实施例中，当目标对象的标识不同时，就可以认为两个标识不同的目标对象为不同的目标对象。

[0046] 其中，根据两个目标对象之间关联的相同关联对象的数量，以及每个目标对象各自关联的关联对象的数量，确定两个目标对象之间相似度的方式可以有多种，在此不加以限制。

[0047] S303，构建包括用于表示第一目标对象的第一节点以及用于表示第二目标对象的第二节点的有向网络图，并设定有向网络图中由该第二节点指向第一节点的有向边的权重为该第一相似度，由第一节点指向第二节点的有向边的权重为第二相似度。

[0048] 可以理解的是，有向网络图中包括节点以及节点之间的有向边。在本申请实施例中，构建出的有向网络图中包含有分别表示该多个目标对象的多个节点，而为了便于描述，本申请是以任意两个目标对象为例进行说明。相应的，对于有向网络图中任意的第一节点和第二节点，将第一节点表示的第一目标对象指向第二节点所表示的第二目标对象的相似度作为第一节点指向所述第二节点的有向边的权重；而将第二目标对象指向第一目标对象的相似度作为第二节点指向第一节点的有向边的权重。

[0049] 可以理解的是，如果一个目标对象指向另一个目标对象的相似度为零，则说明这

两个目标对象之间不具有相同的关联对象，在该种情况下，有向网络图中表征这两个目标对象的节点之间可以不具有相连的有向边。

[0050] 需要说明的是，在本申请实施例的以上步骤S302和步骤S303，仅仅是基于任意两个目标对象之间关联的关联对象的相似程度，确定出有向网络图中表征两个目标对象的两个节点之间有向边的权重的一种实现方式，但是可以理解的是，体现目标对象之间关联的关联对象的相似程度的方式可以有多种，相应的，基于目标对象之间关联的关联对象的相似程度，确定表征目标对象的节点之间的有向边的权重的方式也可以有多种，在此不加以限制。

[0051] S304，为该有向网络图中的每个节点分别分配唯一的类别标识。

[0052] 其中，类别标识用于表征节点所归属的类别，由于在基于有向网络图进行聚类之前，不清楚哪些节点所表示的目标对象可以聚类为一个类别，因此，可以认为每一个节点分别属于一个类别，从而为每个节点分配唯一的类别标识。后续基于有向网络图进行聚类的过程中，部分或者全部节点的类别标识会不断发生变化，直至所有节点的类别标识均不发生变化时，则完成聚类。

[0053] S305，依次将有向网络图中的每个节点作为当前需处理的目标节点，并将目标节点的类别标识更新为该目标节点对应的至少一个入度节点组中，入度边总权重最大的目标入度节点组的类别标识，直至该有向网络图中所有节点的类别标识不再发生变化。

[0054] 其中，对于有向网络图中任意一个节点，有向网络图中指向该节点的其他节点称为该节点的入度节点，节点的入度节点也可以理解为有向网络图中，该节点的邻居节点。对于有向网络图中的任意一个节点而言，该节点的入度节点至少有一个。

[0055] 为了便于描述，本申请实施例将从入度节点指向该节点的有向边可以称为入度边，可见，一个入度节点对应着一条入度边。

[0056] 其中，入度节点组包括：有向边指向该目标节点且具有相同类别标识的至少一个入度节点，入度节点组的入度边总权重为该入度节点组中所有入度节点对应的入度边的权重总和。

[0057] 可以理解的是，由于入度边的权重表示该入度边对应的入度节点与该待更新节点的相似度，因此，如果入度边的权重越高，该入度边对应的入度节点所表示的目标对象与该目标节点对应的目标对象属于一个类别的可能性最大，相应的，如果一个入度节点组中所有入度边的权重之和最大，则该目标节点与该入度节点组中所有入度节点属于同一个类别的可能性最大，从而可以将该目标节点的类别标识与该入度边权重最大的目标入度节点组的类别标识进行统一，本实施例将类别标识统一为该目标入度节点组的类别标识。也就是说，需要从目标节点对应的多个入度节点组中，确定出有向边的总权重最大的目标入度节点组，并目标节点的类别标识更新为该目标入度节点组的类别标识。

[0058] 需要说明的是，该步骤S305为不断循环迭代的过程，每完成一次迭代，都需要判断本次迭代过程中是否存在类别标识发生变化的节点，如果存在类别标识发生变化的节点，则仍需要从向网络图中重新选取节点作为目标节点，并重新进行迭代。

[0059] 可以理解的是，如果完成最近一次迭代之后，在该最近一次迭代过程中，如果节点的类别标识没有发生变化，则说明该类别标识与其入度节点之间的聚类已经完成，即使后续再重新迭代，该节点的类别标识也不会发生变化。为了避免对这些类别标识没有发生变

化的节点的重复聚类，以减少数据处理量，可选的，可以仅仅在首次迭代时，将有向网络图中的所有节点均作为待处理节点；如果存在未处理的待处理节点，从未处理的待处理节点中选取当前需处理的目标节点，并将目标节点的类别标识更新为所述目标节点对应的至少一个入度节点组中，入度边总权重最大的目标入度节点组的类别标识，直至所有待处理节点均作为目标节点被处理为止；且，当所有待处理节点均作为目标节点之后，如果存在更新后的类别标识与更新前的类别标识不同的节点，则可以仅仅将更新前的类别标识与更新后的类别标识不同的节点确定为待处理节点，并返回执行如果存在未处理的待处理节点，从未处理的所述待处理节点中选取当前需处理的目标节点等操作；如果不存在更新后的类别标识与更新前的类别标识不同的节点，则表示聚类结束，可以执行后续步骤S306。

[0060] S306，将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别。

[0061] 可见，本申请实施例中，根据待聚类的任意两个目标对象之间的共同关联的关联对象的数量以及目标对象各自关联的对象的数量，计算目标对象之间的相似度之后，依据目标对象之间的相似度，构建出能够表示不同目标对象之相似度的有向网络图，并基于有向网络图中各个节点之间的相似度(权重)，对有向网络图中不同节点进行类别聚类，从而有利于从全局角度挖掘目标对象的相似度并对目标对象进行聚类，进而可以有效提高目标对象聚类的精准度。

[0062] 下面以待聚类的目标对象为用户设备以及文档的场景为例，对本申请实施例的对象聚类方法进行介绍。

[0063] 首先，以对用户设备的聚类为例进行介绍。结合图1和图2，参见图4，其示出了一种对象聚类方法一个实施例的流程示意图，本实施例的方法应用于如上所提到的计算机设备或者分布式计算系统中，本实施例的方法是以登录社交网络的多台用户设备进行聚类，以将同一用户的设备聚类到一起为例进行介绍。在本申请实施例的用户设备可以为手机、平板电脑、台式机等等终端。

[0064] 如图4，本实施例的方法可以包括：

[0065] S401，获取网络中的待分析数据集，该待分析数据集包括多个数据关系，每个数据关系中包括用户标识与用户设备的标识之间的对应关系。

[0066] 其中，每个数据关系中，用户标识表示访问(或者说登录)预设网络系统的用户的标识；用户设备的标识表示该用户访问该预设网络系统所采用的用户设备的标识。例如，数据关系可以表示为(用户U，用户设备Ue)的形式。

[0067] 其中，用户标识可以是用户在该预设网络系统中的用户名、用户的网络账号、用户的电话号码等等。用户设备的标识用户唯一标识一台用户设备，如该用户设备的标识可以为用户设备的IP地址、设备标识码等等。

[0068] 可以理解的是，由于进行聚类分析过程中，可以对针对一个或多个预设网络进行分析，因此，预先设定的预设网络可以是一个或多个，如，预设网络系统可以多个社交网络，例如多个不同的即时通讯系统、论坛系统等等。

[0069] 需要说明的是，虽然预设网络可以有多个，由于一个数据关系中是用户标识与用户设备的标识之间的对应关系，在该数据关系确定的情况下，该数据关系中的预设网络系统是确定的唯一的一个预设网络。

[0070] 举例说明，如，用户A以即时通讯用户U1的身份，通过手机M1登录即时通讯系统，则即时通讯用户U1的用户名与手机M1的标识构成一对数据关系；又如，用户A以即时通讯用户U2的身份，通过手机M2登录即时通讯系统，则会得到即时通讯用户U2的用户名与手机M2的标识所构成的一对数据关系；又如，用户B以即时通讯用户U1的身份，通过手机M2登录即时通讯系统，则又会得到即时通讯用户U1的用户名与手机M2的标识所构建的一对数据关系。

[0071] 可以理解的是，本申请实施例的目的是为了确定出哪些用户设备是同一个用户经常使用的设备，以将同一个用户所使用的用户设备聚类到一起，因此，该待分析数据集不包括完成相同的数据关系。

[0072] S402，对于任意两个不同用户设备的标识所表征的第一用户设备及第二用户设备，获取包含第一用户设备的标识的至少一个第一数据关系，以及包含第二用户设备的标识的至少一个第二数据关系。

[0073] 如，假设第一用户设备的用户设备的标识为Ue1，如果数据关系中包含Ue1，则该数据关系为包含第一用户设备的标识Ue1的第一数据关系。

[0074] 其中，第一数据关系与第二数据关系的数量可以不同。

[0075] 需要说明的是，本申请实施例仅仅是为了便于描述，而将任意两个用户设备中的一个称为第一用户设备，而将又一个用户设备称为第二应用设备，其中，第一用户设备与第二用户设备所具有的用户设备的标识不同。

[0076] S403，从该至少一个第一数据关系以及该至少一个第二数据关系中，确定出包含有相同用户标识的至少一对数据关系对。

[0077] 其中，每对数据关系对中包括：具有相同用户标识的第一数据关系以及第二数据关系。

[0078] 可以理解的是，一对数据关系对表示同一个用户既使用过第一用户设备，又使用过第二用户设备登录过预设网络系统。

[0079] 举例说明，假设第一用户设备的标识为Ue1，第二用户设备的标识为Ue2，如果第一数据关系为(userA,Ue1)，而第二数据关系为(userA,Ue2)，则说明该第一数据关系与第二数据关系为具有相同用户标识userA的一对数据关系对，同时可以说明用户标识为userA的用户使用过第一用户设备Ue1以及第二应用设备Ue2登录预设的网络系统。

[0080] S404，将数据关系对的总数量与该至少一个第一数据关系对应的第一数量的比值，确定为第二用户设备与第一用户设备的相似度。

[0081] 其中，数据关系对的总数量也就是同时使用过第一用户设备以及第二用户设备登录预设网络系统的用户的总数。

[0082] 为了便于区分，本申请实施例中，将第一数据关系的总个数称为第一数量，而将第二数据关系的总个数称为第二数量。其中，第一数量表示使用过或者说通过该第一用户设备登录预设网络系统的用户的总数；而第二数量表示使用过或者说通过该第二用户设备登录预设网络系统的用户的总数。

[0083] 具体的，该第二用户设备Ue2与第一用户设备Ue1的相似度W_{Ue2Ue1}可以表示为如下：

[0084]

$$W_{Ue2Ue1}=\frac{|N(Ue1)\cap N(Ue2)|}{|N(Ue1)|}$$ （公式一）；

[0085] 其中，|N(Ue1)∩N(Ue2)|表示同时使用过第一用户设备Ue1以及第二用户设备Ue2登录预设网络系统的用户的总数量，即数据关系对的总数量；而|N(Ue1)|表示使用过第一用户设备Ue1登录预设网络系统的用户的总数，即第一数据关系的第一数量。

[0086] S405，将数据关系对的总数量与该多个第二数据关系对应的第二数量的比值，确定为第一用户设备与该第二用户设备的相似度。

[0087] 具体的，该第一用户设备Ue1与第二用户设备Ue2的相似度W_{Ue1Ue2}可以表示为如下：

[0088]

$$W_{Ue1Ue2}=\frac{|N(Ue1)\cap N(Ue2)|}{|N(Ue2)|}$$ （公式二）；

[0089] 其中，|N(Ue1)∩N(Ue2)|为数据关系对的总数量；而|N(Ue2)|表示使用过第二用户设备Ue2登录预设网络系统的用户的总数，即第二数据关系的第二数量。

[0090] 需要说明的是，在同时使用过第一用户设备以及第二用户设备登录预设网络系统的用户的总数量，即数据关系对的总数量；使用过第一用户设备登录预设网络系统的用户的数量，即第一数据关系的第一数量；以及，使用过第二用户设备登录预设网络系统的用户的数量，即第二数据关系的第二数量确定的确定下，计算第二用户设备与第一用户设备的相似度，以及第一用户设备与第二用户设备的相似度的方式并不限于步骤S304以及步骤S305所示的方式，在实际应用中还可以有其他计算两个设备之间相似度的方式，在此不加以限制。

[0091] 可以理解的是，在本申请实施例中步骤S401和步骤S402仅仅为一种获取用户设备对应的用户标识的一种可选的实现方式，在实际应用中，该计算机设备或分布式计算系统也可以是直接获取待聚类的多个用户设备各自对应的用户标识。也就是说，获取到待聚类的每个用户设备各自对应的用户标识，每个用户设备可以对应一个或多个用户标识，用户设备所对应的用户标识表示通过该用户设备登录预设网络系统的用户的标识。如，对于待聚类的任意两个第一用户设备和第二用户设备，可以获取到通过第一用户设备登录社交网络所有用户的用户标识，以及获取到通过第二用户设备登录社交网络的所有用户的用户标识。

[0092] 相应的，通过第一用户设备登录社交网络的用户的第一数据可以为第一用户设备对应的用户标识的数量。通过第二用户设备登录社交网络的用户的第二数量可以为该第二用户设备对应的用户标识的数量。

[0093] 本实施例中，第二用户设备与第一用户设备的相似度相当于前面实施例中所提到的第一相似度，而第二用户设备与第一用户设备的相似度相对于前面实施例所提到的第二相似度。

[0094] S406，将第一用户设备以及第二用户设备作为有向网络图的节点，并将第一用户设备与第二用户设备的相似度作为有向网络图中由第一用户设备指向第二用户设备的边的权重，将第二用户设备与第一用户设备的相似度作为由第二用户设备指向第一用户设备

的边的权重。

[0095] 可以理解的是，有向网络图中任意两个节点之间的边均具有方向和权重，而且这两节点之间不同方向的边所具有的权重可以不同。如，对于有向网络图中任意两个节点：节点A与节点B，节点A指向节点B的边所具有的权重可以为权重1，而节点B指向节点A的边所具有的权重可以为权重2，权重1和权重2可以不同。

[0096] 在本申请实施例中，以将每一个用户设备均作为有向网络图中的节点，来构建有向网络图。对于任意两个用户设备，即第一用户设备和第二用户设备，在有向网络图中，该第一用户设备的节点指向该第二用户设备的节点的边(也称为有向边)所具有的权重表征该第一用户设备与第二用户设备的相似度；相应的，第二用户设备的节点指向该第一用户设备的节点的边(也称为有向边)所具有的权重表征该第二用户设备与第一用户设备的相似度。

[0097] 如，参见图5a，其示出了本申请实施例中构建出的一种有向网络图的部分结构示意图，在图5a中每个节点表示唯一的用户设备的标识所对应的用户设备，图2中每条边的上方所标出的数字为该边所对应的权重。由图5a可知，节点Ue1与节点Ue2之间具有两条不同指向的有向边，其中，由Ue1指向Ue2的有向边的权重为2/3，而由Ue2指向Ue1的有向边的权重2/5。

[0098] S407，将有向网络图中节点所表示的用户设备的标识初始化该节点的社区标识。

[0099] 社区标识用于表示节点所代表的用户设备所聚类到的社区，一个社区也可以认为一个聚类类别。

[0100] 如图5a所示，在该图5a的有向网络图中，每个节点均对应着一个社区标识，该社区标识为该节点对应的用户设备的标识，如图5a中，每个节点旁边表示有节点的社区标识，其中括号内的标识表示该节点对应的用户设备的标识。

[0101] 可以理解的是，节点对应的用户设备的标识作为节点对应的社区标识仅仅是一种实现方式。由于聚类之前，无法确定哪些用户设备可以聚类到一个社区，因此，只要是为每一个节点分配一个唯一的社区标识即可，因此，该步骤S407也可以是为每个节点分配一个该有向网络图中唯一的一个社区标识。

[0102] S408，依次将有向网络图中的每个节点作为待更新节点，并从与该待更新节点相连的多个入度节点中，确定出权重最大的入度边所对应的目标入度节点，将该目标入度节点的社区标识作为该待更新节点的社区标识。

[0103] 可以理解的是，由于入度边的权重越高，该入度边对应的入度节点对应的用户设备与该待更新节点对应的用户设备属于同一个用户经常使用的用户设备的可能性最大，因此，可以将该待更新节点与该权重最大的入度边所对应的目标入度节点划归到一个类别，即，待更新节点与该目标入度节点属于同一个社区。

[0104] 为了标识出属于同一个社区的节点(或者说用户设备)，需要将该目标入度节点与该待更新节点的社区标识进行统一。如，参见图5b，其示出了在图5a所示的有向网络图的基础上，将一个节点的社区标识更新为节点对应的权重最大的入度边所对应的目标入度节点的社区标识之后的示意图。如，对比图5a和图5b中用户设备Ue1对应的节点可知，在图5a中用户设备Ue1对应的节点的社区标识为Ue1，而由于该节点的入度节点中，表征用户设备Ue2的节点(当前的社区标识也为Ue2)指向该用户设备Ue1的节点的入度边的权重最大，因此，

将用户设备Ue1对应的节点的社区标识变更为Ue2。更新为该用户设备Ue1对应的节点的社区标识之后，可以依次对其他节点的社区标识进行更新。

[0105] 在本申请实施例中，是以将待更新节点的社区标识更新为与该待更新节点对应的目标入度节点的社区标识为例进行介绍，但是可以理解的是，如果将目标入度节点的社区标识更新为该待更新节点的社区标识也同样适用于本申请实施例。

[0106] S409，将有向网络图中所有节点作为待处理节点。

[0107] 在第一轮循环中，将有向网络图中的所有节点均作为待处理节点，以便依次更新有向网络图中各个节点的社区标识。

[0108] S410，从有向网络图的待处理节点中，选取一个未经处理的待处理节点作为当前需要处理的目标节点。

[0109] 其中，本申请实施例中依次对有向网络图中的每一个节点执行如下步骤S411至S412的操作，为了便于区分，将当前需要处理的节点称为目标节点，由于步骤S411以及S412是一个循环执行的过程，如果在本轮循环中，该有向网络图中的待处理节点已经作为目标节点，则在本轮处理中无需重复作为目标节点。

[0110] S411，将该目标节点的入度节点中，相同社区标识的入度节点确定为一个入度节点组，并针对每一个入度节点组，计算入度节点组中所有入度节点对应的入度边的权重的总和，得到每个入度节点组的入度边总权重；

[0111] 可以理解的是，经过步骤S408，一个目标节点的入度节点中，可能会存在两个或多个具有相同社区标识的入度节点，本申请实施例中，将具有相同社区标识的入度节点划归为一个入度节点组。当然，如果目标节点的一个入度节点的社区标识与其他入度节点的社区标识均不相同，则该入度节点可以单独归为一个入度节点组，且该入度节点组对应的入度边总权重就是该入度节点指向该目标节点的入度边的权重。可见，一个入度节点组中包括至少一个入度节点。

[0112] 为了确定出目标节点与哪些入度节点组内的入度节点属于同一用户经常使用的用户设备所对应的节点，需要计算入度节点组内所有入度节点对应的入度边的权重之和，在本申请实施例中，将入度节点组中所有入度节点对应的入度边的权重之和称为入度边总权重。其中，该入度边总权重反映出该入度节点组内所有入度节点与该目标节点的相似程度。

[0113] S412，将入度边总权重最大的一个入度节点组的社区标识作为该目标节点的社区标识。

[0114] 可以理解的是，由于同一个入度节点组内的入度节点实际上属于同一个社区(即聚类为一个类别)，而且，如果入度节点组的入度边总权重内最大，说明该目标节点与该入度节点组内所有入度节点的相似度最高，因此可以将该目标节点与该入度节点组内的入度节点聚类为一个社区，并将该目标节点的社区标识更新为该入度边总权重最大的入度节点组的社区标识。

[0115] S413，如果该目标节点的社区标识发生变化，则将目标节点记录为发生社区标识更新的节点。

[0116] 可以理解的是，如果入度边总权重最大的入度节点组的社区标识，与该目标节点更新前的社区标识不同，则将入度边总权重最大的入度节点组的社区标识作为该目标节点

的社区标识之后，该目标节点的社区标识发生更新，则需要标记出该目标节点，以便后续将该目标节点作为下一轮需要更新的待处理节点。

[0117] 可选的，在步骤S410选取出该目标节点时，可以将该目标节点的社区标识作为更新前社区标识，并将步骤S412中对该目标节点更新后的社区标识作为更新后社区标识，进而比较该目标节点的更新前社区标识与更新后社区标识是否一致，如果不一致，则说明该目标节点的社区标识发生变化。

[0118] 需要说明的是，步骤S413为可选步骤，在实际应用中也可以在将本轮中所有待处理节点均作为目标节点进行处理之后，再确定这些目标节点中是否存在社区标识发生更新的节点，即可以直接在后续步骤S415中直接判断是否存在社区标识发生更新的目标节点。

[0119] S414，检测有向网络图中是否存在未作为目标节点的待处理节点，如果是，则返回步骤S410；如果否，则执行步骤415；

[0120] S415，判断有向网络图中是否存在社区标识发生更新的节点，如果是，将社区标识发生更新的节点作为有向网络图中的待处理节点，并返回执行步骤S410；如果否，则将具有相同社区标识的节点所对应的用户设备确定为属于同一个社区，得到聚类出的多个社区。

[0121] 其中，属于同一个社区的用户设备可以认为是：聚类出的属于同一个用户经常使用的用户设备集合。

[0122] 可以理解的是，如果有向网络图中每个节点的社区标识均更新为与该节点的相似度最大的入度节点组所对应的社区标识，那么再经过步骤S410至步骤S414的重复迭代，该有向网络图中每个节点的社区标识也不会再发生变化，因此，如果有向网络图中不存在社区标识发生更新的节点时，则说明完成对有向网络图中所有节点的聚类，在该种情况下，具有相同社区标识的节点被聚类为一个社区。

[0123] 需要说明的是，在本申请实施例中，步骤S408为可选步骤，其仅仅是考虑到刚构建出的有向网络图中，每个节点的社区标识都是唯一的，对应一个节点而言，该节点的入度节点中，不存在两个或多个社区标识相同的入度节点，因此，为了便于理解，而直接将权重最大的入度边所对应的目标入度节点的社区标识作为待更新节点的社区标识。但是可以理解的是，如果没有该步骤S408，而直接执行步骤S409至415的操作同样也是可行的，在该种情况下，在第一轮循环时，由于每个节点的社区标识都是唯一的，因此可以认为是每一个入度节点属于一个入度节点组，这样，入度边总权重最大的入度节点组实际上也就是权重最大的入度边所对应的目标入度节点。

[0124] 下面以对多篇包含有不同社群标识的文档进行主题发现为例，对本申请实施例的对象聚类方法进行介绍。

[0125] 结合图1和图2，参见图6，其示出了本申请一种对象聚类方法又一个实施例的流程示意图，本实施例的方法适用于如上所提到的计算机设备或者分布式计算系统中，本实施例的方法可以包括

[0126] S601，获取网络中的待分析数据集，该待分析数据集包括多个数据关系，每个数据关系中包括：社群标识与文档的标识之间的对应关系。

[0127] 其中，每个数据关系中，社群标识表示一些预设网络系统的标识，如社交网络的标识，例如，即时通讯系统相关的标识；文档的标识表示从网络中获取到的用于唯一表示该文档的标识。

[0128] 可以理解的是，本申请实施例的目的是为了确定出哪些文档是同一个社群网络经常使用的文档，以将同一个社群经常使用到的文档聚类到一起，以提取出该文档的主题。

[0129] S602，对于任意两个不同文档的标识所表征的第一文档及第二文档，获取包含第一文档的标识的至少一个第三数据关系，以及包含第二文档的标识的至少一个第四数据关系。

[0130] S603，从该至少一个第三数据关系以及该至少一个第四数据关系中，确定出包含有相同社群标识的至少一对数据关系对。

[0131] 其中，每对数据关系对中包括：具有相同社群标识的第三数据关系以及第四数据关系。

[0132] 可以理解的是，一对数据关系对表示同一个社群既使用过第一文档，又使用过第二文档。

[0133] S604，将数据关系对的总数量与该至少一个第三数据关系对应的第三数量的比值，确定为第二文档与第一文档的相似度。

[0134] 其中，数据关系对的总数量也就是同时使用过第一文档以及第二文档的社群的总数。

[0135] 为了便于区分，本申请实施例中，将第三数据关系的总个数称为第三数量，而将第四数据关系的总个数称为第四数量。其中，第三数量表示使用过第一文档的社群的总数；而第四数量表示使用过第四文档的社群的总数。

[0136] S605，将数据关系对的总数量与该多个第四数据关系对应的第四数量的比值，确定为第一文档与该第二文档的相似度。

[0137] 其中，计算第二文档与第一文档的相似度，以及第一文档与第二文档的相似度的具体方式可以参见前面实施例的相关介绍，在此不再赘述。

[0138] S606，将第一文档以及第二文档作为有向网络图的节点，并将第一文档与第二文档的相似度作为有向网络图中由第一文档指向第二文档的边的权重，将第二文档与第一文档的相似度作为由第二文档指向第一文档的边的权重。

[0139] S607，将有向网络图中节点所表示的文档的标识初始化该节点的主题标识。

[0140] 主题标识用于表示节点所代表的文档所对应的主题，一个主题也可以认为一个聚类类别。

[0141] S608，依次将有向网络图中的每个节点作为待更新节点，并从与该待更新节点相连的多个入度节点中，确定出权重最大的入度边所对应的目标入度节点，将该目标入度节点的主题标识作为该待更新节点的主题标识。

[0142] S609，将有向网络图中所有节点作为待处理节点。

[0143] S610，从有向网络图的待处理节点中，选取一个未经处理的待处理节点作为当前需要处理的目标节点。

[0144] S611，将该目标节点的入度节点中，相同主题标识的入度节点确定为一个入度节点组，并针对每一个入度节点组，计算入度节点组中所有入度节点对应的入度边的权重的总和，得到每个入度节点组的入度边总权重。

[0145] S612，将入度边总权重最大的一个入度节点组的主题标识作为该目标节点的主题标识。

[0146] S613,如果该目标节点的主题标识发生变化,则将目标节点记录为发生主题标识更新的节点。

[0147] S614,检测有向网络图中是否存在未作为目标节点的待处理节点,如果是,则返回步骤S610;如果否,则执行步骤615;

[0148] S615,判断有向网络图中是否存在主题区标识发生更新的节点,如果是,将主题标识发生更新的节点作为有向网络图中的待处理节点,并返回执行步骤S610;如果否,则将具有相同主题标识的节点所对应的文档确定为属于同一个主题,得到聚类出的多个主题。

[0149] 另一方面,本申请实施例还提供了一种对象聚类装置。

[0150] 参见图7,其示出了本申请一种对象聚类装置一个实施例的组成结构示意图,本实施例的装置可以包括:

[0151] 数据获取单元701,用于获取待聚类的多个目标对象各自关联的关联对象集合,所述关联对象集合包括至少一个关联对象;

[0152] 有向图构建单元702,用于基于任意两个目标对象之间关联的关联对象的相似程度,确定待构建的有向网络图中表征所述两个目标对象的两个节点之间有向边的权重,并构建所述有向网络图;

[0153] 类别初始化单元703,用于为所述有向网络图中的每个节点分别分配唯一的类别标识;

[0154] 聚类分析单元704,用于依次将每个所述节点作为当前需处理的目标节点,从所述目标节点对应的至少一个入度节点组中,确定出指向该目标节点的有向边的总权重最大的目标入度节点组,并所述节点的类别标识更新为所述目标入度节点组的目标标识,直至所述有向网络图中所有节点的类别标识不再发生变化,其中,所述入度节点组包括有向边指向所述目标节点且具有相同类别标识的至少一个入度节点;

[0155] 类别提取单元705,用于将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别,得到该多个目标对象对应的多个聚类类别。

[0156] 可选的,所述有向图构建单元,包括:

[0157] 第一权重确定单元,用于对于所述多个目标对象中任意的第一目标对象以及第二目标对象,根据与所述第一目标对象和所述第二目标对象都关联的关联对象的总数量,以及所述第一目标对象关联的关联对象的第一数量,确定出待构建的有向网络图中,从表示所述第二目标对象的第二节点指向表示所述第一目标对象的第一节点的有向边的权重;

[0158] 第二权重确定单元,用于根据所述总数量,以及所述第二目标对象关联的关联对象的第二数量,确定出从所述第一节点指向所述第二节点的有向边的权重;

[0159] 有向图构建子单元,用于构建所述有向网络图。

[0160] 进一步的,所述第一权重确定单元,具体用于,将与所述第一目标对象以及所述第二目标对象都关联的关联对象的总数量,与所述第一对象关联的关联对象的第一数量的比值,确定为从表示所述第二目标对象的第二节点指向表示所述第一目标对象的第一节点的有向边的权重;

[0161] 所述第二权重确定单元,具体用于将所述总数量,与所述第二对象关联的关联对象的第二数量的比值,确定为从所述第一节点指向所述第二节点的有向边的权重。

[0162] 可选的,数据获取单元,具体为,用于获取待聚类的多个目标对象各自对应的至少

一个数据关系，所述数据关系包括所述目标对象的标识与所述目标对象关联的关联对象之间的对应关系；

[0163] 所述类别初始化单元，包括：

[0164] 类别初始化子单元，用于将有向网络图中每个节点表示的目标对象的标识作为所述节点的类别标识。

[0165] 可选的，所述聚类分析单元，包括：

[0166] 节点初始处理单元，用于将所述有向网络图中的所有节点均作为待处理节点；

[0167] 聚类分析子单元，用于如果存在未处理的待处理节点，从未处理的所述待处理节点中选取当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所有待处理节点均作为目标节点被处理为止；

[0168] 循环控制单元，用于如果存在更新后的类别标识与更新前的类别标识不同的节点，则将更新前的类别标识与更新后的类别标识不同的节点确定为待处理节点，并触发返回执行所述聚类分析子单元的操作；

[0169] 所述类别提取单元具体为，用于如果不存在更新后的类别标识与更新前的类别标识不同的节点，则将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别。

[0170] 可选的，所述数据获取单元具体为，用于获取待聚类的多个用户设备各自关联的用户标识集合，所述用户标识集合包括多个用户标识，其中，用户设备关联的用户标识为通过所述用户设备访问预设网络的用户的标识。

[0171] 本申请实施例还提供了一种计算机设备，该计算机设备可以包括上述所述的任一种对象聚类装置。该计算机设备的组成结构可以参见图1所示，在本申请实施例中的计算机设备中，该存储器中所存储的程序具体用于：

[0172] 获取待聚类的多个目标对象各自关联的关联对象集合，所述关联对象集合包括至少一个关联对象；

[0173] 基于任意两个目标对象之间关联的关联对象的相似程度，确定待构建的有向网络图中表征所述两个目标对象的两个节点之间有向边的权重，并构建所述有向网络图；为所述有向网络图中的每个节点分别分配唯一的类别标识；

[0174] 依次将每个所述节点作为当前需处理的目标节点，从所述目标节点对应的至少一个入度节点组中，确定出指向该目标节点的有向边的总权重最大的目标入度节点组，并所述节点的类别标识更新为所述目标入度节点组的目标标识，直至所述有向网络图中所有节点的类别标识不再发生变化，其中，所述入度节点组包括有向边指向所述目标节点且具有相同类别标识的至少一个入度节点；

[0175] 将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别。

[0176] 需要说明的是，本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。对于装置类实施例而言，由于其与方法实施例基本相似，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。

[0177] 最后，还需要说明的是，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。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0178] 对所公开的实施例的上述说明，使本领域技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

[0179] 以上仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

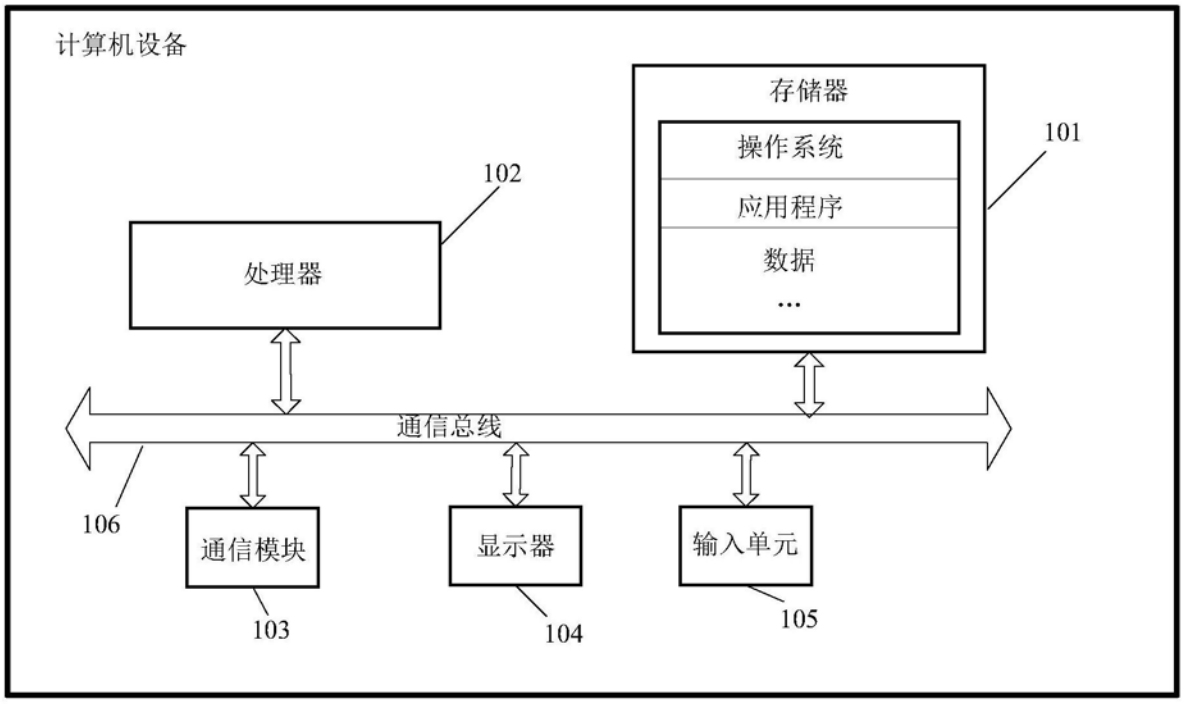

计算机设备
存储器
操作系统
应用程序
数据
...
101
102
处理器
通信总线
106
通信模块
显示器
输入单元
103
104
105

图1

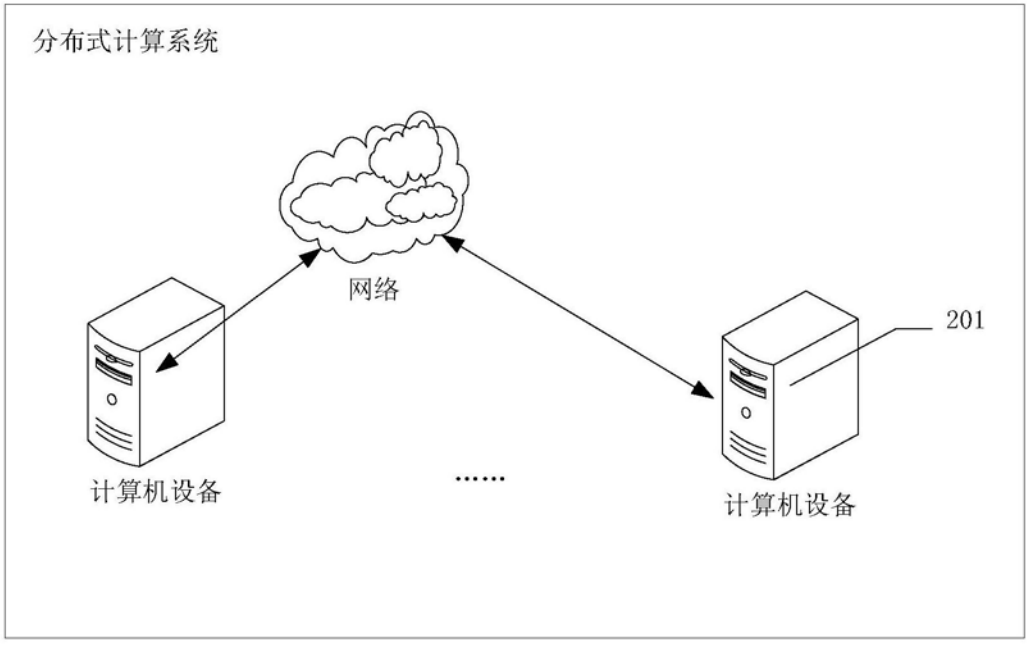

分布式计算系统
网络
201
计算机设备
.....
计算机设备

图2

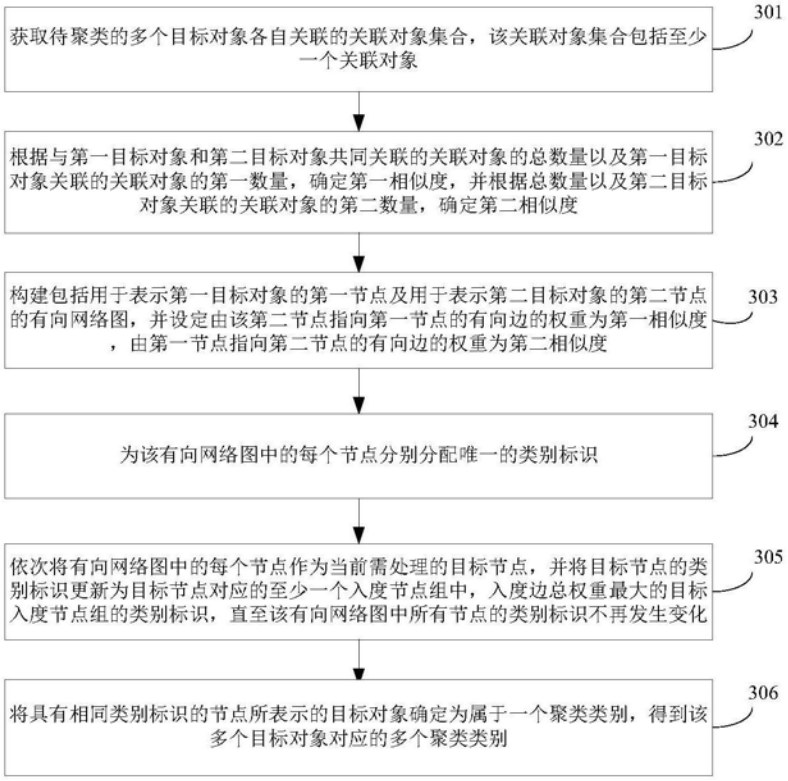

301
获取待聚类的多个目标对象各自关联的关联对象集合，该关联对象集合包括至少一个关联对象
302
根据与第一目标对象和第二目标对象共同关联的关联对象的总数量以及第一目标对象关联的关联对象的第一数量，确定第一相似度，并根据总数量以及第二目标对象关联的关联对象的第二数量，确定第二相似度
303
构建包括用于表示第一目标对象的第一节点及用于表示第二目标对象的第二节点的有向网络图，并设定由该第二节点指向第一节点的有向边的权重为第一相似度，由第一节点指向第二节点的有向边的权重为第二相似度
304
为该有向网络图中的每个节点分别分配唯一的类别标识
305
依次将有向网络图中的每个节点作为当前需处理的目标节点，并将目标节点的类别标识更新为目标节点对应的至少一个入度节点组中，入度边总权重最大的目标入度节点组的类别标识，直至该有向网络图中所有节点的类别标识不再发生变化
306
将具有相同类别标识的节点所表示的目标对象确定为属于一个聚类类别，得到该多个目标对象对应的多个聚类类别

图3

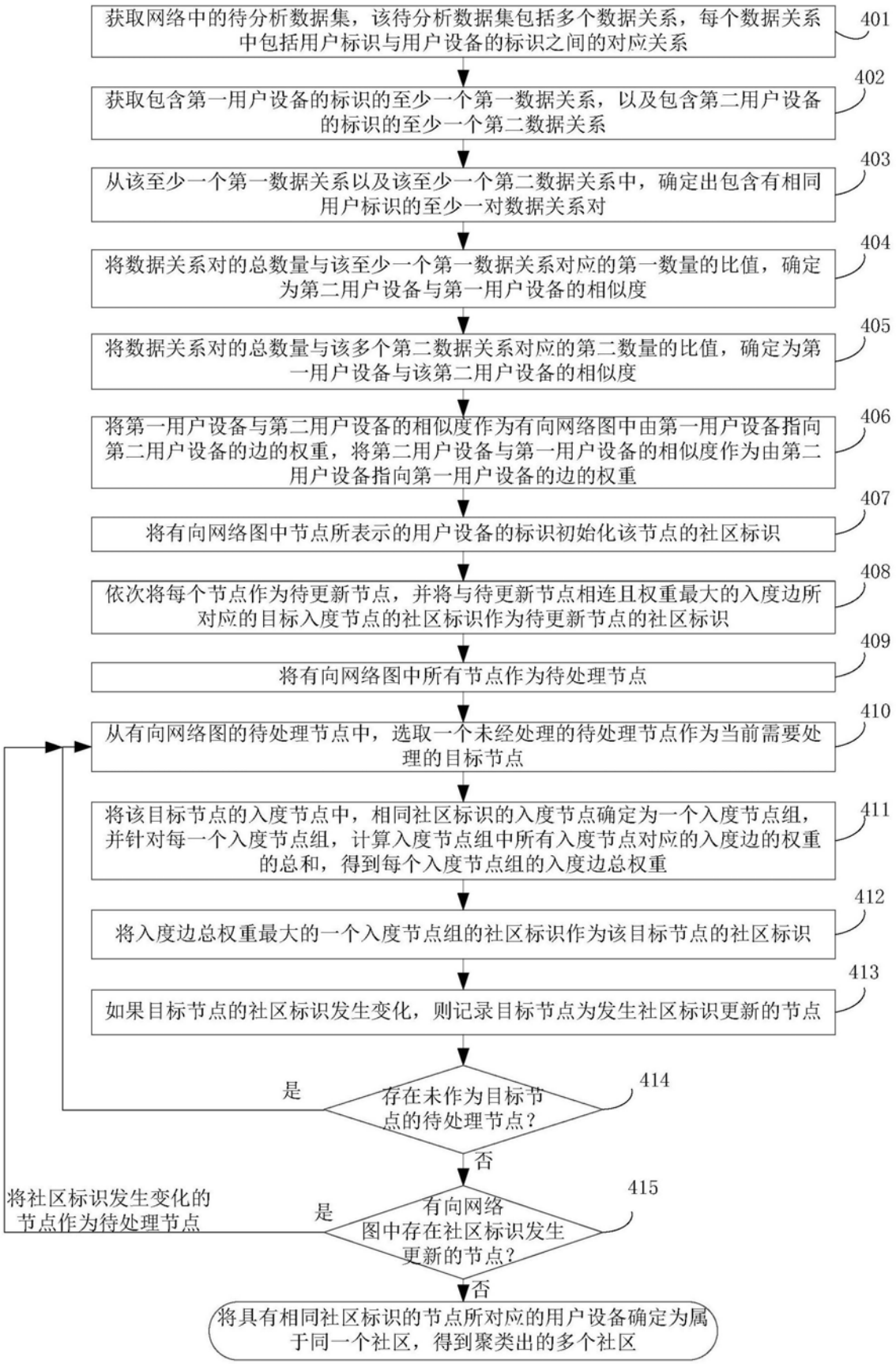

获取网络中的待分析数据集，该待分析数据集包括多个数据关系，每个数据关系中包括用户标识与用户设备的标识之间的对应关系
401
获取包含第一用户设备的标识的至少一个第一数据关系，以及包含第二用户设备的标识的至少一个第二数据关系
402
从该至少一个第一数据关系以及该至少一个第二数据关系中，确定出包含有相同用户标识的至少一对数据关系对
403
将数据关系对的总数量与该至少一个第一数据关系对应的第一数量的比值，确定为第二用户设备与第一用户设备的相似度
404
将数据关系对的总数量与该多个第二数据关系对应的第二数量的比值，确定为第一用户设备与该第二用户设备的相似度
405
将第一用户设备与第二用户设备的相似度作为有向网络图中由第一用户设备指向第二用户设备的边的权重，将第二用户设备与第一用户设备的相似度作为由第二用户设备指向第一用户设备的边的权重
406
将有向网络图中节点所表示的用户设备的标识初始化该节点的社区标识
407
依次将每个节点作为待更新节点，并将与待更新节点相连且权重最大的入度边所对应的目标入度节点的社区标识作为待更新节点的社区标识
408
将有向网络图中所有节点作为待处理节点
409
从有向网络图的待处理节点中，选取一个未经处理的待处理节点作为当前需要处理的目标节点
410
将该目标节点的入度节点中，相同社区标识的入度节点确定为一个入度节点组，并针对每一个入度节点组，计算入度节点组中所有入度节点对应的入度边的权重的总和，得到每个入度节点组的入度边总权重
411
将入度边总权重最大的一个入度节点组的社区标识作为该目标节点的社区标识
412
如果目标节点的社区标识发生变化，则记录目标节点为发生社区标识更新的节点
413
存在未作为目标节点的待处理节点？
414
是
否
有向网络图中存在社区标识发生更新的节点？
415
是
将社区标识发生变化的节点作为待处理节点
否
将具有相同社区标识的节点所对应的用户设备确定为属于同一个社区，得到聚类出的多个社区

图4

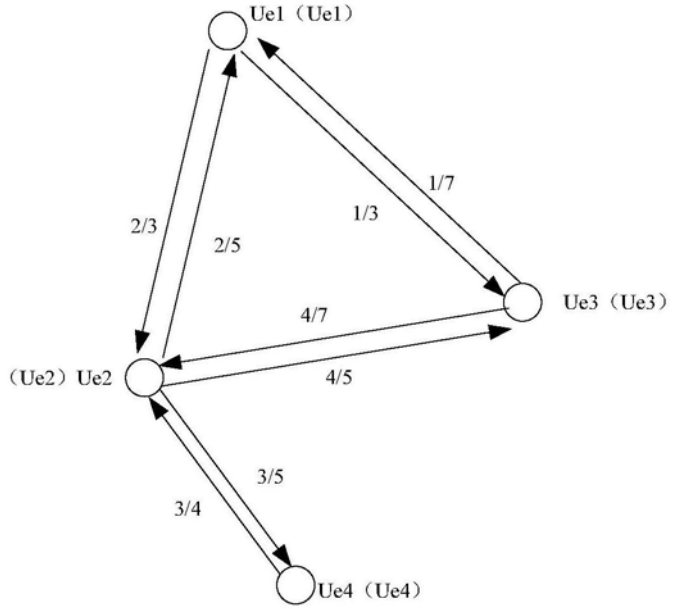

Ue1（Ue1）
1/7
1/3
2/3
2/5
Ue3（Ue3）
4/7
（Ue2）Ue2
4/5
3/5
3/4
Ue4（Ue4）

图5a

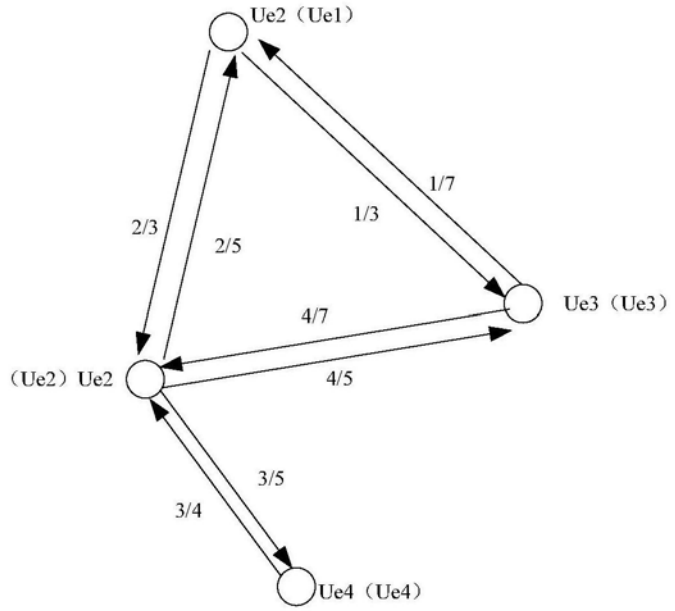

Ue2（Ue1）
1/7
1/3
2/3
2/5
Ue3（Ue3）
4/7
（Ue2）Ue2
4/5
3/5
3/4
Ue4（Ue4）

图5b

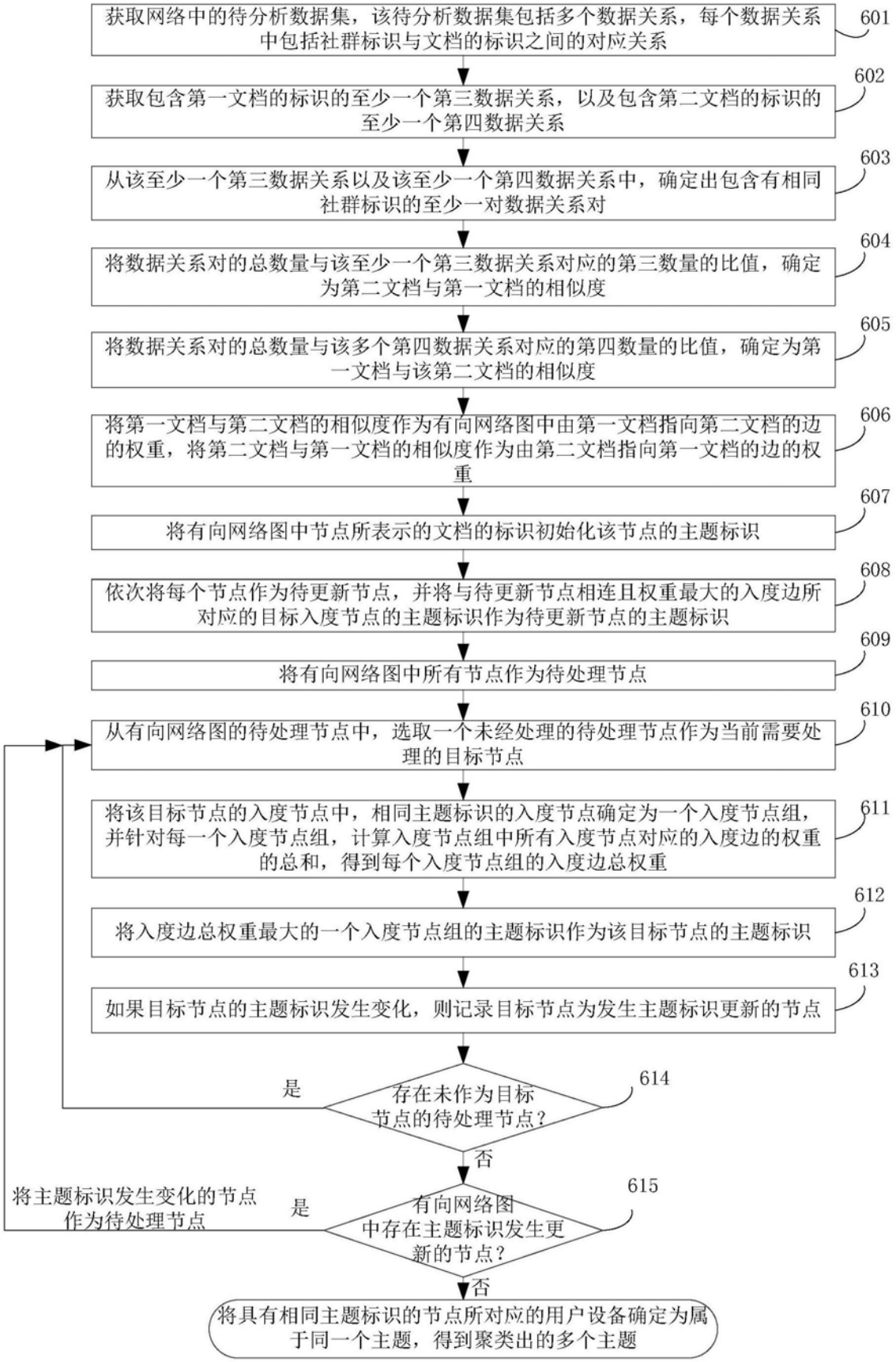

获取网络中的待分析数据集，该待分析数据集包括多个数据关系，每个数据关系中包括社群标识与文档的标识之间的对应关系
601
获取包含第一文档的标识的至少一个第三数据关系，以及包含第二文档的标识的至少一个第四数据关系
602
从该至少一个第三数据关系以及该至少一个第四数据关系中，确定出包含有相同社群标识的至少一对数据关系对
603
将数据关系对的总数量与该至少一个第三数据关系对应的第三数量的比值，确定为第二文档与第一文档的相似度
604
将数据关系对的总数量与该多个第四数据关系对应的第四数量的比值，确定为第一文档与该第二文档的相似度
605
将第一文档与第二文档的相似度作为有向网络图中由第一文档指向第二文档的边的权重，将第二文档与第一文档的相似度作为由第二文档指向第一文档的边的权重
606
将有向网络图中节点所表示的文档的标识初始化该节点的主题标识
607
依次将每个节点作为待更新节点，并将与待更新节点相连且权重最大的入度边所对应的目标入度节点的主题标识作为待更新节点的主题标识
608
将有向网络图中所有节点作为待处理节点
609
从有向网络图的待处理节点中，选取一个未经处理的待处理节点作为当前需要处理的目标节点
610
将该目标节点的入度节点中，相同主题标识的入度节点确定为一个入度节点组，并针对每一个入度节点组，计算入度节点组中所有入度节点对应的入度边的权重的总和，得到每个入度节点组的入度边总权重
611
将入度边总权重最大的一个入度节点组的主题标识作为该目标节点的主题标识
612
如果目标节点的主题标识发生变化，则记录目标节点为发生主题标识更新的节点
613
存在未作为目标节点的待处理节点？
614
是
否
有向网络图中存在主题标识发生更新的节点？
615
是
将主题标识发生变化的节点作为待处理节点
否
将具有相同主题标识的节点所对应的用户设备确定为属于同一个主题，得到聚类出的多个主题

图6

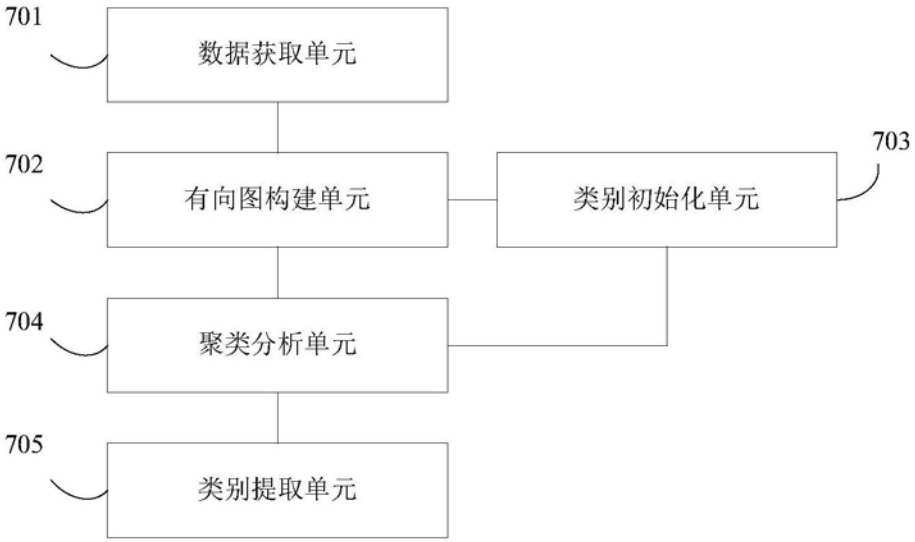


图7