(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 110674413 B
(45)授权公告日 2022.03.25

(21)申请号 201910857778.7

(22)申请日 2019.09.09

(65)同一申请的已公布的文献号
申请公布号 CN 110674413 A

(43)申请公布日 2020.01.10

(73)专利权人 平安科技(深圳)有限公司
地址 518033 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 邓强 张娟 屠宁 赵之砚 施奕明

(74)专利代理机构 北京市京大律师事务所 11321
代理人 刘挽澜

(51)Int.Cl.
G06F 16/9536(2019.01)
G06V 10/762(2022.01)
G06K 9/62(2022.01)

(56)对比文件
CN 109951377 A,2019.06.28
US 2019050898 A1,2019.02.14
CN 109190033 A,2019.01.11
EP 2778960 A1,2014.09.17
祝周等.基于多维混合图和核心节点的社团发现算法.《网络空间安全》.2019,(第02期),

审查员 陈丽娜

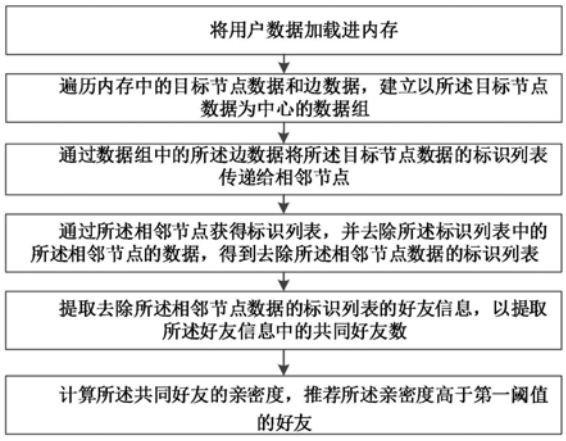
权利要求书3页 说明书10页 附图2页

(54)发明名称

用户关系挖掘方法、装置、设备和存储介质

(57)摘要

本申请涉及数据分析领域,提供了用户关系挖掘的方法、装置、设备和存储介质,方法包括:将用户数据加载进内存。遍历内存中的目标节点数据和边数据,建立以所述目标节点数据为中心的数据组。利用数据组中的边数据将目标节点数据的标识列表传递给相邻节点。通过相邻节点获得标识列表,并去除所述标识列表中的相邻节点的数据,得到去除相邻节点数据的标识列表。提取去除相邻节点数据的标识列表的好友信息,以提取好友信息中的共同好友。计算共同好友的亲密度,推荐所述亲密度高于第一阈值的好友。本申请提供了一种模型而避免了将节点以及节点属性复制多次带来的存储资源的浪费。根据简单的数学公式获得亲密度关系,使得计算的繁杂度减少。



CN 110674413 B

1.一种用户关系挖掘的方法,其特征在于,所述方法包括:

将用户数据加载进内存;所述用户数据包括节点数据以及边数据;所述节点数据用于记录节点数据的属性;所述节点数据包括好友信息;所述边数据是指边的属性,用于记录人与人之间的关系;

遍历内存中的目标节点数据和边数据,建立以所述目标节点数据为中心的数据组;所述目标节点数据为多个节点数据中的任一节点数据;

利用数据组中的所述边数据将所述目标节点数据的标识列表传递给相邻节点;所述标识列表用于存储所有相邻节点的数据;

通过所述相邻节点获得标识列表,并去除所述标识列表中的所述相邻节点的数据,得到去除所述相邻节点数据的标识列表;所述去除所述相邻节点数据的标识列表是指所述目标节点的二度关系;所述二度关系是指两个节点数据中间隔一个节点数据;

提取去除所述相邻节点数据的标识列表的好友信息,以提取所述好友信息中的共同好友;好友信息的共同好友数通过以下数学公式进行执行:

$$Score(x,y)=\frac{|Neighbor(i)\cap Neighbor(j)|}{|Neighbor(i)\cup Neighbor(j)|}$$

其中,Neighbor(i)表示第i个所述目标节点的好友;Score表示共同好友的数量;

计算所述共同好友的亲密度,推荐所述亲密度高于第一阈值的好友;所述亲密度计算公式通过以下数学公式进行执行:

$$v(fof)=\sum_{f_i}\frac{(\delta_{u,f_i}*\delta_{f_i,fof})^{0.3}}{\sqrt{friends_{f_i}}}$$

其中,$\delta_{u,fi}$为u与fi好友之间的所述亲密度,$\delta_{fi,fof}$为fi与fof建立好友之间的所述亲密度,0.3为惩罚因子,$friends_{fi}$为fi的好友数量。

2.根据权利要求1所述的方法,其特征在于,所述用户数据在用户进行运算之前,所述方法还包括:

收集完成的用户数据,每个所述节点数据都有唯一标识号ID,并存储在外存储器上;所述用户数据是收集完成的用户数据的一部分。

3.根据权利要求2所述的方法,其特征在于,所述存储在外存储器上,包括:

将数据库中的用户数据储存到文本文件;

根据所述文本文件生成SparkRDD,在进行计算时将所述SparkRDD转换成数据文件;

通过所述数据文件将数据读入Spark,使用GraphX进行图计算。

4.根据权利要求1所述的方法,其特征在于,所述提取所述好友信息中的共同好友之后,所述方法还包括:

采集GraphX所处的网络环境的资源,以生成资源集合;

根据预设n维属性对所述资源集合的资源进行属性标记,以生成新的资源集合;

通过提取函数对所述新的资源集合进行特征提取,以得到特征向量,并获取初始样本空间;

通过参数自助法得到K值,并运用K均值聚类对所述初始样本空间进行分类,以将所述初始样本空间分为K类资源;

以及，将所述K类资源归入到每一类的聚类中心所对应的资源图谱类中，以完资源图谱。

5.根据权利要求4所述的方法，其特征在于，所述通过参数自助法方法得到K值，包括：

将所述K值取一个预设值，并通过K-means方法得到K种类的统计值，并且获取统计量的模型；

通过所述统计量的模型的生成数据样本集合；

获取估计聚类好坏的指标，从K+1开始起，每次递增1，以逐一模拟生成的样本的聚类的总体类内误差WSS；

在所述聚类的WSS满足预设条件时，接受K+1类，且后面依次增加k，直到不满足所述预设条件，以确定所述K值。

6.根据权利要求4所述的方法，其特征在于，所述运用K均值聚类对所述初始样本空间进行分类，包括：

从所述初始样本空间中任意选择K个特征向量作为初始聚类中心；

获取所述初始样本空间中的其他对象与所述聚类中心的距离；

将每个类别中的所有对象所对应的均值作为类别的聚类中心，并获取目标函数的值，以更新所述聚类中心，直到更新后的聚类中心与前聚类中心相等或差值小于预设阈值。

7.根据权利要求1所述的方法，其特征在于，所述将用户数据加载进内存之前，所述方法还包括：

利用结构化查询语言选取所述信息，同一监测点位的所述用户数据作为一组，进行重复的所述用户数据查找，并删除相同属性的重复所述用户数据；

或者，通过三倍标准差法确定上限值与下限值，根据所述上限值和下限值构造所述用户数据范围，若所述用户数据不落在所述范围，则判断所述用户数据为异常值并进行剔除。

8.一种用户关系挖掘的装置，其特征在于，所述装置包括：

输入输出模块，将用户数据加载进内存；所述用户数据包括节点数据以及边数据；所述节点数据用于记录节点数据的属性；所述节点数据包括好友信息；所述边数据是指边的属性，用于记录人与人之间的关系；

处理模块，遍历内存中的目标节点数据和边数据，建立以所述目标节点数据为中心的数据组；所述目标节点数据为多个节点数据中的任一节点数据；利用数据组中的所述边数据将所述目标节点数据的标识列表传递给相邻节点；所述标识列表用于存储所有相邻节点的数据；通过所述相邻节点获得标识列表，并去除所述标识列表中的所述相邻节点的数据，得到去除所述相邻节点数据的标识列表；所述去除所述相邻节点数据的标识列表是指所述目标节点的二度关系；所述二度关系是指两个节点数据中间隔一个节点数据；提取去除所述相邻节点数据的标识列表的好友信息，以提取所述好友信息中的共同好友；好友信息的共同好友数通过以下数学公式进行执行：

$$Score(x,y)=\frac{|Neighbor(i)\cap Neighbor(j)|}{|Neighbor(i)\cup Neighbor(j)|}$$

其中，Neighbor(i)表示第i个所述目标节点的好友；Score表示共同好友的数量；

计算所述共同好友的亲密度，推荐所述亲密度高于第一阈值的好友；所述亲密度计算

公式通过以下数学公式进行执行：

$$v(fof)=\sum_{f_i}\frac{(\delta_{u,f_i}*\delta_{f_i,fof})^{0.3}}{\sqrt{friends_{f_i}}}$$

其中，$\delta_{u,fi}$为u与fi好友之间的所述亲密度，$\delta_{fi,fof}$为fi与fof建立好友之间的所述亲密度，0.3为惩罚因子，$friends_{fi}$为fi的好友数量。

9.一种计算机设备，其特征在于，所述计算机设备包括：

至少一个处理器、存储器和输入输出单元；

其中，所述存储器用于存储程序代码，所述处理器用于调用所述存储器中存储的程序代码来执行如权利要求1-7中任一项所述的方法。

10.一种计算机存储介质，其特征在于，其包括指令，当其在计算机上运行时，使得计算机执行如权利要求1-7中任一项所述的方法。

用户关系挖掘方法、装置、设备和存储介质

技术领域

[0001] 本申请涉及数据分析领域，尤其涉及一种用户关系挖掘的方法、装置、设备和存储介质。

背景技术

[0002] 在社交领域中，图数据挖掘是关系挖掘和群体画像中的重要方法。图数据由节点数据和边组成，图中的节点数据用于表示发生连接的主体，边用来表示主体之间的关联，边越密集，边权重越大，表示关联越强。目前图计算的典型环境是Spark项目中的GraphX环境，其核心是对Pregel图计算模型的实现。

[0003] 图数据主要由节点数据属性和边属性构成。在社交网络的图数据中，节点数据属性的量远远超过边属性。现有的GraphX计算模型将图数据拆分成节点-边-节点模式的以边为中心的数据组，和多条边相关联的某个节点会在每一条边的数据组中复制，导致节点数据的大量冗余存储，为计算带来大量的资源消耗。同时GraphX中的参数配置缺乏灵活性，出现计算瓶颈时难以寻找解决方案，使得就算获得很好的数据，也无法做到比较好的推荐，严重限制了其可用性。

发明内容

[0004] 本申请提供了一种通过配置用户关系挖掘的方法，能够解决现有技术中计算带来大量的资源消耗的问题。

[0005] 第一方面，本申请提供一种用户关系挖掘的方法，包括：

[0006] 将用户数据加载进内存。所述用户数据包括节点数据以及边数据。所述节点数据用于记录节点数据的属性。所述节点数据至少包括好友信息。所述边数据是指边的属性，用于记录人与人之间的关系。

[0007] 遍历内存中的目标节点数据和边数据，建立以所述目标节点数据为中心的数据组。所述目标节点数据为所述多个节点数据中的任一节点数据。

[0008] 利用数据组中的所述边数据将所述目标节点数据的标识列表传递给相邻节点。所述标识列表用于存储所有相邻节点的数据。

[0009] 通过所述相邻节点获得标识列表，并去除所述标识列表中的所述相邻节点的数据，得到去除所述相邻节点数据的标识列表。所述去除所述相邻节点数据的标识列表是指所述目标节点的二度关系。所述两度关系是指二度关系是指两个节点数据中间隔一个节点数据。

[0010] 提取去除所述相邻节点数据的标识列表的好友信息，以提取所述好友信息中的共同好友。所述提取好友信息的共同好友数通过以下数学公式进行执行：

[0011] $$Score(x,y)=\frac{|Neighbor(i)\cap Neighbor(j)|}{|Neighbor(i)\cup Neighbor(j)|}$$

[0012] 其中，Neighbor(i)表示第i个所述目标节点的好友。Score表示共同好好的数量。

[0013] 计算所述共同好友的亲密度，推荐所述亲密度高于第一阈值的好友。所述亲密度计算公式通过以下数学公式进行执行：

[0014] $$v(fof)=\sum_{fi}\frac{(\delta_{u,fi}*\delta_{fi,of})^{0.3}}{\sqrt{friends_{fi}}}$$

i

[0015] 其中，$\delta_{u,fi}$为u与fi好友之间的所述亲密度，$\delta_{u,fi}$为fi与fof建立好友之间的所述亲密度，0.3为惩罚因子。

[0016] 相较于现有技术，本申请提供了一种计算模型。处理的基本单元是节点及其相连的所有边，使用节点-边的以节点为中心的基本单元，当某个节点收集其边属性时，仅需要对包含所述节点的基本单元进行操作。由于基本单元保留了节点相连的所有边，避免了边遍历以及其带来的大量聚合操作。所提方法避免了GraphX中使用节点-边-节点的以边为中心的基本单元，从而避免了将节点以及节点属性复制多次带来的大量存储资源的浪费。根据简单的数学公式获得亲密度关系，使得计算的繁杂度减少。

[0017] 在一些可能的设计中，所述用户数据在用户进行运算之前，所述方法还包括：

[0018] 收集完成的用户数据，每个所述节点数据都有唯一标识号ID，并存储在外存储器上；所述用户数据是收集完成的用户数据的一部分。

[0019] 在一些可能的设计中，所述并存储在外存储器上，所述方法还包括：

[0020] 将数据库中的用户数据储存到文本文件。

[0021] 根据所述文本文件生成SparkRDD，在进行计算时将所述SparkRDD转换成数据文件。

[0022] 通过所述数据文件将数据读入Spark，利于GraphX进行图计算。

[0023] 在一些可能的设计中，所述提取所述好友信息中的共同好友之后，所述方法还包括：

[0024] 采集所述GraphX所处的网络环境的资源，以生成资源集合。

[0025] 根据预设n维属性对所述资源集合的资源进行属性标记，以生成新的资源集合。

[0026] 通过提取函数对所述新的资源集合进行特征提取，以得到特征向量，并获取初始样本空间。

[0027] 通过参数自助法得到K值，并运用K均值聚类对所述初始样本空间进行分类，以将所述初始样本空间分为K类资源。

[0028] 以及，将所述K类资源归入到每一类的聚类中心所对应的资源图谱类中，以完资源图谱。

[0029] 在一些可能的设计中，所述通过参数自助法方法得到K值，包括：

[0030] 将所述K值取一个预设值，并通过K-means方法得到K种类的统计值，并且获取统计量的模型。

[0031] 通过所述统计量的模型的生成数据样本集合。

[0032] 获取估计聚类好坏的指标，从K+1开始起，每次递增1，以逐一模拟生成的样本的聚类的总体类内误差WSS。

[0033] 在所述聚类的WSS满足预设条件时，接受K+1类，且后面依次增加k，直到不满足所述预设条件，以确定所述K值。

[0034] 在一些可能的设计中，所述运用K均值聚类对所述初始样本空间进行分类，包括：

[0035] 从所述初始样本空间中任意选择K个特征向量作为初始聚类中心。

[0036] 获取所述初始样本空间中的其他对象与所述聚类中心的距离。

[0037] 将每个类别中的所有对象所对应的均值作为类别的聚类中心，并获取目标函数的值，以更新所述聚类中心，直到更新后的聚类中心与前聚类中心相等或差值小于预设阈值。

[0038] 在一些可能的设计中，所述将用户数据加载进内存之前，所述方法还包括：

[0039] 利用结构化查询语言选取所述信息，同一监测点位的所述用户数据作为一组，进行重复的所述用户数据查找，并删除相同属性的重复所述用户数据。

[0040] 或者，通过三倍标准差法确定上限值与下限值，根据所述上限值和下限值构造所述用户数据范围，若所述用户数据不落在所述范围，则判断所述用户数据为异常值并进行剔除。

[0041] 第二方面，本申请提供一种用户关系挖掘的装置，具有实现对应于上述第一方面提供的用户关系挖掘的平台的方法的功能。所述功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。硬件或软件包括一个或多个与上述功能相对应的模块，所述模块可以是软件和/或硬件。

[0042] 所述用户关系挖掘装置包括：

[0043] 输入输出模块，用于将用户数据加载进内存。所述用户数据包括节点数据以及边数据。所述节点数据用于记录节点数据的属性。所述节点数据至少包括好友信息。所述边数据是指边的属性，用于记录人与人之间的关系。

[0044] 处理模块，用于遍历内存中的目标节点数据和边数据，建立以所述目标节点数据为中心的数据组。所述目标节点数据为所述多个节点数据中的任一节点数据。

[0045] 利用数据组中的所述边数据将所述目标节点数据的标识列表传递给相邻节点。所述标识列表用于存储所有相邻节点的数据。

[0046] 通过所述相邻节点获得标识列表，并去除所述标识列表中的所述相邻节点的数据，得到去除所述相邻节点数据的标识列表。所述去除所述相邻节点数据的标识列表是指所述目标节点的二度关系。所述两度关系是指二度关系是指两个节点数据中间隔一个节点数据。

[0047] 提取去除所述相邻节点数据的标识列表的好友信息，以提取所述好友信息中的共同好友。所述提取好友信息的共同好友数通过以下数学公式进行执行：

[0048] $$Score(x,y)=\frac{|Neighbor(i)\cap Neighbor(j)|}{|Neighbor(i)\cup Neighbor(j)|}$$

[0049] 其中，Neighbor(i)表示第i个所述目标节点的好友。Score表示共同好好的数量。

[0050] 计算所述共同好友的亲密度，推荐所述亲密度高于第一阈值的好友。所述亲密度计算公式通过以下数学公式进行执行：

[0051] $$v(fof)=\sum_{fi}\frac{(\delta_{u,fi}*\delta_{fi,of})^{0.3}}{\sqrt{friends_{fi}}}$$

i

[0052] 其中，δ_{u,f_i}为u与fi好友之间的所述亲密度，δ_{u,f_i}为fi与fof建立好友之间的所述

亲密度,0.3为惩罚因子。

[0053] 在一些可能的设计中,所述处理模块还用于:

[0054] 收集完成的用户数据,每个所述节点数据都有唯一标识号ID,并存储在外存储器上。

[0055] 在一些可能的设计中,所述处理模块还用于:

[0056] 将数据库中的用户数据储存到文本文件。

[0057] 根据所述文本文件生成SparkRDD,在进行计算时将所述SparkRDD转换成数据文件。

[0058] 通过所述数据文件将数据读入Spark,利于GraphX进行图计算。

[0059] 在一些可能的设计中,所述处理模块还用于:

[0060] 采集所述GraphX所处的网络环境的资源,以生成资源集合。

[0061] 根据预设n维属性对所述资源集合的资源进行属性标记,以生成新的资源集合。

[0062] 通过提取函数对所述新的资源集合进行特征提取,以得到特征向量,并获取初始样本空间。

[0063] 通过参数自助法得到K值,并运用K均值聚类对所述初始样本空间进行分类,以将所述初始样本空间分为K类资源。

[0064] 以及,将所述K类资源归入到每一类的聚类中心所对应的资源图谱类中,以完资源图谱。

[0065] 在一些可能的设计中,所述处理模块还用于:

[0066] 将所述K值取一个预设值,并通过K-means方法得到K种类的统计值,并且获取统计量的模型。

[0067] 通过所述统计量的模型的生成数据样本集合。

[0068] 获取估计聚类好坏的指标,从K+1开始起,每次递增1,以逐一模拟生成的样本的聚类的总体类内误差WSS。

[0069] 在所述聚类的WSS满足预设条件时,接受K+1类,且后面依次增加k,直到不满足所述预设条件,以确定所述K值。

[0070] 在一些可能的设计中,所述处理模块还用于:

[0071] 从所述初始样本空间中任意选择K个特征向量作为初始聚类中心。

[0072] 获取所述初始样本空间中的其他对象与所述聚类中心的距离。

[0073] 将每个类别中的所有对象所对应的均值作为类别的聚类中心,并获取目标函数的值,以更新所述聚类中心,直到更新后的聚类中心与前聚类中心相等或差值小于预设阈值。

[0074] 在一些可能的设计中,所述处理模块还用于:

[0075] 利用结构化查询语言选取所述信息,同一监测点位的所述用户数据作为一组,进行重复的所述用户数据查找,并删除相同属性的重复所述用户数据。

[0076] 或者,通过三倍标准差法确定上限值与下限值,根据所述上限值和下限值构造所述用户数据范围,若所述用户数据不落在所述范围,则判断所述用户数据为异常值并进行剔除。

[0077] 本申请又一方面提供了一种创建贷款页面的设备,其包括至少一个连接的处理器、存储器、输入输出单元,其中,所述存储器用于存储程序代码,所述处理器用于调用所述

存储器中的程序代码来执行上述各方面所述的方法。

[0078] 本申请又一方面提供了一种计算机存储介质，其包括指令，当其在计算机上运行时，使得计算机执行上述各方面所述的方法。

附图说明

[0079] 图1为本申请实施例中用户关系挖掘的方法的流程示意图；

[0080] 图2为本申请实施例中用户关系挖掘的装置的结构示意图；

[0081] 图3为本申请实施例中计算机设备的结构示意图。

[0082] 本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0083] 应当理解，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或模块的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或模块，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或模块，本申请中所出现的模块的划分，仅仅是一种逻辑上的划分，实际应用中实现时可以有另外的划分方式，例如多个模块可以结合成或集成在另一个系统中，或一些特征可以忽略，或不执行。

[0084] 请参照图1，以下对本申请提供一种用户关系挖掘的方法进行举例说明，所述方法包括：

[0085] 101、将用户数据加载进内存。

[0086] 所述用户数据包括节点数据以及边数据。所述节点数据用于记录节点数据的属性。所述节点数据至少包括好友信息。所述边数据是指边的属性，用于记录人与人之间的关系。

[0087] 所述边数据至少包括亲属关系，朋友关系以及公司部门关系。所述节点数据包括身份证号、手机号、性别、好友信息、年龄以及爱好。

[0088] 102、遍历内存中的目标节点数据和边数据，建立以所述目标节点数据为中心的数据组。

[0089] 所述目标节点数据为所述多个节点数据中的任一节点数据。

[0090] 解决拷贝通过GraphX建立以边为中心的数据单元，所述数据单元包含所述边数据以及所述边数据关联的两个节点数据。因此GraphX保存的数据单元中，同一个节点数据会出现在不同边为中心的数据组中，从而造成节点数据的冗余存储，占用额外资源。

[0091] 103、利用数据组中的所述边数据将所述目标节点数据的标识列表传递给相邻节点。

[0092] 所述标识列表用于存储所有相邻节点的数据。

[0093] 所述传递通过边数据进行传递。例如a的边记录了与之相邻的节点数据，a有三个

边,分别指向b,c,d三个节点数据。因此,通过遍历a的三个边,即可将[b,c,d]这个列表传递到a的三个相邻节点数据上。

[0094] 104、通过所述相邻节点获得标识列表,并去除所述标识列表中的所述相邻节点的数据,得到去除所述相邻节点数据的标识列表。

[0095] 所述去除所述相邻节点数据的标识列表是指所述目标节点的二度关系。所述两度关系是指二度关系是指两个节点数据中间隔一个节点数据。

[0096] a的标识列表是[b,c,d],将[b,c,d]传递给b节点和、c节点和d节点,以传递给b节点为例,去除标识列表中的b节点信息标识列表变成了[c,d],因此b节点与c和d节点形成二度关系。同理若传递给c节点,则去除标识列表中的c节点信息,标识列表变成了[b、d],因此c节点与b、d节点形成二度关系。

[0097] 在本例中,通过使用以节点为中心的数据组,对列表数据进行了高效的传递,且避免了节点数据的复制。相比之下,GraphX以边为中心的数据组处理方式不仅造成节点数据的复制,也造成节点收集的数据标识列表的复制,造成大量数据冗余,浪费存储资源。

[0098] 105、提取去除所述相邻节点数据的标识列表的好友信息,以提取所述好友信息中的共同好友。

[0099] 所述提取好友信息的共同好友数通过以下数学公式进行执行:

[0100] $$Score(x,y)=\frac{|Neighbor(i)\cap Neighbor(j)|}{|Neighbor(i)\cup Neighbor(j)|}$$

[0101] 其中,Neighbor(i)表示第i个所述目标节点的好友。Score表示共同好好的数量。

[0102] 相当于对每个共同好友一视同仁,都贡献1分,但是共同好友中,有些人好友多,有些好友少,当某个共同好友的所述好友数较少时,这个共同好友应该更加重要,所以可以通过除以每个共同好友的所述好友数进行加权。

[0103] 如果所述好友数相差过大,需要通过开方、对数等方式进行处理。具体如下:

[0104] $$Score(x,y)=\sum_{k\in Neighbor(i)\cap Neighbor(j)}\frac{1}{\sqrt{|Neighbor(k)|}}$$

[0105] $$Score(x,y)=\sum_{k\in Neighbor(i)\cap Neighbor(j)}\frac{1}{\log_2|Neighbor(k)|}$$

[0106] 106、计算所述共同好友的亲密度,推荐所述亲密度高于第一阈值的好友。

[0107] 所述亲密度计算公式通过以下数学公式进行执行:

[0108] $$v(fof)=\sum_{fi}\frac{(\delta_{u,fi}*\delta_{fi,of})^{0.3}}{\sqrt{friends_{fi}}}$$

[0109] 其中,$\delta_{u,fi}$为u与fi好友之间的所述亲密度,$\delta_{u,fi}$为fi与fof建立好友之间的所述亲密度,0.3为惩罚因子。

[0110] 所述亲密度相差越大,权重越小。惩罚因子可以视情况进行调整。通过增加增加了亲密度特征$\delta_{u,fi}*\delta_{fi,of}$,来使得推荐的是认识的人概率更大。

[0111] 一些实施方式中,所述用户数据在用户进行运算之前,所述方法还包括:收集完成

的用户数据,每个所述节点数据都有唯一标识号ID,并存储在外存储器上;所述用户数据是收集完成的用户数据的一部分。

[0112] 一些实施方式中,所述存储在外存储器上,所述方法还包括:

[0113] 将数据库中的用户数据储存到文本文件。

[0114] 根据所述文本文件生成SparkRDD,在进行计算时将所述SparkRDD转换成数据文件。

[0115] 通过所述数据文件将数据读入Spark,利于GraphX进行图计算。

[0116] 所述文本文件可以是excel,txt,json等格式的文件,用于保存获取到的数据库的用户数据。

[0117] 例如将保存在MySQL中的元数据导出到txt文件中。文件信息保存在txt中,则可以通过SparkContext读取所述txt文件生成SparkRDD,并通过转化格式的接口将SparkRDD转换成DataFrame,方便下一步计算。

[0118] 一些实施方式中,所述提取所述好友信息中的共同好友之后,所述方法还包括:

[0119] 对所述共同好友采用kmeans进行聚类操作;所述对共同好友采用kmeans进行聚类操作包括:

[0120] 采集所述GraphX所处的网络环境的资源,以生成资源集合。

[0121] 根据预设n维属性对所述资源集合的资源进行属性标记,以生成新的资源集合。

[0122] 通过提取函数对所述新的资源集合进行特征提取,以得到特征向量,并获取初始样本空间。

[0123] 通过参数自助法得到K值,并运用K均值聚类对所述初始样本空间进行分类,以将所述初始样本空间分为K类资源。

[0124] 以及,将所述K类资源归入到每一类的聚类中心所对应的资源图谱类中,以完资源图谱。

[0125] 将相似好友信息的好友进行聚类,认为他们的各项信息相似度较高,推荐出来的朋友更容易认识,有更多的类似的兴趣爱好等,解决无法推荐较为满意的好友的问题。

[0126] 一些实施方式中,所述通过参数自助法方法得到K值,包括:

[0127] 将所述K值取一个预设值,并通过K-means方法得到K种类的统计值,并且获取统计量的模型。

[0128] 通过所述统计量的模型的生成数据样本集合。

[0129] 获取估计聚类好坏的指标,从K+1开始起,每次递增1,以逐一模拟生成的样本的聚类的总体类内误差WSS。

[0130] 在所述聚类的WSS满足预设条件时,接受K+1类,且后面依次增加k,直到不满足所述预设条件,以确定所述K值。

[0131] 获得一个相对较好的聚类K值,可以获得更好的聚类结果。

[0132] 一些实施方式中,所述运用K均值聚类对所述初始样本空间进行分类,包括:

[0133] 从所述初始样本空间中任意选择K个特征向量作为初始聚类中心。

[0134] 获取所述初始样本空间中的其他对象与所述聚类中心的距离。

[0135] 将每个类别中的所有对象所对应的均值作为类别的聚类中心,并获取目标函数的值,以更新所述聚类中心,直到更新后的聚类中心与前聚类中心相等或差值小于预设阈值。

[0136] 随机选取聚类中心,以防止刚开始的聚类中心过近的问题。

[0137] 一些实施方式中,所述将用户数据加载进内存之前,所述方法还包括:

[0138] 对获取到的所述用户数据进行数据清洗;所述对获取到的所述用户数据进行数据清洗包括:

[0139] 利用结构化查询语言选取所述信息,同一监测点位的所述用户数据作为一组,进行重复的所述用户数据查找,并删除相同属性的重复所述用户数据。

[0140] 或者,通过三倍标准差法确定上限值与下限值,根据所述上限值和下限值构造所述用户数据范围,若所述用户数据不落在所述范围,则判断所述用户数据为异常值并进行剔除。

[0141] 在数据处理之前将异常的数据进行剔除,以防止错误数据对模型的干扰。

[0142] 如图2所示的一种用户关系挖掘的装置20的结构示意图,其可应用于用户关系挖掘。本申请实施例中的用户关系挖掘的装置能够实现对应于上述图1所对应的实施例中所执行的用户关系挖掘的方法的步骤。用户关系挖掘的装置20实现的功能可以通过硬件实现,也可以通过硬件执行相应的软件实现。硬件或软件包括一个或多个与上述功能相对应的模块,所述模块可以是软件和/或硬件。所述用户关系挖掘的装置可包括输入输出模块201和处理模块202,所述处理模块202和输入输出模块201的功能实现可参考图1所对应的实施例中所执行的操作,此处不作赘述。输入输出模块201可用于控制所述输入输出模块201的输入、输出以及获取操作。

[0143] 一些实施方式中,所述输入输出模块201可用于将用户数据加载进内存。所述用户数据包括节点数据以及边数据。所述节点数据用于记录节点数据的属性。所述节点数据至少包括好友信息。所述边数据是指边的属性,用于记录人与人之间的关系。

[0144] 所述处理模块202可用于遍历内存中的目标节点数据和边数据,建立以所述目标节点数据为中心的数据组。所述目标节点数据为所述多个节点数据中的任一节点数据。

[0145] 利用数据组中的所述边数据将所述目标节点数据的标识列表传递给相邻节点。所述标识列表用于存储所有相邻节点的数据。

[0146] 通过所述相邻节点获得标识列表,并去除所述标识列表中的所述相邻节点的数据,得到去除所述相邻节点数据的标识列表。所述去除所述相邻节点数据的标识列表是指所述目标节点的二度关系。所述两度关系是指二度关系是指两个节点数据中间隔一个节点数据。

[0147] 提取去除所述相邻节点数据的标识列表的好友信息,以提取所述好友信息中的共同好友。所述提取好友信息的共同好友数通过以下数学公式进行执行:

[0148] $$Score(x,y)=\frac{|Neighbor(i)\cap Neighbor(j)|}{|Neighbor(i)\cup Neighbor(j)|}$$

[0149] 其中,Neighbor(i)表示第i个所述目标节点的好友。Score表示共同好好的数量。

[0150] 计算所述共同好友的亲密度,推荐所述亲密度高于第一阈值的好友。所述亲密度计算公式通过以下数学公式进行执行:

[0151] $$v(fof)=\sum_{fi}\frac{(\delta_{u,fi}*\delta_{fi,of})^{0.3}}{\sqrt{friends_{fi}}}$$ $_i$

[0152] 其中，δ_{u,f_i}为u与fi好友之间的所述亲密度，δ_{u,f_i}为fi与fof建立好友之间的所述亲密度，0.3为惩罚因子。

[0153] 一些实施方式中，所述处理模块202还用于：

[0154] 收集完成的用户数据，每个所述节点数据都有唯一标识号ID，并存储在外存储器上。

[0155] 一些实施方式中，所述处理模块202还用于：

[0156] 将数据库中的用户数据储存到文本文件。

[0157] 根据所述文本文件生成SparkRDD，在进行计算时将所述SparkRDD转换成数据文件。

[0158] 通过所述数据文件将数据读入Spark，利于GraphX进行图计算。

[0159] 一些实施方式中，所述处理模块202还用于：

[0160] 采集所述GraphX所处的网络环境的资源，以生成资源集合。

[0161] 根据预设n维属性对所述资源集合的资源进行属性标记，以生成新的资源集合。

[0162] 通过提取函数对所述新的资源集合进行特征提取，以得到特征向量，并获取初始样本空间。

[0163] 通过参数自助法得到K值，并运用K均值聚类对所述初始样本空间进行分类，以将所述初始样本空间分为K类资源。

[0164] 以及，将所述K类资源归入到每一类的聚类中心所对应的资源图谱类中，以完资源图谱。

[0165] 一些实施方式中，所述处理模块202还用于：

[0166] 将所述K值取一个预设值，并通过K-means方法得到K种类的统计值，并且获取统计量的模型。

[0167] 通过所述统计量的模型的生成数据样本集合。

[0168] 获取估计聚类好坏的指标，从K+1开始起，每次递增1，以逐一模拟生成的样本的聚类的总体类内误差WSS。

[0169] 在所述聚类的WSS满足预设条件时，接受K+1类，且后面依次增加k，直到不满足所述预设条件，以确定所述K值。

[0170] 一些实施方式中，所述处理模块202还用于：

[0171] 从所述初始样本空间中任意选择K个特征向量作为初始聚类中心。

[0172] 获取所述初始样本空间中的其他对象与所述聚类中心的距离。

[0173] 将每个类别中的所有对象所对应的均值作为类别的聚类中心，并获取目标函数的值，以更新所述聚类中心，直到更新后的聚类中心与前聚类中心相等或差值小于预设阈值。

[0174] 一些实施方式中，所述处理模块202还用于：

[0175] 利用结构化查询语言选取所述信息，同一监测点位的所述用户数据作为一组，进行重复的所述用户数据查找，并删除相同属性的重复所述用户数据。

[0176] 或者，通过三倍标准差法确定上限值与下限值，根据所述上限值和下限值构造所述用户数据范围，若所述用户数据不落在所述范围，则判断所述用户数据为异常值并进行剔除。

[0177] 上面从模块化功能实体的角度分别介绍了本申请实施例中的创建装置，以下从硬

件角度介绍一种计算机设备，如图3所示，其包括：处理器、存储器、输入输出单元(也可以是收发器，图3中未标识出)以及存储在所述存储器中并可在所述处理器上运行的计算机程序。例如，该计算机程序可以为图1所对应的实施例中用户关系挖掘的方法对应的程序。例如，当计算机设备实现如图2所示的用户关系挖掘的装置20的功能时，所述处理器执行所述计算机程序时实现上述图2所对应的实施例中由用户关系挖掘的装置20执行的用户关系挖掘的方法中的各步骤。或者，所述处理器执行所述计算机程序时实现上述图2所对应的实施例的用户关系挖掘的装置20中各模块的功能。又例如，该计算机程序可以为图1所对应的实施例中用户关系挖掘的方法对应的程序。

[0178] 所称处理器可以是中央处理单元(Central ProcessingUnit，CPU)，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，所述处理器是所述计算机装置的控制中心，利用各种接口和线路连接整个计算机装置的各个部分。

[0179] 所述存储器可用于存储所述计算机程序和/或模块，所述处理器通过运行或执行存储在所述存储器内的计算机程序和/或模块，以及调用存储在存储器内的数据，实现所述计算机装置的各种功能。所述存储器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(比如声音播放功能、图像播放功能等)等；存储数据区可存储根据手机的使用所创建的数据(比如音频数据、视频数据等)等。此外，存储器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，例如硬盘、内存、插接式硬盘，智能存储卡(SmartMedia Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)、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。

[0180] 所述输入输出单元也可以用接收器和发送器代替，可以为相同或者不同的物理实体。为相同的物理实体时，可以统称为输入输出单元。该输入输出可以为收发器。

[0181] 所述存储器可以集成在所述处理器中，也可以与所述处理器分开设置。

[0182] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质(如ROM/RAM)中，包括若干指令用以使得一台终端(可以是手机，计算机，服务器或者网络设备等)执行本申请各个实施例所述的方法。

[0183] 上面结合附图对本申请的实施例进行了描述，但是本申请并不局限于上述的具体实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本申请的启示下，在不脱离本申请宗旨和权利要求所保护的范围情况下，还可做出很多形式，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，这些均属于本申请的保护之内。

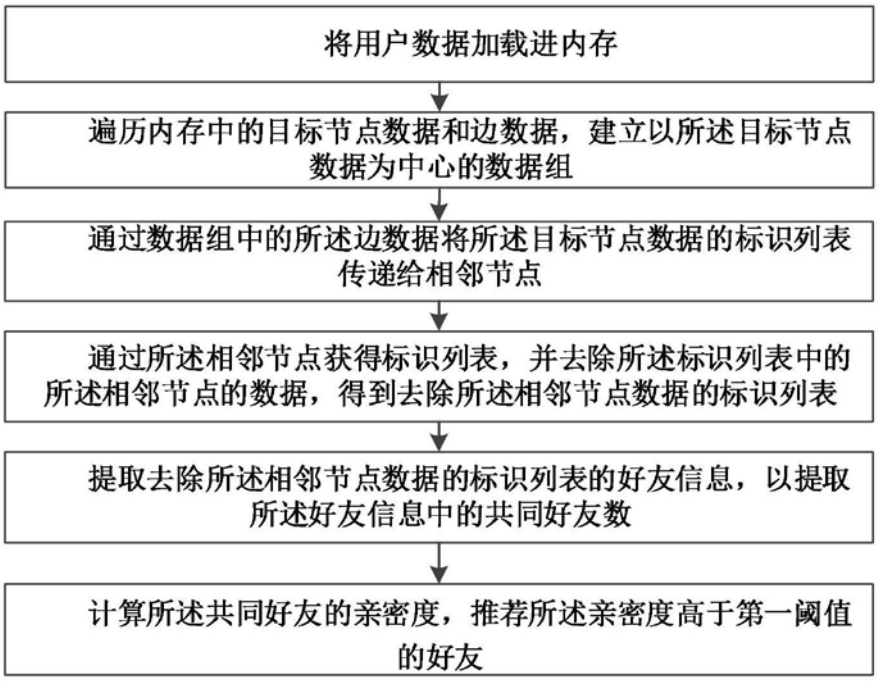

将用户数据加载进内存
遍历内存中的目标节点数据和边数据，建立以所述目标节点数据为中心的数据组
通过数据组中的所述边数据将所述目标节点数据的标识列表传递给相邻节点
通过所述相邻节点获得标识列表，并去除所述标识列表中的所述相邻节点的数据，得到去除所述相邻节点数据的标识列表
提取去除所述相邻节点数据的标识列表的好友信息，以提取所述好友信息中的共同好友数
计算所述共同好友的亲密度，推荐所述亲密度高于第一阈值的好友

图1

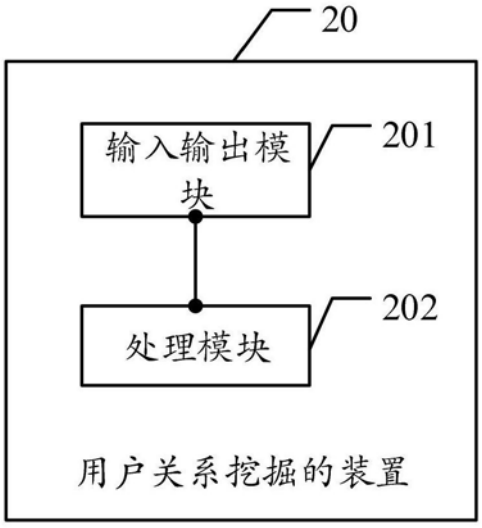

20
输入输出模块
201
处理模块
202
用户关系挖掘的装置

图2

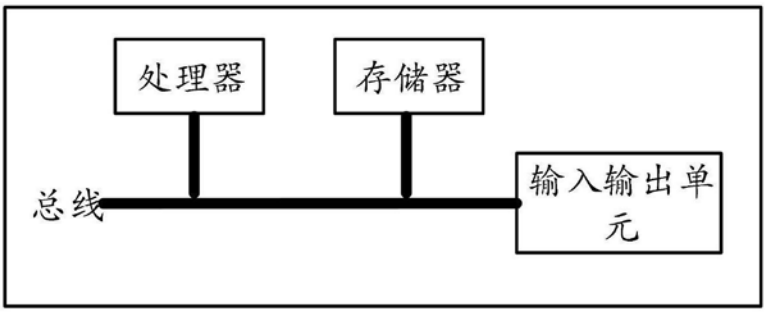


图3