

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 10 月 8 日 (08.10.2020)



(10) 国际公布号
WO 2020/199745 A1

(51) 国际专利分类号:
G06F 16/28 (2019.01)

(21) 国际申请号: PCT/CN2020/073872

(22) 国际申请日: 2020 年 1 月 22 日 (22.01.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910250748.X 2019年3月29日 (29.03.2019) CN

(71) 申请人: 创新先进技术有限公司 (ADVANCED NEW TECHNOLOGIES CO., LTD.) [—/CN]; 开曼群岛大开曼岛乔治镇医院路27号开曼企业中心, GD CAYMAN KY1-9008 (KY)。

(72) 发明人: 曹绍升 (CAO, Shaosheng); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 崔卿 (CUI, Qing); 中国浙江省杭州市余杭区文一西路969号3号楼5楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街9号嘉华大厦B座409, Beijing 100085 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: SAMPLE CLUSTERING METHOD AND DEVICE

(54) 发明名称: 一种样本聚类方法及装置

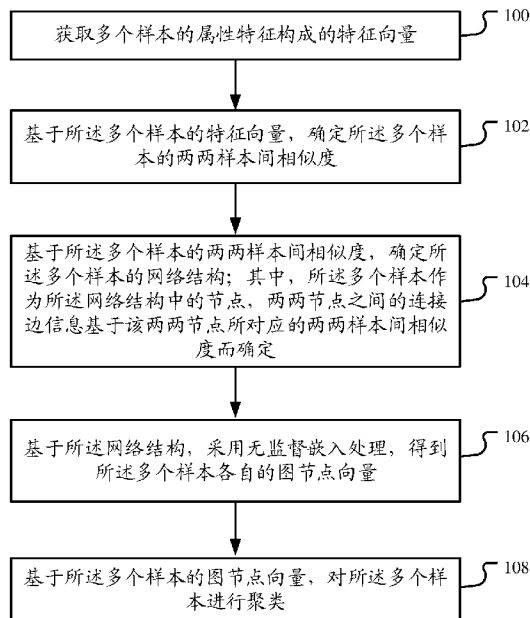


图 1

- 100 Obtain feature vectors formed by attribute features of a plurality of samples
- 102 Determine a similarity between every two samples of the plurality of samples on the basis of the feature vectors of the plurality of samples
- 104 Determine a network structure of the plurality of samples on the basis of the similarity between every two samples of the plurality of samples, wherein the plurality of samples serve as nodes in the network structure, and connection edge information between every two nodes is determined on the basis of the similarity between every two samples corresponding to every two nodes
- 106 On the basis of the network structure, use unsupervised embedding processing to obtain respective graph node vectors of the plurality of samples
- 108 Cluster the plurality of samples on the basis of the graph node vectors of the plurality of samples

(57) Abstract: A sample clustering method and device. The method comprises: first, obtaining feature vectors formed by attribute features of a plurality of samples (100); second, determining a similarity between every two samples of the plurality of samples on the basis of the feature vectors of the plurality of samples (102); next, determining a network structure of the plurality of samples on the basis of the similarity between every two samples of the plurality of samples, wherein the plurality of samples serve as nodes in the network structure, and connection edge information between every two nodes is determined on the basis of the similarity between

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

every two samples corresponding to every two nodes (104); then, on the basis of the network structure, using unsupervised embedding processing to obtain respective graph node vectors of the plurality of samples (106); and then, clustering the plurality of samples on the basis of the graph node vectors of the plurality of samples (108).

(57) 摘要: 一种样本聚类方法及装置。所述方法包括: 首先, 获取多个样本的属性特征构成的特征向量(100); 其次, 基于所述多个样本的特征向量, 确定所述多个样本的两两样本间相似度(102); 接着, 基于所述多个样本的两两样本间相似度, 确定所述多个样本的网络结构; 其中, 所述多个样本作为所述网络结构中的节点, 两两节点之间的连接边信息基于该两两节点所对应的两两样本间相似度而确定(104); 然后, 基于所述网络结构, 采用无监督嵌入处理, 得到所述多个样本各自的图节点向量(106); 之后, 基于所述多个样本的图节点向量, 对所述多个样本进行聚类(108)。

一种样本聚类方法及装置

技术领域

[01]本说明书一个或多个实施例涉及计算机信息处理领域，尤其涉及一种样本聚类方法及装置。

5 背景技术

[02]在进行数据分析，特别是大数据分析时，聚类是一种常用的分析方法。直观上讲，聚类是指将相关高的样本聚在一起，从而形成一个样本类簇（cluster）。聚类可以把相关高的样本划分到一个样本类簇，使得同一个类簇内的样本相关度较高，而不同类簇之间的样本相关度较低。

- 10 [03]在很多场景中，都需要对样本进行聚类，例如，对电商用户进行聚类，得到不同类型的人群，进而为特定类型的人群提供针对性的服务，可以有效的改善用户体验。因此，提供一种能够对样本进行准确聚类的方法具有重要意义。

发明内容

- 15 [04]本说明书一个或多个实施例描述了一种样本聚类方法及装置，可以提高样本聚类结果的准确度。

[05]根据第一方面，提供了一种样本聚类方法，所述方法包括：

[06]获取多个样本的属性特征构成的特征向量；

[07]基于所述多个样本的特征向量，确定所述多个样本的两两样本间相似度；

- 20 [08]基于所述多个样本的两两样本间相似度，确定所述多个样本的网络结构；其中，所述多个样本作为所述网络结构中的节点，两两节点之间的连接边信息基于该两两节点所对应的两两样本间相似度而确定；

[09]基于所述网络结构，采用无监督嵌入处理，得到所述多个样本各自的图节点向量；

[10]基于所述多个样本的图节点向量，对所述多个样本进行聚类。

- 25 [11]在一个实施例中，所述样本包括用户样本，所述属性特征包括用户样本的注册信息和/或行为信息。

[12]在一个实施例中,所述基于所述多个样本的特征向量,确定所述多个样本的两两样本间相似度,包括:

[13]对所述多个样本的特征向量进行归一化处理;

5 [14]基于归一化处理后的所述多个样本的特征向量,计算所述多个样本的两两样本间余弦相似度或欧式距离,获得所述多个样本的两两样本间相似度。

[15]在一个实施例中,所述基于所述多个样本的两两样本间相似度,确定所述多个样本的网络结构,包括:

[16]按照从大到小的顺序,对所述多个样本的两两样本间相似度进行排序;

10 [17]在排序结果中前 M 个或前百分之 M' 的两两样本间相似度对应的样本之间构建连接边,以得到所述多个样本的网络结构;其中,所述前 M 个或前百分之 M' 的两两样本间相似度作为各自对应的连接边的权值。

[18]在一个实施例中,所述基于所述多个样本的两两样本间相似度,确定所述多个样本的网络结构,包括:

15 [19]从所述多个样本的两两样本间相似度中确定出第一两两样本间相似度,所述第一两两样本间相似度大于预设阈值;

[20]在所述第一两两样本间相似度对应的样本之间构建连接边,以得到所述多个样本的网络结构;其中,所述第一两两样本间相似度作为对应的连接边的权值。

[21]在一个示例中,所述基于所述网络结构,采用无监督嵌入处理,得到所述多个样本各自的图节点向量,包括:

20 [22]对于所述多个样本中的第一样本,基于第一样本和其邻居样本各自的特征向量,分别确定第一样本和其邻居样本各自的初次迭代的图节点向量;

[23]对于第一样本,执行多级向量迭代,将多级向量迭代的结果作为该第一样本的图节点向量;其中所述多级向量迭代中的每级向量迭代包括,基于第一样本的上一级迭代的图节点向量和第一样本的邻居样本的上一级迭代的图节点向量,确定第一样本的本次迭代的图节点向量。
25

[24]在一个示例中,所述基于所述网络结构,采用无监督嵌入处理,得到所述多个样本各自的图节点向量,包括:

[25]对于所述多个样本中的第一样本,基于第一样本和其邻居样本各自的特征向量,分

别确定第一样本和其邻居样本各自的初次迭代的图节点向量;

[26]对于所述第一样本,执行多级向量迭代,将多级向量迭代的结果作为该第一样本的图节点向量;其中所述多级向量迭代中的每级向量迭代包括,基于第一样本的邻居样本的上一级迭代的图节点向量,确定第一样本的本次迭代的图节点向量。

5 [27]在一个实施例中,所述基于所述网络结构,采用无监督嵌入处理,得到所述多个样本各自的图节点向量,包括:

[28]采用 deepwalk 图嵌入算法或 GraRep 图嵌入算法,进行嵌入处理,得到所述多个样本各自的图节点向量。

[29]根据第二方面,提供了一种样本聚类装置,所述装置包括:

10 [30]获取单元,配置为获取多个样本的属性特征构成的特征向量;

[31]第一确定单元,配置为基于所述多个样本的特征向量,确定所述多个样本的两两样本间相似度;

[32]第二确定单元,配置为基于所述多个样本的两两样本间相似度,确定所述多个样本的网络结构;其中,所述多个样本作为所述网络结构中的节点,两两节点之间的连接边
15 信息基于该两两节点所对应的两两样本间相似度而确定;

[33]处理单元,配置为基于所述网络结构,采用无监督嵌入处理,得到所述多个样本各自的图节点向量;

[34]聚类单元,配置为基于所述多个样本的图节点向量,对所述多个样本进行聚类。

[35]根据第三方面,提供了一种计算机可读存储介质,其上存储有计算机程序,当所述
20 计算机程序在计算机中执行时,令计算机执行第一方面的方法。

[36]根据第四方面,提供了一种计算设备,包括存储器和处理器,所述存储器中存储有可执行代码,所述处理器执行所述可执行代码时,实现第一方面的方法。

[37]根据本说明书实施例提供的方法和装置,通过本说明书实施例的方案,可以基于多个样本的图节点向量,对多个样本进行聚类;样本的图节点向量刻画了样本在整个网络
25 中的拓扑性质,反映了样本之间的间接相似度,因此,提高了聚类结果的准确性。

附图说明

[38]为了更清楚地说明本发明实施例的技术方案,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发明的一些实施例，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。

[39]图 1 示出根据一个实施例的样本聚类方法的流程图；

5 [40]图 2 示出根据一个实施例的多个样本的网络结构；

[41]图 3 示出根据一个实施例的样本聚类装置的示意性框图。

具体实施方式

[42]下面结合附图，对本说明书提供的方案进行描述。

[43]如前所述，在很多场景中，都需要对样本进行聚类，例如，对用户进行聚类，从而
10 将用户划分到不同人群。对于用户样本而言，可以基于用户之间的通信关系、设备共用关系、转账关系等交互关系，对用户样本进行聚类。例如，可以根据样本间的有无交互关系以及交互的频次构建关系网络图。关系网络图中的节点表示样本，关系网络图中的连接边表示样本之间具有交互关系，连接边的长短表示交互的频次。然后，可以采用某种聚类算法，对关系网络图中的节点进行合并，从而实现对样本的聚类。基于这种聚类
15 方案，可以把交互频次较高的样本聚到一个样本类簇中。例如，在这种聚类方案下，往往会将一个家庭的成员聚为一个样本类簇。而很多情况下，在某一方面，例如兴趣爱好、消费习惯等，一个家庭的各成员之间的并不必然相似。因此，基于样本之间的交互关系进行聚类，并不直接反映出样本之间的相似度。甚至在同一样本类簇中，两个样本之间的相似度很低。

20 [44]而在很多应用场景中，将相似度较高的样本聚到同一样本类簇中，具有重要意义。例如，对于电商而言，将相似度较高的用户划分到同一个用户群体，进而可以为该用户群体，提供相同或相似的服务，从而可以提高用户的体验。

[45]根据一种方案，可以基于样本的属性特征，得到样本两两间的相似度，对样本进行聚类。样本一般具有多个属性特征。以用户样本为例，属性特征可以包括注册信息和/
25 或行为信息；其中，注册信息具体可以包括年龄、性别、教育程度、职业，行为信息可以包括购物信息、娱乐信息等，购物信息可以包括购买的商品类型、对商品的评价等，运动信息可以包括娱乐的类型、娱乐的时间等。以商品样本为例，属性特征可以包括商品的名称、类别、价格、上架时间等等。

[46]但在很多情况下，样本的属性特征并不全面。一般的做法是，简单的利用平均值或0填充缺失的属性特征。因此，在样本的属性特征缺失比较严重的情况下，识别样本两两间的相似度的结果可能并不准确。以A、B、C三个样本为例，可以设定，每个样本应有20个属性特征，但对于A样本，只获得了其前10个属性特征；对于B样本，只获得了其后10个属性特征；对于C样本，获得了其20个属性特征。A样本和C样本的前10个属性特征很接近，因此，两者具有很高的相似度。B样本和C样本的后10个属性特征很接近，因此，两者具有很高的相似度。但A样本和B样本之间，由于信息的缺失，导致两者的相似度较低，使得在聚类时，A样本和B样本可能不被聚在一个样本类簇中。然而，实际上，A、B样本通过样本C存在一定的间接相似性，而直接基于属性特征计算相似度的方案并不能考虑到这样的间接相似性。

[47]在本申请说明书实施例中，可以先基于两两样本间的相似度构建网络结构，基于各样本的邻居样本，得到刻画了各样本在网络结构中的拓扑性质的图节点向量。样本的图节点向量可以反映样本与其高阶邻居之间的间接相似度；基于多个样本的图节点向量，对多个样本进行聚类，提高了聚类结果的准确性。

[48]接下来，参考图1，对本说明书实施例提供的样本聚类方法进行具体介绍。所述方法可以由任何具有计算、处理能力的装置、设备、平台、设备集群执行。如图1所示，所述方法包括如下步骤。步骤100，获取多个样本的属性特征构成的特征向量；步骤102，基于所述多个样本的特征向量，确定所述多个样本的两两样本间相似度；步骤104，基于所述多个样本的两两样本间相似度，确定所述多个样本的网络结构，其中，所述多个样本作为所述网络结构中的节点，两两节点之间的连接边信息基于该两两节点所对应的两两样本间相似度而确定；步骤106，基于所述网络结构，采用无监督嵌入处理，得到所述多个样本各自的图节点向量；步骤108，基于所述多个样本的图节点向量，对所述多个样本进行聚类。

[49]接下来，结合具体例子，对上述步骤进行具体描述。

[50]首先，在步骤100中，获取多个样本的属性特征构成的特征向量。

[51]样本的属性特征是指可用于描述样本特性的信息。可以理解，样本可以有多个属性特征。在一个例子中，样本具体为用户样本，其属性特征信息可以包括注册信息和/或行为信息。注册信息是指用户在注册帐户时，填写的信息，可以包括年龄、性别、教育程度、职业等。行为信息是指用户在相应网站上发生的行为的信息，例如，购物信息、信息行为等。购物信息购买的商品类型、对商品的评价等，运动信息可以包括娱乐的类型、

娱乐的时间等。

[52] 在一个例子中，样本具体为商品样本，属性特征可以包括商品的名称、类别、价格、上架时间、购买人群、评价数量、好评率等等。

[53] 可以将样本的多个属性特征映射为特征向量，以用于后续样本分析。

5 [54] 接着，在步骤 102 中，基于所述多个样本的特征向量，确定所述多个样本的两两样本间相似度。

[55] 如上所述，特征向量为将样本的属性特征进行映射得到的向量，而样本的属性特征为可用于描述样本特性的信息。容易理解，对于多个样本中的任意两个样本，可以基于这两个样本的特征向量，计算这两个样本间的相似度。

10 [56] 在本说明书实施例中，样本间的相似度，是指两个样本在一方面或多方面的相似程度，例如，消费偏好的相似程度、理财投资的相似程度、兴趣爱好的相似程度等等。容易理解，为了得到两个样本在某一方面或多方面的相似度，可以选择相应的属性特征进行计算分析。

15 [57] 在一个实施例中，可对所述多个样本的特征向量进行归一化处理。然后，对于所述多个样本中的任意两个样本，计算两者特征向量夹角的余弦值，得到两者的余弦相似度，得到的余弦相似度用作两者间相似度。采用此种方式，可以得到所述多个样本中各个两两样本间相似度。

20 [58] 在一个实施例中，可对所述多个样本的特征向量进行归一化处理。然后，对于所述多个样本中的任意两个样本，基于两者的特征向量计算欧式距离，得到的欧式距离用作两者间相似度。采用此种方式，可以得到所述多个样本中各个两两样本间相似度。

[59] 其次，在步骤 104 中，基于所述多个样本的两两样本间相似度，确定所述多个样本的网络结构。

25 [60] 所述网络结构也可以称为拓扑结构，其描述了多个样本之间的关系。多个样本作为网络结构中的节点，两两节点之间的连接边信息是基于该两两节点所对应的两两样本间相似度而确定。连接边信息可以包括连接边的权值。可以设定样本与其自身的相似度为 1，样本与任一个其他样本之间的相似度介于 0 和 1 之间。

[61] 在一个实施例中，可以不对所述多个样本的两两样本间相似度进行过滤，即在所述网络结构中，任意两个节点之间存在连接边。任意两个节点之间的连接边的权重基于其

对应的两个样本的样本间相似度而确定。

[62] 在一个实施例中，可以在样本间相似度较高的两个样本之间构建连接边，以用于后续的聚类分析。具体的，可以对所述多个样本的两两样本间相似度进行过滤。然后，基于过滤结果，确定所述网络结构。在该实施例中，对于任意两两样本，当其相似度满足
5 预设条件时，在其对应的两两节点之间构建连接边，并将其相似度作为构建的连接边的权重。

[63] 在该实施例的一个示例中，按照从大到小的顺序，对所述多个样本的两两样本间相似度进行排序；在排序结果中前 M 个的两两样本间相似度对应的样本之间构建连接边，以得到所述多个样本的网络结构；其中，所述前 M 个两两样本间相似度作为各自对应的
10 的连接边的权值。

[64] M 的具体值为预先设定，具体可以根据经验确定 M 的具体值，也可以根据试验确定出的 M 的具体值。对于不同的网络结构， M 的取值可以不同。

[65] 在一个实施例中，按照从大到小的顺序，按照从大到小的顺序，对所述多个样本的两两样本间相似度进行排序；在排序结果中前百分之 M' 的两两样本间相似度对应的样本之间构建连接边，以得到所述多个样本的网络结构；其中，所述前百分之 M' 的两两
15 样本间相似度作为各自对应的连接边的权值。

[66] M' 的具体值为预先设定，具体可以根据经验确定 M' 的具体值，也可以根据试验确定出的 M' 的具体值。对于不同的网络结构， M' 的取值可以不同。

[67] 在一个实施例中，对于任意两个样本，可以将其样本间相似度和预设阈值比较，如果大于预设阈值，则在这两个样本间构建连接边，并将该两个样本间的样本间相似度作为构建的连接边的权值。如此，可以在各个样本间相似度大于预设阈值的两两样本间构建连接边，以确定所述网络结构。换言之，可以从所述多个样本的两两样本间相似度中确定出第一两两样本间相似度，所述第一两两样本间相似度大于预设阈值；在所述第一两两样本间相似度对应的样本之间构建连接边，以得到所述多个样本的网络结构；其中，
20 所述第一两两样本间相似度作为对应的连接边的权值。

[68] 在该实施例中，以包括上述 A、B、C 三个样本的多个样本为例。假定 A、C 样本的样本间相似度为 0.5，B、C 样本的样本间相似度为 0.5，以及预设阈值为 0.3。容易理解，A、B 样本的样本间相似度低于 0.3。因此，通过该实施例的方案，可以构建如图 2 所示的网络结构。如图 2 所示，A、C 样本之间具有连接边，权值为 0.5；B、C 样本之间具

有连接边，权值为 0.5；而 A、B 样本之间没有连接边。

[69]然后，在步骤 106 中，基于所述网络结构，采用无监督嵌入处理，得到所述多个样本各自的图节点向量。

[70]在一个实施例中，对于所述多个样本中的任一样本，例如，第一样本，可以基于第一样本的邻居样本，来确定第一样本的图节点向量。第一样本的邻居样本是指与第一样本直接相连的样本。

[71]在该实施例的一个示例中，如前所述对所述多个样本的两两样本间相似度进行过滤以得到网络结构。如此得到的网络结构中，样本间相似度较高的两两样本之间具有连接边，而样本间相似度较低的两两样本之间没有连接边。此时，对于所述多个样本中的任一样本，例如样本 i，可以先基于样本 i 和其邻居样本各自的特征向量，分别确定样本 i 和其邻居样本各自的初次迭代的图节点向量。具体的，可以将特征向量作归一化处理，然后作为初次迭代的图节点向量。也可以将特征向量作归一化处理后，乘以非线性变换函数，以得到初次迭代的图节点向量。上述非线性变换函数可以使用 Relu 函数，Sigmoid 函数，Tanh 函数等等。然后，对于样本 i，执行多级向量迭代。对于所述多级向量迭代中的每级向量迭代，基于样本 i 的邻居样本的上一级迭代的图节点向量，确定样本 i 的本次迭代的图节点向量。将所述多级向量迭代中的最后一级向量迭代得到的图节点向量作为最终的图节点向量，并用于后续的分析。

[72]需要说明的是，如果样本的属性特征很多（例如，一万多个属性特征），则可以使用主成分分析法（Principal Component Analysis, PCA）等，提取主要属性特征。然后，基于主要属性特征构成的特征向量，确定初次迭代的图节点向量。

[73]对于样本 i，其可以具有多个邻居样本。可以设定 N_i 为样本 i 的邻居样本构成的集合，样本 j 为 N_i 中的任一个元素，即样本 j 为样本 i 的一个邻居样本。

[74]在该实施例的一个示例中，样本 i 的 t 级迭代的图节点向量 h_i^t 通过下式计算：

$$h_i^t = \sum_{j \in N_i} h_j^{t-1} \quad (1)$$

[75]即，对样本 i 各个邻居样本的上一级迭代（t-1 级迭代）的图节点向量 h_j^{t-1} ，进行求和，得到样本 i 本次迭代（t 级迭代）的图节点向量 h_i^t 。

[76]在该实施例的一个示例中, 样本 i 的 t 级迭代的图节点向量 h_i^t 通过下式计算:

$$h_i^t = \sum_{j \in N_i} a_{ij} h_j^{t-1} \quad (2)$$

[77]式(2)在式(1)的基础上, 在样本 i 各个邻居样本的上一级迭代的图节点向量 h_j^{t-1} , 进行求和时, 引入权重因子 a_{ij} 。 a_{ij} 可以为样本 i 和样本 j 之间的样本间相似度, 或者, 样本 i 和样本 j 之间连接边的权值。

[78]在该实施例的另一个示例中, 在执行多级向量迭代中的每级向量迭代时, 基于样本 i 的上一级迭代的图节点向量和样本 i 的邻居样本的上一级迭代的图节点向量, 确定样本 i 的本次迭代的图节点向量。将所述多级向量迭代中的最后一级向量迭代得到的图节点向量作为最终的图节点向量, 并用于后续的分析。

10 [79]在该实施例的又一个示例中, 样本 i 的 t 级迭代的图节点向量 h_i^t 通过下式计算:

$$h_i^t = h_i^{t-1} + \sum_{j \in N_i} h_j^{t-1} \quad (3)$$

[80]即, 对样本 i 上一级迭代的图节点向量 h_i^{t-1} , 以及样本 i 各个邻居样本的上一级迭代的图节点向量 h_j^{t-1} , 进行求和, 得到样本 i 本次迭代 (t 级迭代) 的图节点向量 h_i^t 。

[81]在该实施例的一个示例中, 样本 i 的 t 级迭代的图节点向量 h_i^t 通过下式计算:

$$h_i^t = h_i^{t-1} + \sum_{j \in N_i} a_{ij} h_j^{t-1} \quad (4)$$

[82]式(4)在式(3)的基础上, 在本 i 各个邻居样本的上一级迭代的图节点向量 h_j^{t-1} , 进行求和时, 引入权重因子 a_{ij} 。 a_{ij} 介于 0 和 1 之间, 具体可以为样本 i 和样本 j 之间的样本间相似度。

[83]在该实施例中, 继续参考图 2, C 样本为 A 样本的邻居样本, 因此, A 样本的图节点向量反映了 C 样本的特征向量。C 样本为 B 样本的邻居样本, 因此, B 样本的图节点向量也反映了 C 样本的特征向量。因此, 通过样本 C 可以建立了 A 样本和 B 样本间的间接相似度, 而这种间接相似度通过 A 样本的图节点向量和 B 样本的图节点向量来反映。

[84]在一个实施例中，可以采用 deepwalk 图嵌入算法或 GraRep 图嵌入算法，对网络结构上的各个节点进行嵌入处理，得到所述多个样本各自的图节点向量。

[85]通过上述实施例得到的样本的图节点向量，刻画了样本在整个网络中的拓扑性质，这种拓扑性质反应了样本之间的间接相似度，例如，样本与邻居的邻居之间的相似度。

5 [86]之后，在步骤 108 中，基于所述多个样本的图节点向量，对所述多个样本进行聚类。

[87]可以把各个样本的图节点向量，看作各个样本的新的特征向量。利用各个样本的新的特征向量，利用聚类算法（例如，k-means）对各个样本进行聚类。

[88]综上，本说明书实施例的方案，可以基于多个样本的图节点向量，对多个样本进行聚类；样本的图节点向量刻画了样本在整个网络中的拓扑性质，反映了样本之间的间接
10 相似度，因此，提高了聚类结果的准确性。

[89]仍以上述 A、B、C 三个样本为例，通过本说明实施例的方案，可以得到 A、B、C 三个样本各自的图节点向量。其中，A 样本的图节点向量反映了 A 样本和 B 样本之间的间接相似度，B 样本的图节点向量反映了 B 样本和 A 样本之间的间接相似度。在采用包括了 A、B、C 三个样本的多个样本的图节点向量进行聚类时，考虑了 A 样本和 B
15 样本之间的间接相似度，从而可以提高聚类结果的准确性。

[90]根据第二方面，提供了一种样本聚类装置 300。参考图 3，所述装置 300 包括：

[91]获取单元 310，配置为获取多个样本的属性特征构成的特征向量；

[92]第一确定单元 320，配置为基于所述多个样本的特征向量，确定所述多个样本的两两样本间相似度；

20 [93]第二确定单元 330，配置为基于所述多个样本的两两样本间相似度，确定所述多个样本的网络结构；其中，所述多个样本作为所述网络结构中的节点，两两节点之间的连接边信息基于该两两节点所对应的两两样本间相似度而确定；

[94]处理单元 340，配置为基于所述网络结构，采用无监督嵌入处理，得到所述多个样本各自的图节点向量；

25 [95]聚类单元 350，配置为基于所述多个样本的图节点向量，对所述多个样本进行聚类。

[96]在一个实施例中，所述样本包括用户样本，所述属性特征包括用户样本的注册信息和/或行为信息。

[97]在一个实施例中，所述第一确定单元 320 包括归一化子单元 321 和计算子单元 322；

[98] 所述归一化子单元 321 配置为对所述多个样本的特征向量进行归一化处理;

[99] 所述计算子单元 322 配置为基于归一化处理后的所述多个样本的特征向量, 计算所述多个样本的两两样本间余弦相似度或欧式距离, 获得所述多个样本的两两样本间相似度。

5 [100] 在一个实施例中, 所述第二确定单元 330 包括排序子单元 331 和构建子单元 332;

[101] 所述排序子单元 331 配置为按照从大到小的顺序, 对所述多个样本的两两样本间相似度进行排序;

[102] 所述构建子单元 332 配置为在排序结果中前 M 个或前百分之 M' 的两两样本间相似度对应的样本之间构建连接边, 以得到所述多个样本的网络结构; 其中, 所述前 M
10 个或前百分之 M' 的两两样本间相似度作为各自对应的连接边的权值。

[103] 在一个实施例中, 所述第二确定单元 330 包括第三确定子单元 333 和构建子单元 332;

[104] 所述第三确定子单元 333 配置为从所述多个样本的两两样本间相似度中确定出第一两两样本间相似度, 所述第一两两样本间相似度大于预设阈值;

15 [105] 所述构建子单元 332 配置为在所述第一两两样本间相似度对应的样本之间构建连接边, 以得到所述多个样本的网络结构; 其中, 所述第一两两样本间相似度作为对应的连接边的权值。

[106] 在一个示例中, 所述处理单元 340 包括第四确定子单元 341 和执行子单元 342;

[107] 所述第四确定子单元 341 配置为对于所述多个样本中的第一样本, 基于第一样本和其邻居样本各自的特征向量, 分别确定第一样本和其邻居样本各自的初次迭代的图节点向量;
20

[108] 所述执行子单元 342 配置为对于第一样本, 执行多级向量迭代, 将多级向量迭代的结果作为该第一样本的图节点向量; 其中所述多级向量迭代中的每级向量迭代包括, 基于第一样本的上一级迭代的图节点向量和第一样本的邻居样本的上一级迭代的图节点向量, 确定第一样本的本次迭代的图节点向量。
25

[109] 在一个示例中, 所述处理单元 340 包括第四确定子单元 341 和执行子单元 342;

[110] 所述第四确定子单元 341 配置为对于所述多个样本中的第一样本, 基于第一样本和其邻居样本各自的特征向量, 分别确定第一样本和其邻居样本各自的初次迭代的图

节点向量;

[111] 所述执行子单元 342 配置为对于所述第一样本, 执行多级向量迭代, 将多级向量迭代的结果作为该第一样本的图节点向量; 其中所述多级向量迭代中的每级向量迭代包括, 基于第一样本的邻居样本的上一级迭代的图节点向量, 确定第一样本的本次迭代的图节点向量。

[112] 在一个实施例中, 所述处理单元 340 还配置为采用 deepwalk 图嵌入算法或 GraRep 图嵌入算法, 进行嵌入处理, 得到所述多个样本各自的图节点向量。

[113] 装置 300 的各功能单元可以参考图 1 所示的方法实施例实现, 在此不再赘述。

[114] 另一方面, 本说明书的实施例提供了一种计算机可读存储介质, 其上存储有计算机程序, 当所述计算机程序在计算机中执行时, 令计算机执行图 1 所示的方法。

[115] 另一方面, 本说明书的实施例提供了一种计算设备, 包括存储器和处理器, 所述存储器中存储有可执行代码, 所述处理器执行所述可执行代码时, 实现图 1 所示的方法。

[116] 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。

[117] 以上所述的具体实施方式, 对本发明的目的、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, 以上所述仅为本发明的具体实施方式而已, 并不用于限定本发明的保护范围, 凡在本发明的技术方案的基础之上, 所做的任何修改、等同替换、改进等, 均应包括在本发明的保护范围之内。

权利要求书

1.一种样本聚类方法，所述方法包括：

获取多个样本的属性特征构成的特征向量；

基于所述多个样本的特征向量，确定所述多个样本的两两样本间相似度；

5 基于所述多个样本的两两样本间相似度，确定所述多个样本的网络结构；其中，所述多个样本作为所述网络结构中的节点，两两节点之间的连接边信息基于该两两节点所对应的两两样本间相似度而确定；

基于所述网络结构，采用无监督嵌入处理，得到所述多个样本各自的图节点向量；

基于所述多个样本的图节点向量，对所述多个样本进行聚类。

10 2.根据权利要求 1 所述的方法，所述样本包括用户样本，所述属性特征包括用户样本的注册信息和/或行为信息。

3.根据权利要求 1 所述的方法，所述基于所述多个样本的特征向量，确定所述多个样本的两两样本间相似度，包括：

对所述多个样本的特征向量进行归一化处理；

15 基于归一化处理后的所述多个样本的特征向量，计算所述多个样本的两两样本间余弦相似度或欧式距离，获得所述多个样本的两两样本间相似度。

4.根据权利要求 1 所述的方法，所述基于所述多个样本的两两样本间相似度，确定所述多个样本的网络结构，包括：

按照从大到小的顺序，对所述多个样本的两两样本间相似度进行排序；

20 在排序结果中前 M 个或前百分之 M' 的两两样本间相似度对应的样本之间构建连接边，以得到所述多个样本的网络结构；其中，所述前 M 个或前百分之 M' 的两两样本间相似度作为各自对应的连接边的权值。

5.根据权利要求 1 所述的方法，所述基于所述多个样本的两两样本间相似度，确定所述多个样本的网络结构，包括：

25 从所述多个样本的两两样本间相似度中确定出第一两两样本间相似度，所述第一两两样本间相似度大于预设阈值；

在所述第一两两样本间相似度对应的样本之间构建连接边，以得到所述多个样本的网络结构；其中，所述第一两两样本间相似度作为对应的连接边的权值。

30 6.根据权利要求 4 或 5 所述的方法，所述基于所述网络结构，采用无监督嵌入处理，得到所述多个样本各自的图节点向量，包括：

对于所述多个样本中的第一样本，基于第一样本和其邻居样本各自的特征向量，分

别确定第一样本和其邻居样本各自的初次迭代的图节点向量;

对于第一样本,执行多级向量迭代,将多级向量迭代的结果作为该第一样本的图节点向量;其中所述多级向量迭代中的每级向量迭代包括,基于第一样本的上一级迭代的图节点向量,第一样本的邻居样本的上一级迭代的图节点向量,确定第一样本的本次迭代的图节点向量。

7.根据权利要求4或5所述的方法,所述基于所述网络结构,采用无监督嵌入处理,得到所述多个样本各自的图节点向量,包括:

对于所述多个样本中的第一样本,基于第一样本和其邻居样本各自的特征向量,分别确定第一样本和其邻居样本各自的初次迭代的图节点向量;

对于所述第一样本,执行多级向量迭代,将多级向量迭代的结果作为该第一样本的图节点向量;其中所述多级向量迭代中的每级向量迭代包括,基于第一样本的邻居样本的上一级迭代的图节点向量,确定第一样本的本次迭代的图节点向量。

8.根据权利要求1所述的方法,所述基于所述网络结构,采用无监督嵌入处理,得到所述多个样本各自的图节点向量,包括:

采用 deepwalk 图嵌入算法或 GraRep 图嵌入算法,进行嵌入处理,得到所述多个样本各自的图节点向量。

9.一种样本聚类装置,所述装置包括:

获取单元,配置为获取多个样本的属性特征构成的特征向量;

第一确定单元,配置为基于所述多个样本的特征向量,确定所述多个样本的两两样本间相似度;

第二确定单元,配置为基于所述多个样本的两两样本间相似度,确定所述多个样本的网络结构;其中,所述多个样本作为所述网络结构中的节点,两两节点之间的连接边信息基于该两两节点所对应的两两样本间相似度而确定;

处理单元,配置为基于所述网络结构,采用无监督嵌入处理,得到所述多个样本各自的图节点向量;

聚类单元,配置为基于所述多个样本的图节点向量,对所述多个样本进行聚类。

10.根据权利要求9所述的装置,所述样本包括用户样本,所述属性特征包括用户样本的注册信息和/或行为信息。

11.根据权利要求9所述的装置,所述第一确定单元包括归一化子单元和计算子单元;所述归一化子单元配置为对所述多个样本的特征向量进行归一化处理;

所述计算子单元配置为基于归一化处理后的所述多个样本的特征向量,计算所述多

个样本的两两样本间余弦相似度或欧式距离，获得所述多个样本的两两样本间相似度。

12.根据权利要求 9 所述的装置，所述第二确定单元包括排序子单元和构建子单元；

所述排序子单元配置为按照从大到小的顺序，对所述多个样本的两两样本间相似度进行排序；

5 所述构建子单元配置为在排序结果中前 M 个或前百分之 M' 的两两样本间相似度对应的样本之间构建连接边，以得到所述多个样本的网络结构；其中，所述前 M 个或前百分之 M' 的两两样本间相似度作为各自对应的连接边的权值。

13.根据权利要求 9 所述的装置，所述第二确定单元包括第三确定子单元和构建子单元；

10 所述第三确定子单元配置为从所述多个样本的两两样本间相似度中确定出第一两两样本间相似度，所述第一两两样本间相似度大于预设阈值；

所述构建子单元配置为在所述第一两两样本间相似度对应的样本之间构建连接边，以得到所述多个样本的网络结构；其中，所述第一两两样本间相似度作为对应的连接边的权值。

15 14.根据权利要求 12 或 13 所述的装置，所述处理单元包括第四确定子单元和执行子单元；

所述第四确定子单元配置为对于所述多个样本中的第一样本，基于第一样本和其邻居样本各自的特征向量，分别确定第一样本和其邻居样本各自的初次迭代的图节点向量；

20 所述执行子单元配置为对于第一样本，执行多级向量迭代，将多级向量迭代的结果作为该第一样本的图节点向量；其中所述多级向量迭代中的每级向量迭代包括，基于第一样本的上一级迭代的图节点向量和第一样本的邻居样本的上一级迭代的图节点向量，确定第一样本的本次迭代的图节点向量。

15.根据权利要求 12 或 13 所述的装置，所述处理单元包括第四确定子单元和执行子单元；

25 所述第四确定子单元配置为对于所述多个样本中的第一样本，基于第一样本和其邻居样本各自的特征向量，分别确定第一样本和其邻居样本各自的初次迭代的图节点向量；

30 所述执行子单元配置为对于所述第一样本，执行多级向量迭代，将多级向量迭代的结果作为该第一样本的图节点向量；其中所述多级向量迭代中的每级向量迭代包括，基于第一样本的邻居样本的上一级迭代的图节点向量，确定第一样本的本次迭代的图节点向量。

16.根据权利要求 9 所述的装置，所述处理单元还配置为采用 deepwalk 图嵌入算法

或 GraRep 图嵌入算法, 进行嵌入处理, 得到所述多个样本各自的图节点向量。

17. 一种计算机可读存储介质, 其上存储有计算机程序, 当所述计算机程序在计算机中执行时, 令计算机执行权利要求 1-8 中任一项的所述的方法。

18. 一种计算设备, 包括存储器和处理器, 其特征在于, 所述存储器中存储有可执行
5 代码, 所述处理器执行所述可执行代码时, 实现权利要求 1-8 中任一项所述的方法。

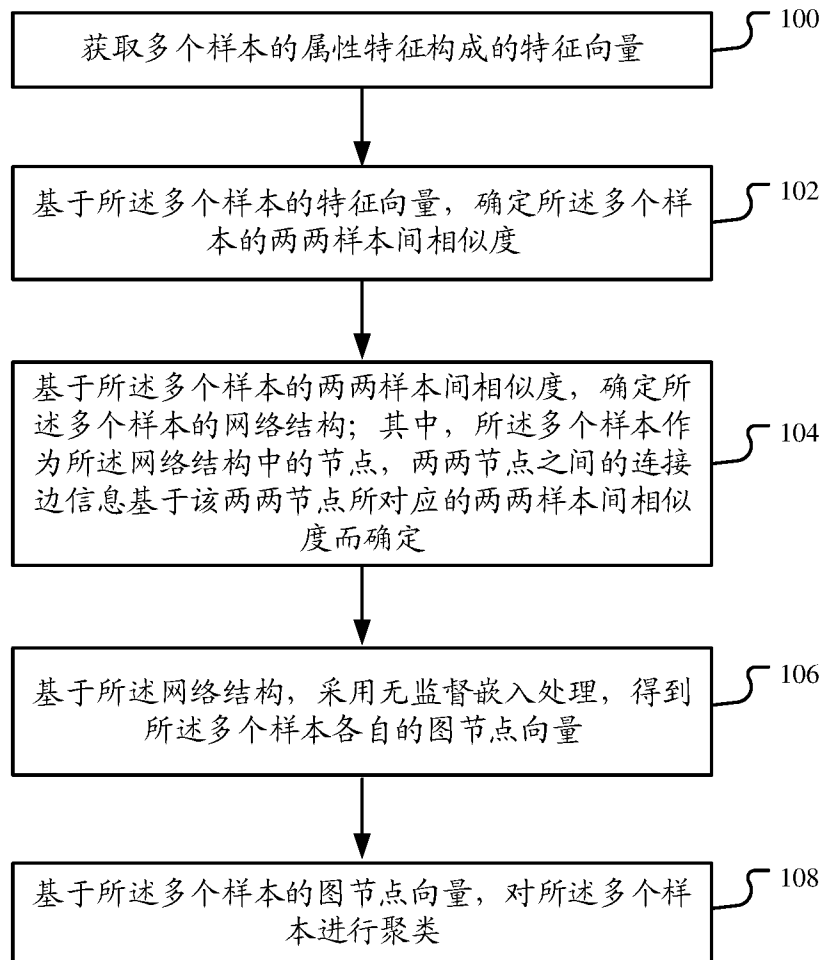


图 1

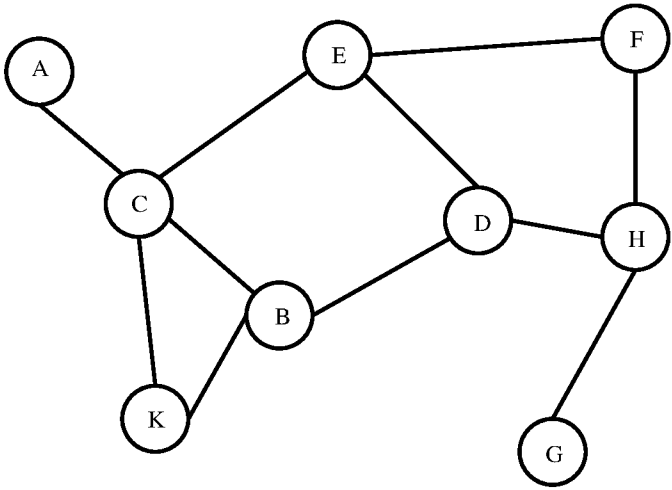


图 2

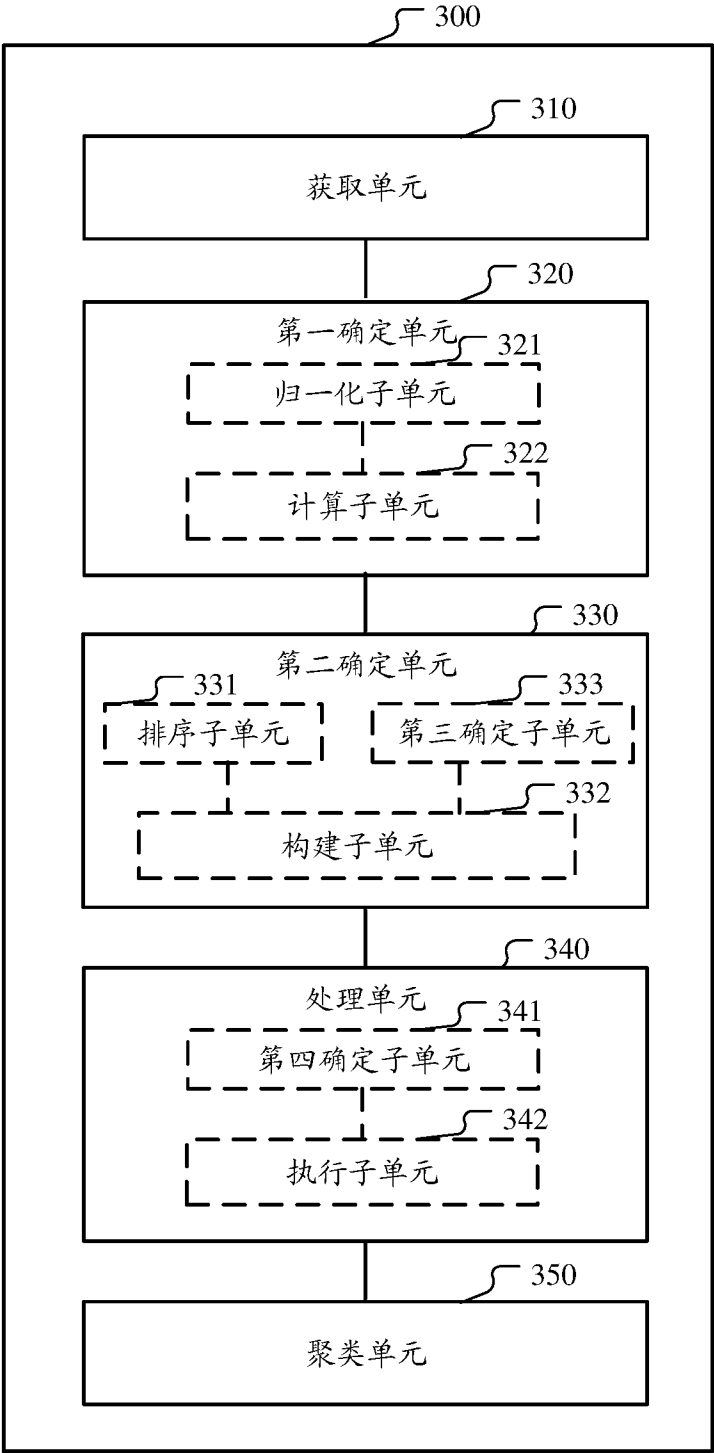


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/073872

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/28(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPODOC, WPI, CNPAT, CNKI, IEEE: 样本, 数据, 聚类, 特征, 向量, 属性, 相似度, 相似程度, 距离, 网络, 节点, 边, 连接, 图, 嵌入, sample, data, clustering, feature, vector, attribute, similarity, network, structure, node, edge, graph, deepwalk, GraRep

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110032606 A (ALIBABA GROUP HOLDING LIMITED) 19 July 2019 (2019-07-19) claims 1-18	1-18
Y	CN 107103329 A (ALIBABA GROUP HOLDING LIMITED) 29 August 2017 (2017-08-29) description, paragraphs [0020]-[0065]	1-18
Y	CN 109447261 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 08 March 2019 (2019-03-08) description, paragraph [0003]	1-18
A	CN 102982342 A (XIAMEN UNIVERSITY) 20 March 2013 (2013-03-20) entire document	1-18
A	US 2012263376 A1 (SHARP LABORATORIES OF AMERICA, INC.) 18 October 2012 (2012-10-18) entire document	1-18

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

07 April 2020

Date of mailing of the international search report

22 April 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/073872

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110032606	A	19 July 2019	None			
CN	107103329	A	29 August 2017	None			
CN	109447261	A	08 March 2019	None			
CN	102982342	A	20 March 2013	None			
US	2012263376	A1	18 October 2012	WO	2012141332	A1	18 October 2012

国际检索报告

国际申请号

PCT/CN2020/073872

A. 主题的分类

G06F 16/28 (2019.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

EPDOC, WPI, CNPAT, CNKI, IEEE: 样本, 数据, 聚类, 特征, 向量, 属性, 相似度, 相似程度, 距离, 网络, 节点, 边, 连接, 图, 嵌入, sample, data, clustering, feature, vector, attribute, similarity, network, structure, node, edge, graph, deepwalk, GraRep

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 110032606 A (阿里巴巴集团控股有限公司) 2019年 7月 19日 (2019 - 07 - 19) 权利要求1-18	1-18
Y	CN 107103329 A (阿里巴巴集团控股有限公司) 2017年 8月 29日 (2017 - 08 - 29) 说明书第[0020]-[0065]段	1-18
Y	CN 109447261 A (北京邮电大学) 2019年 3月 8日 (2019 - 03 - 08) 说明书第[0003]段	1-18
A	CN 102982342 A (厦门大学) 2013年 3月 20日 (2013 - 03 - 20) 全文	1-18
A	US 2012263376 A1 (SHARP LABORATORIES OF AMERICA, INC.) 2012年 10月 18日 (2012 - 10 - 18) 全文	1-18

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 4月 7日

国际检索报告邮寄日期

2020年 4月 22日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

高民芳

电话号码 86-(10)-53961520

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/073872

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	110032606	A	2019年 7月 19日	无	
CN	107103329	A	2017年 8月 29日	无	
CN	109447261	A	2019年 3月 8日	无	
CN	102982342	A	2013年 3月 20日	无	
US	2012263376	A1	2012年 10月 18日	W0 2012141332 A1	2012年 10月 18日