

(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
2017 年 10 月 12 日 (12.10.2017)



(10) 国际公布号
WO 2017/173929 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2017/077997
- (22) 国际申请日: 2017 年 3 月 24 日 (24.03.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610205887.7 2016 年 4 月 5 日 (05.04.2016) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Cayman (KY)。
- (72) 发明人: 黄光远 (HUANG, Guangyuan); 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)。 兰金鹤 (LAN, Jinhe); 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)。 麦港林 (MAI, Ganglin); 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: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY

AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))。

(54) Title: UNSUPERVISED FEATURE SELECTION METHOD AND DEVICE

(54) 发明名称: 无监督的特征选择方法、装置

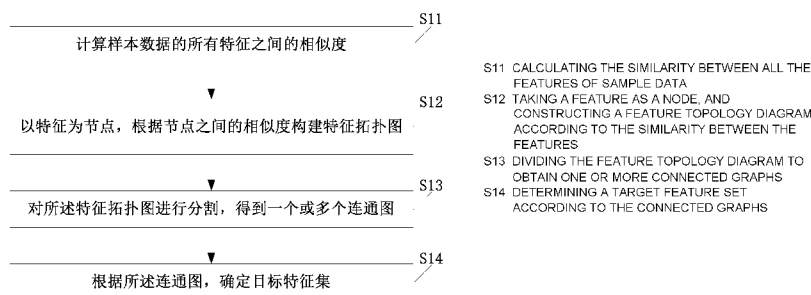


图 1

(57) Abstract: Disclosed are an unsupervised feature selection method and device. In the method, a feature topology diagram is firstly constructed according to the similarity between features, and then the feature topology diagram is divided, so that feature nodes with a relatively high similarity are divided into the same connected graph to realize the feature clustering of sample data, and thus the features of the sample data can also be selected in a scenario where a classification mark of the sample data cannot be pre-determined; after the feature clustering is completed through the connected graph, a node can be further selected from each connected graph, and the features corresponding to the node are written into a target feature set as representative features, so as to obtain comprehensive and non-repetitive representative features corresponding to the entire sample data. Therefore, in the present application, without needing to depend on a classification mark of sample data, unsupervised feature selection can be achieved, and it is ensured that two or more similar features do not appear in a target feature set, so that the target feature set can describe the sample data more intuitively, thereby avoiding information redundancy.

(57) 摘要:

[见续页]



WO 2017/173929 A1

本申请公开一种无监督的特征选择方法、装置，其首先根据特征之间的相似度构建特征拓扑图，然后将该特征拓扑图进行分割，使得相似度较高的特征节点划分在同一个连通图中，实现对样本数据的特征聚类，从而在无法预先确定样本数据的分类标示的场景下也可以对样本数据的特征进行选择；在通过连通图完成特征聚类后，进一步可以从每个连通图中选出一个节点，并将该节点对应的特征为代表特征，记入目标特征集，从而得到整个样本数据对应的全面且不重复的代表特征。因此，本申请不需要依赖样本数据的分类标示，可以实现无监督的特征选择，且保证目标特征集中不会出现两个或两个以上相类似的特征，使得目标特征集可以更直观地描述样本数据，避免信息冗余。

无监督的特征选择方法、装置

本申请要求 2016 年 04 月 05 日递交的申请号为 201610205887.7、发明名称为“无监督的特征选择方法、装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及计算机数据处理技术领域，尤其涉及无监督的特征选择方法、装置。

背景技术

10 目前，基于大数据分析的应用不断涌现，如信息检索、挖掘用户的消费倾向、消费模式等；通过对样本数据进行分析，提取出能够描述样本数据的属性信息，即样本的特征。例如，从某商品的网络销售数据中，可以提取出购买用户等级、购买次数、购买金额等特征。一般的，样本数据都具有多个特征，但未必每个特征都对数据的分析有用处，因此需要从大量特征中选出有效的特征，以减少特征集中特征的个数，进而减小具体应
15 用中基于特征集的计算过程的复杂度，提高预测精度。

针对样本数据的特征选择问题，相关技术中已经给出多种实现方法，这些实现方法大多都是需要依赖样本数据分类标示的有监督方式，即各条样本数据的类别应当为已知，从而根据这些已知类别调整特征选择模型的参数，以保证特征选择的准确率。

20 但是有些应用场景下样本数据不具备分类标示，从而难以通过现有特征选择方法进行特征选择。另外，应用现有特征选择方法会出现所选出的两个或两个以上特征所代表的信息相似的现象，即最终得到的特征集存在信息冗余。

发明内容

为克服相关技术中存在的问题，本申请提供一种无监督的特征选择方法、装置。

25 本申请第一方面，提供一种无监督的特征选择方法，包括：

计算样本数据的所有特征之间的相似度；

以特征为节点，根据节点之间的相似度构建特征拓扑图；

对所述特征拓扑图进行分割，得到一个或多个连通图；

根据所述连通图，确定目标特征集。

30 结合第一方面，在第一方面第一种可行的实施方式中，根据节点之间的相似度构建

特征拓扑图，包括：

在任意相似度大于零的两个节点之间添加一条边，得到所述特征拓扑图。

结合第一方面，或者第一方面第一种可行的实施方式，在第一方面第二种可行的实施方式中，对所述特征拓扑图进行分割，得到一个或多个连通图，包括：

5 分别将所述特征拓扑图中每条边对应的相似度与预设相似度阈值进行比较；

删除所述特征拓扑图中相似度小于所述预设相似度阈值的边；

将通过边连接的节点划分在同一个连通图中，得到一个或多个连通图。

结合第一方面，或者第一方面第一种可行的实施方式，在第一方面第三种可行的实施方式中，对所述特征拓扑图进行分割，得到一个或多个连通图，包括：

10 为所述特征拓扑图中的每个节点设置唯一的标签；

将每个节点的标签传播至相应的邻居节点；

对于每个节点，将自己的标签和接收到的所有来自邻居的标签进行比较，选其中值最小的标签作为相应节点的新标签；

判断各个节点的标签是否发生变化；

15 如果存在至少一个节点的标签发生变化，则返回执行所述将每个节点的标签传播至相应的邻居节点的步骤；

如果所有节点的标签均未发生变化，则将标签相同的节点划分在同一个连通图中，得到一个或多个连通图。

结合第一方面，或者第一方面第一种可行的实施方式，在第一方面第四种可行的实施方式中，根据所述连通图，确定目标特征集，包括：

20 分别确定每个连通图中的核心节点，并将各个核心节点对应的特征记入目标特征集。

结合第一方面第四种可行的实施方式，在第一方面第五种可行的实施方式中，分别确定每个连通图中的核心节点，包括：

25 根据各个节点的节点度值、介数值和 K_{core} 系数中的任意一种确定每个连通图中的核心节点。

本申请第二方面，提供一种无监督的特征选择装置，包括：

相似度计算单元，用于计算样本数据的所有特征之间的相似度；

拓扑图构建单元，用于以特征为节点，根据节点之间的相似度构建特征拓扑图；

连通图挖掘单元，用于对所述特征拓扑图进行分割，得到一个或多个连通图；

30 目标特征选择单元，用于根据所述连通图，确定目标特征集。

结合第二方面，在第二方面第一种可行的实施方式中，所述拓扑图构建单元具体被配置为：

在任意相似度大于零的两个节点之间添加一条边，得到所述特征拓扑图。

结合第二方面，或者第二方面第一种可行的实施方式，在第二方面第二种可行的实

5 施方式中，所述连通图挖掘单元包括：第一挖掘模块；

所述第一挖掘模块被配置为：

分别将所述特征拓扑图中每条边对应的相似度与预设相似度阈值进行比较；

删除所述特征拓扑图中相似度小于所述预设相似度阈值的边；

将通过边连接的节点划分在同一个连通图中，得到一个或多个连通图。

10 结合第二方面，或者第二方面第一种可行的实施方式，在第二方面第三种可行的实施方式中，所述连通图挖掘单元包括：第二挖掘模块；

所述第二挖掘模块被配置为：

为所述特征拓扑图中的每个节点设置唯一的标签；

将每个节点的标签传播至相应的邻居节点；

15 对于每个节点，将自己的标签和接收到的所有来自邻居的标签进行比较，选其中值最小的标签作为相应节点的新标签；

判断各个节点的标签是否发生变化；

如果存在至少一个节点的标签发生变化，则返回执行所述将每个节点的标签传播至相应的邻居节点的步骤；

20 如果所有节点的标签均未发生变化，则将标签相同的节点划分在同一个连通图中，得到一个或多个连通图。

结合第二方面，或者第二方面第一种可行的实施方式，在第二方面第四种可行的实施方式中，所述目标特征选择单元具体被配置为：

分别确定每个连通图中的核心节点，并将各个核心节点对应的特征记入目标特征集。

25 结合第二方面第四种可行的实施方式，在第二方面第五种可行的实施方式中，所述目标特征选择单元具体被配置为：根据各个节点的节点度值、介数值和 K_{core} 系数中的任意一种确定每个连通图中的核心节点。

由以上技术方案可知，本申请实施例基于图论知识完成对样本数据的分析，首先根据特征之间的相似度构建特征拓扑图，然后将该特征拓扑图进行分割，使得相似度较高的特征节点划分在同一个连通图中，实现对样本数据的特征聚类，从而在无法预先确定

30

样本数据的分类标示的场景下也可以对样本数据的特征进行选择；在通过连通图完成特征聚类后，进一步从每个连通图中选出一个节点，并将该节点对应的特征为代表特征，记入目标特征集，从而得到整个样本数据对应的全面且不重复的代表特征。因此，本实施例不需要依赖样本数据的分类标示，可以实现无监督的特征选择，且保证目标特征集中不会出现两个或两个以上相类似的特征，使得目标特征集可以更直观地描述样本数据，避免信息冗余。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

10 附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并与说明书一起用于解释本发明的原理。

图 1 是本申请实施例提供的一种无监督的特征选择方法的流程图；

图 2 是本申请实施例提供的无监督的特征选择方法中挖掘连通图的原理示意图；

15 图 3 是本申请实施例提供的无监督的特征选择方法中分割特征拓扑图的一种方法示意图；

图 4 是本申请实施例提供的一种无监督的特征选择装置的结构框图。

具体实施方式

20 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

25 图 1 示出的一种无监督的特征选择方法的流程图。如图 1 所示，该方法包括以下步骤。

S11、计算样本数据的所有特征之间的相似度。

具体的，本申请实施例可以根据欧氏距离、余弦距离和杰卡德相似系数中的任意一种计算两个特征之间的相似度。

30 其中，欧氏距离表征空间两点之间的绝对距离。假设两个特征分别为 X 、 Y ，可以视为多维空间中的两个点，坐标分别为 $X(x_1, x_2, \dots)$ 和 $Y(y_1, y_2, \dots)$ ，则其欧氏距离计

算公式为: $Eu(X,Y)=\sqrt{\sum (x_i-y_i)^2}$, 计算结果 $Eu(X,Y)$ 即可作为特征 X 和 Y 之间的相似度; 其中 x_i 表示特征 X 在多维坐标系中第 i 维的坐标, y_i 表示特征 Y 在多维坐标系中第 i 维的坐标, $i=1, 2, \dots$ 。

余弦距离, 也称为余弦相似度, 是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。特征 X 和 Y 的余弦距离, 即 X 和 Y 对应的向量 $\vec{X}$ 和 $\vec{Y}$ 的夹角余弦值, 计算公式为: $\cos(X,Y)=\frac{\vec{X}*\vec{Y}}{|\vec{X}||\vec{Y}|}$, 计算结果 $\cos(X,Y)$ 即可作为特征 X 和 Y 之间的相似度。相比欧氏距离, 余弦距离更加注重两个特征在方向上的差异。

杰卡德相似系数, 两个集合的交集元素个数与这两个集合的并集元素个数之间的比值, 称为这两个集合的杰卡德系数; 对于特征 X 和 Y, 其对应的向量 $\vec{X}$ 和 $\vec{Y}$ 为多维向量, 将每个维度的数值分别视为一个元素, 多维向量的各个维度值构成一个集合, 故特征 X 和 Y 也可以视为两个集合, 进而可以计算其杰卡德相似系数, 公式为: $J(X,Y)=\frac{|\vec{X} \cap \vec{Y}|}{|\vec{X} \cup \vec{Y}|}$, 计算结果 $J(X,Y)$ 即可作为特征 X 和 Y 之间的相似度。

当然, 除了上述欧氏距离、余弦距离和杰卡德相似系数以为, 本申请实施例还可以采用其他方法计算两个特征之间的相似度, 实际应用中可以根据具体场景的不同, 选择不同的公式来计算特征之间的相似性。

S12、以特征为节点, 根据节点之间的相似度构建特征拓扑图。

S13、对所述特征拓扑图进行分割, 得到一个或多个连通图。

在拓扑图中, 若任意两个节点间均有路径相连 (既包括通过一条边直接相连, 也包括通过至少一个节点和至少两条边间接相连), 则称为连通图。根据上述建图的逻辑可知, 相对于连通图外的节点, 处于同一连通图中的各节点之间的连接关系更紧密, 即处于同一连通图中的各节点对应的特征之间的相似度更高, 也即步骤 S13 实际是将相似特征对应的节点聚合到同一个连通图中, 实现对样本数据的特征聚类, 从而不需要预知样本数据的分类标示。

S14、根据所述连通图, 确定目标特征集。

由于每个连通图都相当于一类相似特征的集合, 故本实施例步骤 S14 可以从每个连通图中选择一个节点, 将该节点对应的特征作为该连通图所对应的特征集合的代表特征, 记入目标特征集, 使得目标特征集中每种类型的特征仅包含一个, 保证目标特征集中的

特征之间的差异性。

由以上技术方案可知，本申请实施例基于图论知识完成对样本数据的分析，首先根据特征之间的相似度构建特征拓扑图，然后将该特征拓扑图进行分割，使得相似度较高的特征节点划分在同一个连通图中，实现对样本数据的特征聚类，从而在无法预先确定样本数据的分类标示的场景下也可以对样本数据的特征进行选择；在通过连通图完成特征聚类后，进一步从每个连通图中选择一个节点，并将该节点对应的特征为代表特征，记入目标特征集，从而得到整个样本数据对应的全面且不重复的代表特征。因此，本实施例不需要依赖样本数据的分类标示，可以实现无监督的特征选择，且保证目标特征集中不会出现两个或两个以上相类似的特征，使得目标特征集可以更直观地描述样本数据，避免信息冗余。

在本申请一个可行的实施例中，步骤 S12 所述的根据节点之间的相似度构建特征拓扑图，具体可以包括：

在任意相似度大于零的两个节点之间添加一条边，得到所述特征拓扑图。

如图 2 所示，假设有标号为 1 至 8 的八个特征，分别表示某购物网站中某个商品 7 天浏览次数、14 天浏览次数、30 天浏览次数、60 天浏览次数、7 天购买次数、14 天购买次数、30 天购买次数、60 天购买次数。构建这八个特征对应的特征拓扑图的过程为：经过相似度计算得知，标号为 1 和 2 的两个特征之间的相似度为 0.7，则将二者对应的节点通过一条边连接，标号为 3 和 7 的两个特征之间的相似度为 0.2，则将二者对应的节点通过一条边连接……依此类推，可以得到图 2 所示特征拓扑图。另外，为了更直观的展现特征之间的相似度高低，可以设定相似度越小，相应的边长度越大，使得相应的两个节点之间距离越大（例如图 2 中，由于 $0.7 > 0.2$ ，故标号 1 和 2 之间的边比标号 3 和 7 之间的边短），从而可以根据节点密集程度确定相应特征的相似度高低，即节点之间越密集度，相应的特征之间相似度越高。

在本申请一个可行的实施例中，上述步骤 S14 所述的根据所述连通图，确定目标特征集，具体可以包括：

分别确定每个连通图中的核心节点，并将各个核心节点对应的特征记入目标特征集。

仍参照图 2，根据上述步骤 S13 对八个特征对应的特征拓扑图进行分割，可以得到标号分别为 A 和 B 的两个连通图（虚线圈表示）；进而，在步骤 S14 中可以确定连通图 A 的核心节点为节点 3，连通图 B 的核心节点为节点 7，并将该核心节点 3 和 7 对应的特征作为代表特征记入目标特征集，即核心节点 3 对应的特征为节点 1 至 4 对应的四个特

征的代表特征，核心节点 7 对应的特征为节点 5 至 8 对应的四个特征的代表特征。

可选的，上述确定各个连通图的核心节点的方法有多种，至少包括：根据各个节点的节点度值（Node Degree）、介数值和 Kcore 系数中的任意一种确定所述核心节点；具体可以根据应用场景的不同，选择不同的确定方法。

- 5 其中，某个节点的节点度值，指在拓扑图中该节点相关联的边的条数；相应的，可以选择节点度值最大的节点作为核心节点。

某个节点的介数值，指拓扑图中经过该节点的最短路径的数目与拓扑图中最短路径的总数之间的比值；相应的，可以选择介数值最大的节点作为核心节点。

- 10 根据 KCore 算法计算各节点在连通图结构上的 Kcore 系数，本质上是过滤连通图中的边缘节点，找出连通图中相对核心位置的点并发现他们的关联。

仍以图 2 所示情境为例，根据节点度值确定核心节点的步骤为：连通图 A 中，节点 3 相关联的边共 4 条，即节点 3 的节点度值为 4，且连通图 A 中其他节点的节点度值都小于 4，故可以确定节点 3 为连通图 A 的核心节点；相应的，节点 3 对应的特征作为连通图 A 的代表特征，被记入目标特征集。

- 15 由以上技术方案可知，本申请基于图论知识对样本数据特征进行聚类、选择，不需要依赖样本数据的分类标示（即无监督），且最终得到的目标特征集中的各个代表特征差异都较大，不会存在信息冗余。以电商数据为例，对于“用户的星级”、“15 天浏览次数”、“15 天购买次数”、“15 天购买总金额”等用户特征，通过本实施例可以识别出“15 天浏览次数”、“15 天购买次数”和“15 天购买总金额”三者为相似特征，目
20 标特征集中最多存在三者中的一个。

上述步骤 S13 所述的对特征拓扑图进行分割，实际是实现特征聚类，也即实现对所有特征的无监督分类。实际应用中，步骤 S13 的具体实施方法可以有多种。在本申请一个可行的实施例中，可以直接根据特征之间的相似度进行特征拓扑图的分割，具体步骤如下：

- 25 S1311、分别将所述特征拓扑图中每条边对应的相似度与预设相似度阈值进行比较；
S1312、删除所述特征拓扑图中相似度小于所述预设相似度阈值的边；
S1313、将通过边连接的节点划分在同一个连通图中，得到一个或多个连通图。

- 具体的，两个特征之间的相似度取值范围为 $[0, 1]$ ，相似度为 0 表示两个特征完全不相似，相似度为 1 表示两个特征完全相同。相应的，上述预设相似度阈值可以根据具体
30 应用场景设置，预设相似度阈值越高，特征拓扑图中节点分布越分散，从而可以挖掘出

的连通图个数越多且每个连通图的规模越小，从而使得目标特征集中可以得到更多的代表特征；相反，预设相似度阈值越低，特征拓扑图中节点分布越密集，可以挖掘出的连通图个数越少且每个连通图的规模越大，从而最终得到的目标特征集中代表特征越少。一般的，预设相似度阈值不宜小于 0.5。

5 仍参照图 2，对标号为 1 至 8 的八个特征对应的特征拓扑图进行分割，假设预设相似度阈值设置为 0.5 则将特征拓扑图中每条边对应的相似度分别与 0.5 进行比较，删除相似度小于 0.5 的边，如节点 3 和 7 之间的边（以虚线表示），从而可以得到 A 和 B 两个连通图。

参照图 3，在本申请另一个可行的实施例中，步骤 S13 中还可以基于标签传播算法
10 实现对特征拓扑图的分割，具体步骤如下：

S1321、为所述特征拓扑图中的每个节点设置一个唯一的标签；具体的，可以将节点 ID 作为其标签。

S1322、将每个节点的标签传播至相应的邻居节点。

S1323、对于每个节点，将自己的标签和接收到的所有来自邻居的标签进行比较，选
15 其中值最小的标签作为相应节点的新标签。

S1324、判断各个节点的标签是否发生变化（即判断是否存在标签发生变化的节点），如果存在至少一个节点的标签发生变化，则返回至步骤 S1322，如果所有节点的标签均未发生变化，则执行步骤 S1325。

S1325、将标签相同的节点划分在同一个连通图中，得到一个或多个连通图。

20 仍参照图 2 所示情境，对于节点 1，将其标签 label1 分别传播至其邻居节点 2 和 3，并接收邻居节点 2、3 传播来的标签 label2 和 label3，将 label1、label2 和 label3 进行比较，得到的比较结果为 label3 的值最小，则将节点 1 的标签由 label1 更改为 label3；其他节点也依照与节点 1 相同的方式进行标签传播、比较及更新。由于至少节点 1 对应的标签发生了变化，故返回步骤 S1322，继续在八个节点之间进行标签传播、比较及更
25 新；依此循环，直至相对于上一次循环结果，八个节点的标签都无变化，停止循环并执行步骤 S1325；例如，假设此时节点 1 至 4 对应的标签都为 label3，节点 5 至 8 对应的标签都为 label7，则将标签同为 label3 的节点 1 至 4 划分在同一个连通图 A 中（实际意义为表征浏览次数的特征都划分在了连通图 A 中），标签同为 label7 的节点 5 至 8 划分在同一个连通图 B 中（实际意义为表征购买次数的特征都划分在了连通图 B 中）。

30 需要说明的是，上述步骤 S1323 中选择值最小的标签作为节点的新标签仅仅是一种

具体实施方式，也可以选择值最大的标签，只要保证同一特征拓扑图中的所有节点采用的选择标准一致即可。

上述标签传播算法尤其适用于特征拓扑图较简单的场景，标签传播、比较工作量不会太大，且简单易行。在本申请其他实施例中，对于较复杂的特征拓扑图，也可以采用社区发现算法完成对特征拓扑图的分割（特征聚类），得到连通图。

另外，本申请实施例还提供了一种计算机存储介质，例如可以是 ROM、随机存取存储器（RAM）、CD-ROM、磁带、软盘和光数据存储设备等；该计算机存储介质中存储有程序，当所述存储介质中的程序由数据分析平台的相应处理器执行时，使得该数据分析平台能够执行上述方法实施例中记载的无监督的特征选择方法的部分或全部步骤。

相应于上述无监督的特征选择方法，本申请实施例还提供了一种无监督的特征选择装置。图 4 为该装置的一种结构框图，具体包括：相似度计算单元 301、拓扑图构建单元 302、连通图挖掘单元 303 和目标特征选择单元 304。

其中，该相似度计算单元 301 用于，计算样本数据的所有特征之间的相似度。

该拓扑图构建单元 302 用于，以特征为节点，根据节点之间的相似度构建特征拓扑图。

该连通图挖掘单元 303 用于，对所述特征拓扑图进行分割，得到一个或多个连通图。

该目标特征选择单元 304 用于，根据所述连通图，确定目标特征集。

由以上装置结构可知，本申请实施例首先根据特征之间的相似度构建特征拓扑图，然后将该特征拓扑图进行分割，使得相似度较高的特征节点划分在同一个连通图中，实现对样本数据的特征聚类，从而在无法预先确定样本数据的分类标示的场景下也可以对样本数据的特征进行选择；在通过连通图完成特征聚类后，进一步从每个连通图中选出一个节点，并将该节点对应的特征为代表特征，记入目标特征集，从而得到整个样本数据对应的全面且不重复的代表特征。因此，本实施例不需要依赖样本数据的分类标示，可以实现无监督的特征选择，且保证目标特征集中不会出现两个或两个以上相类似的特征，使得目标特征集可以更直观地描述样本数据，避免信息冗余。

在本申请一个可行的实施例中，上述拓扑图构建单元 302 具体被配置为：在任意相似度大于零的两个节点之间添加一条边，得到所述特征拓扑图。

在本申请一个可行的实施例中，上述连通图挖掘单元 303 可以包括第一挖掘模块；

该第一挖掘模块具体被配置为：分别将所述特征拓扑图中每条边对应的相似度与预设相似度阈值进行比较；删除所述特征拓扑图中相似度小于所述预设相似度阈值的边；将通过边连接的节点划分在同一个连通图中，得到一个或多个连通图。

在本申请另一个可行的实施例中，上述连通图挖掘单元 303 也可以包括第二挖掘模块；该第二挖掘模块具体被配置为：为所述特征拓扑图中的每个节点设置唯一的标签；将每个节点的标签传播至相应的邻居节点；对于每个节点，将自己的标签和接收到的所有来自邻居的标签进行比较，选其中值最小的标签作为相应节点的新标签；判断各个节点的标签是否发生变化；如果存在至少一个节点的标签发生变化，则返回执行所述将每个节点的标签传播至相应的邻居节点的步骤；如果所有节点的标签均未发生变化，则将
5 10 标签相同的节点划分在同一个连通图中，得到一个或多个连通图。

在本申请一个可行的实施例中，上述目标特征选择单元 304 具体被配置为：分别确定每个连通图中的核心节点，并将各个核心节点对应的特征记入目标特征集。

可选的，为确定每个连通图中的核心节点，所述目标特征选择单元具体被配置为：根据各个节点的节点度值、介数值和 Kcore 系数中的任意一种确定每个连通图中的核心
15 节点。

关于上述实施例中的装置，其中各个单元执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本发明的其它实施方案。本申请旨在涵盖本发明的任何变型、用途或者适应性变化，这些变型、用途
20 或者适应性变化遵循本发明的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本发明的真正范围和精神由下面的权利要求指出。

应当理解的是，本发明并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本发明的范围仅由所附的权利要求来限制。

权 利 要 求 书

1、一种无监督的特征选择方法，其特征在于，包括：

计算样本数据的所有特征之间的相似度；

以特征为节点，根据节点之间的相似度构建特征拓扑图；

5 对所述特征拓扑图进行分割，得到一个或多个连通图；

根据所述连通图，确定目标特征集。

2、根据权利要求 1 所述的方法，其特征不在于，根据节点之间的相似度构建特征拓扑图，包括：

10 在任意相似度大于零的两个节点之间添加一条边，得到所述特征拓扑图。

3、根据权利要求 1 或 2 所述的方法，其特征不在于，对所述特征拓扑图进行分割，得到一个或多个连通图，包括：

分别将所述特征拓扑图中每条边对应的相似度与预设相似度阈值进行比较；

15 删除所述特征拓扑图中相似度小于所述预设相似度阈值的边；

将通过边连接的节点划分在同一个连通图中，得到一个或多个连通图。

4、根据权利要求 1 或 2 所述的方法，其特征不在于，对所述特征拓扑图进行分割，得到一个或多个连通图，包括：

20 为所述特征拓扑图中的每个节点设置唯一的标签；

将每个节点的标签传播至相应的邻居节点；

对于每个节点，将自己的标签和接收到的所有来自邻居的标签进行比较，选其中值最小的标签作为相应节点的新标签；

判断各个节点的标签是否发生变化；

25 如果存在至少一个节点的标签发生变化，则返回执行所述将每个节点的标签传播至相应的邻居节点的步骤；

如果所有节点的标签均未发生变化，则将标签相同的节点划分在同一个连通图中，得到一个或多个连通图。

30 5、根据权利要求 1 或 2 所述的方法，其特征不在于，根据所述连通图，确定目标特

征集，包括：

分别确定每个连通图中的核心节点，并将各个核心节点对应的特征记入目标特征集。

- 5 6、根据权利要求5所述的方法，其特征在于，分别确定每个连通图中的核心节点，包括：

根据各个节点的节点度值、介数值和 Kcore 系数中的任意一种确定每个连通图中的核心节点。

- 10 7、一种无监督的特征选择装置，其特征在于，包括：

相似度计算单元，用于计算样本数据的所有特征之间的相似度；

拓扑图构建单元，用于以特征为节点，根据节点之间的相似度构建特征拓扑图；

连通图挖掘单元，用于对所述特征拓扑图进行分割，得到一个或多个连通图；

目标特征选择单元，用于根据所述连通图，确定目标特征集。

15

- 8、根据权利要求7所述的装置，其特征在于，所述拓扑图构建单元具体被配置为：在任意相似度大于零的两个节点之间添加一条边，得到所述特征拓扑图。

- 9、根据权利要求7或8所述的装置，其特征在于，所述连通图挖掘单元包括：第一挖掘模块；

所述第一挖掘模块被配置为：

分别将所述特征拓扑图中每条边对应的相似度与预设相似度阈值进行比较；

删除所述特征拓扑图中相似度小于所述预设相似度阈值的边；

将通过边连接的节点划分在同一个连通图中，得到一个或多个连通图。

25

- 10、根据权利要求7或8所述的装置，其特征在于，所述连通图挖掘单元包括：第二挖掘模块；

所述第二挖掘模块被配置为：

为所述特征拓扑图中的每个节点设置唯一的标签；

- 30 将每个节点的标签传播至相应的邻居节点；

对于每个节点，将自己的标签和接收到的所有来自邻居的标签进行比较，选其中值最小的标签作为相应节点的新标签；

判断各个节点的标签是否发生变化；

如果存在至少一个节点的标签发生变化，则返回执行所述将每个节点的标签传播至相应的邻居节点的步骤；

如果所有节点的标签均未发生变化，则将标签相同的节点划分在同一个连通图中，得到一个或多个连通图。

11、根据权利要求 7 或 8 所述的装置，其特征在于，所述目标特征选择单元具体被配置为：

分别确定每个连通图中的核心节点，并将各个核心节点对应的特征记入目标特征集。

12、根据权利要求 11 所述的装置，其特征在于，为确定每个连通图中的核心节点，所述目标特征选择单元具体被配置为：

根据各个节点的节点度值、介数值和 K_{core} 系数中的任意一种确定每个连通图中的核心节点。

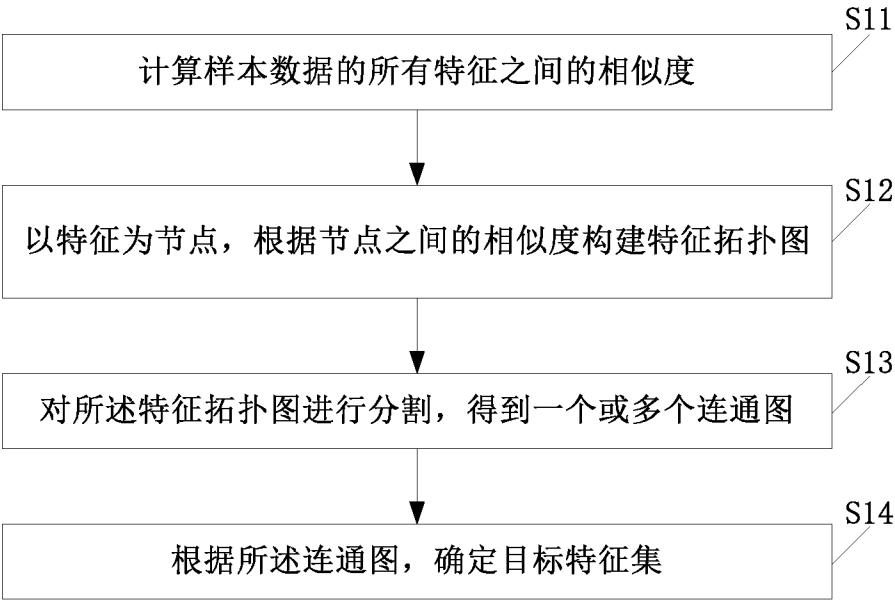


图 1

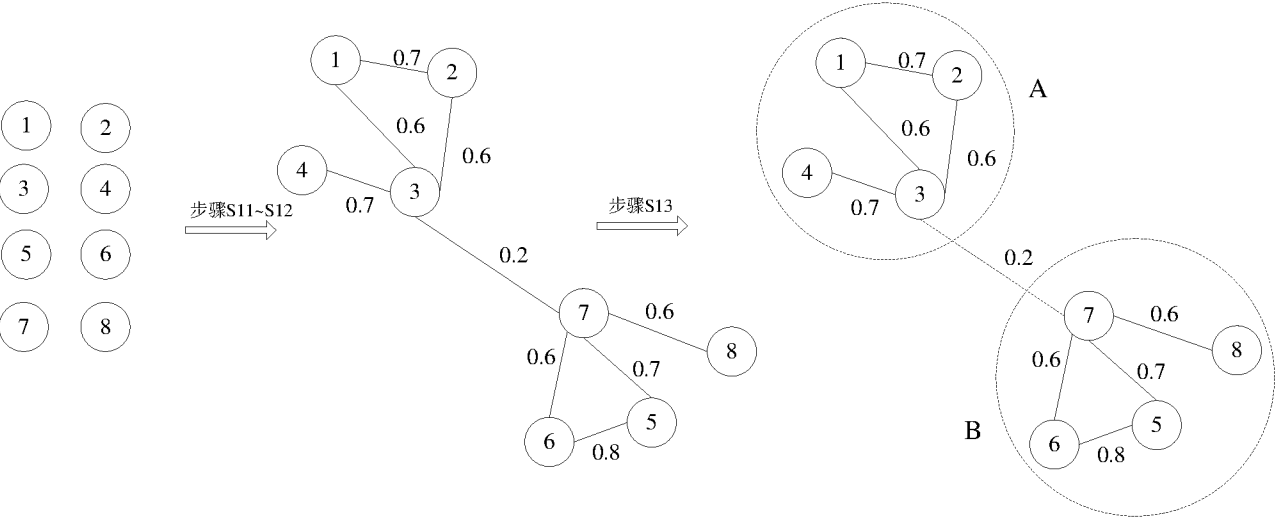


图 2

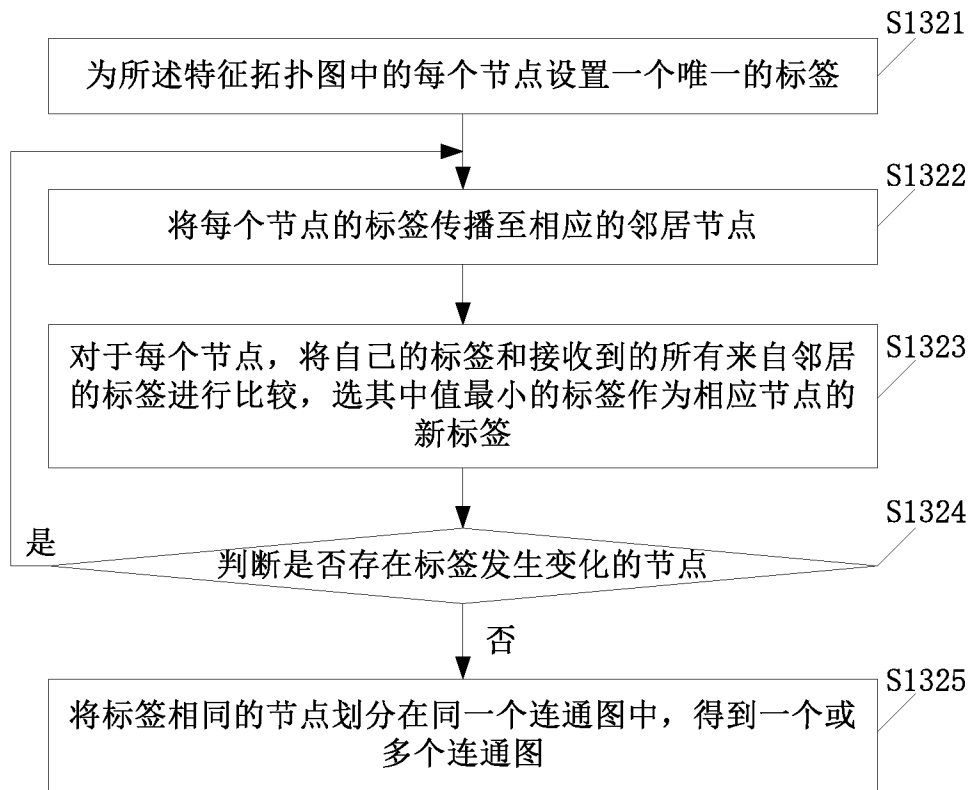


图 3



图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/077997

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) 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)

USTXT; EPTXT; CNTXT; CNABS; WOTXT; DWPI; SIPOABS: connected graph, Alibaba; HUANG, Guangyuan; MAI, Ganglin; LAN, Jinhe; feature, select, property, attribute, similar, topology, graph, cut, partition, divide, cluster, node, decrease, omit, avoid, delete, discard, redundancy, dimension, complexity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103853831 A (THE 28TH RESEARCH INSTITUTE OF CHINA ELECTRONICS TECHNOLOGY GROUP CORP.), 11 June 2014 (11.06.2014), the whole document	1-12
A	US 2013177244 A1 (GENERAL ELECTRIC COMPANY), 11 July 2013 (11.07.2013), the whole document	1-12
A	CN 101470731 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES), 01 July 2009 (01.07.2009), the whole document	1-12
Y	ZHANG, Lijun, "Research and Application of Large-Scale Machine Learning Theory", ELECTRONIC TECHNOLOGY & INFORMATION SCIENCE, CHINA DOCTORAL DISSERTATIONS FULL-TEXT DATABASE, no. 12, 15 December 2013 (15.12.2013), ISSN: 1674-022X, sections 2.2.2-2.3.1	1-12

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 25 May 2017 (25.05.2017)	Date of mailing of the international search report 16 June 2017 (15.06.2017)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer WU, Shifen Telephone No.: (86-10) 62089928

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/077997

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	XIA, Lei et al., "Applied Research of Node Similarity Label Propagation in Social Networks", COMPUTER ENGINEERING AND APPLICATIONS, no. 14, 31 July 2014 (31.07.2014), ISSN: 1002-8331, the whole document	1-12
Y	MITRA, P. et al. "Unsupervised Feature Selection Using Feature Similarity", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 24, no. 3, 31 March 2002 (31.03.2002), ISSN: 0162-8828, sections 2-3	1-12

International application No.
PCT/CN2017/077997

Form PCT/ISA/210 (patent family annex) (July 2009)

A. 主题的分类 G06F 17/30 (2006.01) 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)) USTXT; EPTXT; CNTXT; CNABS; WOTXT; DWPI; SIPOABS: 阿里巴巴, 黄光远, 麦港林, 兰金鹤, 特征, 属性, 选择, 相似, 拓扑, 连通图, 分割, 拆分, 聚类, 节点, 减少, 省略, 避免, 删除, 丢弃, 维度, 冗余, Alibaba, Huang Guangyuan, Mai Ganglin, Lan Jinhe, feature, select, property, attribute, similar, topology, graph, cut, partition, divide, cluster, node, decrease, omit, avoid, delete, discard, redundancy, dimension, complexity																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>CN 103853831 A (中国电子科技集团公司第二十八研究所) 2014年 6月 11日 (2014 - 06 - 11) 全文</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>US 2013177244 A1 (GENERAL ELECTRIC COMPANY) 2013年 7月 11日 (2013 - 07 - 11) 全文</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>CN 101470731 A (中国科学院自动化研究所) 2009年 7月 1日 (2009 - 07 - 01) 全文</td> <td>1-12</td> </tr> <tr> <td>Y</td> <td>张利军. "大规模机器学习理论研究与应用" 中国博士学位论文全文数据库 信息科技辑, 第12期, 2013年 12月 15日 (2013 - 12 - 15), ISSN: 1674-022X, 第2.2.2-2.3.1节</td> <td>1-12</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	A	CN 103853831 A (中国电子科技集团公司第二十八研究所) 2014年 6月 11日 (2014 - 06 - 11) 全文	1-12	A	US 2013177244 A1 (GENERAL ELECTRIC COMPANY) 2013年 7月 11日 (2013 - 07 - 11) 全文	1-12	A	CN 101470731 A (中国科学院自动化研究所) 2009年 7月 1日 (2009 - 07 - 01) 全文	1-12	Y	张利军. "大规模机器学习理论研究与应用" 中国博士学位论文全文数据库 信息科技辑, 第12期, 2013年 12月 15日 (2013 - 12 - 15), ISSN: 1674-022X, 第2.2.2-2.3.1节	1-12
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
A	CN 103853831 A (中国电子科技集团公司第二十八研究所) 2014年 6月 11日 (2014 - 06 - 11) 全文	1-12															
A	US 2013177244 A1 (GENERAL ELECTRIC COMPANY) 2013年 7月 11日 (2013 - 07 - 11) 全文	1-12															
A	CN 101470731 A (中国科学院自动化研究所) 2009年 7月 1日 (2009 - 07 - 01) 全文	1-12															
Y	张利军. "大规模机器学习理论研究与应用" 中国博士学位论文全文数据库 信息科技辑, 第12期, 2013年 12月 15日 (2013 - 12 - 15), ISSN: 1674-022X, 第2.2.2-2.3.1节	1-12															
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。 <table border="0"> <tr> <td> * 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 </td> <td> "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 </td> </tr> </table>			* 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													
* 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 2017年 5月 25日		国际检索报告邮寄日期 2017年 6月 16日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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 吴士芬 电话号码 (86-10) 62089928															

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
A	夏磊等. “节点相似度标签传播在社会网络中的应用研究” 计算机工程与应用, 第14期, 2014年 7月 31日 (2014 - 07 - 31), ISSN: 1002-8331, 全文	1-12
Y	MITRA, P. 等. “Unsupervised Feature Selection Using Feature Similarity” IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 第24卷, 第3期, 2002年 3月 31日 (2002 - 03 - 31), ISSN: 0162-8828, 第2-3节	1-12

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/077997

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	103853831	A	2014年 6月 11日	CN	103853831	B	2017年 2月 1日
US	2013177244	A1	2013年 7月 11日	US	8774513	B2	2014年 7月 8日
CN	101470731	A	2009年 7月 1日	CN	101470731	B	2012年 6月 20日

表 PCT/ISA/210 (同族专利附件) (2009年7月)