(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110880006 A
(43)申请公布日 2020.03.13

(21)申请号 201811034036.6

(22)申请日 2018.09.05

(71)申请人 广州视源电子科技股份有限公司
地址 510530 广东省广州市黄埔区云埔四路6号

(72)发明人 方建生

(74)专利代理机构 广州华进联合专利商标代理有限公司 44224
代理人 黄晓庆

(51)Int.Cl.
G06K 9/62(2006.01)
G06F 16/35(2019.01)
G06Q 30/02(2012.01)
G06Q 50/00(2012.01)

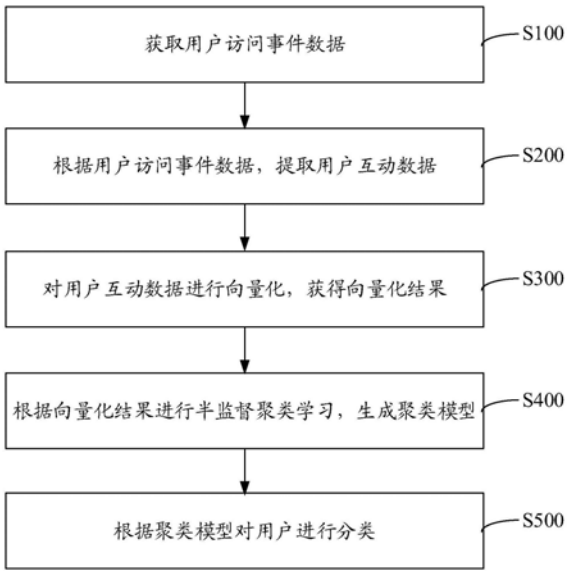
权利要求书2页 说明书8页 附图5页

(54)发明名称

用户分类方法、装置、计算机设备和存储介质

(57)摘要

本申请提供一种用户分类方法、装置、计算机设备和存储介质，其中，方法包括：获取用户访问事件数据，从用户访问事件数据中，提取用户互动数据，对用户互动数据进行向量化，并且根据向量化结果进行半监督聚类学习，生成聚类模型，再根据聚类模型对用户进行分类。整个过程中，基于用户访问事件，准确获取用户访问事件数据，并且采用用户互动数据向量化+半监督聚类学习，充分考虑互动数据中数据特征进行聚类操作，构建聚类模型，因此，能够对用户精准分类。



CN 110880006 A

1.一种用户分类方法,所述方法包括:

获取用户访问事件数据;

根据所述用户访问事件数据,提取用户互动数据;

对用户互动数据进行向量化,获得向量化结果;

根据所述向量化结果进行半监督聚类学习,生成聚类模型;

根据所述聚类模型对用户进行分类。

2.根据权利要求1所述的方法,其特征在于,所述对用户互动数据进行向量化,获得向量化结果包括:

根据当前预设文本关键词,对所述用户互动数据进行TF-IDF计算,将所述TF-IDF计算值作为向量化结果值。

3.根据权利要求2所述的方法,其特征在于,所述根据所述向量化结果进行半监督聚类学习,生成聚类模型之后,还包括:

对所述聚类模型进行DBI评估,记录当前预设文本关键词下DBI评估值;

更新所述当前预设文本关键词,返回所述根据当前预设文本关键词,对所述用户互动数据进行TF-IDF计算的步骤。

4.根据权利要求3所述的方法,其特征在于,所述根据所述聚类模型对用户进行分类包括:

查找DBI评估值最小对应的聚类模型,根据查找到的所述聚类模型对用户进行分类。

5.根据权利要求1所述的方法,其特征在于,所述对用户互动数据进行向量化,获得向量化结果之前,还包括:

标记所述用户互动数据中已关联客户标识的用户互动数据以及未关联用户标识的用户互动数据。

6.根据权利要求5所述的方法,其特征在于,所述对用户互动数据进行向量化,获得向量化结果包括:

分别获取已关联客户的用户互动数据向量化结果和未关联客户的用户互动数据向量化结果;

所述根据所述向量化结果进行半监督聚类学习,生成聚类模型包括:

使用所述已关联客户的用户互动数据向量化结果对所述未关联客户的用户互动数据向量化结果进行约束种子k均值聚类学习,生成聚类模型。

7.根据权利要求1所述的方法,其特征在于,所述用户访问事件包括菜单点击、页面浏览以及互动文本。

8.一种用户分类装置,其特征在于,所述装置包括:

获取模块,用于获取用户访问事件数据;

提取模块,用于根据所述用户访问事件数据,提取用户互动数据;

向量化模块,用于对用户互动数据进行向量化,获得向量化结果;

聚类模块,用于根据所述向量化结果进行半监督聚类学习,生成聚类模型;

分类模块,用于根据所述聚类模型对用户进行分类。

9.一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,其特征在于,所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述方法的步骤。

10.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的方法的步骤。

用户分类方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及数据处理技术领域，特别是涉及一种用户分类方法、装置、计算机设备和存储介质。

背景技术

[0002] 企业用户(粉丝)是关注和参与网络社交平台企业号的虚拟群体，如QQ企业号、企业论坛、微信公众号、微博企业号等，是企业维系和营销的对象。用户在企业号上的行为，包括企业号功能菜单点击、页面浏览以及互动的文本记录，是用户数据挖掘的重要信息，可发现用户对企业产品的关注点或兴趣点。

[0003] 针对用户的精准维系和营销，前提是有效识别出用户的偏好并分类，一般是根据用户的活跃度以及菜单和页面访问轨迹综合分析而定性。如果用户关联了客户标识，则可明确知道用户是企业的客户，针对其使用的产品提供维系服务和增值营销。然而，传统的用户分类方法无法实现精准分类。

发明内容

[0004] 基于此，有必要针对上述技术问题，提供一种能够精准分类的用户分类方法、装置、计算机设备和存储介质。

[0005] 一种用户分类方法，所述方法包括：

[0006] 获取用户访问事件数据；

[0007] 根据所述用户访问事件数据，提取用户互动数据；

[0008] 对用户互动数据进行向量化，获得向量化结果；

[0009] 根据所述向量化结果进行半监督聚类学习，生成聚类模型；

[0010] 根据所述聚类模型对用户进行分类。

[0011] 在其中一个实施例中，所述对用户互动数据进行向量化，获得向量化结果包括：

[0012] 根据当前预设文本关键词，对所述用户互动数据进行TF-IDF(词频-逆向文件频率)计算，将所述TF-IDF计算值作为向量化结果值。

[0013] 在其中一个实施例中，所述根据所述向量化结果进行半监督聚类学习，生成聚类模型之后，还包括：

[0014] 对所述聚类模型进行DBI(基于聚类性能度量指标)评估，记录当前预设文本关键词下DBI评估值；

[0015] 更新所述当前预设文本关键词，返回所述根据当前预设文本关键词，对所述用户互动数据进行TF-IDF计算的步骤。

[0016] 在其中一个实施例中，所述根据所述聚类模型对用户进行分类包括：

[0017] 查找DBI评估值最小对应的聚类模型，根据查找到的所述聚类模型对用户进行分类。

[0018] 在其中一个实施例中，所述对用户互动数据进行向量化，获得向量化结果之前，还

包括：

[0019] 标记所述用户互动数据中已关联客户标识的用户互动数据以及未关联用户标识的用户互动数据。

[0020] 在其中一个实施例中，所述对用户互动数据进行向量化，获得向量化结果包括：

[0021] 分别获取已关联客户的用户互动数据向量化结果和未关联客户的用户互动数据向量化结果；

[0022] 所述根据所述向量化结果进行半监督聚类学习，生成聚类模型包括：

[0023] 使用所述已关联客户的用户互动数据向量化结果对所述未关联客户的用户互动数据向量化结果进行约束种子k均值聚类学习，生成聚类模型。

[0024] 在其中一个实施例中，所述用户访问事件包括菜单点击、页面浏览以及互动文本。

[0025] 一种用户分类装置，所述装置包括：

[0026] 获取模块，用于获取用户访问事件数据；

[0027] 提取模块，用于根据所述用户访问事件数据，提取用户互动数据；

[0028] 向量化模块，用于对用户互动数据进行向量化，获得向量化结果；

[0029] 聚类模块，用于根据所述向量化结果进行半监督聚类学习，生成聚类模型；

[0030] 分类模块，用于根据所述聚类模型对用户进行分类。

[0031] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如上述方法的步骤。

[0032] 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如上述的方法的步骤。

[0033] 上述用户分类方法、装置、计算机设备和存储介质，获取用户访问事件数据，从用户访问事件数据中，提取用户互动数据，对用户互动数据进行向量化，并且根据向量化结果进行半监督聚类学习，生成聚类模型，再根据聚类模型对用户进行分类。整个过程中，基于用户访问事件，准确获取用户访问事件数据，并且采用用户互动数据向量化+半监督聚类学习，充分考虑互动数据中数据特征进行聚类操作，构建聚类模型，因此，能够对用户精准分类。

附图说明

[0034] 图1为一个实施例中用户分类方法的应用环境图；

[0035] 图2为一个实施例中用户分类方法的流程示意图；

[0036] 图3为另一个实施例中用户分类方法的流程示意图；

[0037] 图4为一个实施例中用户分类装置的结构框图；

[0038] 图5为另一个实施例中用户分类装置的结构框图；

[0039] 图6为一个实施例中计算机设备的内部结构图。

具体实施方式

[0040] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0041] 本申请提供的用户分类方法，可以应用于如图1所示的应用环境中。其中，终端102通过网络与服务器104通过网络进行通信。用户通过终端102进行包括菜单点击、页面浏览以及互动文本的访问操作，终端102采集访问事件数据，并将访问事件数据发送至服务器104，服务器104获取单个或者多个终端104上传的用户访问事件数据，从用户访问事件数据中，提取用户互动数据，对用户互动数据进行向量化，并且根据向量化结果进行半监督聚类学习，生成聚类模型，再根据聚类模型对用户进行分类。其中，终端102可以但不限于是各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器104可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

[0042] 在一个实施例中，如图2所示，提供了一种用户分类方法，以该方法应用于图1中的服务器为例进行说明，包括以下步骤：

[0043] S100：获取用户访问事件数据。

[0044] 用户访问事件包括用户通过终端访问企业客户的官网、与企业客户在线服务热线进行交流以及在企业客户公众号下留言等。具体的，可以是用户在终端上进行上述访问操作，终端将用户访问事件数据发送至服务器，服务器获取用户访问事件数据，或服务器在与终端进行数据交互时，采集获取用户访问事件数据。

[0045] S200：根据用户访问事件数据，提取用户互动数据。

[0046] 用户访问企业客户的过程中是一个互动的过程，提取这个互动过程的用户互动数据。具体来说，若用户是点击企业客户APP(Application，应用程序)中菜单或浏览企业客户官网，则可以直接提取预设与菜单点击或网页浏览对应的数据，得到用户互动数据；若用户是在企业客户公众号或微博上留言，则可以直接将留言的文字数据作为用户互动数据。非必要的，提取用户互动数据可以理解为将访问事件数据汇聚成一段短文本。

[0047] S300：对用户互动数据进行向量化，获得向量化结果。

[0048] 用户互动数据中具有一定聚合和相似性，采用对用户互动数据进行向量化处理，充分考虑、挖掘用户互动数据中的聚合与相似性，饿到向量化结果。非必要的，对用户互动数据进行向量化的过程可以是对用户互动数据进行TF-IDF计算，将TF-IDF计算值作为向量化结果值。

[0049] S400：根据向量化结果进行半监督聚类学习，生成聚类模型。

[0050] 在向量化结果中携带有用户互动数据之间的聚合和相似性，在这里对向量化结果进行半监督聚类学习，进一步挖掘用户互动数据之间的聚合性，生成聚类模型。具体来说，可以采用半监督聚类方法中约束种子k均值(Constrained Seed k-Means)算法进行半监督聚类学习，生成聚类模型。k均值算法定义：给定样本集$D=\{x_1,x_2,\cdots,x_n\}$，k均值(k-means)算法针对聚类所得簇划分$C=\{C_1,C_2,\cdots,C_k\}$，最小化平方误差$E=\sum_{i=1}^{k}\sum_{x\in C_i}(x-\mu_i)^2$，其中$\mu_i=\frac{1}{C_i}\sum_{x\in C_i}x$是簇$C_i$的均值向量。E刻画了簇内样本围绕均值向量的紧密程度，越小则簇内样本相似度越高。

[0051] S500：根据聚类模型对用户进行分类。

[0052] 在聚类模型中已经充分挖掘用户之间的聚合性和相似性，因此，可以根据聚类模型精准对用户进行分类。

[0053] 上述用户分类方法，获取用户访问事件数据，从用户访问事件数据中，提取用户互动数据，对用户互动数据进行向量化，并且根据向量化结果进行半监督聚类学习，生成聚类模型，再根据聚类模型对用户进行分类。整个过程中，基于用户访问事件，准确获取用户访问事件数据，并且采用用户互动数据向量化+半监督聚类学习，充分考虑互动数据中数据特征进行聚类操作，构建聚类模型，因此，能够对用户精准分类。

[0054] 如图3所示，在其中一个实施例中，对用户互动数据进行向量化，获得向量化结果包括：

[0055] 根据当前预设文本关键词，对用户互动数据进行TF-IDF计算，将TF-IDF计算值作为向量化结果值。

[0056] 当前预设文本关键词是指针对当前时刻的预设文本关键词，预设文本关键词是预先设定的关键词，一般来说，预设的关键词会与企业客户之间存在一定相关性。TF-IDF计算是指词频-逆向文件频率计算，词频指的是某一个给定的词语在该文本中出现的次数。定义$TF_i = \frac{n_i}{\sum_k n_k}$，其中分子表示文本中第i个词语出现的次数，分母表示文本中所有词语出现的次数总和。逆向文件频率是由总文件数目除以包含词语的文本数目，再将得到的商取对数得到。定义$IDF_i = \log \frac{|D|}{|j:i \in j|}$，其中对数的分子表示文本总数，对数的分母是包含第i个词语的文本总数j。关键词在整体训练文本数的占比以及单个文本中的频次体现了文本的特征。基于如下的假设：用文本关键词的TF-IDF值作为文本向量。假定用k个关键词来提取文本特征，定义$x=\{x_1,x_2,...,x_k\}$表示关键词和具体文本相关性的向量，并定义$x_i=TF_i \cdot IDF_i$ $(i=1,2,...,k)$。

[0057] 如图3所示，在其中一个实施例中，步骤S400之后，还包括：

[0058] S420：对聚类模型进行DBI评估，记录当前预设文本关键词下DBI评估值。

[0059] S440：更新当前预设文本关键词，返回根据当前预设文本关键词，对用户互动数据进行TF-IDF计算的步骤。

[0060] 聚类性能度量指标用于评估训练出的模型好坏，由于用户数据簇间定义比较明确，所以选用DBI指数来度量。DBI定义：

[0061] $$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \left(\frac{avg(C_i) + avg(C_j)}{d_{cen}(\mu_i, \mu_j)} \right)$$

[0062] 其中avg(C)表示簇C内样本间的平均距离，$dcen(\mu_i,\mu_j)$表示簇C_i和簇C_j心点间的距离，k是聚类簇的数量。根据DBI定义，DBI值越小，训练出的模型性能越好。对于同一训练集，结合分类标签的定义，可以设置不同的关键词，文本向量化的特征值将会不同，训练出的模型泛化性能就有优劣之分，通过DBI评价可选择最优模型。

[0063] 在其中一个实施例中，根据聚类模型对用户进行分类包括：

[0064] 查找DBI评估值最小对应的聚类模型，根据查找到的聚类模型对用户进行分类。

[0065] 如之前，模型对应的DBI值越小，其性能越好。在这里，查找到DBI评估值最小对应的聚类模型，根据查找到的聚类模型对用户进行分类。

[0066] 如图3所示，在其中一个实施例中，步骤S300之前还包括：

[0067] S220:标记用户互动数据中已关联客户标识的用户互动数据以及未关联用户标识的用户互动数据。

[0068] 对关联客户标识的用户打上所属类别的标签,这里的类别结合企业实际的情况来定义,比如按照企业的产品来分类、也可按照企业客户的等级来分类,取决于分类后要开始实际维系和营销的场景。

[0069] 在其中一个实施例中,对用户互动数据进行向量化,获得向量化结果包括:分别获取已关联客户的用户互动数据向量化结果和未关联客户的用户互动数据向量化结果;

[0070] 根据向量化结果进行半监督聚类学习,生成聚类模型包括:

[0071] 使用已关联客户的用户互动数据向量化结果对未关联客户的用户互动数据向量化结果进行约束种子k均值聚类学习,生成聚类模型。

[0072] 具体来说,可以使用少量有标记样本对无标记样本的聚类过程进行约束和指导,即使用少量已关联客户的用户互动数据向量化结果对未关联客户的用户互动数据向量化结果进行约束种子k均值聚类学习,生成聚类模型。约束种子k均值利用监督信息中有标识样本作为种子,初始化k均值算法的k个聚类中心,且在聚类簇迭代更新过程中不改变种子样本的簇隶属关系。

[0073] 应该理解的是,虽然图2-图3的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,图2-图3中的至少一部分步骤可以包括多个子步骤或者多个阶段,这些子步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些子步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0074] 为更进一步详细解释上述用户分类方法的技术方案及其效果,下面将采用具体应用实例进行说明。

[0075] 在具体应用实例中,针对已完成用户互动数据并为已关联客户标识的用户打上分类标签,定义企业四类产品为类别标签作为聚类簇$y=\{1,2,3,4\}$,并初步设置k个关键词$\{w_1,w_2,\cdots,w_k\}$用于文本向量化。n个未标记样本集,表示为:

[0076] $D_u=\{x_1,x_2,\cdots,x_n\}$,$x_i=\{x_{i1},x_{i2},\cdots,x_{ik}\}$ $(i=1,2,...,n)$。

[0077] m个有标记样本,表示为:

[0078] $D_l=\{(x_1,y),(x_2,y),\cdots,(x_m,y)\}$,$x_i=\{x_{i1},x_{i2},\cdots,x_{ik}\}$ $(i=1,2,\cdots,m)$。

[0079] 文本向量化。文本向量化算法主要三个步骤是:计算关键词词频、计算关键词逆文件频率、样本向量构造。不同的关键词设置,输出的样本向量值不一样,在经过约束种子k均值算法训练出模型后,通过DBI比较性能,选择最佳的模型用于分类。半监督聚类约束种子k均值算法。算法主要过程是将有标记样本作为种子计算簇的均值向量,然后为未标记样本聚类簇,迭代到簇的均值向量不再更新为止。

[0080] 如图4所示,一种用户分类装置,装置包括:

[0081] 获取模块100,用于获取用户访问事件数据;

[0082] 提取模块200,用于根据用户访问事件数据,提取用户互动数据;

[0083] 向量化模块300,用于对用户互动数据进行向量化,获得向量化结果;

[0084] 聚类模块400,用于根据向量化结果进行半监督聚类学习,生成聚类模型;

[0085] 分类模块500,用于根据聚类模型对用户进行分类。

[0086] 上述用户分类装置,获取模块100获取用户访问事件数据,提取模块200从用户访问事件数据中,提取用户互动数据,向量化模块300对用户互动数据进行向量化,聚类模块400根据向量化结果进行半监督聚类学习,生成聚类模型,分类模块500根据聚类模型对用户进行分类。整个过程中,基于用户访问事件,准确获取用户访问事件数据,并且采用用户互动数据向量化+半监督聚类学习,充分考虑互动数据中数据特征进行聚类操作,构建聚类模型,因此,能够对用户精准分类。

[0087] 在其中一个实施例中,向量化模块300还用于根据当前预设文本关键词,对用户互动数据进行TF-IDF计算,将TF-IDF计算值作为向量化结果值。

[0088] 如图5所示,在其中一个实施例中,上述用户分类装置还包括:

[0089] DBI评估模块420,用于对聚类模型进行DBI评估,记录当前预设文本关键词下DBI评估值;

[0090] 循环计算模块440,用于更新当前预设文本关键词,控制向量化模块300以及聚类模块进行相应操作。

[0091] 在其中一个实施例中,分类模块500还用于查找DBI评估值最小对应的聚类模型,根据查找到的聚类模型对用户进行分类。

[0092] 如图5所示,在其中一个实施例中,上述用户分类装置还包括:

[0093] 标记模块220,用于标记用户互动数据中已关联客户标识的用户互动数据以及未关联用户标识的用户互动数据。

[0094] 在其中一个实施例中,对向量化模块300还用于分别获取已关联客户的用户互动数据向量化结果和未关联客户的用户互动数据向量化结果;聚类模块400还用于使用已关联客户的用户互动数据向量化结果对未关联客户的用户互动数据向量化结果进行约束种子k均值聚类学习,生成聚类模型。

[0095] 在其中一个实施例中,用户访问事件包括菜单点击、页面浏览以及互动文本。

[0096] 关于用户分类装置的具体限定可以参见上文中对于用户分类方法的限定,在此不再赘述。上述用户分类装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行以上各个模块对应的操作。

[0097] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是服务器,其内部结构图可以如图6所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储预设关键词等数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种用户分类方法。

[0098] 本领域技术人员可以理解,图6中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0099] 在一个实施例中，提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，处理器执行计算机程序时实现以下步骤：

[0100] 获取用户访问事件数据；

[0101] 根据用户访问事件数据，提取用户互动数据；

[0102] 对用户互动数据进行向量化，获得向量化结果；

[0103] 根据向量化结果进行半监督聚类学习，生成聚类模型；

[0104] 根据聚类模型对用户进行分类。

[0105] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：

[0106] 根据当前预设文本关键词，对用户互动数据进行TF-IDF计算，将TF-IDF计算值作为向量化结果值。

[0107] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：

[0108] 对聚类模型进行DBI评估，记录当前预设文本关键词下DBI评估值；更新当前预设文本关键词，返回根据当前预设文本关键词，对用户互动数据进行TF-IDF计算的步骤。

[0109] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：

[0110] 查找DBI评估值最小对应的聚类模型，根据查找到的聚类模型对用户进行分类。

[0111] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：

[0112] 标记用户互动数据中已关联客户标识的用户互动数据以及未关联用户标识的用户互动数据。

[0113] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：

[0114] 分别获取已关联客户的用户互动数据向量化结果和未关联客户的用户互动数据向量化结果；使用已关联客户的用户互动数据向量化结果对未关联客户的用户互动数据向量化结果进行约束种子k均值聚类学习，生成聚类模型。

[0115] 在其中一个实施例中，用户访问事件包括菜单点击、页面浏览以及互动文本。

[0116] 在一个实施例中，提供了一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现以下步骤：

[0117] 获取用户访问事件数据；

[0118] 根据用户访问事件数据，提取用户互动数据；

[0119] 对用户互动数据进行向量化，获得向量化结果；

[0120] 根据向量化结果进行半监督聚类学习，生成聚类模型；

[0121] 根据聚类模型对用户进行分类。

[0122] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：

[0123] 根据当前预设文本关键词，对用户互动数据进行TF-IDF计算，将TF-IDF计算值作为向量化结果值。

[0124] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：

[0125] 对聚类模型进行DBI评估，记录当前预设文本关键词下DBI评估值；更新当前预设文本关键词，返回根据当前预设文本关键词，对用户互动数据进行TF-IDF计算的步骤。

[0126] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：

[0127] 查找DBI评估值最小对应的聚类模型，根据查找到的聚类模型对用户进行分类。

[0128] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：

[0129] 标记用户互动数据中已关联客户标识的用户互动数据以及未关联用户标识的用户互动数据。

[0130] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：

[0131] 分别获取已关联客户的用户互动数据向量化结果和未关联客户的用户互动数据向量化结果；使用已关联客户的用户互动数据向量化结果对未关联客户的用户互动数据向量化结果进行约束种子k均值聚类学习，生成聚类模型。

[0132] 在其中一个实施例中，用户访问事件包括菜单点击、页面浏览以及互动文本。

[0133] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0134] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0135] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

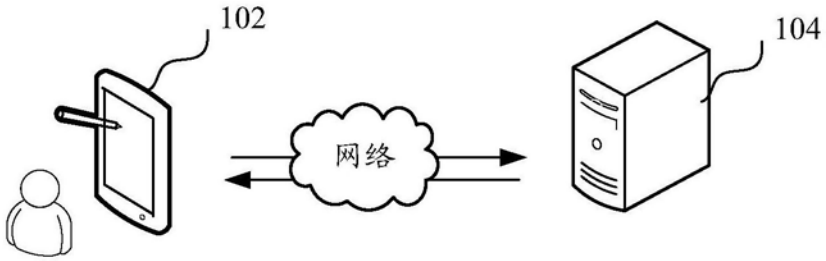

102
104
网络

图1

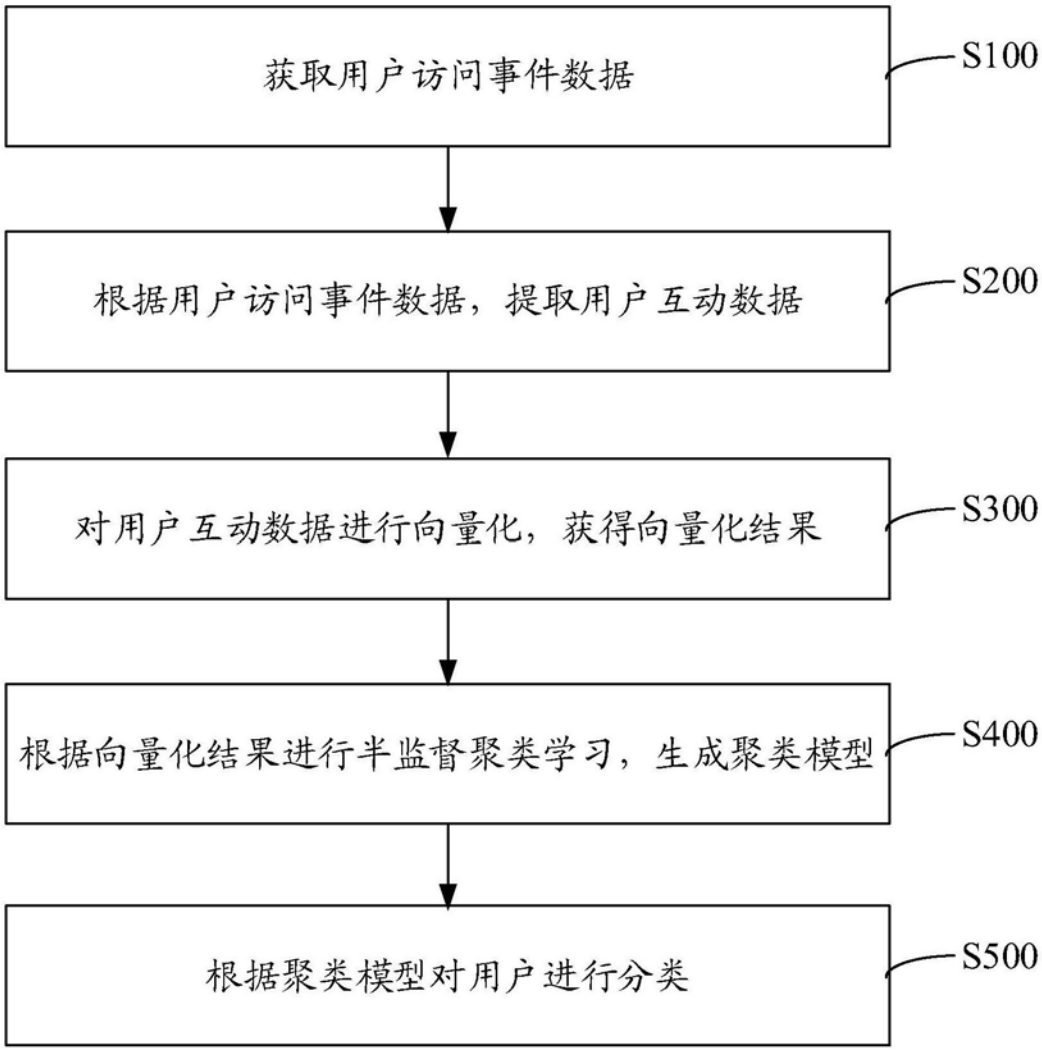

获取用户访问事件数据
S100
根据用户访问事件数据，提取用户互动数据
S200
对用户互动数据进行向量化，获得向量化结果
S300
根据向量化结果进行半监督聚类学习，生成聚类模型
S400
根据聚类模型对用户进行分类
S500

图2

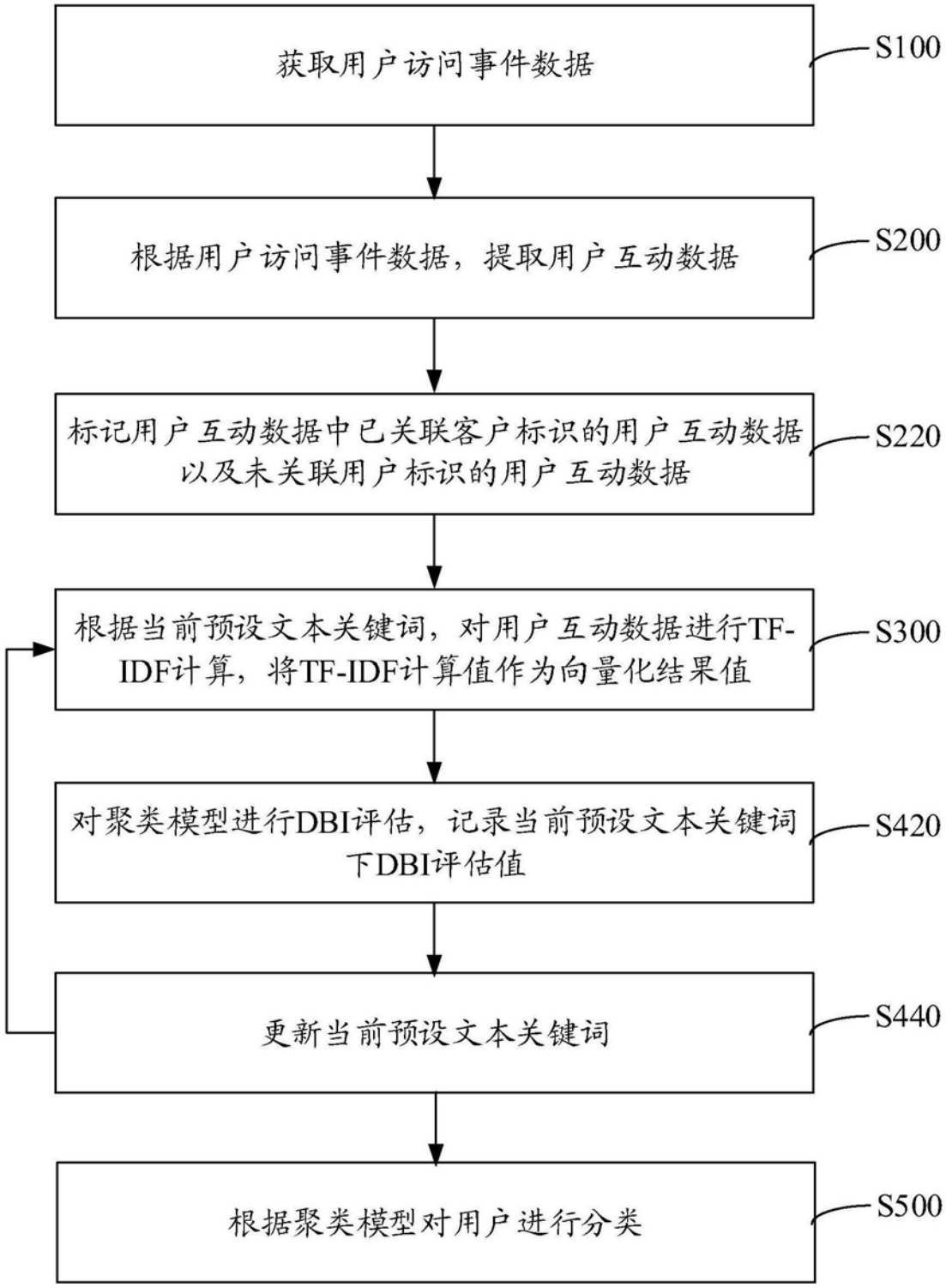

获取用户访问事件数据
S100
根据用户访问事件数据，提取用户互动数据
S200
标记用户互动数据中已关联客户标识的用户互动数据以及未关联用户标识的用户互动数据
S220
根据当前预设文本关键词，对用户互动数据进行TF-IDF计算，将TF-IDF计算值作为向量化结果值
S300
对聚类模型进行DBI评估，记录当前预设文本关键词下DBI评估值
S420
更新当前预设文本关键词
S440
根据聚类模型对用户进行分类
S500

图3

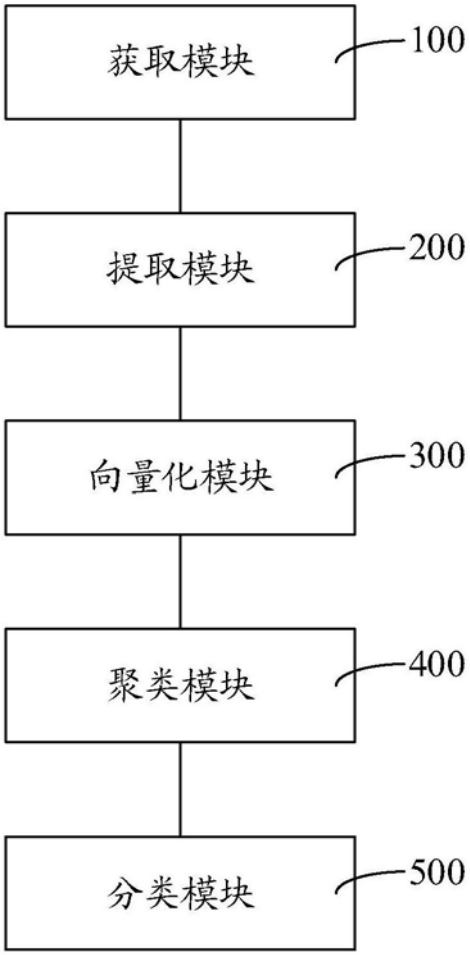


图4

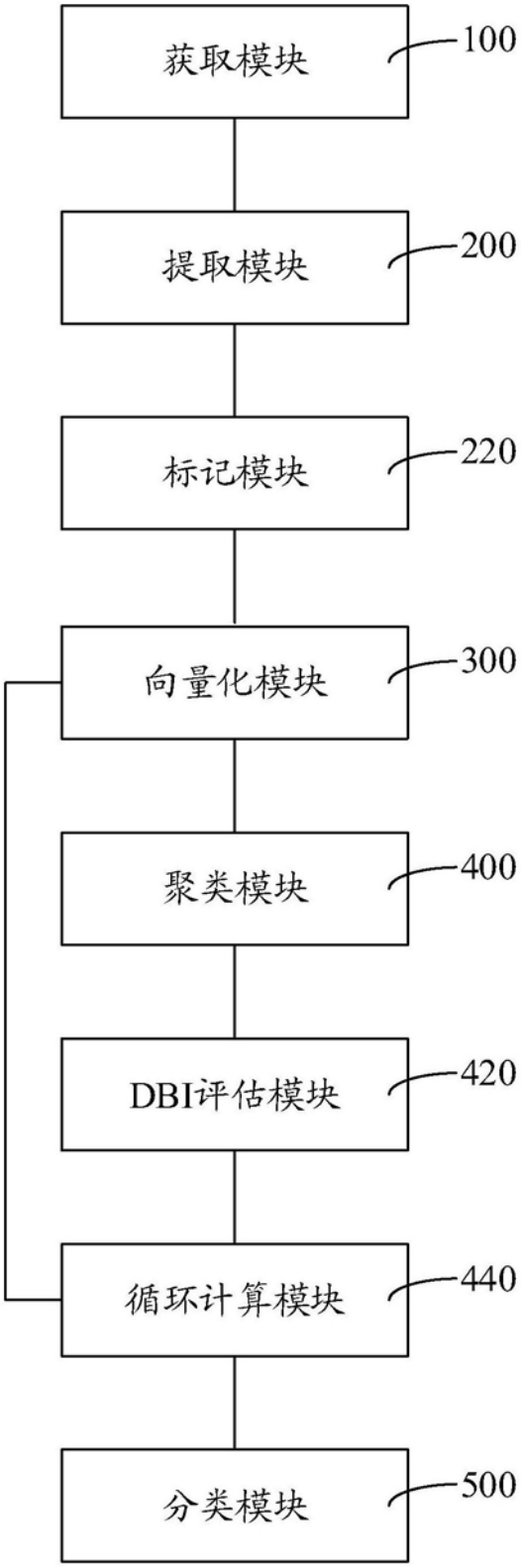

获取模块
100
提取模块
200
标记模块
220
向量化模块
300
聚类模块
400
DBI评估模块
420
循环计算模块
440
分类模块
500

图5

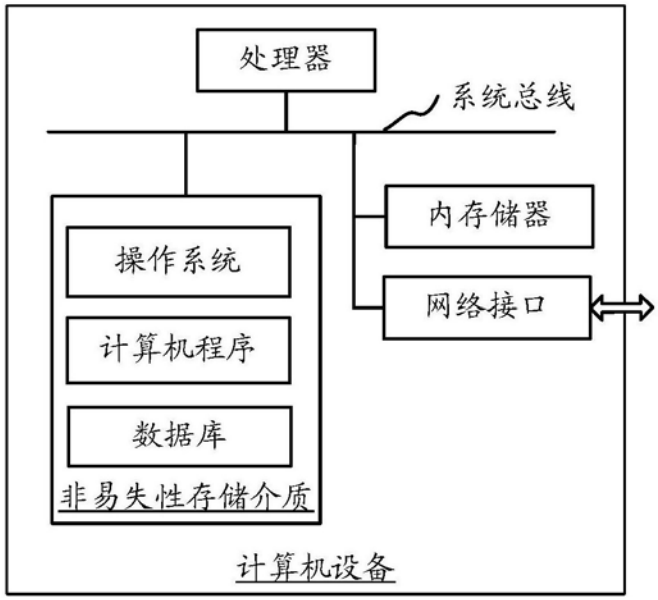


图6