

(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 15 日 (15.10.2020)



(10) 国际公布号
WO 2020/207196 A1

(51) 国际专利分类号:
G06F 16/955 (2019.01)

(21) 国际申请号: PCT/CN2020/079445

(22) 国际申请日: 2020 年 3 月 16 日 (16.03.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910284354.6 2019年4月10日 (10.04.2019) CN

(71) 申请人: 腾讯科技(深圳)有限公司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。

(72) 发明人: 陈鑫 (CHEN, Xin); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。 闫肃 (YAN, Su); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。

(74) 代理人: 北京三高永信知识产权代理有限公司 (BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市海淀区学院路蓟门里和景园 A 座 1 单元 102 室, Beijing 100088 (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: METHOD AND APPARATUS FOR GENERATING USER TAG, STORAGE MEDIUM AND COMPUTER DEVICE

(54) 发明名称: 用户标签生成方法、装置、存储介质和计算机设备

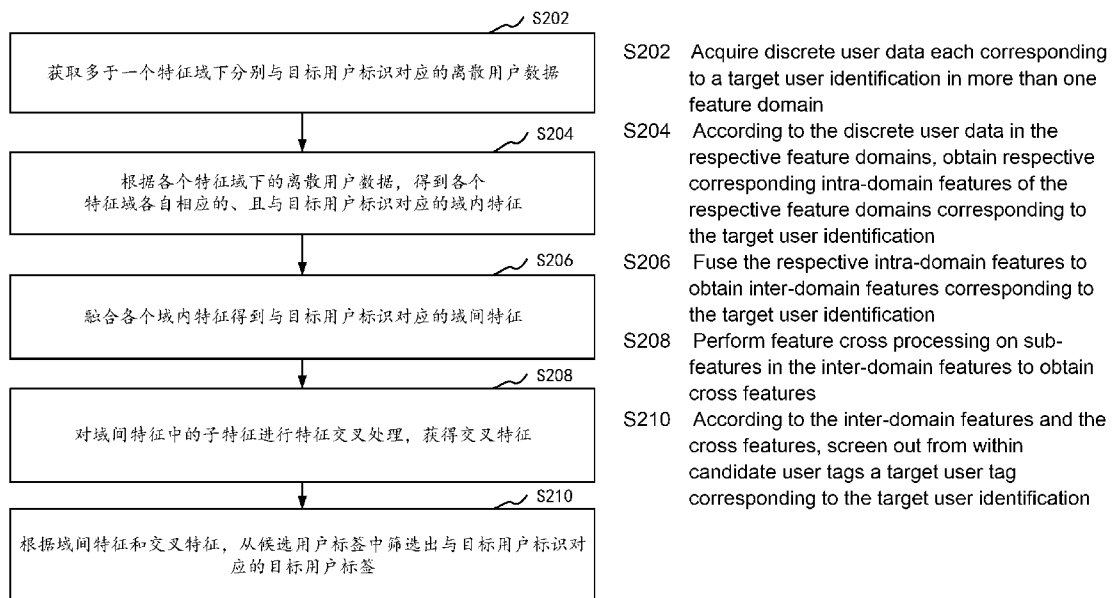


图 2

(57) Abstract: A method and apparatus for generating a user tag, a storage medium, and a computer device, which relate to the field of neural networks. The method comprises: acquiring discrete user data respectively corresponding to a target user identification in more than one feature domain (S202); according to the discrete user data in the respective feature domains, obtaining respective corresponding intra-domain features of the respective feature domains corresponding to the target user identification (S204); fusing the respective intra-domain features to obtain inter-domain features corresponding to the target user identification (S206); performing feature cross

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))。

processing on sub-features in the inter-domain features to obtain cross features (S208); and according to the inter-domain features and the cross features, screening out from within candidate user tags a target user tag corresponding to the target user identification (S210). The accuracy of the generated user tags may be improved.

(57) 摘要: 一种用户标签生成方法、装置、存储介质和计算机设备, 涉及神经网络领域, 所述方法包括: 获取多于一个特征域下分别与目标用户标识对应的离散用户数据(S202); 根据各个所述特征域下的离散用户数据, 得到各个所述特征域各自相应的、且与所述目标用户标识对应的域内特征(S204); 融合各个所述域内特征得到与所述目标用户标识对应的域间特征(S206); 对所述域间特征中的子特征进行特征交叉处理, 获得交叉特征(S208); 根据所述域间特征和所述交叉特征, 从候选用户标签中筛选出与所述目标用户标识对应的目标用户标签(S210)。可以提高生成的用户标签的准确性。

用户标签生成方法、装置、存储介质和计算机设备

本申请要求于2019年04月10日提交的申请号为201910284354.6、发明名称为“用户标签生成方法、装置、存储介质和计算机设备”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。

技术领域

本申请涉及神经网络领域，特别涉及一种用户标签生成方法、装置、存储介质和计算机设备。

背景技术

随着计算机技术的发展，人们越来越多地依赖互联网来获取各方面的信息，为了达到向用户及时推荐各种有用信息又尽量避免推荐无用信息的目的，通常根据用户的用户标签确定接受信息的目标人群。

然而，生成用户标签的方式主要是基于统计的用户画像构建，这种方式很难对稀疏行为进行准确的评判，难以找到用户真正感兴趣的信息，从而导致得到的用户标签的准确率较低。

发明内容

本申请实施例提供了一种用户标签生成方法、装置、存储介质和计算机设备，能够有效地提高生成的用户标签的准确性。所述技术方案如下：

一种用户标签生成方法，所述方法应用于终端中，包括：

获取多于一个特征域下分别与目标用户标识对应的离散用户数据；

根据各个特征域下的离散用户数据，得到各个特征域各自相应的、且与目标用户标识对应的域内特征；

融合各个域内特征得到与目标用户标识对应的域间特征；

对域间特征中的子特征进行特征交叉处理，获得交叉特征；

根据域间特征和交叉特征，从候选用户标签中筛选出与目标用户标识对应的目标用户标签。

一种用户标签生成装置，所述装置设置在终端中，包括：

获取模块，用于获取多于一个特征域下分别与目标用户标识对应的离散用户数据；

融合处理模块，用于根据各个特征域下的离散用户数据，得到各个特征域各自相应的、且与目标用户标识对应的域内特征；融合各个域内特征得到与目标用户标识对应的域间特征；

交叉处理模块，用于对域间特征中的子特征进行特征交叉处理，获得交叉特征；

筛选模块，用于根据域间特征和交叉特征，从候选的用户标签中筛选出与目标用户标识对应的用户标签。

一种计算机可读存储介质，存储有计算机程序，计算机程序被处理器执行时，使得处理器执行上述用户标签生成方法的步骤。

一种计算机设备，包括存储器和处理器，存储器存储有计算机程序，计算机程序被处理器执行时，使得处理器执行上述用户标签生成方法的步骤。

本申请实施例提供的技术方案带来的有益效果至少包括：

在获取到多于一个特征域下分别与目标用户标识对应的离散用户数据后，一方面对这些离散的用户数据在特征域内以及特征域间进行融合，这样可以很好地对稀疏行为进行融合，得到融合了各种稀疏行为的域间特征；另一方面还对该域间特征中的子特征进行交叉处理，这样可以很好地对各个子特征进行融合，得到融合了域间特征中子特征的交叉特征。由于域间特征融合了各种稀疏行为，交叉特征融合了域间特征中的子特征，根据域间特征和交叉特

征从候选用户标签中筛选与目标用户标识对应的目标用户标签，能够有效地提高生成的用户标签的准确性。

附图说明

图 1 为一个实施例中用户标签生成方法的应用环境图；

图 2 为一个实施例中用户标签生成方法的流程示意图；

图 3 为一个实施例中机器学习模型的结构示意图；

图 4 为一个实施例中机器学习模型的域内融合层基于注意力机制对一个特征域下的离散特征向量进行融合的原理示意图；

图 5 为一个实施例中机器学习模型的特征交叉层基于因子分解机原理对域间特征向量中的子特征进行特征交叉处理的原理示意图；

图 6 为一个实施例中基于用户标签进行待推荐内容推荐的示意图；

图 7 为一个实施例中用户标签生成装置的结构框图；

图 8 为另一个实施例中用户标签生成装置的结构框图；

图 9 为另一个实施例中用户标签生成装置的结构框图；

图 10 为一个实施例中计算机设备的结构框图。

具体实施方式

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，此处所描述的实施例仅用以解释本申请，并不用于限定本申请。

图 1 为一个实施例中用户标签生成方法的应用环境图。参照图 1，该用户标签生成方法应用于用户标签生成系统。该用户标签生成系统包括终端 110 和服务器 120。终端 110 和服务器 120 通过网络连接。终端 110 可以是台式终端或移动终端，移动终端可以手机、平板电脑、笔记本电脑等中的至少一种。服务器 120 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。终端 110 和服务器 120 均可单独用于执行本申请实施例中提供的用户标签生成方法。终端 110 和服务器 120 也可协同用于执行本申请实施例中提供的用户标签生成方法。

需要说明的是，本申请实施例是用于对应用户生成用户标签的技术方案。用户标签可以是指可用于有效反映用户与其它用户之间特性和共性的标记。比如恐怖、海鲜、喜剧或者数码等。用户标签可以是对用户感兴趣内容的标记，也可以是对用户不感兴趣内容的标记。

在一个实施例中，计算机设备可通过机器学习模型来生成目标用户的目标用户标签。该机器学习模型先基于目标用户的离散特征数据生成目标用户的域间特征和交叉特征，再基于域间特征和交叉特征从用户标签池所包括的候选用户标签中筛选出目标用户的目标用户标签。其中，计算机设备也可对域间特征和交叉特征进行融合，得到一个最终反映用户特性的目标特征，再基于目标特征从用户标签池中筛选出目标用户的目标用户标签。

需要说明的是，首先，目标用户的离散特征数据来源于多于一个的特征域。特征域是指不同属性的用户数据所属的领域，比如年龄领域、性别领域、视频领域或者图文领域等。这样，基于多于一个特征域的离散特征数据来生成特征，便可以尽可能扩宽生成特征的数据来源，得到真正反映用户特性的特征。其次，机器学习模型基于目标用户的离散特征数据生成目标用户的特征时，先根据各个特征域下的离散用户数据得到各个特征域下的域内特征，然后融合各个域内特征得到域间特征，再对域间特征中的子特征进行特征交叉处理获得交叉特征，便可以得到融合了各种稀疏行为的域间特征（稀疏行为是指用户在某个特征域中产生数量极少的行为，如用户浏览了 1000 条娱乐类新闻，只浏览了 10 条金融类新闻，用户浏览 10 条金融类新闻的行为是稀疏行为），以及融合了域间特征中的子特征的交叉特征，从而可以根据域间特征和交叉特征来从候选用户标签中筛选出与目标用户标识对应的目标用户标签，有

效地提高了用户标签生成的准确性。

另外，每个用户都有对应的一个或者多于一个的用户标签，这个（些）用户标签可以认为是对用户真实兴趣特征的勾勒，是真实用户的综合模型。那么也可以称用户对应的这一个或多于一个的用户标签为该用户的用户画像。基于用户画像，计算机设备可以在各种领域对用户进行信息推荐。比如，新闻应用程序中的新闻推荐、视频应用程序中的视频推荐、社交应用程序中的文章推荐以及各种应用程序中的推广信息推荐等。例如，在新闻推荐的场景下，基于本申请实施例的技术方案对目标用户生成的用户标签为娱乐和金融，那么则可以向该目标用户推荐娱乐类新闻和金融类新闻。这样，计算机设备便可以在用户数据稀疏度高的情况下仍然进行准确的信息推荐。

如图 2 所示，在一个实施例中，提供了一种用户标签生成方法。本实施例以该方法应用于计算机设备来举例说明。该计算机设备可以是上述图 1 中的终端 110 或服务器 120。参照图 2，该用户标签生成方法包括如下步骤：

S202，获取多于一个特征域下分别与目标用户标识对应的离散用户数据。

其中，离散用户数据是指用户的稀疏行为下所产生的离散的用户数据。可以理解，通常情况下每个用户一般只会对自己感兴趣的零散内容产生行为。那么，基于这些零散内容产生的行为相应地也会产生一些零散的用户数据，计算机设备得到的也就是各个用户的离散用户数据。

特征域（field）是指不同属性的用户数据所属的领域。比如年龄领域、性别领域、视频领域或者图文领域等。本申请实施例中，在一个特征域下，与目标用户标识对应的离散用户数据的数量可以是一个也可以多于一个。比如，在性别领域下，与目标用户标识对应的离散用户数据的数量为一个，即要么为男、要么为女。而在视频领域下，与目标用户标识对应的离散用户数据的数量则可以多于一个，如，A 观看了《XXX》电视剧和 A 观看了《XX》电影等。

在一个实施例中，特征域的划分可根据实际需要自定义划分。在一种特征域划分方式下划分得到的多于一个的特征域，可以是另一种特征域划分方式下划分得到的一个特征域。比如，在 A 方式下划分得到年龄领域、性别领域和地域领域，在 B 方式下划分得到基础信息领域。那么，可以认为年龄领域、性别领域和地域领域可以是基础信息领域的子领域。在一种特征域划分方式下划分得到的一个特征域，也可以是另一种特征域划分方式下划分得到的多于一个的特征域。比如，在 A 方式下划分得到视频领域，在 B 方式下划分得到视频标签领域和视频类目领域。那么，可以认为视频标签领域和视频类目领域可以是视频领域的子领域。其中，视频标签领域中视频标签的数量大于视频类目领域中视频类目的数量。也就是说视频标签的划分比视频类目的划分更细。

举例说明，视频类目领域中的视频类目可以划分出体育、美食、娱乐和汽车等。而实际应用过程中，也可以对视频类目进行细分得到视频标签，如体育可以细分为球类比赛、田径比赛、赛车比赛和游泳比赛，而球类比赛又可以进一步分为篮球比赛、足球比赛和乒乓球比赛等。

在一些实施例中，多于一个的特征域可以是年龄领域、性别领域、地域领域、图文标签领域、视频标签领域、公众标识领域、图文类目领域和视频类目领域，这八个领域。

在一些实施例中，计算机设备在确定目标用户标识，也就是能够唯一标识待生成用户标签的用户的用户数据后，可以在网络上爬取多于一个特征域下分别与目标用户标识对应的离散用户数据。

举例说明，假设终端上运行有视频应用程序，该视频应用程序所对应的服务器在获取到终端当前登录视频应用程序所通过的用户标识后，即将该用户标识作为目标用户标识，继而在网络上爬取多于一个特征域下分别与目标用户标识对应的离散用户数据。此种应用场景，可以理解为提供推荐信息的对象即具备生成用户标签的能力。

在另外的实施例中，计算机设备也可以获取其他计算机设备传输的用户标识，以及多于一个特征域下分别与该用户标识对应的离散用户数据，从而将该用户标识作为目标用户标识，并获得了多于一个特征域下分别与目标用户标识对应的离散用户数据。

举例说明，假设终端上运行有视频应用程序，该终端在获取到当前登录视频应用程序所通过的用户标识后，可获取多于一个特征域下分别与该用户标识对应的离散用户数据，再将该用户标识和获取的离散用户数据对应发送至具备用户标签生成能力的计算机设备，该计算机设备即获取到目标用户标识，以及多于一个特征域下分别与目标用户标识对应的离散用户数据。此种应用场景，可以理解为：具备用户标签生成能力的计算机设备为提供推荐信息的对象提供生成用户标签的服务。

S204，根据各个特征域下的离散用户数据，得到各个特征域各自相应的、且与目标用户标识对应的域内特征。

其中，一个特征域对应的域内特征，是指融合了该特征域内的离散用户数据后得到的反映用户在该特征域内特性的数据。

可以理解，通常情况下，一个特征域下的用户数据的数量多于一个，也就是离散用户数据。多于一个的用户数据的数据量大且离散，计算机设备基于这样的数据生成用户标签时需要进行复杂运算，而且不一定能够得到较为准确的用户标签。在本申请实施例中，对于每一个特征域，基于该特征域下的所有离散用户数据，生成一个域内特征，通过该域内特征来表征目标用户在该特征域内特性，既可以抽象出用户特性，又可以去除数据的离散性，便于后续的用户标签生成过程。

在一个实施例中，计算机设备将每一个特征域下的每一个离散用户数据分别向量化，得到各个离散用户数据各自相应的离散特征向量，再将每个特征域下的各个离散特征数据相应的离散特征向量融合，得到该特征域相应的域内特征向量。这里的融合可以是组合、拼接或者按权重加和等。

其中，向量化是指将其他形式的数据以数学形式进行表达。比如，文本形式的“XXX”表示为数学形式“[0 0 0 1 0 0 0 0 0 0 0...]”，此时，“[0 0 0 1 0 0 0 0 0 0 0...]”即为对“XXX”进行向量化的结果，即为“XXX”的向量。可以理解，这里不限定将其他形式的数据转换为何种表示的向量，只要能够将其他形式的数据进行数学化表示即可。

举例说明，假设目标用户1在视频领域下的离散用户数据包括“点击查看了A电影”、“点击查看了B电视剧”、“点击查看了C动画片”和“点击查看了D短片”，向量化各个离散用户数据得到各自相应的离散特征向量X1、X2、X3、X4。可将X1、X2、X3和X4拼接得到视频领域下与目标用户1相应的域内特征向量 $X1+X2+X3+X4$ ，也可将X1、X2、X3和X4求平均得到视频领域下与目标用户1相应的域内特征向量 $(X1+X2+X3+X4)/4$ 等。

在一个实施例中，计算机设备可将每个特征域下的离散用户数据一起向量化，得到各个特征域各自相应的、且与目标用户标识对应的域内特征。

举例说明，目标用户1在视频领域下的离散用户数据包括“点击查看了A电影”、“点击查看了B电视剧”、“点击查看了C动画片”和“点击查看了D短片”。计算机设备可将这些数据一起向量化，得到视频领域下与目标用户1相应的域内特征向量Q。

需要说明的是，离散特征数据通常为文本形式的数据。域内特征可以是文本形式的数据，也可以是向量形式的数据，还可以是其他形式的数据。域内特征是向量形式的数据时，即为将每个特征域下的各个离散特征数据相应的离散特征向量融合后得到的结果。

S206，融合各个域内特征得到与目标用户标识对应的域间特征。

其中，域间特征是指融合了多于一个特征域相应的域内特征后得到的、既能反映用户各个特征域内特性、又能反映用户在这些特征域之间的特性的数据。融合是指通过一个数据表示多于一个数据，并包含这多于一个数据表达的信息。融合可以是组合、拼接或者按权重加和等。

可以理解，通常情况下，一个用户存在多于一个特征域的用户数据。即使每个特征域对

应一个域内特征，这些域内特征的数据量也较大且离散，计算机设备基于这样的数据生成用户标签时需要进行复杂运算，而且不一定能够得到较为准确的用户标签。在本申请实施例中，将多于一个特征域的域内特征，融合成生成一个域间特征，既可以抽象出用户特性，又可以去除数据的离散性，便于后续的用户标签生成过程。

在一个实施例中，计算机设备将每一个特征域下的每一个离散用户数据分别向量化，得到各个离散用户数据各自相应的离散特征向量，再将每个特征域下的各个离散特征数据相应的离散特征向量融合，得到该特征域相应的域内特征向量后，可以继续将这些域内特征向量进行融合，得到域间特征向量。这里的融合可以是组合、拼接或者按权重加和等。

举例说明，假设目标用户 1 在四个特征域各自对应的域内特征向量为 Y_1 、 Y_2 、 Y_3 、 Y_4 ，可将 Y_1 、 Y_2 、 Y_3 和 Y_4 拼接得到与目标用户 1 相应的域间特征向量 $Y_1+Y_2+Y_3+Y_4$ ，也可将 Y_1 、 Y_2 、 Y_3 和 Y_4 求平均得到与目标用户 1 相应的域间特征向量 $(Y_1+Y_2+Y_3+Y_4)/4$ 等。

需要说明的是，域间特征可以是文本形式的数据，也可以是向量形式的数据，还可以是其他形式的数据。域间特征是向量形式的数据时，即为将每个特征域相应的域内特征向量融合后得到的结果。

S208，对域间特征中的子特征进行特征交叉处理，获得交叉特征。

其中，域间特征可以包括多于一个的子特征。交叉特征是指对域内特征中的子特征进行融合后得到的，反映各个子特征间相关性的数据。

在一些实施例中，计算机设备在可以对域间特征中任意的两个或两个以上的子特征进行交叉处理，经过交叉处理后便可得到多于一个可以表示子特征之间相关性的交叉子特征，然后将这些交叉子特征融合，即可得到与目标用户标识对应的交叉特征。这里的融合可以是组合、拼接或者按权重加和等。

其中，当域间特征为向量形式的数据，即域间特征向量时，域间特征向量的每个向量元素即为域间特征向量的一个子特征。

举例说明，假设域间特征的特征子特征用向量元素 $(t$ 为大于或等于 0 的整数) 表示。域间特征中包括有是否喜欢球类比赛、是否喜欢美食、是否喜欢娱乐和是否喜欢汽车等方面的信息。对于用户是否喜欢球类比赛，还可以细分为是否喜欢篮球比赛、是否喜欢足球比赛和是否喜欢乒乓球比赛等赛事的信息。假设用户是否喜欢篮球比赛用向量元素 X_i 表示，用户是否喜欢足球比赛用向量 X_m 表示，则交叉处理可得到交叉子特征 X_iX_m 。因此，用户对不同球类比赛的爱好可以通过这个交叉子特征建立关联，以实现更加精准的用户特性表达。

S210，根据域间特征和交叉特征，从候选用户标签中筛选出与目标用户标识对应的目标用户标签。

在一些实施例中，计算机设备将域间特征和交叉特征分别与各个候选用户标签进行比对，计算域间特征和各个候选用户标签的差异度或者相似度，以及交叉特征和各个候选用户标签的差异度或者相似度，然后从候选用户标签中根据差异度或者相似度筛选出与域间特征和交叉特征共同匹配的候选用户标签。其中，匹配条件可以是差异度低于或者不超过预设差异度阈值，或者，相似度高于或者达到预设相似度阈值等。

在一个实施例中，计算机设备也可先融合域间特征和交叉特征得到目标特征，再根据目标特征从候选用户标签中筛选出与目标用户标识对应的目标用户标签。其中，目标特征是最终得到的，能够更精准地对用户特性进行表达的数据。可以理解，目标用户的离散特征数据来源于多于一个的特征域。这样，基于多于一个特征域的离散特征数据来生成目标特征，便可以尽可能扩宽生成目标特征的数据来源。而且，先根据各个特征域下的离散用户数据得到各个特征域下的域内特征，然后融合各个域内特征得到域间特征，再对域间特征中的子特征进行特征交叉处理获得交叉特征，便可以得到融合了各种稀疏行为的域间特征，和融合了域间特征中的子特征的交叉特征，从而融合了域间特征和交叉特征的目标特征，便可以更加精准地对用户特性进行表达。

在一些实施例中，计算机设备在得到域间特征和交叉特征后，将这两个特征进行融合，

得到目标特征。此后，计算机设备则可继续将该目标特征和各候选用户标签进行比对，计算目标特征和各个候选用户标签的差异度或者相似度，从而从候选用户标签中根据相似度或者差异度筛选出与目标特征匹配的候选用户标签，作为与目标用户标识对应的目标用户标签。

比如，计算机设备可将域间特征和交叉特征分别向量化，得到域间特征向量和交叉特征向量，再将域间特征向量和交叉特征向量融合，得到目标特征向量。这里的融合可以是组合、拼接或者按权重加和等。

上述用户标签生成方法，在获取到多于一个特征域下分别与目标用户标识对应的离散用户数据后，一方面对这些离散的用户数据在特征域内以及特征域间进行融合，这样可以很好地对稀疏行为进行融合，得到融合了各种稀疏行为的域间特征；另一方面还对该域间特征中的子特征进行交叉处理，这样可以很好地对各个子特征进行融合，得到融合了域间特征中子特征的交叉特征。由于域间特征融合了各种稀疏行为，交叉特征融合了域间特征中的子特征，那么根据域间特征和交叉特征从候选用户标签中筛选与目标用户标识对应的目标用户标签，能够有效地提高生成的用户标签的准确性。

在一个实施例中，S204包括：将各个离散用户数据输入机器学习模型，机器学习模型包括输入层和域内融合层；通过输入层向量化各个离散用户数据，得到各个离散用户数据各自相应的离散特征向量；通过域内融合层融合各个特征域下的离散特征向量，得到各个特征域相应的、且与目标用户标识对应的域内特征向量。

其中，机器学习模型是通过样本学习具备某种能力的数学模型。机器学习英文全称为Machine Learning，简称ML。机器学习模型可采用神经网络模型、支持向量机、逻辑回归模型、随机森林模型或者梯度提升树模型等。机器学习的学习方式可以是监督学习、非监督学习或者强化学习等。

在本申请实施例中的机器学习模型是通过样本学习具备用户标签生成能力的模型。该机器学习模型可以是端到端的模型，输入端为目标用户原始的离散用户数据，输出端为预测的候选用户标签为目标用户的目标用户标签的评分。机器学习模型为多层网络结构，不同的网络层对输入其的数据进行不同的处理，并输出处理结果至下一层网络层。

在一些实施例中，计算机设备在获取到多于一个特征域下分别与目标用户标识对应的离散用户数据后，可将文本形式的这些离散用户数据直接输入机器学习模型。机器学习模型的输入层则将每个离散用户数据分别映射为相应的离散特征向量。比如，机器学习模型的输入层将离散用户数据“女”映射为“[0 2 0 1 0 3 0 0 0 0...]”，将离散用户数据“21岁”映射为“[1 2 0 1 0 4 0 0 0 0...]”，将离散用户数据“北京朝阳区”映射为“[1 0 2 0 0 0 0 4 0...]”等。

在一个实施例中，机器学习模型的输入层可将每个离散用户数据分别映射为相应的、固定长度的离散特征向量。该固定长度可以认为是离散特征向量的维度。这样将离散特征向量均统一到同一个维度便于后续的融合处理。可以理解，由于不同的特征域下离散用户数据的数量各不相同，部分特征域下的离散用户数据的数量级为百万级别，而部分特征域下的离散用户数据的数量仅为十级别；需要将离散特征向量的维度设置的尽可能大，以保留足够多的信息。比如，离散特征向量的维度是128或者256等。

举例说明，图3示出了一个实施例中机器学习模型的结构示意图。参考图3，计算机设备可将多于一个特征域下分别与目标用户标识对应的离散用户数据直接输入机器学习模型，机器学习模型的输入层（Feature-input layer）将离散用户数据映射为相应的离散特征向量。例如，将任一个特征域下的离散用户数据 $x_1, x_2, \dots, x_n$ ，分别映射为相应的离散特性向量 $\vec{t_1}, \vec{t_2}, \dots, \vec{t_n}$ 。

在一个实施例中，通过域内融合层融合各个特征域下的离散特征向量，得到各个特征域相应的、且与目标用户标识对应的域内特征向量，包括：通过域内融合层分别获取各个特征域下的各个离散特征向量各自对应的注意力分配权重；通过域内融合层，将各个特征域下的

各个离散特征向量按照各自对应的注意力分配权重进行线性融合，得到各个特征域各自相应的、且与目标用户标识对应的域内特征向量。

其中，域内融合层是本申请实施例的机器学习模型中，按特征域对离散用户数据相应的离散特征向量在域内进行融合的网络层。通过域内融合层，可将输入的、且属于相同特征域的离散特征向量融合为一个域内特征向量，从而得到多于一个的域内特征向量。域内特征向量的数量与特征域的数量相同，也就是每个特征域对应一个域内特征向量。比如，计算机设备获取了N个特征域下的离散用户数据，则在经过域内融合层后，可以得到N个域内特征向量。

可以理解，一个特征域内可以有数量较多的离散用户数据，这些离散用户数据对用户特性体现的重要程度各不相同。比如，某用户在视频领域中有上万条体育视频的浏览记录，但仅有几条娱乐视频的浏览记录。该用户在体育方面的兴趣明显高于娱乐方面的兴趣，在用户特性体现上体育视频比娱乐视频要重要。由此可见，在对一个特征域内的各个离散用户数据相应的离散特征向量进行融合时，需要将对用户特性体现重要的离散特征向量体现出来，以得到更能表征用户特性的域内特征向量。在本实施例中，则通过注意力机制（Attention）对不同的离散特征向量分配不同的注意力分配权重，通过注意力分配权重来体现离散特征向量的重要程度，从而凸显出对用户特性体现重要的离散特征向量。

在一个实施例中，离散特征向量对应的注意力分配权重的计算公式如下式所示：

$$\alpha_i = \frac{e^{\vec{u}_t \cdot \text{relu}(\vec{t}_i \cdot \vec{W}_t + \vec{b}_t)}}{\sum_j^H e^{\vec{u}_t \cdot \text{relu}(\vec{t}_j \cdot \vec{W}_t + \vec{b}_t)}} \quad (1)$$

其中， $\vec{t}_i$ 和 $\vec{t}_j$ 均为离散特性数据对应的离散特征向量。 α_i 为 $\vec{t}_i$ 的注意力分配权重。空间变化矩阵 $\vec{W}_t$ 、偏置 $\vec{b}_t$ 和域内注意力向量 $\vec{u}_t$ 为域内融合层的模型参数，经过模型训练学习得到。非线性函数relu为域内融合层中的激活函数。 $\vec{t}_i$ 经过空间变化矩阵 $\vec{W}_t$ 、偏置 $\vec{b}_t$ 和非线性函数relu可变换到注意力空间；然后与域内注意力向量 $\vec{u}_t$ 相乘，即能得到 $\vec{t}_i$ 在注意力空间、且在特征域内的注意力分配信息；再经过softmax进行权重计算，即可得到 $\vec{t}_i$ 在注意力空间、且在特征域内的注意力分配权重。H为特征域内离散特征数据的数量。

在一个实施例中，计算机设备可在特征域内离散用户数据的数据量较大时，对特征域内的离散用户数据进行筛选，过滤掉重要程度较低的离散用户数据。重要程度较低的离散用户数据可以是出现次数较少的离散用户数据。比如，某用户在视频领域中有上万条体育视频的浏览记录，但仅有几条娱乐视频的浏览记录，此时则可将仅有的几条娱乐视频浏览记录过滤掉。这样，计算机设备可以将各个特征域内的离散用户数据控制在预设数量之内。

那么，特征域k对应的域内特征向量即可通过下式得到：

$$\vec{f}_k = \sum_i^H \alpha_i * \vec{t}_i \quad (2)$$

继续参考图3，输入层输出的离散特征向量被输入域内融合层（Attention-pooling layer），域内融合层基于注意力机制分别对每个特征域下的离散特征向量进行融合，得到各个特征域各自相应的域内特征向量（ $\vec{f}_1$ 、 $\vec{f}_2$ 、...、 $\vec{f}_M$ ）输出至下一层。其中，M为特征域的数量。域内融合层基于注意力机制对一个特征域下的离散特征向量进行融合的原理如图4所示。

参考图4，机器学习模型的域内融合层可通过自身的模型参数，基于注意力机制为各个离散特征向量计算出一个注意力分配权重。比如，某一特征域k下的离散特征向量 $\vec{t}_1$ 的注意

力分配权重为 α_1 、离散特征向量 $\vec{t}_2$ 的注意力分配权重为 $\alpha_2 \cdots$ 离散特征向量 $\vec{t}_n$ 的注意力分配权重为 α_n 。域内融合层再将该特征域内的离散特征向量按照相应的注意力分配权重进行加权求平均，得到该特征域相应的域内特征向量 $\vec{f}_k$ 。

在一个实施例中，计算机设备将各个离散用户数据相应的离散特征向量统一到相同的维度，得到的各个特征域的域内特征向量的维度也相同、且与离散特征向量的维度相同。

上述实施例中，在对特征域内的离散特征向量进行融合时，采用注意力机制融合的方式，为不同的离散特征向量分配不同的融合权重，这样可以从大量的离散特征向量中有选择地选取更重要的信息作为辅助，更充分地对用户特性进行表示，大大提高了得到的域内特征向量的准确性和有效性。

这样，计算机设备通过机器学习模型来进行数据处理，由于本申请实施例中的机器学习模型是通过样本有监督地训练得到的模型。在选择设置好模型结构、训练样本和相应的训练标签后，即可通过样本自学习到模型参数，从而利用机器学习模型强大的特征表达能力和学习能力，学习到更好地对用户特征进行向量表达的能力，有利于后续的处理以及用户标签的生成。

在一个实施例中，机器学习模型还包括域间融合层。S206 包括：通过域间融合层分别获取各个域内特征向量各自对应的注意力分配权重；通过域间融合层，将各个域内特征向量按照各自对应的注意力分配权重进行线性融合，得到与目标用户标识对应的域间特征向量。

其中，域间融合层是本申请实施例的机器学习模型中，对各个特征域的域内特征向量在域间进行融合的网络层。通过域内融合层，可将输入其的、各个特征域的域内特征向量融合为一个域间特征向量。比如，域内融合层输出了 N 个域内特征向量至域间融合层，域间融合层则将这 N 个域内特征向量融合为一个域间特征向量。

可以理解，一个用户可以在数量较多的特征域中对应离散用户数据，这些不同特征域的离散用户数据对用户特性体现的重要程度各不相同。比如，某用户在图文领域中有上万条浏览记录，但在视频领域仅有几条浏览记录。那么，该用户在图文领域的兴趣明显高于视频领域的兴趣，在用户特性体现上图文领域比视频领域重要。由此可见，在对多于一个特征域各自相应的域内特征向量进行融合时，需要将用户对用户特性体现重要的特征域体现出来，以得到更能表征用户特性的域间特征向量。在本实施例中，则通过注意力机制对不同的域内特征向量分配不同的注意力分配权重，通过注意力分配权重来体现特征域的重要程度，从而凸显出对用户特性体现重要的特征域。

在一个实施例中，域内特征向量对应的注意力分配权重的计算公式如下式所示：

$$\beta_k = \frac{e^{\vec{u}_f \cdot \text{relu}(\vec{f}_k \cdot W_f + \vec{b}_f)}}{\sum_j^M e^{\vec{u}_f \cdot \text{relu}(\vec{f}_j \cdot W_f + \vec{b}_f)}} \quad (3)$$

其中， $\vec{f}_k$ 和 $\vec{f}_j$ 均为域内特征向量。 β_k 为 $\vec{f}_k$ 的注意力分配权重。空间变化矩阵 W_f 、偏置 $\vec{b}_f$ 和域间注意力向量 $\vec{u}_f$ 为域间融合层的模型参数，经过模型训练学习得到。非线性函数 relu 为域间融合层中的激活函数。 $\vec{f}_k$ 经过空间变化矩阵 W_f 、偏置 $\vec{b}_f$ 和非线性函数 relu 可变换到注意力空间；然后与域间注意力向量 $\vec{u}_f$ 相乘，即能得到 $\vec{f}_k$ 在注意力空间、且在特征域间的注意力分配信息；再经过 softmax 进行权重计算，即可得到 $\vec{f}_k$ 在注意力空间、且在特征域间的注意力分配权重。M 为域内特征向量的数量，也就是特征域的数量。

那么，域间特征向量即可通过下式得到：

$$\vec{X} = \sum_k^M \beta_i * \vec{f}_k \quad (4)$$

继续参考图 3，输入层输出的域内特征向量被输入域间融合层（Attention-merge layer），域间融合层基于注意力机制对每个特征域相应的域内特征向量进行融合，得到域间特征向量 $\vec{X}$ 输出至下一层。

在一个实施例中，计算机设备将各个离散用户数据相应的离散特征向量统一到相同的维度，得到的各个特征域的域内特征向量的维度也相同、且与离散特征向量的维度相同。那么，对域内特征向量融合得到的域间特征向量的维度也与离散特征向量的维度相同。

上述实施例中，在对各个特征域相应的域内特征向量进行融合时，采用注意力机制融合的方式，为不同的域内特征向量分配不同的融合权重，这样可以从大量的域内特征向量中有选择地选取更重要的信息作为辅助，更充分地对用户特性进行表示，大大提高了得到的域间特征向量的准确性和有效性。

在一个实施例中，机器学习模型还包括特征交叉层。S208 包括：通过特征交叉层，对域间特征向量进行去稀疏处理得到稠密化的域间特征向量；对稠密化的域间特征向量中的子特征进行二阶交叉处理，获得交叉特征。

可以理解，经过前述实施例，计算机设备已通过机器学习模型将获取的大量的目标用户的离散用户数据映射为一个域间特征向量。该域间特征向量经过基于注意力机制的域内融合和域间融合，在一定程度上能够比较好地对用户特性进行表达。但是，对于特征域间的用户数据的相关性缺少表达，那么可以对域间特征进行交叉处理，得到能够表达用户数据之间的相关性的交叉特征。

在一些实施例中，计算机设备在利用机器学习模型将离散用户数据映射为离散特征向量时，离散特征向量为了更多地保留信息，通常将离散特征向量的维度设置的较大，而基于离散特征向量经过域内融合和域间融合后得到的域间特征向量的维度也较大，可以看出域间特征向量的特征表达是较稀疏的。计算机设备为了通过机器学习模型更好地对域间特征向量的子特征进行交叉处理，将因子分解机的处理原理巧妙地融入其中，基于因子分解机的处理原理对域间特征向量进行去稀疏处理得到稠密化的域间特征向量，以基于稠密化的域间特征向量进行特征交叉处理。

在一个实施例中，通过特征交叉层，对域间特征向量进行去稀疏处理得到稠密化的域间特征向量，包括：通过特征交叉层，将域间特征向量中的子特征分别映射为预设维度的隐层空间向量。对稠密化的域间特征向量中的子特征进行二阶交叉处理，获得交叉特征，包括：对于域间特征向量中任意的两个子特征，将两个子特征以及两个子特征所映射得到的隐层空间向量的乘积，作为两个子特征的二阶交叉特征向量；组合各个二阶交叉特征向量，得到交叉特征向量。

可以理解，域间特征向量中的子特征在域间特性空间是稀疏的，不便于运算；但可以将这些子特征映射到另一空间，在这一空间里，这些子特征所映射得到的结果可以是稠密的，这一空间可以称为隐层向量空间，这些子特征所映射得到的结果可以是隐层空间向量。

在一些实施例中，计算机设备通过特征交叉层计算两个子特征之间的相关性时，可以使用一个隐层空间向量作为子特征的代表，再通过隐层特征向量之间的运算来得到子特征之间的交叉子特征。可以理解，这里是将域间特征向量中稀疏的子特征映射到隐层向量空间中，利用稠密的隐层空间向量进行去稀疏处理。

举例说明，假设本申请中机器学习模型的域间融合层所输出的域间特征向量为 $\vec{X} = \{X_1, X_2, \dots, X_i, \dots, X_j, \dots\}$ ，对于 $\vec{X}$ 中的任意两个子特征 X_i 和 X_j 之间的相关性，特征交叉层可以将子特征 X_i 映射为隐层空间向量 $\vec{V}_i$ 作为表达，将子特征 X_j 映射为隐层空间向

量 $\vec{V}_j$ 作为表达,再通过计算 $(\vec{V}_i \cdot \vec{V}_j) * X_i * X_j$,得到子特征 X_i 和 X_j 之间的二阶交叉特征向量 $\vec{C}_{i,j}$ 。

基于此,特征交叉层可以组合各个二阶交叉特征向量,得到交叉特征向量如下式所示:

$$\begin{aligned} \vec{C} &= \{C_{1,2}, C_{1,3}, \dots, C_{i,j}, \dots\} \\ &= \{(\vec{V}_1 \cdot \vec{V}_2) * X_1 * X_2, (\vec{V}_1 \cdot \vec{V}_3) * X_1 * X_3, \dots, (\vec{V}_i \cdot \vec{V}_j) * X_i * X_j, \dots\} \end{aligned} \quad (5)$$

其中,假设域间特征向量 $\vec{X}$ 的维度为N,则交叉特征向量的维度为 $N*(N-1)/2$ 。

在本实施例中,提供了基于本申请实施例对应的场景下得到的域间特征向量,结合因子分解机的原理计算交叉特征的方式,能够得到更为准确的交叉特征向量。

继续参考图3,域间融合层输出的域间特征向量被输入特征交叉层(Feature-cross layer),特征交叉层基于因子分解机原理对域间特征向量中的子特征进行特征交叉处理,得到交叉特征向量 $\vec{C}$ 输出至下一层。其中,特征交叉层基于因子分解机原理对域间特征向量中的子特征进行特征交叉处理的原理如图5所示。

参考图5,特征交叉层可通过自身的模型参数,将域间特征向量($X_1, X_2, \dots, X_n$)中的子特征各自映射为一个隐层空间向量: $\vec{V}_1, \vec{V}_2, \dots, \vec{V}_n$ 。那么,对于域间特征向量中的每一个子特征,都可以得到其本身与隐层空间向量的乘积: $X_1 * \vec{V}_1, X_2 * \vec{V}_2, \dots, X_n * \vec{V}_n$ 。特征交叉层则对任意两个乘积进行点乘运算,得到多个交叉子特征 $X_1 * X_2 * (\vec{V}_1 \cdot \vec{V}_2)$, $X_1 * X_n * (\vec{V}_1 \cdot \vec{V}_n)$, $\dots$, $X_2 * X_n * (\vec{V}_2 \cdot \vec{V}_n)$ 。这样,特性交叉层即可拼接这些交叉子特征得到交叉特征向量 $\vec{C}$ 。

在另外的实施例中,计算机设备还可以通过机器学习模型的特征交叉层对多于两个的子特征进行特征交叉处理,得到更高阶的交叉子特征。比如,对任意三个子特征或者四个子特征交叉进行处理。

可以理解的是,域内融合层和域间融合层进行的是一阶的融合操作,特征交叉层则进行的是二阶的融合操作。

上述实施例中,基于因子分解机的原理对域间特征向量中的子特征进行交叉处理,可以避免因人工难以手动设计交叉特征的问题,从而即便是在特征分布复杂的场景下,也可以很容易地对特征进行拟合,得到二阶甚至高于二阶的交叉特征,以便通过二阶以及更高阶的交叉特征来实现不同特征域间的用户特性的相关性的表达,提高最终表征用户特性的特征向量的准确性。

在一个实施例中,机器学习模型还包括第一全连接层和第二全连接层。S210包括:通过第一全连接层,将交叉特征向量的维度调整至与域间特征向量的维度一致后,与域间特征向量拼接得到中间特征向量;通过第二全连接层,将中间特征向量映射为目标特征向量;根据目标特征向量,从候选用户标签中筛选出与目标用户标识对应的目标用户标签。

通常情况下,全连接层的每一个结点都与上一层的所有结点相连,用来把前序层提取到的特征综合起来。在本申请实施例中,特征交叉层在对域间特征向量进行特征交叉处理得到交叉特征向量后,将其输出至第一全连接层(Fully connected layers, FC)。特征交叉层还可对域间特征向量进行线性变化后输出至第一全连接层。由于域间特征向量反映了用户在域内和域间的特征,交叉特征向量则反映了用户域间特征向量各个子特征之间的相关性,都是前序

层提取到的能够体现用户特性的特征，那么第一全连接层则可将这些特征综合起来协同表征用户特性。

根据前述实施例可知，对域间特征向量进行特征交叉处理得到的交叉特征向量与域间特征向量的维度不同。第一全连接层可以先将交叉特征向量的维度调整至与域间特征向量的维度一致后，再将交叉特征向量与域间特征向量拼接得到中间特征向量，将得到的中间特征向量输出至第二全连接层，第二全连接层则进一步对前序层提取到的特征所拼接的中间特征向量进行融合，最终得到能够更充分体现用户特性的目标特征向量。

继续参考图 3，特征交叉层输出的交叉特征向量被输入第一全连接层，且域间融合层输出的域间特征向量也被输入第一全连接层。第一全连接层则将交叉特征向量的维度调整至与域间特征向量的维度一致后，将交叉特征向量与域间特征向量拼接得到中间特征向量，再输出到第二全连接层。第二全连接层则将中间特征向量映射为目标特征向量 $\vec{z}$ 。

在一个实施例中，全连接层可以通过 relu 非线性激活函数对输入的数据进行处理得到处理结果。

在一个实施例中，机器学习模型还包括预测层。根据目标特征向量，从候选用户标签中筛选出与目标用户标识对应的目标用户标签，包括：通过预测层，输出目标特征向量与各个候选用户标签所对应的标签向量的相关评分；筛选出相关评分靠前的预设数量的候选用户标签，作为与目标用户标识对应的目标用户标签。

其中，预测层是本申请实施例的机器学习模型中，对目标特征向量与各个候选用户标签所对应标签向量的相关度进行评分的网络层。通过预测层，可对输入的目标特征向量和候选用户标签所对应的标签向量进行运算，得到相关评分。用户标签向量通过对文本形式的候选用户标签向量化得到。

在一些实施例中，用户标签池中可以包括大量的候选用户标签，每个候选用户标签均可对应存在一个用户标签向量。计算机设备在通过机器学习模型得到目标用户的目标特征向量后，可以将该目标特征向量和候选用户标签的用户标签向量输入预测层，通过该预测层输出目标特征向量与各个候选用户标签所对应的标签向量的相关评分。计算机设备可再按照相关评分对相应的候选用户标签进行排序。与目标特征向量的相关评分高的标签向量所对应的候选用户标签排序靠前，与目标特征向量的相关评分低的标签向量所对应的候选用户标签排序靠后。这样计算机设备即可从排序的首个候选用户标签起，选取预设数量的候选用户标签作为与目标用户标识对应的目标用户标签。

继续参考图 3，第二全连接层则将中间特征向量映射为目标特征向量后，传递至预测层 (Predicting layer)。预测层再接收候选用户标签的标签向量 (Tag-embedding: Tag1、Tag2、...、Tagk) 作为输入，得到目标特征向量与各个标签向量的相关评分 (y_1 、 y_2 、...、 y_k)。

在一个实施例中，计算机设备可通过预测层，得到目标特征向量和用户标签池中所有候选用户标签所对应标签向量的相关评分，也可从用户标签池中初步筛选出一部分候选用户标签，仅计算目标特征向量和这部分候选用户标签所对应标签向量的相关评分。

需要说明的是，从机器学习网络层所实现的功能的层面来说，预测层可以看作是一个多用户标签分类器。目标特征向量与各个候选用户标签所对应标签向量的相关评分，可以看作是目标特征向量被分类至各个候选用户标签的概率。

在本实施例中，通过预测层来对用户的标签分类进行评分，能够利用机器学习模型的强大的学习能力提高评分的准确率。

上述实施例中，在得到机器学习模型前序层所提取的特征：域间特征向量和交叉特征向量后，将这些特征综合起来得到能够充分反映用户特性的目标特征向量，再依据该目标特征向量能够更准确地筛选目标用户的用户标签。

在一个实施例中，该用户标签生成方法还包括：收集多于一个特征域下分别与样本用户标识对应的离散用户数据样本；确定离散用户数据样本分别对应的各个候选用户标签的训练

标签；将离散用户数据样本输入机器学习模型，经过机器学习模型所包括的各个层依次处理后，输出离散用户数据样本与各个候选用户标签所对应的标签向量的相关评分样本；根据相关评分样本和对应的训练标签的差异，训练机器学习模型。

可以理解，计算机设备训练机器学习模型是对用户基于用户特征进行用户标签的分类。也就是说，预测用户标签是否应该被标记给用户。计算机设备可事先设置候选用户标签，然后收集多于一个特征域下分别与样本用户标识对应的离散用户数据样本，并确定这些离散用户数据样本分别对应各个候选用户标签的训练标签，有监督地训练机器学习模型。

在一些实施例中，机器学习模型分类的依据则是用户特性与用户标签的相关度。这样，机器学习模型的一部分可被设置为用于对离散的用户数据进行特征提取和表征，得到一个高阶的、能充分反映用户特性的目标特征向量。机器学习模型的剩余部分可再被设置为用于对目标特征向量与各个分类（候选用户标签）所对应的向量（标签向量）进行相关评分。在设计机器学习模型的训练数据时，可将每个样本用户标识所对应的所有离散用户数据样本共同作为机器学习模型的训练输入，然后确定每个样本用户标识所对应的所有离散用户数据样本分别对应各个分类（候选用户标签）的训练标签。这样，机器学习模型的中间层在得到各个样本用户标识所对应的目标特征向量样本后，即可对目标特征向量样本进行用户标签分类，而此时目标特征向量样本分别对应各个分类（候选用户标签）均存在训练标签，就可以进行有监督的机器学习模型训练。

进一步地，计算机设备可根据训练输入对应各个候选用户标签的相关评分样本和对应的训练标签的差异来构建损失函数，再按照最小化该损失函数的方向调整机器学习模型的模型参数并继续训练，直至满足训练结束条件。

其中，计算机设备将离散用户数据样本输入机器学习模型，经过机器学习模型所包括的各个层依次处理后，输出离散用户数据样本与各个候选用户标签所对应标签向量的相关评分样本的处理过程，可参考前述实施例中使用机器学习模型时的数据处理过程。

在一个实施例中，计算机设备可通过使用 sigmoid 交叉熵定义损失函数，此时机器学习模型的预测层的损失函数如下式：

$$Loss = -\frac{1}{K} * \sum_{k=1}^K \left(\hat{y}_k \log y_k + (1 - \hat{y}_k) \log (1 - y_k) \right) \quad (6)$$

$$y_k = \frac{1}{1 + e^{-\vec{e} \cdot \vec{t}_k}} \quad (7)$$

其中，K 为候选用户标签的数量。 $\hat{y}_k$ 为样本用户标识所对应的所有离散用户数据样本对应第 k 个候选用户标签的训练标签。 y_k 为机器学习模型的预测层实际输出的、样本用户标识所对应的所有离散用户数据样本与第 k 个候选用户标签的相关评分。 $\vec{e}$ 为样本用户标识所对应的所有离散用户数据样本经过机器学习模型处理得到的目标特征向量样本， $\vec{t}_k$ 为第 k 个候选用户标签的标签向量。

需要说明的是，由于计算机设备事先设置了大量的候选用户标签，而对于某一个特定的用户，这些候选用户标签通常只有一部分是符合该特定的用户的特性，可以被标记给该特定的用户；另外一部分则不符合该特定的用户的特性，不应当被标记给该特定的用户。计算机设备所确定的离散用户数据样本分别对应各个候选用户标签的训练标签，可以包括正训练标签和负训练标签。

在一个实施例中，计算机设备可获取与样本用户标识对应的历史浏览记录，查询历史浏览记录中的点击内容相应的候选用户标签，确定离散用户数据样本分别对应的这些候选用户标签的训练标签为正训练标签。再在用户标签池中随机选取一部分候选用户标签，确定离散用户数据样本分别对应的这些候选用户标签的训练标签为负训练标签。这样即可综合正例和

反例来训练机器学习模型。

可以理解，用户在计算机设备上产生用户行为时通常会形成浏览记录。比如，用户在视频应用程序中观看视频后会形成历史浏览记录，在新闻应用程序中查看新闻后也会形成历史浏览记录。其中，用户行为所作用的对象可以存在对应的候选用户标签。比如，用户行为所作用的对象为 XX 电影，该电影是一部惊悚动作类型的电影，可以存在对应的候选用户标签“惊悚”和“动作”。基于这些内容即可得到用于训练机器学习模型的训练数据，而且采用训练完的机器学习模型对用户生成用户标签后，又可基于生成的用户标签向用户推荐用户行为所作用的对象，如视频、新闻、图文文章或者推广内容等。

举例说明，假设与样本用户标识对应的历史浏览记录中存在一条点击内容（视频或新闻等）：“XX…”，该点击内容所对应的候选用户标签为“娱乐”，那么与样本用户标识对应的离散用户数据样本所对应的“娱乐”这一候选用户标签的训练标签为正训练标签。计算机设备在用户标签池中随机选取一个候选用户标签“体育”，那么与样本用户标识对应的离散用户数据样本所对应的“体育”这一候选用户标签的训练标签设置为负训练标签。

在一个实施例中，确定离散用户数据样本分别对应各个候选用户标签的训练标签，包括：获取与样本用户标识对应的历史浏览记录；查询历史浏览记录中的点击内容相应的第一候选用户标签，以及历史浏览记录中的未点击内容相应的第二候选用户标签；将离散用户数据样本分别对应的各个第一候选用户标签的训练标签设置为正训练标签，并将离散用户数据样本分别对应的各个第二候选用户标签的训练标签设置为负训练标签。

可以理解，用户在计算机设备上产生行为时通常会形成浏览记录。比如，用户在视频应用程序中观看视频后会形成历史浏览记录，在新闻应用程序中查看新闻后也会形成历史浏览记录。历史浏览记录中可以包括曝光未点击内容记录和曝光且点击内容记录。其中，曝光未点击内容记录所记录的是展示给用户，但用户未进行进一步点击查看的内容。曝光且点击内容记录所记录的是展示给用户，且用户进行了进一步点击查看的内容。

通常情况下，当用户对曝光的内容感兴趣时才会进一步点击查看该内容，而对曝光的内容不感兴趣时则不会进一步点击查看该内容。基于此，可以认为曝光且点击内容是用户感兴趣的内容，那么曝光且点击内容对应的候选用户标签可以反映该用户的特性而将其标记给该用户。曝光未点击内容则可以被认为是用户不感兴趣的内容，那么曝光未点击内容对应的候选用户标签不能反映该用户的特性而不应当将其标记给该用户。

在一些实施例中，计算机设备在获取与样本用户标识对应的历史浏览记录后，可查询历史浏览记录中的曝光且点击内容和曝光未点击内容，并确定曝光且点击内容相应的第一候选用户标签，以及曝光未点击内容相应的第二候选用户标签。由于，第一候选用户标签是可以标记给该用户的用户标签，那么即将该用户的离散用户数据样本分别对应的各个第一候选用户标签的训练标签设置为正训练标签。而第二候选用户标签是不应当标记给该用户的用户标签，那么即将该用户的离散用户数据样本分别对应的各个第二候选用户标签的训练标签设置为负训练标签。

在本实施例中，选用样本用户的历史浏览记录中曝光且点击内容的用户标签为正例，曝光未点击内容的用户标签为反例，可以得到更加合理的正反例来训练模型。

在一个实施例中，该用户标签生成方法还包括：选取异于样本用户标识的参考用户标识，以及所述参考用户标识对应的离散用户数据样本；获取离散用户数据样本的正训练标签所对应的第三候选用户标签；将各个第三候选用户标签的训练标签，设置为样本用户标识对应的离散用户数据样本的负训练标签。

其中，参考用户标识所标识的用户可以是另一个样本用户。在一些实施例中，计算机设备可选取异于样本用户标识的参考用户标识，获取参考用户标识所对应离散用户数据样本的正训练标签所对应的那些候选用户标签。然后将样本用户标识的离散用户数据样本分别对应的这些候选用户标签的训练标签，设置为负训练标签。也就是说，样本用户的数量多于一个，其中一个样本用户的正例，可以作为另一个样本用户的反例。

举例说明，比如（数字代表对应的用户标签索引） a 用户：正例：0, 2, 3；反例：4, 5；b 用户：正例：6, 11, 7；反例：12, 9。可以将 a 的正例补充到 b 的反例里面，即 b 的反例变成 0, 2, 3, 12, 9。这样处理的意义在于可以在出现比较少的稀疏用户标签上面仍保持足够的准确性，体现每个用户的个性化兴趣。

在一个实施例中，在训练机器学习模型时，分别采用不同的样本组织方式组织了三组不同的正反例样本。分别通过这三组正反例样本对模型进行了线下测试。测试结果如下表一所示：

表一：

样本组织方式	Prec@10	Prec@20	Prec@50
Model V1	0.3566	0.3401	0.3239
Model V2	0.3779	0.3566	0.3372
Model V3	0.3824	0.3617	0.3422

其中，Model V1 正例采用曝光且点击内容的用户标签，反例采用随机用户标签；Model V2 正例采用曝光且点击内容的用户标签，反例采用曝光未点击内容的用户标签；Model V2 正例采用曝光且点击内容的用户标签，反例采用曝光未点击内容的用户标签和上一个样本用户的正例。Prec@K 表示机器学习模型预估的相关评分最高的 K 个标签有多少比例是正例的用户标签。从表一可以看出，正负样例组织的两步优化在 prec@K 的各个位置都有显著提高。也就是说，机器学习模型预测的靠前的用户标签都是非常精准的。

需要说明的是，表一所测试的机器学习模型在进行特征处理时，未基于注意力机制进行域内融合、未基于注意力机制进行域内融合且未基于因子分解机原理进行特征交叉。

在一个实施例中，在训练机器学习模型时，逐步增加基于注意力机制进行域内融合、基于注意力机制进行域内融合和基于因子分解机原理进行特征交叉处理；再采用曝光且点击内容的用户标签作为正例，采用曝光未点击内容的用户标签和上一个样本用户的正例作为反例进行了线下测试。测试结果如下表二所示：

表二：

模型改进方式	Loss	Prec@10	Prec@20	Prec@50
Model V3	0.326	0.3824	0.3617	0.3422
Model V4	0.323	0.3829	0.3623	0.3427
Model V5	0.319	0.3836	0.3622	0.3425
Model V6	0.314	0.3838	0.3634	0.3434

其中，Model V3 所测试的机器学习模型未基于注意力机制进行域内融合、未基于注意力机制进行域内融合且未基于因子分解机原理进行特征交叉；Model V4 所测试的机器学习模型增加了基于注意力机制进行域内融合的处理；Model V5 所测试的机器学习模型增加了基于注意力机制进行域内融合和基于因子分解机原理进行特征交叉处理；Model V6 所测试的模型增加了基于注意力机制进行域内融合、基于注意力机制进行域内融合的处理、以及基于因子分解机原理进行特征交叉的处理。从表二可以看出随着模型的逐步改进，模型的损失函数逐渐降低，Prec@K 逐步升高（除 Model V4 在 Prec@20 与 Prec@50 略低外，可能由于模型过拟合导致）。

在另外的实施例中，计算机设备还对训练得到的机器学习模型进行了线上验证。计算机设备可定义衡量指标：标签有点数(QC, Quantity Click)、标签点击率(CTR, Click Through Rate)和标签有点率(DCTR, Document-based Click Through Rate)。这些指标以曝光和点击的推荐内容所对应的用户标签作为数据，标签有点数是人均用户标签点击个数，标签有点数越多，说明提供的用户标签覆盖用户更广的兴趣；标签点击率表示用户标签点击的次数除以用户标签曝光的次数，该指标越高说明曝光的效率越高；标签有点率表示每个人去重后的点击的用户标签的个数除以去重后的曝光用户标签个数，该指标越高表示用户标签的准确率越高。其中，用户标签曝光的次数是指用户标签所对应的推荐内容的曝光次数，用户标签点击的次数是指用户标签所对应的推荐内容的点击次数。

在一个实施例中，对 Model V3 和 Model V6 所测试的模型，进行了标签有点数(QC)、标签点击率(CTR)和标签有点率(DCTR)的线上测试，测试结果如下表三所示：

表三：

模型改进方式	QC	CTR	DCTR
Model V3	0.78%	3.89%	7.42%
Model V6	2.36%	4.24%	7.94%

从表三可以看出，机器学习模型在增加了基于注意力机制进行域内融合、基于注意力机制进行域内融合的处理、以及基于因子分解机原理进行特征交叉的处理后，标签有点数(QC)、标签点击率(CTR)和标签有点率(DCTR)都有较大提升。

上述实施例中，计算机设备通过样本有监督地训练机器学习模型来进行数据处理，在选择设置好模型结构、训练样本和相应的训练标签后，即可通过样本自学习到模型参数，从而利用机器学习模型强大的特征表达能力和学习能力，学习到更好地对用户特征进行向量表达的能力，提高了用户标签生成准确率。

在一个实施例中，该用户标签生成方法还包括：获取多于一个的待推荐内容；确定各个待推荐内容各自相应的用户标签；将各自相应的用户标签与目标用户标签匹配的待推荐内容，推荐至目标用户标识所对应的终端。

其中，待推荐内容是待向用户推荐的内容。待推荐内容可以是推广信息、应用程序、视频、音频、新闻、文章或者商品等。每个待推荐内容可存在对应的用户标签。待推荐内容对应的用户标签，可以根据待推荐内容的内容特征从用户标签池中选取得到。待推荐内容对应的用户标签可以反映待推荐内容的内容特性。用户的用户标签可以反映用户的兴趣特性。那么，可以基于用户标签对用户进行内容推荐。当一个待推荐内容对应的用户标签与一个用户的用户标签相匹配时，可以认为该待推荐内容符合该用户的兴趣偏好，可以将该待推荐内容推荐给该用户。

在一些实施例中，计算机设备可以获取多于一个的待推荐内容，分别确定各个待推荐内容各自相应的用户标签，然后分别将各个待推荐内容相应的用户标签与目标用户标签进行比较，也就是与目标用户的用户标签进行比较。当各个待推荐内容相应的用户标签与目标用户标签匹配时，则将该待推荐内容推荐至目标用户标识所对应的终端。

在一个实施例中，目标用户标签可以多于一个，且存在先后顺序。那么相应的用户标签与目标用户标签匹配的待推荐内容在推荐时，也存在推荐顺序。相应的用户标签与靠前的目标用户标签匹配的待推荐内容的推荐顺序靠前，相应的用户标签与靠后的目标用户标签匹配的待推荐内容的推荐顺序靠后。

举例说明，如图 6 所示，假设计算机设备通过机器学习模型生成的目标用户标签包括：用户标签 3、用户标签 7 和用户标签 1。计算机设备获取的待推荐内容包括：待推荐内容 A、待推荐内容 B、待推荐内容 C、待推荐内容 D 和待推荐内容 E。其中，待推荐内容 A 对应用户标签 4、待推荐内容 B 对应用户标签 1、待推荐内容 C 对应用户标签 12、待推荐内容 D 对应用户标签 3、待推荐内容 E 对应用户标签 1。那么，待推荐内容 B、D 和 E 对应的用户标签与目标用户标签匹配，可以被推荐至目标用户，且推荐顺序为：待推荐内容 D→待推荐内容 E→待推荐内容 B。该推荐顺序也就是推荐内容被展示给目标用户时的展示顺序。

上述实施例中，在生成目标用户的目标用户标签后，可基于目标用户标签进行推荐场景的推荐应用，提高了推荐的准确性与有效性。

应该理解的是，虽然上述各个实施例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，上述各实施例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。

如图 7 所示，在一个实施例中，提供了一种用户标签生成装置 700，该装置设置在计算机设备中。参照图 7，该用户标签生成装置 700 包括：获取模块 701、融合处理模块 702、交叉处理模块 703 和筛选模块 704。

获取模块 701，用于获取多于一个特征域下分别与目标用户标识对应的离散用户数据。

融合处理模块 702，用于根据各个特征域下的离散用户数据，得到各个特征域各自相应的、且与目标用户标识对应的域内特征；融合各个域内特征得到与目标用户标识对应的域间特征。

交叉处理模块 703，用于对域间特征中的子特征进行特征交叉处理，获得交叉特征。

筛选模块 704，用于根据域间特征和交叉特征，从候选的用户标签中筛选出与目标用户标识对应的用户标签。

在一个实施例中，融合处理模块 702 还用于将各个离散用户数据输入机器学习模型，机器学习模型包括输入层和域内融合层；通过输入层向量化各个离散用户数据，得到各个离散用户数据各自相应的离散特征向量；通过域内融合层融合各个特征域下的离散特征向量，得到各个特征域相应的、且与目标用户标识对应的域内特征向量。

在一个实施例中，融合处理模块 702 还用于通过域内融合层分别获取各个特征域下的各离散特征向量各自对应的注意力分配权重；通过域内融合层，将各个特征域下的各个离散特征向量按照各自对应的注意力分配权重进行线性融合，得到各个特征域各自相应的、且与目标用户标识对应的域内特征向量。

在一个实施例中，机器学习模型还包括域间融合层。融合处理模块 702 还用于通过域间融合层分别获取各个域内特征向量各自对应的注意力分配权重；通过域间融合层，将各个域内特征向量按照各自对应的注意力分配权重进行线性融合，得到与目标用户标识对应的域间特征向量。

在一个实施例中，机器学习模型还包括特征交叉层。交叉处理模块 703 还用于通过特征交叉层，对域间特征向量进行去稀疏处理得到稠密化的域间特征向量；对稠密化的域间特征向量中的子特征进行二阶交叉处理，获得交叉特征。

在一个实施例中，交叉处理模块 703 还用于通过特征交叉层，将域间特征向量中的子特征分别映射为预设维度的隐层空间向量；对于域间特征向量中任意的两个子特征，将两个子特征以及两个子特征所映射得到的隐层空间向量的乘积，作为两个子特征的二阶交叉特征向量；组合各个二阶交叉特征向量，得到交叉特征向量。

在一个实施例中，机器学习模型还包括第一全连接层和第二全连接层。筛选模块 704 还用于通过第一全连接层，将交叉特征向量的维度调整至与域间特征向量的维度一致后，与域间特征向量拼接得到中间特征向量；通过第二全连接层，将中间特征向量映射为目标特征向量；根据目标特征向量，从候选用户标签中筛选出与目标用户标识对应的目标用户标签。

在一个实施例中，机器学习模型还包括预测层。筛选模块 704 还用于通过预测层，输出目标特征向量与各个候选用户标签所对应的标签向量的相关评分；筛选出相关评分靠前的预设数量的候选用户标签，作为与目标用户标识对应的目标用户标签。

如图 8 所示，在一个实施例中，用户标签生成装置 700 还包括：训练模块 705，用于收集多于一个特征域下分别与样本用户标识对应的离散用户数据样本；确定离散用户数据样本分别对应的各个候选用户标签的训练标签；将离散用户数据样本输入机器学习模型，经过机器学习模型所包括的各个层依次处理后，输出离散用户数据样本与各个候选用户标签所对应的标签向量的相关评分样本；根据相关评分样本和对应的训练标签的差异，训练机器学习模型。

在一个实施例中，训练模块 705 还用于获取与样本用户标识对应的历史浏览记录；查询

历史浏览记录中的点击内容相应的第一候选用户标签，以及历史浏览记录中的未点击内容相应的第二候选用户标签；将离散用户数据样本分别对应的各个第一候选用户标签的训练标签设置为正训练标签，并将离散用户数据样本分别对应的各个第二候选用户标签的训练标签设置为负训练标签。

在一个实施例中，训练模块 705 还用于选取异于样本用户标识的参考用户标识，以及所述参考用户标识对应的离散用户数据样本；获取离散用户数据样本的正训练标签所对应的第三候选用户标签；将各个第三候选用户标签的训练标签，设置为样本用户标识对应的离散用户数据样本的负训练标签。

如图 9 所示，在一个实施例中，用户标签生成装置 700 还包括：推荐模块 706，用于获取多于一个的待推荐内容；确定各个待推荐内容各自相应的用户标签；将各自相应的用户标签与目标用户标签匹配的待推荐内容，推荐至目标用户标识所对应的终端。

上述用户标签生成装置 700，在获取到多于一个特征域下分别与目标用户标识对应的离散用户数据后，一方面对这些离散的用户数据在特征域内以及特征域间进行融合，这样可以很好地对稀疏行为进行融合，得到融合了各种稀疏行为的域间特征；另一方面还对该域间特征中的子特征进行交叉处理，这样可以很好地对各个子特征进行融合，得到融合了域间特征中子特征的交叉特征。由于域间特征融合了各种稀疏行为，交叉特征融合了域间特征中的子特征，根据域间特征和交叉特征从候选用户标签中筛选与目标用户标识对应的目标用户标签，能够有效地提高生成的用户标签的准确性。

图 10 示出了一个实施例中计算机设备的内部结构图。该计算机设备可以是图 1 中的终端 110 或服务器 120。如图 10 所示，该计算机设备包括该计算机设备通过系统总线连接的处理器、存储器和网络接口。其中，存储器包括非易失性存储介质和内存。该计算机设备的非易失性存储介质存储有操作系统，还可存储有计算机程序，该计算机程序被处理器执行时，可使得处理器实现用户标签生成方法。该内存中也存储有计算机程序，该计算机程序被处理器执行时，可使得处理器执行用户标签生成方法。本领域技术人员可以理解，图 10 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，在一些实施例中，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。

在一个实施例中，本申请提供的用户标签生成装置可以实现为一种计算机程序的形式，计算机程序可在如图 10 所示的计算机设备上运行。计算机设备的存储器中可存储组成该用户标签生成装置的各个程序模块，比如，图 7 所示的获取模块 701、融合处理模块 702、交叉处理模块 703 和筛选模块 704。各个程序模块构成的计算机程序使得处理器执行本说明书中描述的本申请各个实施例的用户标签生成方法中的步骤。

例如，图 10 所示的计算机设备可以通过如图 7 所示的用户标签生成装置中的获取模块 701 获取多于一个特征域下分别与目标用户标识对应的离散用户数据。通过融合处理模块 702 根据各个特征域下的离散用户数据，得到各个特征域各自相应的、且与目标用户标识对应的域内特征；融合各个域内特征得到与目标用户标识对应的域间特征。通过交叉处理模块 703 对域间特征中的子特征进行特征交叉处理，获得交叉特征。通过筛选模块 704 根据域间特征和交叉特征，从候选的用户标签中筛选出与目标用户标识对应的用户标签。

在一个实施例中，提供了一种计算机设备，包括存储器和处理器，存储器存储有计算机程序，计算机程序被处理器执行时，使得处理器执行上述用户标签生成方法的步骤。此处用户标签生成方法的步骤可以是上述各个实施例的用户标签生成方法中的步骤。

在一个实施例中，提供了一种计算机可读存储介质，存储有计算机程序，计算机程序被处理器执行时，使得处理器执行上述用户标签生成方法的步骤。此处用户标签生成方法的步骤可以是上述各个实施例的用户标签生成方法中的步骤。

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，所述的程序可存储于一非易失性计算机可读取存储介质

中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各个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。

以上实施例的各个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。

以上所述仅为本申请的可选实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权利要求书

- 1、一种用户标签生成方法，所述方法应用于计算机设备中，包括：
获取多于一个特征域下分别与目标用户标识对应的离散用户数据；
根据各个所述特征域下的离散用户数据，得到各个所述特征域各自相应的、且与所述目标用户标识对应的域内特征；
融合各个所述域内特征得到与所述目标用户标识对应的域间特征；
对所述域间特征中的子特征进行特征交叉处理，获得交叉特征；
根据所述域间特征和所述交叉特征，从候选用户标签中筛选出与所述目标用户标识对应的目标用户标签。
- 2、根据权利要求1所述的方法，其特征在于，所述根据各个所述特征域下的离散用户数据，得到各个所述特征域各自相应的、且与所述目标用户标识对应的域内特征，包括：
将各个所述离散用户数据输入机器学习模型，所述机器学习模型包括输入层和域内融合层；
通过所述输入层向量化各个所述离散用户数据，得到各个所述离散用户数据各自相应的离散特征向量；
通过所述域内融合层融合各个所述特征域下的离散特征向量，得到各个所述特征域相应的、且与所述目标用户标识对应的域内特征向量。
- 3、根据权利要求2所述的方法，其特征在于，所述通过所述域内融合层融合各个所述特征域下的离散特征向量，得到各个所述特征域相应的、且与所述目标用户标识对应的域内特征向量，包括：
通过所述域内融合层分别获取各个所述特征域下的各个离散特征向量各自对应的注意力分配权重；
通过所述域内融合层，将各个所述特征域下的各个离散特征向量按照各自对应的注意力分配权重进行线性融合，得到各个所述特征域各自相应的、且与所述目标用户标识对应的域内特征向量。
- 4、根据权利要求2所述的方法，其特征在于，所述机器学习模型还包括域间融合层；
所述融合各个所述域内特征得到与所述目标用户标识对应的域间特征，包括：
通过所述域间融合层分别获取各个所述域内特征向量各自对应的注意力分配权重；
通过所述域间融合层，将各个所述域内特征向量按照各自对应的注意力分配权重进行线性融合，得到与所述目标用户标识对应的域间特征向量。
- 5、根据权利要求2所述的方法，其特征在于，所述机器学习模型还包括特征交叉层；
所述对所述域间特征中的子特征进行特征交叉处理，获得交叉特征，包括：
通过所述特征交叉层，对域间特征向量进行去稀疏处理得到稠密化的域间特征向量；
对所述稠密化的域间特征向量中的子特征进行二阶交叉处理，获得交叉特征。
- 6、根据权利要求5所述的方法，其特征在于，所述通过所述特征交叉层，对域间特征向量进行去稀疏处理得到稠密化的域间特征向量，包括：
通过所述特征交叉层，将所述域间特征向量中的子特征分别映射为预设维度的隐层空间向量；
所述对所述稠密化的域间特征向量中的子特征进行二阶交叉处理，获得交叉特征，包括：

对于所述域间特征向量中任意的两个子特征，将所述两个子特征以及所述两个子特征所映射得到的所述隐层空间向量的乘积，作为所述两个子特征的二阶交叉特征向量；

组合各个所述二阶交叉特征向量，得到交叉特征向量。

7、根据权利要求2所述的方法，其特征不在于，所述机器学习模型还包括第一全连接层和第二全连接层；

所述根据所述域间特征和所述交叉特征，从候选用户标签中筛选出与所述目标用户标识对应的目标用户标签，包括：

通过所述第一全连接层，将交叉特征向量的维度调整至与域间特征向量的维度一致后，与所述域间特征向量拼接得到中间特征向量；

通过所述第二全连接层，将所述中间特征向量映射为目标特征向量；

根据所述目标特征向量，从所述候选用户标签中筛选出与所述目标用户标识对应的目标用户标签。

8、根据权利要求7所述的方法，其特征不在于，所述机器学习模型还包括预测层；

所述根据所述目标特征向量，从所述候选用户标签中筛选出与所述目标用户标识对应的目标用户标签，包括：

通过所述预测层，输出所述目标特征向量与各个所述候选用户标签所对应的标签向量的相关评分；

筛选出所述相关评分靠前的预设数量的候选用户标签，作为与所述目标用户标识对应的目标用户标签。

9、根据权利要求2所述的方法，其特征不在于，所述方法还包括：

收集多于一个特征域下分别与样本用户标识对应的离散用户数据样本；

确定所述离散用户数据样本分别对应的各个候选用户标签的训练标签；

将所述离散用户数据样本输入所述机器学习模型，经过所述机器学习模型所包括的各个层依次处理后，输出所述离散用户数据样本与各个候选用户标签所对应的标签向量的相关评分样本；

根据所述相关评分样本和对应的训练标签的差异，训练所述机器学习模型。

10、根据权利要求9所述的方法，其特征不在于，所述确定所述离散用户数据样本分别对应各个候选用户标签的训练标签，包括：

获取与所述样本用户标识对应的历史浏览记录；

查询所述历史浏览记录中的点击内容相应的第一候选用户标签，以及所述历史浏览记录中的未点击内容相应的第二候选用户标签；

将所述离散用户数据样本分别对应的各个所述第一候选用户标签的训练标签设置为正训练标签，并将所述离散用户数据样本分别对应的各个所述第二候选用户标签的训练标签设置为负训练标签。

11、根据权利要求10所述的方法，其特征不在于，所述方法还包括：

选取异于所述样本用户标识的参考用户标识，以及所述参考用户标识对应的离散用户数据样本；

获取所述离散用户数据样本的正训练标签所对应的第三候选用户标签；

将各个所述第三候选用户标签的训练标签，设置为所述样本用户标识对应的离散用户数据样本的负训练标签。

12、根据权利要求 1 至 11 中任一项所述的方法，其特征在于，所述方法还包括：

获取多于一个的待推荐内容；

确定各个所述待推荐内容各自相应的用户标签；

将所述各自相应的用户标签与所述目标用户标签匹配的待推荐内容，推荐至所述目标用户标识所对应的终端。

13、一种用户标签生成装置，所述装置设置在计算机设备中，包括：

获取模块，用于获取多于一个特征域下分别与目标用户标识对应的离散用户数据；

融合处理模块，用于根据各个所述特征域下的离散用户数据，得到各个所述特征域各自相应的、且与所述目标用户标识对应的域内特征；融合各个所述域内特征得到与所述目标用户标识对应的域间特征；

交叉处理模块，用于对所述域间特征中的子特征进行特征交叉处理，获得交叉特征；

筛选模块，用于根据所述域间特征和所述交叉特征，从候选的用户标签中筛选出与所述目标用户标识对应的用户标签。

14、根据权利要求 13 所述的装置，其特征在于，所述融合处理模块，用于将各个所述离散用户数据输入机器学习模型，所述机器学习模型包括输入层和域内融合层；通过所述输入层向量化各个所述离散用户数据，得到各个所述离散用户数据各自相应的离散特征向量；通过所述域内融合层融合各个所述特征域下的离散特征向量，得到各个所述特征域相应的、且与所述目标用户标识对应的域内特征向量。

15、根据权利要求 14 所述的装置，其特征在于，所述融合处理模块，用于通过所述域内融合层分别获取各个所述特征域下的各个离散特征向量各自对应的注意力分配权重；通过所述域内融合层，将各个所述特征域下的各个离散特征向量按照各自对应的注意力分配权重进行线性融合，得到各个所述特征域各自相应的、且与所述目标用户标识对应的域内特征向量。

16、根据权利要求 13 所述的装置，其特征在于，所述机器学习模型还包括域间融合层；

所述融合处理模块，用于通过所述域间融合层分别获取各个所述域内特征向量各自对应的注意力分配权重；通过所述域间融合层，将各个所述域内特征向量按照各自对应的注意力分配权重进行线性融合，得到与所述目标用户标识对应的域间特征向量。

17、根据权利要求 13 所述的装置，其特征在于，所述机器学习模型还包括特征交叉层；

所述交叉处理模块，用于通过所述特征交叉层，对域间特征向量进行去稀疏处理得到稠密化的域间特征向量；对所述稠密化的域间特征向量中的子特征进行二阶交叉处理，获得交叉特征。

18、根据权利要求 17 所述的装置，其特征在于，所述交叉处理模块，用于通过所述特征交叉层，将所述域间特征向量中的子特征分别映射为预设维度的隐层空间向量；对于所述域间特征向量中任意的两个子特征，将所述两个子特征以及所述两个子特征所映射得到的所述隐层空间向量的乘积，作为所述两个子特征的二阶交叉特征向量；组合各个所述二阶交叉特征向量，得到交叉特征向量。

19、根据权利要求 13 所述的装置，其特征在于，所述机器学习模型还包括第一全连接层和第二全连接层；

所述筛选模块，用于通过所述第一全连接层，将交叉特征向量的维度调整至与域间特征向量的维度一致后，与所述域间特征向量拼接得到中间特征向量；通过所述第二全连接层，

将所述中间特征向量映射为目标特征向量；根据所述目标特征向量，从所述候选用户标签中筛选出与所述目标用户标识对应的目标用户标签。

20、根据权利要求 19 所述的装置，其特征在于，所述机器学习模型还包括预测层；

所述筛选模块，用于通过所述预测层，输出所述目标特征向量与各个所述候选用户标签所对应的标签向量的相关评分；筛选出所述相关评分靠前的预设数量的候选用户标签，作为与所述目标用户标识对应的目标用户标签。

21、根据权利要求 14 所述的装置，其特征在于，所述装置包括训练模块；

所述训练模块，用于收集多于一个特征域下分别与样本用户标识对应的离散用户数据样本；确定所述离散用户数据样本分别对应的各个候选用户标签的训练标签；将所述离散用户数据样本输入所述机器学习模型，经过所述机器学习模型所包括的各个层依次处理后，输出所述离散用户数据样本与各个候选用户标签所对应的标签向量的相关评分样本；根据所述相关评分样本和对应的训练标签的差异，训练所述机器学习模型。

22、根据权利要求 21 所述的装置，其特征在于，所述训练模块，用于获取与所述样本用户标识对应的历史浏览记录；查询所述历史浏览记录中的点击内容相应的第一候选用户标签，以及所述历史浏览记录中的未点击内容相应的第二候选用户标签；将所述离散用户数据样本分别对应的各个所述第一候选用户标签的训练标签设置为正训练标签，并将所述离散用户数据样本分别对应的各个所述第二候选用户标签的训练标签设置为负训练标签。

23、根据权利要求 22 所述的装置，其特征在于，所述训练模块，用于选取异于所述样本用户标识的参考用户标识，以及所述参考用户标识对应的离散用户数据样本；获取所述离散用户数据样本的正训练标签所对应的第三候选用户标签；将各个所述第三候选用户标签的训练标签，设置为所述样本用户标识对应的离散用户数据样本的负训练标签。

24、根据权利要求 13 至 22 任一所述的装置，其特征在于，所述装置包括推荐模块；

所述推荐模块，用于获取多于一个的待推荐内容；确定各个所述待推荐内容各自相应的用户标签；将所述各自相应的用户标签与所述目标用户标签匹配的待推荐内容，推荐至所述目标用户标识所对应的终端。

25、一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行如权利要求 1 至 12 中任一项所述方法的步骤。

26、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，所述计算机程序被所述处理器执行时，使得所述处理器执行如权利要求 1 至 12 中任一项所述方法的步骤。

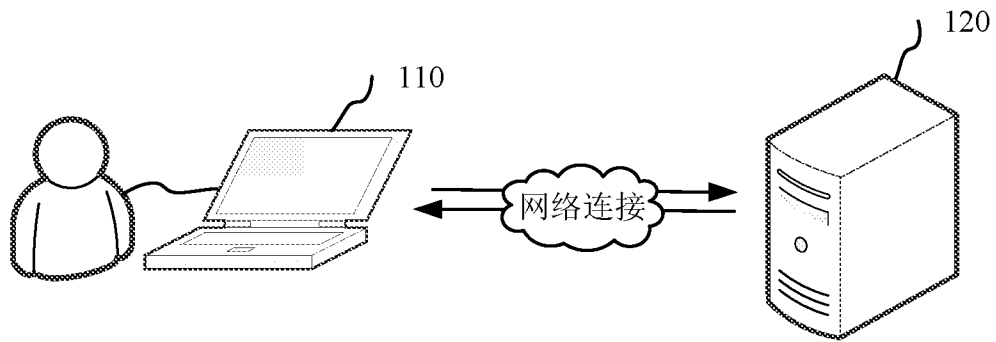


图 1

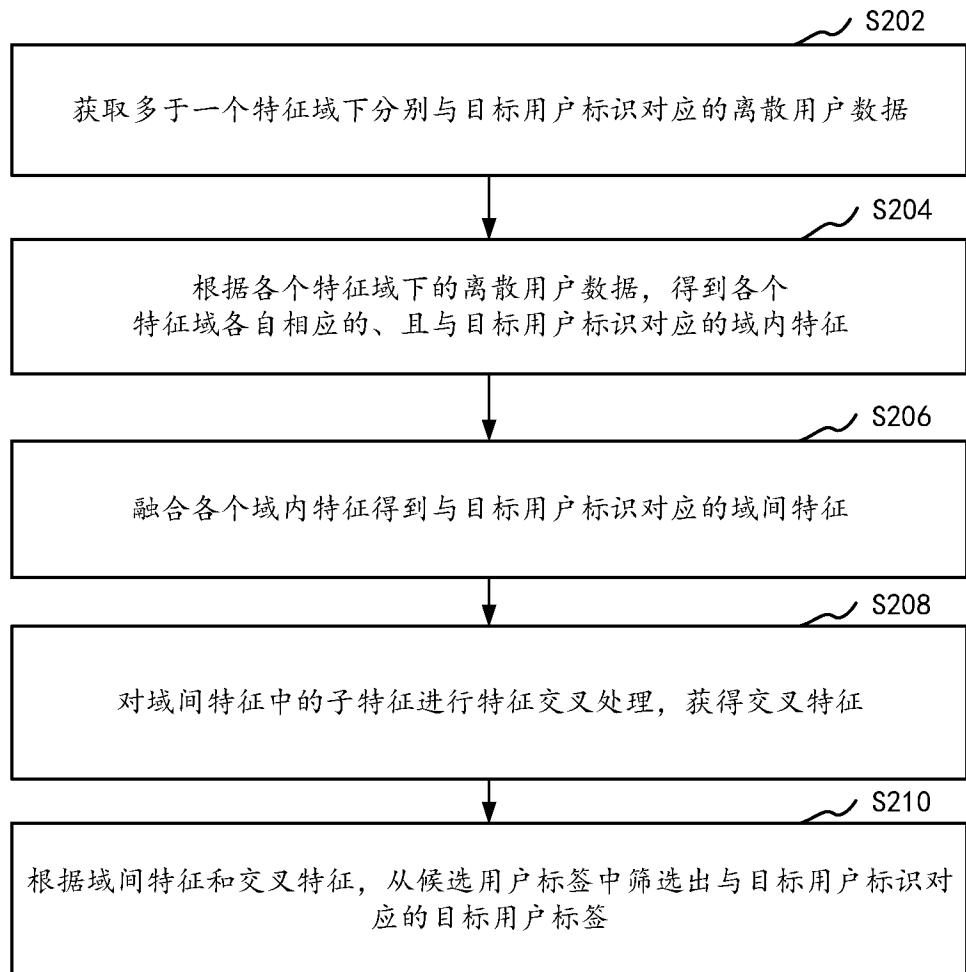
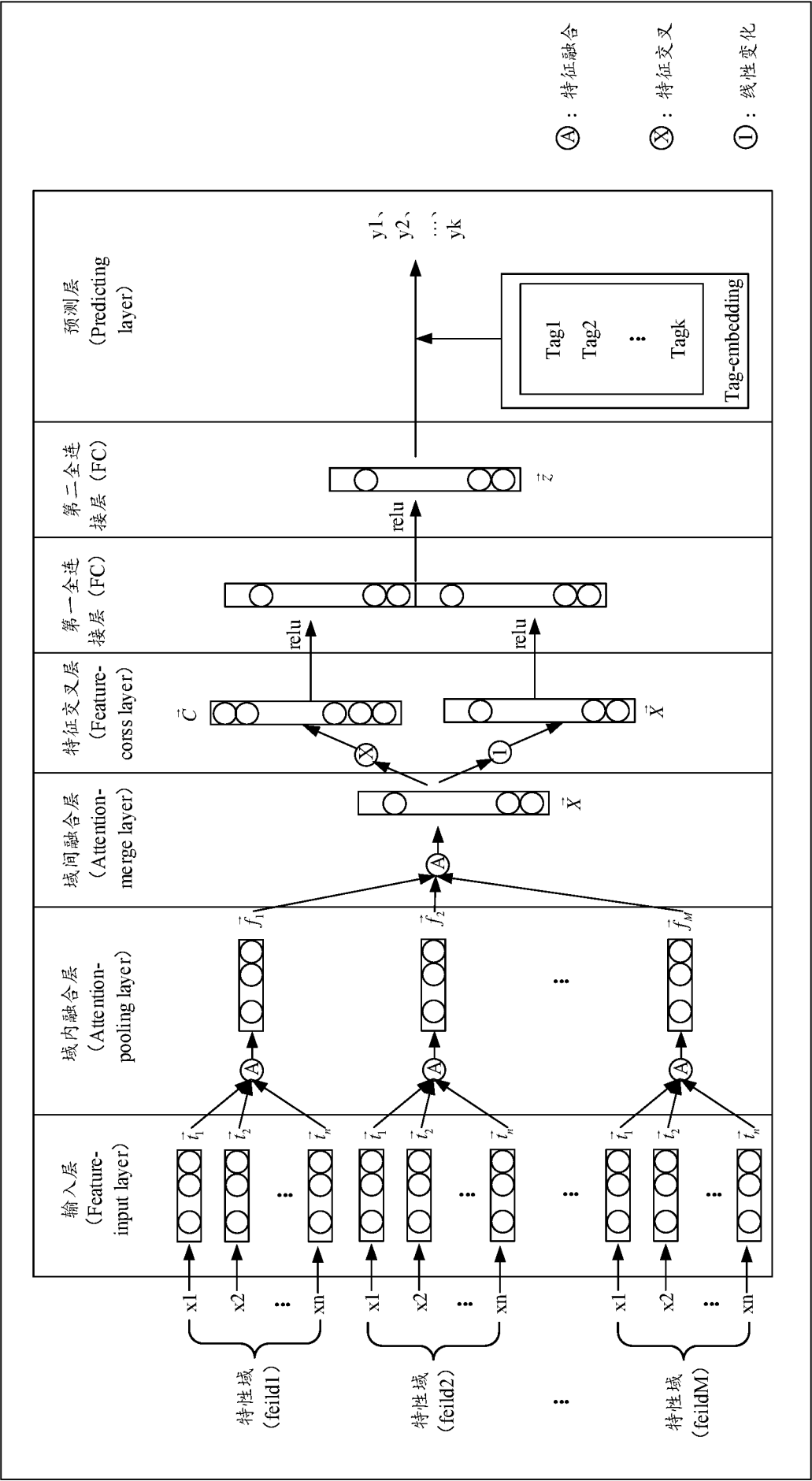


图 2



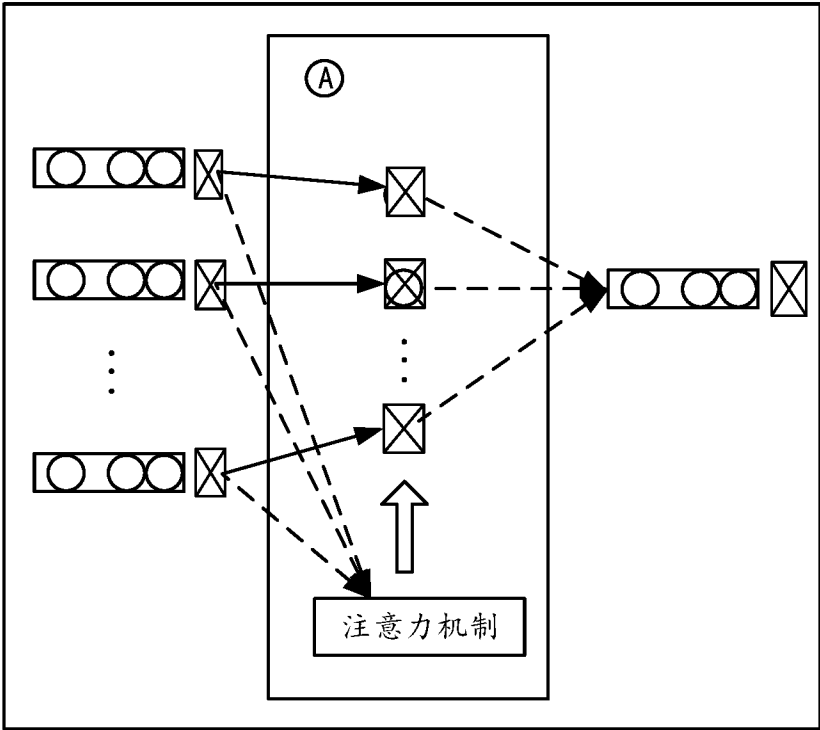


图 4

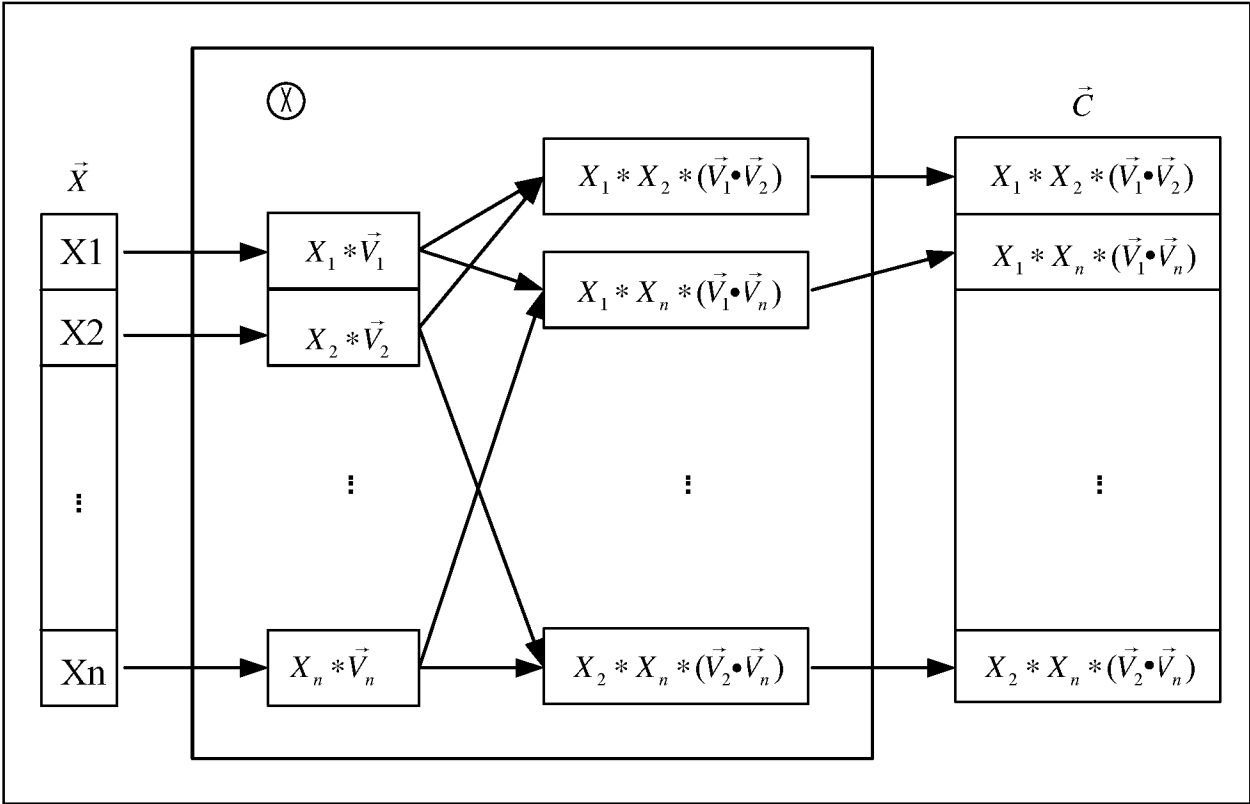


图 5

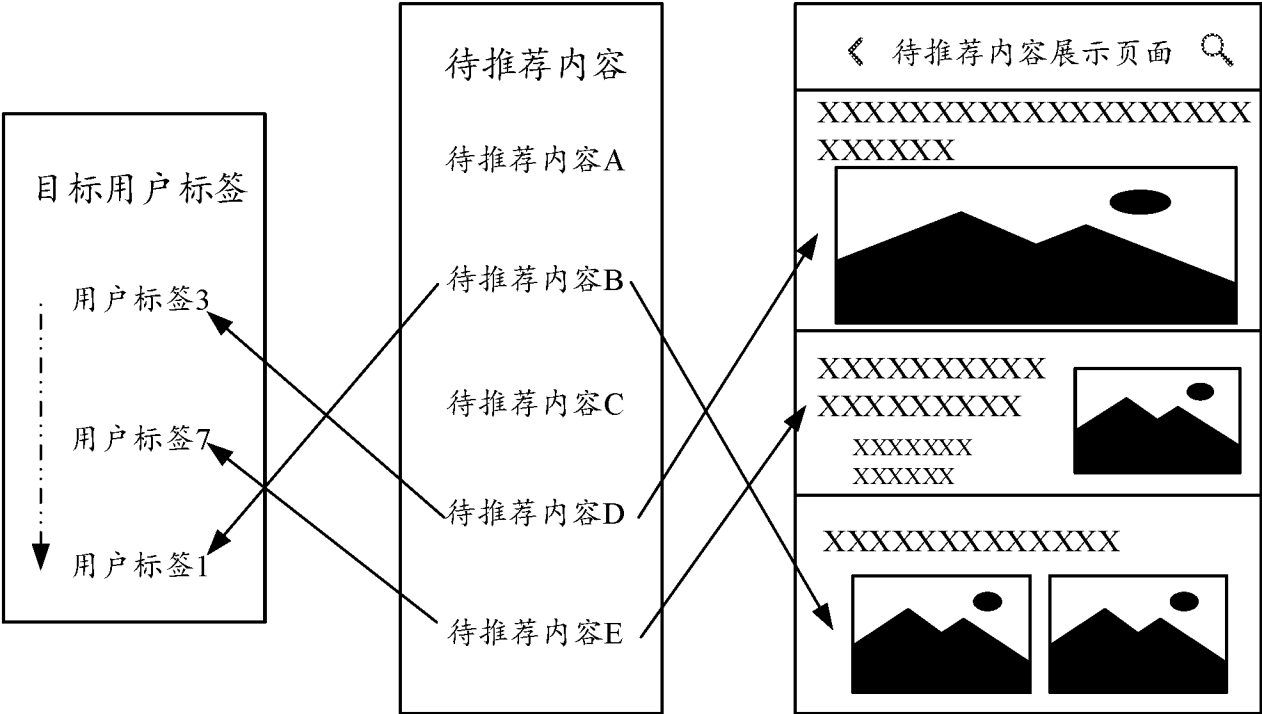


图 6

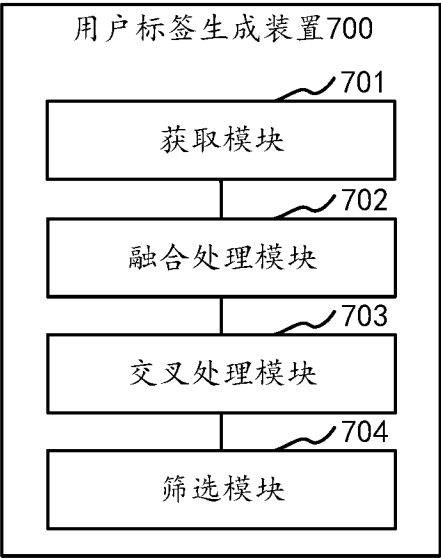


图 7

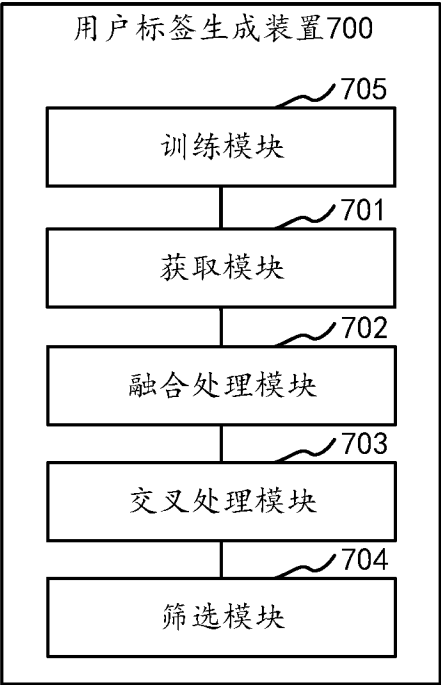


图 8

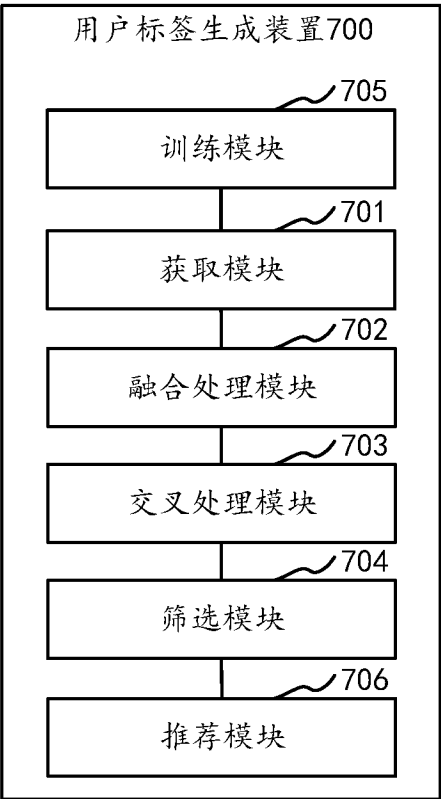


图 9

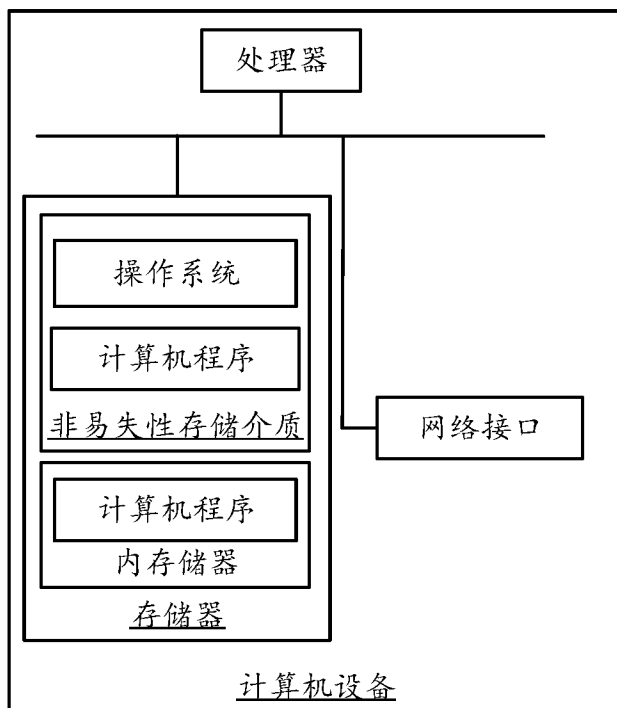


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/079445

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/955(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)

CNPAT, EPODOC, WPI, CNKI, IEEE: 用户画像, 用户标签, 数据, 特征, 提取, 处理, 融合, 交叉, 域, 源, 模型, 学习, user, consumer, figure, label, tag, data, character, feature, pick-up, extract, deal, treatment, fusion, cross, domain, region, source, model, learn

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110263265 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 20 September 2019 (2019-09-20) claims 1-15	1-26
X	CN 105701161 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY et al.) 22 June 2016 (2016-06-22) description, paragraphs [0028]-[0067], abstract	1-26
X	CN 107609116 A (TRANSWARP TECHNOLOGY (SHANGHAI) CO., LTD.) 19 January 2018 (2018-01-19) description, paragraphs [0002]-[0075], abstract	1-26
A	CN 104602042 A (UNION NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 06 May 2015 (2015-05-06) entire document	1-26
A	CN 109345302 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 15 February 2019 (2019-02-15) entire document	1-26



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

30 May 2020

Date of mailing of the international search report

15 June 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/CN2020/079445

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	110263265	A	20 September 2019	None	
CN	105701161	A	22 June 2016	None	
CN	107609116	A	19 January 2018	None	
CN	104602042	A	06 May 2015	None	
CN	109345302	A	15 February 2019	None	

国际检索报告

国际申请号

PCT/CN2020/079445

A. 主题的分类 G06F 16/955 (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)) CNPAT, EPDOC, WPI, CNKI, IEEE: 用户画像, 用户标签, 数据, 特征, 提取, 处理, 融合, 交叉, 域, 源, 模型, 学习, user, consumer, figure, label, tag, data, character, feature, pick-up, extract, deal, treatment, fusion, cross, domain, region, source, model, learn																				
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>PX</td> <td>CN 110263265 A (腾讯科技深圳有限公司) 2019年 9月 20日 (2019 - 09 - 20) 权利要求1-15</td> <td>1-26</td> </tr> <tr> <td>X</td> <td>CN 105701161 A (深圳先进技术研究院 等) 2016年 6月 22日 (2016 - 06 - 22) 说明书第[0028]-[0067]段, 摘要</td> <td>1-26</td> </tr> <tr> <td>X</td> <td>CN 107609116 A (星环信息科技上海有限公司) 2018年 1月 19日 (2018 - 01 - 19) 说明书第[0002]-[0075]段, 摘要</td> <td>1-26</td> </tr> <tr> <td>A</td> <td>CN 104602042 A (合一网络技术北京有限公司) 2015年 5月 6日 (2015 - 05 - 06) 全文</td> <td>1-26</td> </tr> <tr> <td>A</td> <td>CN 109345302 A (腾讯科技深圳有限公司) 2019年 2月 15日 (2019 - 02 - 15) 全文</td> <td>1-26</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	PX	CN 110263265 A (腾讯科技深圳有限公司) 2019年 9月 20日 (2019 - 09 - 20) 权利要求1-15	1-26	X	CN 105701161 A (深圳先进技术研究院 等) 2016年 6月 22日 (2016 - 06 - 22) 说明书第[0028]-[0067]段, 摘要	1-26	X	CN 107609116 A (星环信息科技上海有限公司) 2018年 1月 19日 (2018 - 01 - 19) 说明书第[0002]-[0075]段, 摘要	1-26	A	CN 104602042 A (合一网络技术北京有限公司) 2015年 5月 6日 (2015 - 05 - 06) 全文	1-26	A	CN 109345302 A (腾讯科技深圳有限公司) 2019年 2月 15日 (2019 - 02 - 15) 全文	1-26
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 110263265 A (腾讯科技深圳有限公司) 2019年 9月 20日 (2019 - 09 - 20) 权利要求1-15	1-26																		
X	CN 105701161 A (深圳先进技术研究院 等) 2016年 6月 22日 (2016 - 06 - 22) 说明书第[0028]-[0067]段, 摘要	1-26																		
X	CN 107609116 A (星环信息科技上海有限公司) 2018年 1月 19日 (2018 - 01 - 19) 说明书第[0002]-[0075]段, 摘要	1-26																		
A	CN 104602042 A (合一网络技术北京有限公司) 2015年 5月 6日 (2015 - 05 - 06) 全文	1-26																		
A	CN 109345302 A (腾讯科技深圳有限公司) 2019年 2月 15日 (2019 - 02 - 15) 全文	1-26																		
☐ 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 2020年 5月 30日		国际检索报告邮寄日期 2020年 6月 15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 吴媛媛 电话号码 86-(10)-53961351																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/079445

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	110263265	A	2019年 9月 20日	无	
CN	105701161	A	2016年 6月 22日	无	
CN	107609116	A	2018年 1月 19日	无	
CN	104602042	A	2015年 5月 6日	无	
CN	109345302	A	2019年 2月 15日	无	